Springer Monographs in Mathematics

Pavel Pudlák

Logical Foundations of Mathematics and Computational Complexity

A Gentle Introduction

Springer

Pavel Pudlák
ASCR
Prague, Czech Republic

ISSN 1439-7382 Springer Monographs in Mathematics
ISBN 978-3-319-00118-0 ISBN 978-3-319-00119-7 (eBook)
DOI 10.1007/978-3-319-00119-7
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013936799

Mathematics Subject Classification: 03D15, 03E30, 03E35, 03F03, 03F20, 03F30, 03F40, 68Q15

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 1
Mathematician's World

The real universe arched sickeningly away beneath them. Various pretend ones flitted silently by, like mountain goats. Primal light exploded, splattering space-time as with gobbets of junket. Time blossomed, matter shrank away. The highest prime number coalesced quietly in a corner and hid itself away for ever.

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

FOR an ordinary person, it is a strange, imaginary world. At the entrance we meet very familiar creatures, such as the natural numbers $0, 1, 2, \ldots$, but further on there will appear many strange aliens, like the imaginary unit i, the first uncountable cardinal number $\aleph_1$ and things even stranger than these. In some sense it is like the artificial worlds of science fiction, or like a detective story made up of mysteries with logical solutions, but still in many respects it is very different. The main difference is, perhaps, not in the artificial nature of the things that we encounter in mathematics, which apparently have very little to do with our everyday life, but in the strict rules that they obey. In a good detective story the detective eventually solves a mysterious crime by applying logical deduction. The author usually pretends that you could also have deduced who the murderer was already at the beginning of the story, knowing only the basic data presented on the first few pages. But in fact this is not true; on the contrary, the author chooses the most unlikely person. In mathematics you really can solve problems using only deduction and, in fact, no initial data are needed, except for the statement of the problem; the only things you need are patience and determination.

If you read a good novel or regularly watch a TV series, you enter into the world of the heroes of the story and often forget, at least for a while, that it is not real. In science fiction stories you can even experience a completely different world than ours here on Earth. Science fiction gives writers the opportunity to construct new worlds, even worlds that are in contradiction with firmly established laws of physics. There is nothing wrong with this if it has its own logic. Similarly, mathematicians invent worlds which are sometimes completely alien to ordinary people. In their minds they create mental pictures of the concepts about which they are thinking, as if they could really see numbers, sets, functions, infinitely dimensional spaces and a lot more, and move in this environment arranging these objects until they construct

P. Pudlák, *Logical Foundations of Mathematics and Computational Complexity*, Springer Monographs in Mathematics, DOI 10.1007/978-3-319-00119-7_1,

the one they were looking for. Active mathematicians actually spend a big portion of their lives in this world. The more time they spend there, the more real this world seems to them. Like many teenagers who spend a lot of time in the virtual realities of computer games, mathematicians live part of their lives in what I would call *real virtuality*. Whereas virtual reality is pretend reality, what mathematicians do is the opposite: their worlds seem virtual, but are in some sense very real.

So is the mathematical world real or not? Most mathematicians would defend the true existence of at least some mathematical objects; in fact, most people would agree that the numbers $0, 1, 2, \ldots$ in some abstract sense do exist. As I will explain later, this is not just an important philosophical question, it is a question which is very important for the foundations of mathematics independent of our philosophical view, or our lack of interest in philosophy. But before we discuss such problems we have to know what kind of "things" mathematicians deal with.

1.1 Mathematical Structures

In biology we study animals, plants, bacteria, etc., in astronomy stars, planets, etc. So we can define biology as the science studying living organisms, astronomy as the science of the universe, and so on. But how can we describe mathematics? The answer to this question used to depend on what the main topic in contemporary mathematics was. For ancient Greeks, mathematics was essentially geometry and thus mathematics was the science of space. In the 18th century, when mathematics was tightly connected with physics, an answer to the question '*What is the subject of study of mathematics?*' would most likely be that it is *quantities and the relations between them.* A '*quantity*' was a real number that possibly depended on other numbers. For example, when describing a motion of a physical object, quantities could be position, speed, and momentum, all depending on time. The views on what the subject of mathematics is changed gradually. In roughly the 19th century mathematicians realized that there could be other objects of study on top of the traditional ones. The discovery of non-Euclidean geometries was an important step towards realizing that one does not have to study only objects which occur naturally in real life. An especially dramatic shift happened in algebra, where mathematicians realized that the usual number-theoretic structures are merely special instances from classes of structures sharing properties with the standard ones. Later on, new mathematical fields appeared where the objects studied had little to do with numbers or geometry. A systematic treatment of all mathematical objects became possible only after calculus had been given rigorous foundations and when there was a sufficiently general tool at hand: the concept of set.

I will describe the current standard approach to the question of what mathematical objects are. It is based on the concept of a *mathematical structure*, which gradually developed in the first half of the 20th century and was finally adopted as a key concept by the Bourbakists. *Nicolas Bourbaki* was a pseudonym under which, in 1939, a group of young French mathematicians started publishing an encyclopedic

series of monographs covering the main fields in mathematics. Naturally, an attempt to give a unified treatment to the whole of mathematics needed a general concept such as the concept of mathematical structure.

This is certainly not the only possible view of contemporary mathematics. If I were not interested in foundations and wanted rather to explain the source of ideas which led to the most profound results, I would choose a different vantage point. Quite often it is difficult to formalize general ideas by a single mathematical concept. In fact, the main progress in modern mathematics has in most cases been achieved by realizing that the same idea was present in several fields and thus results and proof techniques could be transferred from one field to another. A prominent example is algebraic geometry, a field which applies geometric ideas to various non-geometric objects, including some discrete structures. Mathematics has always been a never ending struggle to express general ideas in a comprehensible, general and rigorous way and thus it cannot be explained completely by a single concept such as the mathematical structure. Nevertheless, Bourbaki's structuralist approach is the best that we have.

The ancient mathematicians considered only a few structures: the natural numbers, the plane and three dimensional space. Gradually new structures appeared in mathematics, although it was not an easy process to accept them. For instance, the complex numbers turned out to be very useful, but for a long time they were treated as a strange auxiliary means to solve problems about real numbers. We still use the terminology of *real* and *imaginary* numbers, but now we treat these words as purely technical terms and do not attribute more existence to real numbers than to complex ones. In mathematical analysis people realized that functions can be added, multiplied, etc. just as numbers can be, though they are not numbers. An important turning point was when mathematicians realized that they did not have to study only one of the few standard structures, instead they could choose any structure from a large variety. It was as if the objects of study were not given to them, but they could design them according to their own will and need, just following certain rules. (Whether one views it as the possibility to choose, or the possibility to create, depends on one's philosophical standpoint.)

Let us turn to the definition of a structure. Roughly speaking, a mathematical structure is a toy or a gadget that you can play with. You push or turn knobs and something happens. It is also like a painting where a single brush stroke makes no sense, but together the strokes give some meaning. You can also think of a structure as a game. In a game you have certain *objects*, and *rules* that determine what you can do with the objects.

A nice example is Rubik's cube, the well-known toy: the objects are the 26 small cubes and the rules are fixed by the ingenious mechanism of Rubik's cube that allows you to move only certain groups of small cubes together, namely those that form a face of the cube. Though it was important to design the mechanical construction of the cube, so that it worked well and could be mass produced, the essence of it is not the mechanism. The only thing that is important is that you have 26 pieces and particular rules how to move them. You can do "mathematical research" on Rubik's cube by studying what configurations are possible, which are symmetric, how many

steps you need to transform a particular configuration into another one and so on. This is, in fact, what mathematicians actually do with structures.

There are many different structures; some are, in some sense, unique, while some are just members of large classes of similar structures. Let us consider the most familiar structure which is the natural numbers $0, 1, 2, 3, \ldots$. The structuralist point of view is that a single number, say 4, does not have any meaning. It has a meaning only as a part of the structure, namely, that *there are four numbers less than it*. Notice that we need the relation '*less than*', without it we could not distinguish 4 in this way. Furthermore we can add and multiply numbers (this is the 'playing with a toy' alluded to above). Thus we arrive at the following description of the natural numbers as a mathematical structure: they consists of

1. the *set* of nonnegative integers $\{0, 1, 2, 3, \ldots\}$, called the *universe*, or the *base set*, or the *underlying set* of the structure;
2. the *operations* of addition $+$ and multiplication $\cdot$;
3. the *relation* of being less than or equal $\leq$.

Notice the stressed words *set*, *operations*, and *relation*. This is, in fact, the form of all basic structures: they consist of a set on which there are some operations and relations defined. We do not restrict the number of operations and relations, except that their number must be finite. In particular, a structure can have only relations or only operations. For example, we may consider the natural numbers only with the ordering relation, or, on the contrary, we may add more operations. The natural numbers with no operations and $\leq$ as the only relation form a much simpler structure, but they are important when we are interested in a particular class of structures, namely, ordered sets.

In our example above the operations are *binary*, which means that they produce an element from 2 elements. Obviously, one can consider operations with this parameter 2, called the *arity*, replaced by any natural number.[1] In particular, operations of arity 0 are called *constants* and operations of arity 1 are called *functions*. Operations with arity greater than 2 are rare. The arity of a relation can be any number greater or equal than 1. A unary relation, that is a relation of arity 1, is usually called a *predicate*, or a *property*. An example of a ternary relation is the relation '*x is between y and z*' used in the formalization of plane geometry.

It probably required a considerable psychological effort for mathematicians to realize that the underlying set, the universe of a structure, does not determine the relations and operations. For example, originally people thought of the natural numbers as something intrinsically associated with the natural ordering and the two basic operations. The realization that we are completely free to choose operations and relations (and that the resulting structures can be interesting and useful) led to a dramatic development of mathematics in the 19th century, especially in algebra. A similar revolution occurred in physics one century later. In the 20th century theoretical physicists discovered that mathematics offers not only the classical structures of

[1] 'Arity' is not an English word, but it is common in mathematical jargon. The word is derived from the suffix *-ary*.

mathematical analysis, but many more, and they can be very useful in physics. This started with Einstein's use of the tensor calculus on manifolds in general relativity theory and Heisenberg's use of matrices in quantum mechanics.

Now, what happened with quantities? Modern mathematics has replaced this informal term by the concept of *function*. When describing some real phenomenon by two numbers x and y, where the number y is uniquely determined by the number x, we say that y *is a function of* x. This is formally written as

$$y = f(x).$$

We call x a *variable* and y the *value*, and f is a symbol by which we denote the function. The basic functions have names, such as '*square of*', '*sine*', '*exponential*', ..., and they are often expressed using special notation,

$$x^2, \quad \sin x, \quad e^x, \quad \ldots.$$

More generally, y may depend on several variables. Thus, in particular, operations are also functions. We use the word '*operation*' in situations when the function of several variables possesses some "nice" properties. This is the case of the operations of addition and multiplication on the natural numbers: they are commutative and associative (which means that the sums and products do not depend on the order in which they are computed).

If f is, say, a function defined on the real numbers, then it can be studied as the structure consisting of

1. the universe $\mathbb{R}$, which is the set of real numbers, and
2. the function f, as an operation.

It may seem that I am too fastidious about details when mentioning the universe. Isn't the structure already determined by the function? When we are stating theorems about the structure, it must be clear what the elements we are talking about are. We use the universe to determine the range of elements. It is a sort of a universe in which things concerning the structure take place.

I assume that the reader already knows most of these elementary concepts, but it is good to recall the terminology before discussing more difficult ones.

Ordered Sets

Let us now consider an example of a *class* of structures. The structures in the class are called *ordered sets*. This is probably the most ancient kind of structure. As soon as people started to organize their things they made lists by ordering the items that they considered. In fact this structure is imposed on essentially all data people use. We use language which is a sequence of words; written records are also sequences. So things are communicated in some order, whether we want to stress it or not. It is also interesting to note that the word '*ordering*' comes from '*order*' which also

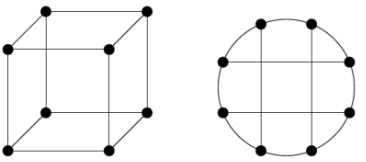


Fig. 1.1 Two drawings of the graph of the cube

means that things are properly organized, the opposite of disorder. And this is in fact the main purpose of mathematical structures, namely, to organize things, to introduce some order into our observations and data, so that we are able to manipulate them efficiently, physically and mentally.

The most obvious example of an ordered set is the set of natural numbers with the ordering relation $\leq$ that I mentioned above. Other familiar examples are the structure of the integers with ordering and the structure of the real numbers with ordering. These three structures are essentially different, not only because they have different universes, but because they have different structures, now using the word in the usual meaning. It does not matter how we represent the natural numbers, the integers and the real numbers, there will always be something different. The natural numbers are distinguished from the integers and the reals by the fact that they contain a smallest element. In the integers there are pairs of numbers such that there is no element between them, for example, there is no element strictly between 0 and 1. This is not true for the structure of reals: for every two elements, there exists an element between them (their mean is such an element).

Graphs

The word '*graph*' is used in two meanings. The traditional one is the diagram of a function, such as the dependence of the price of some commodity on time. It has a different meaning in the modern branch of mathematics that studies discrete structures, *the theory of graphs*. In this theory a *graph* consists of *points* and *arcs* that connect some points. This looks like a geometric concept, and it did originate in geometry, but it has more to do with topology than geometry. Consider for instance a cube. A cube determines in a natural way a graph, where we take the vertices of the cube as points and the edges of the cube as arcs, see Fig. 1.1. In fact, the standard terminology uses '*vertices*' and '*edges*' for all graphs. The reason why graph theory is so different from the classical fields of mathematics is that we completely abstract from the nature of vertices and edges and we only consider facts that depend on information about which vertices are connected and which are not. So if our cube is made of rubber and we twist it, the graph will be the same.

As another example of a graph, let us consider the graph of the flight connections of an airline. You can think of it as cities on a world map connected by arcs. On most such maps the arcs have little to do with the actual routes that an aircraft takes when flying between the two cities. An actual route must follow particular corridors, which is irrelevant for a passenger who only wonders whether there is a direct flight from city X to city Y.

Groups

If mathematicians voted for the most important class of structures, they would probably elect *groups*. The name is just a historical accident, so do not try to guess the meaning from normal use of the word. This concept is slightly more difficult, but worthwhile to learn. A group is a structure with one binary operation which in some sense behaves nicely. What this means precisely can be defined by postulating some simple laws that the operation must meet, which I will state shortly (page 10). Here I will only explain the concept in plain words.

The best way to imagine a group is to think of the elements of the group as *reversible actions* and the group operation as the composition of actions. As usual in mathematics, taking no action also counts as an action, called the *unit element*. Note that there is an important conceptual shift here: the actions themselves are elements, not the objects on which they act. Rubik's cube and similar toys are excellent examples. For Rubik's cube group, an action is, for example, turning the front face clockwise 90°, or turning the top by 180°. These are just some elementary actions. An action, however, may be more complex. For instance, we can compose the first one with the second one and this is also an action. We will get a different action, if we start with the second one and then apply the first one. The trick to solving this puzzle is to have several complex actions which do some particular things, such as turning two neighboring corners in opposite directions while keeping the rest the same. To transform a particular position into the original position is also an action. The goal is to compose this action from the elementary ones.

As you can imagine, the group of Rubik's cube is not a very simple one, it has $2^{12} \cdot 11! \cdot 3^8 \cdot 8!$ elements. There are groups which have infinitely many elements, but whose structure is simpler. Namely, one of the basic groups is the group of integers where the group operation is addition. To visualize it as a group of actions think of it as adding money to and withdrawing money from an account, say, starting with balance 0. Adding money is represented by positive integers, withdrawing by negative ones. This structure is the additive part of the structure of the integers that we considered earlier.

Groups are also essential in the study of *symmetries*. Consider a simple symmetric object, say an equilateral triangle A, B, C. We call a rigid action which transforms the triangle to itself a symmetry. There is a trivial symmetry corresponding to "no action", which we have, in fact, for any geometrical object. A nontrivial symmetry is the rotation where A goes to B, B goes to C and C goes to A. We can describe it by the list $A \to B$, $B \to C$, $C \to A$, or by saying that we rotate counterclockwise by 120°. We have one more rotation for 240°. Then we have another type of symmetry—we can flip the triangle along its axes of symmetry. For instance, flipping along the axis going through A can be described as interchanging B with C while A does not move. Another natural way of representing the same group is by permutations of three elements. The six permutations

$$(A, B, C), \quad (C, A, B), \quad (B, C, A), \quad (B, A, C), \quad (C, B, A), \quad (A, C, B)$$

are the elements of the group. They correspond to the identity, the transformation that does not move anything, and the symmetries denoted by a, b, c, d, e in Fig. 1.2.

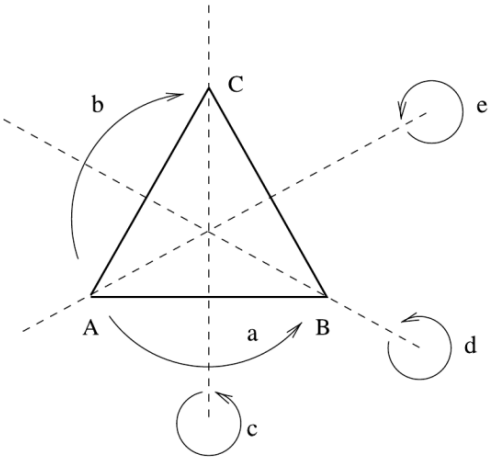


Fig. 1.2 The symmetries of an equilateral triangle

	1	a	b	c	d	e
1	1	a	b	c	d	e
a	a	b	1	e	c	d
b	b	1	a	d	e	c
c	c	d	e	1	a	b
d	d	e	c	b	1	a
e	e	c	d	a	b	1

Fig. 1.3 The multiplication table of the group of symmetries of an equilateral triangle. The unit element of the group is denoted by 1

The identity is the unit element of the group and is denoted by 1. The group operation is the composition of two permutations. For example, (C, A, B) is the transformation $A \mapsto B$, $B \mapsto C$, $C \mapsto A$ and (B, C, A) is the transformation $A \mapsto C$, $B \mapsto A$, $C \mapsto B$. Hence their composition is the identity (A, B, C).

These two representations use specific properties of the group. A general way by which we can represent any binary operation is the multiplication table. The multiplication table of the group of symmetries of an equilateral triangle is in Fig. 1.3.

Finally, we consider a way of representing groups that plays an important role in the study of finite groups—representations by matrices. In this way problems about finite groups can be translated into problems about matrices. Matrices form a very rich structure with a lot of interesting concepts and important theorems. The study of such representations is so useful that it forms a separate field called the *group representation theory*. Here is one such representation of the group of symmetries of an equilateral triangle.

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix},$$
$$\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

(For the definition of the matrix product see page 396.)

We have seen four representations of the same group. Each of them determines the structure of the group, but the group as an abstract object cannot be identified with any of them.

Why do we need various structures, why do we not just use numbers? The examples of graphs and groups show that there are practical situations which cannot be described only by numbers. We can think of structures as *models* of real and potentially realizable situations. Another possible view is that structures give us ways to classify objects. One useful way of classifying collections is to count the number of elements. We count our pieces of luggage to check that we have them all, which clearly does not ensure that we have all *our* luggage. But this test usually works. Numbers are not the only kind of structure used for such a classification. In particular groups are very good for this purpose. They are used in crystallography, to name a practical application. The symmetry group can be used to distinguish various objects, but it does not necessarily determine them completely.

In mathematics such a use of groups is almost ubiquitous. Returning to our example, we can distinguish the triangle from other geometrical objects by its group of symmetries. It is rather awkward here, as the triangle is much simpler than its symmetry group, but for larger objects it makes sense. In this case we would rather use the triangle to define the group.

One of the most beautiful pieces of mathematics, which I will consider in some detail in Chap. 4, is also based on this concept. This is the famous result that algebraic equations of degree 5 are not solvable using radicals. This means that there is no explicit formula using basic arithmetic operations and roots, expressing a solution to the equation in terms of the parameters. Here we have a natural scale given by the degree of the equation. But this gives us no clue why equations should be solvable up to degree 4, and unsolvable from degree 5 on. It was a great insight of Galois that one should assign groups to equations. The kind of groups that can be associated with equations of degree 5 and higher do not occur for equations of lower degree, and this gives the distinction between the solvable and the unsolvable.

Let me finally mention a result which belongs among the major achievements of twentieth century mathematics. The result is interesting also because it is a theorem with the longest proof ever written by mathematicians. It is called *the classification of simple finite groups*. The word "simple" is a little misleading; it is a technical term which specifies groups that are in some sense basic building blocks for constructing all finite groups. Naturally, having a description of them is very important, if we want to understand finite groups. The whole result is contained in a series of papers produced by a number of first rate mathematicians. The total number of pages amounts to several thousands. Some simple groups had to be described explicitly, the smallest one with $2^4 3^2 5 \cdot 11 = 7\,920$ elements, the largest one having $2^{46}3^{20}5^9 7^6 11^2 13^3 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71$ (approximately $8 \cdot 10^{53}$) elements, called *the Monster*. The enormous length of the proof and the huge size of the groups that it describes are certainly remarkable, but what is also interesting is a strange kind of irregularity. We are used to the fact that in mathematics things tend either to be very regular, or to look very random; if there is regularity with some exceptions then the exceptions are small. Here, in contrast, we have 26 exceptions that share very few common properties.

Types of Structures

In order to give a more precise meaning to the concept of a structure, we have to use more technical means of mathematics, some notation, and a few symbols. Formally, a structure is given by a list that consists of several sets. The first set is the universe, the set of objects of the structure. The remaining sets are relations, functions and operations on the universe. Let us denote by $\mathbb{N}$ the set of all natural numbers. Taking $\mathbb{N}$ as the universe, we can define various structures. The universe by no means determines the structure, however, there are some structures with this universe that we like more than the others. On the set $\mathbb{N}$ we usually take the following one $(\mathbb{N}; +, \cdot, \leq)$. To stress the special role of the universe, it is separated from the other sets by a semicolon. In this structure the binary relation $\leq$ is superfluous because we can define it from the operation $+$ (namely, $x \leq y$ if and only if there exists a z such that $x + z = y$), but we may have other reasons for keeping it. This structure has two binary operations and one binary relation—this information is what we call the *type of the structure*. Let $\mathbb{R}$ denote the set of all real numbers. We can define a structure of the same type as the natural numbers by taking $(\mathbb{R}; +, \cdot, \leq)$.

A different example is a directed graph. It is determined by a set of vertices and a general binary relation. Hence we can say that directed graphs are structures of the type consisting of one binary relation.

Structures with one binary operation also have a special name; they are called *magmas*, or *groupoids*.[2] Groups can be defined as those structures with one binary operation that satisfy the following axioms:

1. there exists a *unit element* (an element, usually denoted by 1, that satisfies $x \cdot 1 = 1 \cdot x = x$);
2. the operation is *associative* $((x \cdot y) \cdot z = x \cdot (y \cdot z)$ for all elements x, y, z);
3. every element has its *inverse* (the inverse of x is usually denoted by x^{-1} and satisfies $x \cdot x^{-1} = x^{-1} \cdot x = 1$).

Groupoids and groups belong to a large class of structures, called *algebraic structures*, or *universal algebras*, which are structures that only have functions and operations, but no relations.

All the structures that we have considered so far are *first-order structures*. There are structures that use more complex objects; such structures are called *second order, third order*, etc. In second order structures we have sets of subsets of the universe and relations between such subsets. This can be explained as follows. In a second order structure we have two universes, one consists of the elements that we want to study, the other consists of *sets of elements*, which we call *second order elements*. In a second order structure we also have relations and functions defined on second order elements. In order to imagine second order elements, think of subsets of the universe as properties of elements and sets of these subsets as properties of properties.

[2] Not to be confused with groupoids in category theory.

Example Let us take the color *navy blue* as an example of a property of real objects. Then we can take *dark colors* as an example of a property of properties that contains navy blue as an element.

If we attempt to define second order structures in full generality things become quite complicated. We can consider not only relations between subsets, but also between subsets and elements. Furthermore we should allow talking about properties of relations. But that is still not enough, since functions are also first-order objects, so we should allow relations between functions and so on. It is rather complicated, but it is only a technicality. The essence is that we have certain levels: the zero level are elements, the first level are relations and operations. In a second order structure we can define relations and operations on all objects of the first two levels.

The simplest example of a class of second order structures is the class of topological spaces. Topological space consists of a set of *points* A (this is the universe), and a set of subsets of A, called *open sets* that must satisfy some laws. For instance, the real numbers as topological space (called *the real line*) have the universe $\mathbb{R}$ and the open sets are subsets of reals which are unions of open intervals. (An open interval is the set of numbers between r and s *not* including the endpoints r, s.) The empty set is defined to be open too. Intuitively an open set is a set which does not contain a point on its border.

Let us proceed to the third order. This essentially means that we allow subsets of all subsets of the universe. There is no reason to stop at the third order, but already there it is hard to find nice examples. Let us take the first-order structure of reals $(\mathbb{R}; +, \cdot, \leq)$. Extend it to a second order structure by adding the set of all continuous functions of one variable, denoted by $\mathcal{F}$. Then we would like to consider the limits of the continuous functions, so we add a topology on $\mathcal{F}$ by taking the set of all open subsets of the functions, denoted by $\mathcal{X}$. This results in a third order structure $(\mathbb{R}; +, \cdot, \leq, \mathcal{F}, \mathcal{X})$. In this structure $+$, $\cdot$, $\leq$ are first-order concepts, $\mathcal{F}$ is second order and $\mathcal{X}$ is third order.

The types of structures are associated with certain set-theoretical constructions. The first one is the *Cartesian product* of sets. The Cartesian product of two sets X and Y is denoted by $X \times Y$ and it is the set of all pairs of elements (x, y) where x is an element of X and y is an element of Y. The reason for using $\times$ is that the size of the Cartesian product is the product of the sizes of the two sets; otherwise this set operation shares very little with the corresponding operation on numbers. Clearly, we can iterate this operation to get the product of a finite number of sets. The name '*Cartesian*' is in honor of the French mathematician and philosopher René Descartes (1596–1650), to whom we attribute the invention of coordinates and analytic geometry (although some analytic methods in geometry had already been used in ancient Greece). In modern terms it means that one dimensional space can be identified with the set of real numbers, and higher dimensional spaces are simply the products of copies of one dimensional space. His invention was probably the first step in the process of formalization of mathematical objects by mathematical structures. Mathematicians very often use pictures to visualize structures that they are thinking about. In the case of the Cartesian product $X \times Y$ the picture is the

familiar one with X drawn as the coordinate x, Y the coordinate y and the product being the points on the plane. The Cartesian product corresponds to relations, since we can define relations on a set A as subsets of the products of A with itself. Thus a subset of A is a unary relation, a subset of $A \times A$ is a binary relation, etc.

The second set operation is related to exponentiation and thus it is denoted by Y^X. It is the set of all functions f defined on X and having values in Y. Instead of saying that f is an element of Y^X, we prefer to express it by $f : X \to Y$. The Cartesian product enables us also to define functions with several variables, which we call operations. Thus, for example, a binary operation f on a set A is an element of $A^{A\times A}$, or using the other notation $f : A \times A \to A$ (which is read as 'f *maps* $A \times A$ *into* A'). For higher order structures, we need yet another set operation. Let us denote by $\mathcal{P}(A)$ the *power set* of the set A, the set of all subsets of A. Thus, for example, relations between second order elements are subsets of $\mathcal{P}(A) \times \mathcal{P}(A)$.

This notation can be used to define types of structures, but for this book we do not need a formal definition. Moreover, there are types of structures that do not quite fit into this schema. In classical parts of mathematics real numbers play a key role, thus many structures are somehow connected with them. Consider, for instance, a *real vector space*. It is a set of vectors A and a binary operation on A, usually denoted by $+$, satisfying certain axioms (namely $(A; +)$ is a commutative group). Furthermore, for every real number r, we can multiply any vector a of A by r and thus obtain another vector of A. This does not fit into the above schema, as the real numbers are not in $(A; +)$, they are *external*. In order to define this structure we have to take the union of the two structures—the real numbers and the group of vectors. The resulting object can be denoted by $(\mathbb{R}, A; +_{\mathbb{R}}, \cdot_{\mathbb{R}}, +_A, \cdot_{\mathbb{R},A})$. I have distinguished the two additions and two multiplications by subscripts, (to be more precise, we should write specifications such as $\cdot_{\mathbb{R},A} : \mathbb{R} \times A \to A$ which is multiplication of a vector by a real number, etc.). So we have to generalize the concept of a structure further and allow more than one universe. Also notice that in this particular example the roles of the two universes $\mathbb{R}$ and A are different: while A may vary arbitrarily, $\mathbb{R}$ is fixed for all real vector spaces.

For understanding the foundations of mathematics we do not have to study the whole ramified variety of structures. The most interesting phenomena can be observed in simple first-order structures.

Structures of Structures

In order to understand structures, it is important to realize that only *the form* is important, not the content. This means that the nature of the elements is irrelevant. The word '*structure*' denoting this concept is chosen appropriately, as we would like to identify two objects that have *the same structure*. Thus to get the whole point we only need to define what '*the same structure*' means. Intuitively it means that we can move one structure so that it completely coincides with the other. To move the structure means to move the points of the universe, the rest, the relations

and operations will move along because it is attached to the points. In mathematics structures do not live in space, so the transformations from one structure into another one are not continuous transitions (unless we incidentally study topology). Thus we only need to specify the beginning and the end of the movement. This is done by the concept of *mapping*. (A mapping and a function are the same things; we use different names only because of the different context.) Such a mapping should be *one-to-one*, which means that no two points are mapped onto one, and it should be *onto*, which means that every point of the universe of the second structure is an image of a point of the first structure. The mapping translates in a natural way relations and operations from one structure into the other. If the resulting image is identical with the second structure, we say that the structures are *isomorphic*. Isomorphism is the mathematical concept of having the same form. We often do not distinguish structures that are isomorphic and often say that '*two structures are the same, up to isomorphism*'.

To understand the above definition, think of the problem of comparing two pictures on a film in order to check if they are the same. First you have match them correctly. This means that you need some special points, in this case two are enough, which determine the correct position. If you put the pictures so that the points coincide, then it suffices to check if every line, every spot, etc. coincides.

The study of mappings of one structure into another is not restricted to isomorphisms. Given a class of structures one defines a more general concept, called *homomorphisms* or just *morphisms*, by using more general mappings. In particular, a homomorphism does not have to be a one-to-one mapping, hence it can map several elements on one. In this way some information about the structure on which it is defined may be lost in its image. Homomorphisms enable us to formalize the intuitive concept of similarity. The ability to recognize similarities is one of the most important features of human and animal thinking. Thus it is not surprising that in modern algebra many important results can be stated purely in terms of morphisms. A class of structures and morphisms is in some sense also a structure; it is called a *category*. We can study a class of structures by studying its category.

The Four Color Theorem

I will conclude this section with a couple of mathematical results that will be used as examples in the following chapters.

In 1852 an English mathematician, Francis Guthrie, conjectured that every map can be colored by four colors so that no two neighboring countries have the same color. This is, perhaps, the most famous problem in combinatorics, or at least it had been so until it was solved by Kenneth Appel and Wolfgang Haken in 1975 [5, 6]. The original statement talks about the topology of the plane, but it can be stated as a problem about certain graphs. Given a map, represent countries as vertices, say choose a point inside every country. Then connect by an arc every two vertices that come from neighboring countries. Then, instead of coloring countries, we will

color vertices. The restriction is that two vertices connected by an arc must have different colors. This simple transformation shows why graphs are so useful. We can transform a rather complicated statement to a simple combinatorial one.

This reduction alone does not suffice to translate the problem to graph theory. Not every graph corresponds to a map and it is very easy to construct a graph that is not colorable by four colors (take five vertices and connect every pair of vertices). Thus we need a characterization of graphs that come from maps; these graphs are called *planar*, as they come from maps in the plane. Such a purely combinatorial characterization was found by Kazimierz Kuratowski (1896–1980), a Polish set theorist and topologist; thus the problem has been reduced to finite combinatorics.

Whether or not every map can be colored by four colors has no bearing on the foundations of mathematics. What has is the way the problem was solved. Appel and Haken did not write down a proof of the conjecture, they only tested by computer that a proof exists. Following some earlier results they reduced the problem to a finite number of cases that were possible then to check by computer. Each particular case can be checked "by hand", but the total number of cases is too large for a human, even with the more recent improvements that have reduced the number of cases. This raised a discussion as to whether such proofs are legitimate. Certainly, such a proof conveys less to a mathematician than a usual proof. Typically, a proof is based on a small number of ideas that one can memorize so that it is possible to reconstruct the formal proof when needed. The experience of mathematicians with long proofs is that they are very likely wrong if such a set of basic ideas cannot be extracted from them. I agree with that, as this concerns proofs that are written by people and such proofs are never completely formal. Once the things are done formally, computers are much more precise than people. By now the validity of the theorem has been verified by running the programs on different machines and by using alternative proofs written by different people. What remains a mystery is why we do not have a 'normal' proof, a proof sufficiently short to be understood by people. As we will see later, there are theorems that do not have short proofs. But our mathematical tools are still very limited and thus we are not able to prove for such concrete theorems almost anything about the lengths of their proofs.

Note that there is a generalization of this problem to all orientable surfaces. Interestingly enough, the generalization had been solved for surfaces of all genera, except for the plane, without using a computer and before the original problem was solved.

The Four Color Theorem was not the first case in which an infinite problem was reduced to a finite number of cases. The famous Goldbach Conjecture, probably the oldest unsolved problem in mathematics, says that every even natural number greater than 2 can be expressed as the sum of two prime numbers. A weaker conjecture states that every odd number greater than 7 is a sum of three odd primes. In the 1930s, the Russian number theorist Ivan M. Vinogradov proved the weaker conjecture for all odd numbers starting from some large number N_0 [299]. Thus, theoretically, it sufficed to check all odd numbers less than N_0 in order to complete the proof. Unfortunately the number N_0 was so large (the original estimate was $e^{e^{e^{42}}} \approx 10^{10^{10^{17}}}$) that there was no chance to check the remaining cases by compu-

tation. This is still the case, in spite of the bound on N_0 being substantially reduced and in spite of the possibility to use contemporary powerful computers.[3]

More recently another famous problem has been solved using a computer in a similar way as in the Four Color Theorem. It is the Kepler Conjecture that the densest arrangement of equal balls is, in fact, the one that people have always been using. In 1998 S.P. Ferguson and T.C. Hales announced a proof of the conjecture [112]. It is based on a reduction proposed by L. Fejes Tóth in the 1950s. Since the computations used computer arithmetic, some doubts about the completeness of the proof still persist.

One may expect that computer aided proofs would be quite widespread by now, but it is not so. It turns out that there is a very narrow window where computers may help mathematicians. If ever a proof can be reduced to a finite number of cases, then, usually, either the problem can be solved completely by a mathematician, or the number of cases is so huge that it cannot be checked even by a computer.

Ramsey's Theorem

Frank P. Ramsey (1903–1930), a British mathematician and philosopher, proved a lemma that he needed in order to solve a certain problem in logic (the decidability of a certain part of first order logic) [235]. The lemma was later rediscovered by Paul Erdős and Gyorgy Szekerés working in a totally different field [69]; since then it became one of the main parts of combinatorics. Today this lemma is called *Ramsey's Theorem* and plays an equally important role also in logic and set theory. Therefore this theorem is very useful when we want to illustrate the connections between various fields of mathematics.

The essence of the theorem can easily be explained to anybody. Suppose that you have a symmetric binary relation on a finite set. Such a relation is also called an 'undirected graph', or just a 'graph'. Traditionally, for this theorem, one takes a random group of people and the relation of knowing each other as an example of a graph. The question that this theorem addresses is to what degree the relation can be chaotic, or put positively, must there be at least some order in any such relation? There are many ways to define the degree of order, but the extreme cases are clear: if every pair is connected by the relation, then clearly it has the maximal order; by the same token, if no two are connected it also has the maximal order. Ramsey's theorem, roughly speaking, says that total chaos is impossible. More precisely, we can always find a small subset of vertices where either all elements are connected in the graph, or all elements are not connected. For example, if there are at least 6 people in the group, there must be at least 3 that all know each other or all do not know each other (see Fig. 1.4). Similarly, if the group has at least 18 people, then

[3] The very recent result of T. Tao [289] that every odd integer greater than 1 can be represented as a sum of 5 or fewer primes uses the fact that the Goldbach conjecture has been verified by computation for all numbers up to $4 \cdot 10^{14}$.

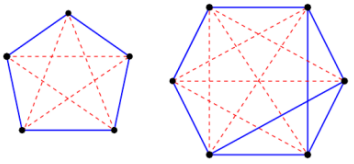


Fig. 1.4 Examples of a colorings of pairs of elements of a 5-element set and a 6-element set. The coloring of the 5-element set shows that $R(3) > 5$ because there is no 3-element monochromatic set. Since $R(3) = 6$, there must exist 3 points connected by lines of the same color in any coloring of a 6-element set. In the example there are two such triples, both form blue (solid line) triangles

there must be at least 4 that all know each other or all do not know each other. For 5, it suffices to have 46 people in the group.

In general, for every number n, we can find a number r, such that a graph on r vertices contains a subset of size n where either all elements are connected or no pair is. The *Ramsey number* $R(n)$ is defined as the least r such that every graph on r vertices contains a subset of size n where either all elements are connected or no pair is. The theorem says that this is a correct definition, such a number exists for every n.

The above examples can be stated as $R(3) \leq 6$, $R(4) \leq 18$ and $R(5) \leq 46$. In fact we know that $R(3) = 6$ and $R(4) = 18$, but we do not know the exact value of $R(5)$. We only know that $43 \leq R(5) \leq 46$. This is remarkable since to determine the value of $R(5)$ is a finite problem, one has "only" to check all the graphs on 43, 44 and 45 vertices. Testing a single graph is not so difficult (though it is quite a time consuming task—there are more than one million subsets of size 5 of a set of size 45), the problem is that there are too many graphs to be tested.

The classical infinite version of the theorem states that for every graph on the natural numbers, there is an *infinite* subset of the natural numbers such that either all elements in the subset are connected, or no pair is. A remarkable fact is that the finite version of the theorem can be derived from the infinite one. The advantage of such a proof of the finite version is that we do not have to bother with counting. The disadvantage, the price for the simplification, is that we do not get any bounds on the Ramsey numbers.

Notes

1. *General structures.* A general structure is defined by an *echelon construction.* The construction starts with base sets (universes) $A_1, \ldots, A_n$. Then we can apply operations of the Cartesian product $\times$, the power set operation $\mathcal{P}$ and the operation of taking the set of all functions from one set into another set B^A. This means that we successively produce sets such that every new set is obtained from $A_1, \ldots, A_n$ and the already produced ones by applying one of the three operations. We identify products of several sets if the order of the sets in them is the same; for instance, we do not distinguish between $(B_1 \times B_2) \times B_3$ and $B_1 \times (B_2 \times B_3)$. Thus we can omit parentheses in the products. A structure is a sequence of the form $(A_1, \ldots, A_n; B_1, \ldots, B_m)$ where $B_1, \ldots, B_m$ are subsets of the sets obtained by the echelon construction or mappings between them.

For example, our third order structure considered above $(\mathbb{R}; +, \cdot, \leq, \mathcal{F}, \mathcal{X})$ is produced from the sequence $\mathbb{R}, \mathbb{R} \times \mathbb{R}, \mathbb{R}^{\mathbb{R}}, \mathcal{P}(\mathbb{R}^{\mathbb{R}})$, where the operations $+$ and $\cdot$ are mappings from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$, the relation $\leq$ is a subset of $\mathbb{R} \times \mathbb{R}$, the set $\mathcal{F}$ is a subset of $\mathbb{R}^{\mathbb{R}}$ (the set of all real functions) and $\mathcal{X}$ is a subset of $\mathcal{P}(\mathbb{R}^{\mathbb{R}})$ (the set of all subsets of real functions).

In a precise definition of a structure we have to associate a *type* to each of the sets. In particular, in first-order structures this means determining if it is a relations or an operation and then its arity. first-order structures are those where neither of the operations $\mathcal{P}(X)$, X^Y is used. In second order structures these operations can be applied, but not iterated, in third order structures they may be iterated once etc.

It is possible to simplify the matter by considering operations and functions as a special kind of relations (for example, a binary operation is a ternary relation). However, quite often, it is an advantage to have operations as a primitive concept.

2. *Higher type functionals.* General structures can use all three operations: Cartesian product, power set operation, and the operation of taking all functions from a given structure to another one. We can get, however, very interesting objects by considering only the last one. This means to concentrate on functions and not to use relations and sets. We start with elements as the basic type of objects; the set of elements is the universe of the structure. The next type consists of functions. A function is a mapping from the universe to itself. Then we can define functionals, which are mappings that map functions to elements. We can use also mappings that assign functions to elements and mappings that assign functions to functions and so on. We will simply call all such objects *functionals* and distinguish them by their *types*. As the types do not have linear structure, we cannot use numbers for denoting types, we need to introduce special notation. The type of elements will be denoted by o ('o' for 'objects'). Given types τ and σ, the type of functionals that map objects of type τ to objects of type σ will be denoted by $\tau \to \sigma$. Thus functions are functionals of type $o \to o$, the lowest level functionals are $(o \to o) \to o$, etc. Note that functionals of type $o \to (o \to o)$ can be identified with binary operations, that is, functions of two variables.

Now we will consider some important classes of structures.

3. *Ordered sets.* An ordered set is a structure with one universe and one binary relation on it denoted usually by $\leq$ (ambiguously, because the relations in different structures are different). By an ordered set we usually mean a *partially* ordered set which means that there may be incomparable elements. The axioms of partially ordered sets are

   a. $x \leq x$—reflexivity,
   b. $x \leq y$ and $y \leq z$ implies $x \leq z$—transitivity,
   c. $x \leq y$ and $y \leq x$ implies $x = y$—antisymmetry.

   The ordered sets where every two elements are comparable are called *linear orderings*; they satisfy also

   d. $x \leq y$ or $y \leq x$.

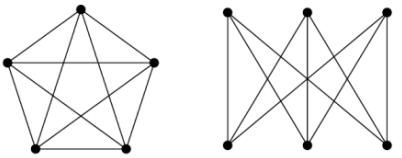


Fig. 1.5 The graphs K_5 and $K_{3,3}$

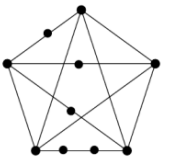


Fig. 1.6 A subdivision of K_5

4. *Graphs.* A general graph is a binary relation on the set of vertices. It is called a *directed graph* because we may have a directed edge (u, v) without having the opposite (v, u). Edges of the form (u, u) are called loops. For instance, partially ordered sets are a subclass of graphs. Graphs in the narrow sense are symmetric, which means (u, v) is an edge if and only if (v, u) is, and loops are prohibited. We denote by (u, v) an ordered pair. For symmetric graphs, we can take unordered pairs which are two-element sets. They are denoted by $\{u, v\}$. (Sometimes a more general concept is considered where there can be more than one arc between two vertices.)

   Kuratowski's characterization of *planar graphs* is based on forbidden subgraphs. He found a set of graphs such that planar graphs are exactly those that do not contain a graph from the set. The set of forbidden graphs consists of the two graphs in Fig. 1.5 *and all their subdivisions.* A *subdivision* of a graph is obtained by refining edges into paths; pictorially, we put several dots on an edge (see Fig. 1.6).
5. *Groups.* A group is usually considered as a structure with one binary operation, one unary operation (a function) and a constant. These are called *multiplication, the inverse element function* and *the unit element.* Thus we write $(G; \cdot, {}^{-1}, 1)$. The inverse element and the unit is definable from multiplication, but having these two additional primitives enables us to write axioms as equations:

   a. $1 \cdot x = x \cdot 1 = 1$,
   b. $x \cdot x^{-1} = x^{-1} \cdot x = 1$,
   c. $x \cdot (y \cdot x) = (x \cdot y) \cdot x$.

   Note that we do not postulate commutativity. You can check that the symmetry group of a triangle is not commutative. The groups where the commutative law $x \cdot y = y \cdot x$ holds are called *commutative* or *Abelian* groups. For commutative groups, one often uses additive notation, thus instead of calling the binary operation '*multiplication*' we call it '*addition*'.

   We will now define some concepts needed to explain the meaning of simple groups.

A *group homomorphism* is a mapping f of a group G_1 into a group G_2 which preserves the operation. This simply means

$$f(x \cdot y) = f(x) \cdot f(y),$$

for every two elements of G_1. (As customary, we use the same dot for both groups, though these are different operations in general.) This condition implies that f preserves 1 and inverses. The *image* of the group G_1 under f, denoted by $Im(f)$, is the set of all elements of G_2 to which some element of G_1 is mapped. This set is, as you can easily check, closed under the operations of multiplication and inverse and contains the unit element. So it is a *subgroup* of G_2. The *kernel* of f, denoted by $Ker(f)$, is the set of elements of G_1 which are mapped onto 1. It is also a subgroup. $Ker(f)$ is the trivial one element subgroup of G_1 if and only if f is a one-to-one mapping, in which case G_1 is isomorphic to $Im(g)$. On the other hand, if the image is the trivial one element subgroup of G_2, then G_1 is equal to $Ker(f)$.

The two groups $Ker(f)$ and $Im(f)$ do not give full information on G_1 in general, but the structure of G_1 can be very well understood if we know them.[4] Take an element g of G_1 which is not mapped to 1. Then the set of all elements which have the same image, namely $f(g)$, is the set of elements of the form $g \cdot h$, where h runs through the elements of $Ker(f)$. Thus G_1 can be decomposed into *cosets* which have a structure similar to $Ker(f)$, every coset corresponding to an element of $Im(f)$. Also, if we know that g_1 and g_2 are from cosets determined by h_1 and h_2, that is, $f(g_1) = h_1$ and $f(g_2) = h_2$, then the element $g_1 \cdot g_2$ is from the coset determined by $h_1 \cdot h_2$.

As an example consider Rubik's cube. We have Rubik's group, let us denote it by G_1, which consists of the transformations of the whole Rubik cube. Note that we consider only transformations that can be physically realized without breaking the cube into pieces (there would be 12 times more transformations, if we allowed decomposing and reassembling the cube). Further we can consider the same transformations, but look only at the small cubes at the edges, which means that we identify the transformations which act in the same way on edges. Let us denote this group by G_2. Then we have a mapping, in fact a group homomorphism $f : G_1 \to G_2$, given by '*forgetting the vertices of Rubik's cube*'. In this case $Im(f) = G_2$, the group of transformations on edges. $Ker(f)$ is the group of the transformations which are mapped on the identity element of G_2 which are transformations which move only the small cubes on vertices while preserving the edges. This decomposition is actually used by Rubik cube solvers.

Now comes the crucial definition. A group G is called *simple*, if for every homomorphism f from G to another group, either f is one-to-one, or f maps G to the trivial one element group. By the remarks above, this is equivalent

[4] In order to get full information about G_1, we need the groups $Ker(f)$ and $Im(f)$ and, furthermore, a homomorphism from $Im(f)$ into the group of automorphisms of $Ker(f)$. It would take us too far afield to explain this relation.

to the condition that either G is equal to $Ker(f)$ and $Im(f)$ is a one element group or G is isomorphic to $Im(f)$ and $Ker(f)$ is a one element group. In other words, we cannot decompose a simple group into smaller groups using a group homomorphism, which makes the study of simple groups more difficult.

6. *Rings and fields.* A *ring* is a structure with two binary operations and one constant on one universe. The operations are usually denoted by $+$ and $\cdot$ (the $\cdot$ is almost always omitted when writing terms), the constant is denoted by 0. The axioms of the rings express that for a given ring $(A; +, \cdot, 0)$ the structure $(A; +, 0)$ is a commutative group, $\cdot$ satisfies the associative law and the two operations are connected by the distributive laws

$$x \cdot (y + z) = x \cdot y + x \cdot z \quad \text{and} \quad (x + y) \cdot z = x \cdot z + y \cdot z.$$

   A ring $(A; +, \cdot, 0)$ is a *field*, if the nonzero elements with the operations $\cdot$ form a group, which means that there is a multiplicative unit element and nonzero elements have multiplicative inverses. The most familiar fields are rational numbers, the real numbers, and the complex numbers. Integers form only a ring.
7. *Universal algebra.* This is the field of mathematics that studies algebraic structures in general, without assuming any special properties of them. The aim is to generalize theorems that are known for various special classes of algebras such as groups, semigroups, rings, fields, lattices, Boolean algebras, etc. Except for fields, these classes can be defined by equations, as we have done for groups and implicitly for rings (the problem with fields is that 0 has to be treated separately). So it is natural to study the equations valid in classes of such structures, the equational theories. Furthermore, there is a natural concept of homomorphisms for universal algebras, namely, as in groups, the mappings which preserve operations.
8. *Topological spaces.* A topological space is a structure of the form $(A; \mathcal{X})$ with $\mathcal{X} \subseteq \mathcal{P}(A)$ where $\mathcal{X}$ contains $\emptyset$ and A and it is closed under arbitrary intersections and finite unions. The sets in $\mathcal{X}$ are called *open sets*. Note that it is a second order structure. Moreover, the condition that the open sets are closed under intersection is even of a higher order (namely the third order) since it talks about arbitrary subsets of sets of subsets of A.
9. *Special structures.* There are several structures that play a special role in mathematics. The reasons for their exceptional status are in that they appear in many problems, or they are in some sense universal, or it is simply the tradition. Examples of such structures are: the ring of integers, the field of real numbers, the ordering of rational numbers, the topology of real numbers, etc. The classes of structures were often defined by choosing some general properties of these special structures.
10. *Real vector spaces.* A real vector space is a structure of the form $(\mathbb{R}, A; +_{\mathbb{R}}, \cdot_{\mathbb{R}}, +_A, \cdot_{\mathbb{R},A})$. It has two types of objects: the real numbers $\mathbb{R}$, called scalars, and vectors A. On the real numbers there are the two basic arithmetical operations $+_{\mathbb{R}}, \cdot_{\mathbb{R}}$; further, there is a binary operation $+_A$ on vectors, called addition, and an operation of multiplying a vector by a scalar $\cdot_{\mathbb{R},A}$. In a real vector space

the real numbers are determined uniquely, so one has to postulate only axioms which determine the structure of vectors and bind it with real numbers. The axioms of the real vector spaces say that $(A; +_A)$ is a commutative group and for every $r, s \in \mathbb{R}$ and $a, b \in A$,

a. $1 \cdot a = a$,
b. $(r + s) \cdot a = r \cdot a + s \cdot a$,
c. $r \cdot (a + b) = r \cdot a + r \cdot b$,
d. $(r \cdot s) \cdot a = r \cdot (s \cdot a)$.

Here 1 stands for the real number 1, (in order to be able to write these axioms as equations, we should include 1 into the definition of the structure as a special constant). I have omitted the subscripts since it is clear from the context which operation is meant. Thus real vector spaces are described by equations, but, clearly, we cannot derive all of their properties from these equations since they say nothing about real numbers. They determine this concept assuming that we know what the real numbers are.

11. *Finite automata.* This is one of the basic concepts of the theory of computation. A finite automaton is a structure of the form $(A, B, Q; q_0, \delta, \sigma)$ where $q_0 \in Q$ is a constant, and $\delta : A \times Q \to Q$, $\sigma : A \times Q \to B$ are operations. The interpretation is that A is the input alphabet, B is the output alphabet, Q is the set of the states of the automaton and q_0 is the initial state. Such an automaton works in discrete steps. In every step it receives a letter a from the input alphabet. It reacts by changing its state from its current state q to the state $\delta(a, q)$ and it writes the output $\sigma(a, q)$. This concept is not only similar to algebras, but it actually is amenable to algebraic methods.
12. *Boolean functions.* A Boolean function is mapping of the form $f : \{0, 1\}^m \to \{0, 1\}^n$, in words, a function that maps strings of zeros and ones to strings of zeros and ones. This is the main structure studied in theoretical computer science. It is the prototype of finite functions, functions with finite domain and finite range.
13. *Boolean algebras.* A Boolean algebra is an algebra with three operations meet $\wedge$, join $\vee$ and complement $'$ and two constants 0 and 1. It satisfies the laws of propositional logic, which can be expressed by the following axioms.

a. commutative and associative laws for $\wedge$ and $\vee$;
b. both distributive laws for $\wedge$ and $\vee$: $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ and $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$;
c. $x \wedge x' = 1$, $x \vee x' = 0$, $x \wedge 1 = x$, $x \vee 0 = x$;
d. De Morgan's laws: $(x \wedge y)' = x' \vee y'$, $(x \vee y)' = x' \wedge y'$.

A Boolean algebra has an ordering which we will denote by $x \leq y$ and which is defined by $x = x \wedge y$; 1 is the top element and 0 is the bottom element.

One can show that this theory is in a certain sense the algebraic theory of the two element set $\{0, 1\}$. Namely, one can define every Boolean function $f : \{0, 1\}^n \to \{0, 1\}$ using the basic operations of the two element Boolean algebra, and all equalities between terms formed from Boolean functions are provable in this theory. However, this theory has also larger finite and infinite models.

14. *Manifolds.* Sometimes it is rather difficult to present a concept in question as a structure, sometimes we even need to generalize further the concept of a structure. A *real manifold* is a topological space where we have locally real coordinates. Intuitively, it is a patchwork assembled from pieces of an n-dimensional Euclidean space. This is formalized by the concept of an *atlas* of *charts*. The charts are homeomorphisms (= mappings preserving the topology) from open subsets of the manifold into a d dimensional real space. A topological space with an atlas is a nice structure, but *it is not a manifold*; it is only one of the infinitely many ways to determine it. To resolve this purely formal problem, one defines a manifold as a topological space with *all admissible charts* for an atlas, charts that are compatible with the charts in the atlas. The atlas itself is not a part of the structure.

    Is it necessary to use such awkward definitions? The point is that there is no problem with an intuitive definition, if we work with typical objects. Once we need to consider extremal cases or when we need to generalize theorems as much as possible, we are in trouble without a rigorous definition.
15. *Categories.* When studying structures we are interested only in their form, but when we prove their existence we need to construct them. This amounts to choose particular elements for the universe and defining which are in the particular relation or what a particular operation does with them. Thus we define the rational numbers as pairs of natural numbers, the real numbers as certain sets of rational numbers, etc. Then, of course, we can forget what the actual elements of the universe are. We are interested only in the shape, but we have to use material to realize it. Can we not avoid the ad hoc part of choosing material and, instead, get the shape directly?

    The theory of categories is at least a partial remedy. In this theory, instead of individual structures, we study *a category*, which is the overarching structure of a *class* of structures. Thus a category is a large structure whose elements are some structures in the usual sense.

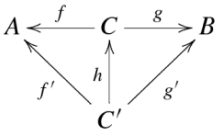
    In a category we have two kinds of basic elements: *objects* and *morphisms* between objects. For every object A, there is an identity morphism i_A from A to A. Given a morphism f from A to B and a morphism g from B to C we can form their composition fg which goes from A to C, otherwise the composition is not defined. The axioms are

    a. $i_A f = f$ and $g i_A = g$, whenever defined,
    b. $f(gh) = f(gh)$, whenever defined.

    (Note that it is not an algebraic structure in the strict sense since the composition is not defined for every pair of morphisms.)

    A typical category is the class of all groups with homomorphisms. In general, the intended meaning of morphisms is some kind of mappings, but there are categories where morphisms are not represented by mappings. For example, any partially ordered set is a category, if we interpret $x \leq y$ as a unique morphism from x to y. A morphism f from A to B is defined as being an *isomorphism*, if there exists a morphism g from B to A such that $fg = i_A$ and $gf = i_B$.

Here is an example of an important concept that can be defined purely using the language of categories. We define that an object C is the product of the objects A and B, if there are morphisms f, g from C to A, B respectively such that for any other object C' and morphisms f', g' from C' to A, B, there exists a *unique* morphism h from C' to C such that $hf = f'$ and $hg = g'$. In the theory of categories, such definitions are often presented in the form of *commutative diagrams*. The 'commutativity' means that if we compose morphism along two paths of arrows that start and end in the same objects, then the resulting two morphisms are equal. Below is a commutative diagram for the definition of the product.



The product does not have to always exist, but when it does then it is unique in the sense that any two such objects are isomorphic. In the category of all sets the product is just the Cartesian product of the two sets. Thus we are able to define it without mentioning pairs! Also in the categories of algebras defined by equations the categorical product is the naturally defined product. It is instructive to realize that the product of two elements in a partially ordered set interpreted as a category is their greatest lower bound.

Categories behave like a special kind of structure except that their universe can be too big to be considered as a set. For instance, all groups do not form a set but a proper class (I will explain this concept later). There is a natural concept of morphisms between categories, they are called *functors*. Functors preserve identity morphisms and the operation of composition of morphisms. The operations used in the construction of structures (product, power-set, the set of all functions from one set to another) can be extended to functors.

For instance, the power-set operation $\mathcal{P}$ can be extended to a functor from the category of sets into itself. As $\mathcal{P}$ is defined for sets, the objects of the category, we only need to define $\mathcal{P}(f)$ for morphisms, which are mappings between sets. Suppose $f : A \to B$, then $\mathcal{P}(f) : \mathcal{P}(A) \to \mathcal{P}(B)$ is defined by putting $\mathcal{P}(f)(X)$ equal to the image of X under the mapping f.

16. *Proof of Ramsey's theorem.* Since the role of edges and non-edges in the theorem is symmetric, one uses colorings of unordered pairs instead of graphs. Assume a coloring of pairs of the set $\{1, \ldots, r\}$ by two colors is given. We want to construct a *monochromatic* subset, which is a subset in which every pair has the same color. We start with the pairs of the form $(1, x)$ with $1 < x \leq r$ and consider their colors. One of the colors has to occur at least $(r-1)/2$-times. We take the subset X_1 of $\{1, \ldots, r\}$ of those elements $x > 1$ for which the pair $(1, x)$ has the prevailing color. (If both colors occur the same number of times, then it does not matter which color we choose.) Let v_1 be the least element of X_1. In the next step we consider all pairs (v_1, x) with $v_1 < x$ and $x \in X_1$. There must be a color that occurs at least $((r-1)/2 - 1)/2$-times. Note that

this color may be different from the color that we selected in the previous step. Take the subset X_2 of those elements $x > v_1$ of X_1 such that the pair (v_1, x) has the prevailing color. We can continue this process until the sets X_i shrink to an empty set. Thus we have chosen elements $1, v_1, \ldots, v_n$ for some n that is approximately the binary logarithm of r. Now look at the colors of the pairs of elements of the set $\{1, v_1, \ldots, v_n\}$. As noted above, both colors can occur, but the coloring has a very special property: the color of every pair (x, y), with $x < y$, depends only on the smaller element x. In the Ramsey theory jargon this property is called '*combed to the right*'. For $x = 1, v_1, \ldots, v_{n-1}$, associate this color with x. Then, again, take a color that occurs at least $(n-1)/2$ times and select the elements associated to this color (and add v_n if you wish to have one more element). Thus you get a monochromatic subset. The size of the subset is approximately $1/2$ of the logarithm of r, hence, if r goes to infinity, then the size of such a subset also goes to infinity. This finishes the proof of the finite version.

The proof of the *infinite version* is almost the same. The only changes are that we take the coloring of pairs of the infinite set $\{1, 2, 3, \ldots\}$, and we do not talk about the prevailing color but a color that occurs infinitely many times. It may happen that both colors considered at some stage occur infinitely many times. In such a case we can choose any of them.

The proof of the finite Ramsey theorem is quite easy, so it is not a matter of economy to deduce it from the infinite version. The reason for doing it is to show how one can get a finite statement from an infinite one. Later we will see nontrivial applications of this proof. The proof is based on the following result. A *tree* is a connected graph without cycles. A *rooted tree* is a tree with a distinguished vertex, called *the root*. (Examples of a rooted trees are in Fig. 4.4 on page 325.) We consider infinite trees. A tree is called *finitely branching*, if the degree of every vertex, that is the number of edges incident with the vertex, is finite. An *infinite branch* is an infinite chain starting in the root (a sequence of pairwise distinct vertices starting in the root in which every two consecutive vertices are connected).

König's Lemma *Every infinite finitely branching tree has an infinite branch.*

Proof Start the construction of an infinite branch with the root of the tree. Since the degree of the root is finite, the subtrees that are connected to it cannot be all finite. So choose as the next vertex the root of an infinite subtree connected to the root. Apply the same argument recursively to subtrees. Thus we obtain an infinite branch.

To derive the finite Ramsey theorem from the infinite one, assume that the finite Ramsey theorem fails. This means that for some n the following holds. For every r, there is a coloring of pairs of elements of $\{1, \ldots, r\}$ by two colors such that no subset of size n is monochromatic. Having these colorings we would like to construct a coloring of the infinite set $\{1, 2, 3, \ldots\}$ that has no monochromatic subsets of size n, hence no infinite monochromatic subsets. Thus we have colorings of pairs on arbitrary large segments that satisfy some property and we would like to construct a coloring of all pairs. König's lemma

is a perfect tool for it. The vertices of the tree will be the colorings for which there is no monochromatic subset of size n. This will include the trivial empty coloring of the empty set of pairs of elements of the one element set $\{1\}$, which will be the root. Two colorings will be connected by an edge in the tree if, for some r, one is on $\{1, \ldots, r\}$, the other is on $\{1, \ldots, r+1\}$ and the second one is an extension of the first one to the larger set. It is quite straightforward to prove that this is a tree. Assuming the finite Ramsey theorem fails, it is an infinite tree. It is finitely branching since there is only a finite number of colorings on a finite set. By König's lemma, there is an infinite branch. This is a sequence of gradually extending colorings. We take their union as the coloring of all pairs. Clearly, if there were a monochromatic subset of size n, it would already be in one of the colorings. □

The above theorems are in fact only special cases of a more general result proved by Ramsey. The general result concern not only pairs, but also k element subsets for every fixed finite k. Furthermore, but less important, one can consider an arbitrary fixed finite number of colors. Here is the general form of the infinite theorem.

The Ramsey Theorem *For every k, m positive integers, and for every coloring of k-element subsets of the natural numbers by m colors, there exists an infinite monochromatic subset X (a subset in which all k-element subsets have the same color).*

The proof of the theorem goes by induction on k. One reduces the case of $k+1$ to the case k by a 'combing' argument similar to the one that we have used above.

1.2 Everything Is a Set

No one shall be able to drive us from the paradise that Cantor created for us.

David Hilbert, *On the infinite*

The concept of a set was introduced in mathematics by Georg Cantor (1845–1918). Similar ideas had been considered before him, in particular in philosophy and logic. Cantor was not a logician and he arrived at the concept working on mathematical problems. Thus he was the first one to realize that sets are not only a good methodological tool but they are also useful for obtaining mathematical results. In spite of his success in proving results using set theory, mathematicians of his time, except for a few, ignored Cantor's results. His first paper on this subject was published in 1878, but it took several decades before set theory was accepted by the mathematical community.

Cantor's first major result in mathematics concerned functions on the real numbers. He proved a theorem about trigonometric series, which are series of sine and cosine functions. In physics this theory is used to decompose sounds into pure sounds. He proved that such a decomposition is unique, if the series of functions

converged at every point. Then he realized that he could weaken the assumption so that the series converged everywhere except for finitely many points. But that was still not the best possible result. He found out that it is also possible to allow certain infinite sets. He went on to describe more and more complex sets. For that he needed to make the concept of a subset of a real line precise. Furthermore, he needed to apply a certain operation (derivation of a set) transfinite number of times, which led to his discovery of transfinite ordinal numbers. (I will explain these concepts in Chap. 3). He realized that set theory was a new, completely unexplored field of research and thus he devoted the rest of his scientific career to this subject.

Georg Cantor
Courtesy of Universität Hamburg

I will not follow Cantor's development of the theory. Instead I will introduce this concept assuming the reader knows almost nothing about it. The concept of a set seems familiar: a set is an arbitrary collection of arbitrary elements. However in order to understand the way this concept is used in mathematics, we have to describe it more precisely. I will state several basic principles that determine the concept of the set. I will state them in plain language so that they are easily understandable. When stated formally they are postulates of set theory, but to obtain an axiomatization that is possible to use in mathematical practice, one needs more postulates. I will present the remaining ones in Chap. 3, which will be devoted to set theory.

The first basic principle is the following one:

The Principle of Extensionality *A set is uniquely determined by the elements that it contains. Thus two sets are considered equal if they have the same elements.*

This means that it does not matter how we specify the elements that belong to a set, what matters is only which elements are selected. We can determine the elements by some property, say described by a formula, or by an algorithm that decides if an element belongs to the set or not, or just list the elements of the set, etc. A particular definition is only a way of specifying the set; the set is simply the collection of elements that satisfy the definition.

This principle is not so easy to accept, unless you already have some experience with set theory since in natural language we tend to identify the sets with their definitions. Suppose I say *the red things in this room*.[5] What I mean is *the set* of the red things in this room. The set consists of a lamp, a pen, a control light on my computer and some books on the shelves. Would you say that the first definition and the list of the objects define the same thing? The principle of extensionality says that it is so: there is an abstract thing, a set, which is defined by both the condition of being red and the list. The reason why in a natural language we do not interpret extensionally definitions in the manner mentioned above is that the

[5] I assume that we agree on what red things are.

same word construction is used in different occasions by different people at different times, etc., thus the actual set of the things it defines varies depending on context.

Let us consider a different example: *the set of people over 60 in the sample of patients that we have treated.* If you write a report on your medical research, you want to give as much information as possible, so you will certainly include the definition of the set. On the other hand, if you process your data on a computer it is a different task. The information on the patients will be most likely stored with their birthdays, thus you can write a program to determine who is over 60. But you can also simply list the patients over 60; the computer does not care and the output will be the same.

At this point it is worthwhile to digress to history. The principle of extensionality is, clearly, the most distinctive feature of the set theoretical approach started by Cantor. It is interesting to compare it with other ideas which appeared or became popular at about the same time. The philosophical doctrine of *logical positivism* is a modern version of positivism developed by the Vienna circle in the 1920s and 30s. According to this doctrine the only meaningful statements are those that talk about observable events. The concept of the *black box* is much more recent, but it can be used to explain the essence of logical positivism. By a black box, we mean a device which we can observe only from outside and we cannot open it in order to see how it functions. Positivism tells us that if we cannot open the box, any theories about what is going on inside are meaningless. We can only make theories about *how it behaves.* A mathematical description of the behavior of a black box is a function. Such a function f tells us that, given an input x to the box, we will get $f(x)$, the value of f on x, as the output. Thus the black box is described by a structure consisting of the set of possible inputs, the set of possible outputs and the function f. A positivist would interpret the concept of function in the same way as contemporary mathematicians: the function is just the set of pairs of input x and the corresponding output $f(x)$. Particular definitions, or algorithms are only auxiliary means of determining the function. Extensional interpretation of functions was probably well established in mathematics before logical positivism appeared, but it cannot be a mere coincidence that similar ideas appeared in different fields of science in a relatively short period of time.

How is this related to sets? Think of a set A as a black box. For a given element x the set A tells us whether or not x belongs to it. If we have another set B which behaves in the same way, then B is equal to A. This is exactly the principle of extensionality.

Extensionality, in a broader sense, is a fundamental principle of all mathematics. It does not concern only set theory because what we call '*abstraction*' can often be explained by extensionality. When we count we only use properties of numbers and we forget about the concrete collections that correspond to the numbers. A number, such as 5, is the same 5 whether it is represented by five apples or five oranges. This concerns every mathematical structure—we abstract from the nature of elements, we only use the shape, the structure that the elements form. In this extensional understanding of structures, a relation is merely a set of pairs, it is *not* the definition that determines which pairs are related. Similarly a function is also a set of pairs, it is *not* a mechanism that produces $f(x)$ from x. It is worthwhile to restate the principle for relations and functions.

The Principle of Extensionality for Relations and Functions *A relation is uniquely determined by the pairs of elements that are related, a function f is uniquely determined by pairs x, an argument, and $f(x)$, the value. Thus two relations are considered equal if they relate the same pairs of elements and two functions are equal if they give the same values for the same arguments.*

We can state such a version of extensionality for every mathematical structure. In set theory the Principle of Extensionality is never postulated for relations, functions or other structures because in set theory all structures are sets, hence the principle for them follows from the principle for sets.

The second basic principle is:

The Principle of Comprehension *For any* reasonable *property, there exists a set which contains exactly those elements that satisfy the property.*

There is a modifier '*reasonable*' in this statement that makes it rather vague. I will explain shortly (in the next section) why we have to use it. For now, let us ignore it. In any case, the meaning is very general: for instance, a property can be determined by an algorithmic procedure, or simply by a list of elements, etc.

This principle seems intuitively completely clear. If we can distinguish some elements, we can name the '*collection*' of these elements. This means that we have a '*name*', or a '*concept*', so we can treat it as an entity. One of the ideas behind sets is to simplify our language by sticking to a single word '*set*' instead of '*a collection*', '*a concept*', '*a class*', etc. (though sometimes we will distinguish between '*classes*' and '*sets*').

Now we can make our first deduction and prove that there exists at least one set, namely an empty set. For that, we simply need a property which is never satisfied (such properties are abundant) and apply the comprehension principle. The extensionality principle, on the other hand, tells us, that the empty set is unique.

To get a good picture of how sets are used we have to stress one more fact, which, perhaps, should not be called a principle, but which is still very important.

The Principle of Being an Element *A set can be an element of another set.*

More formally this means that we do not distinguish elements and sets, thus we have only one *type* of object. Why is it important? If we could not form sets of sets, set theory would be just a kind of descriptive language. We need to produce many different elements to be able to combine them into various structures. If we could not use sets, we would have to postulate the existence of elements somehow. Furthermore, in modern mathematics there are powerful methods that are based on constructions that use the possibility of forming a set from other sets as elements. We build new structures by using parts of, or just whole structures as the elements of the new structures.

Let us note that it follows from the above principles that every set *is* an element of another set. Namely, a set x is an element of a set with the unique element x.

We denote the set with elements $a_1, a_2, \ldots, a_n$ by $\{a_1, a_2, \ldots, a_n\}$. Hence the set with the unique element x is $\{x\}$. (Note that these two sets are in general different: while $\{x\}$ has only one element, x can be empty or have more than one element![6]) In particular this enables us to construct a new set from the empty set. So we have two sets. Then we can form a two element set. This is another set, it has more elements than either of the two. It is not difficult to realize that we can go on and create infinitely many different sets. With infinitely many elements we can construct infinite structures.

It is remarkable that we are creating everything from the empty set. Does this fact have a deeper meaning, or it is just a technical ad hoc trick? In theoretical physics matter dissolves more and more into empty space. Particles are just some probability amplitudes or vibrating strings which themselves have no volume...

Well, so far we have only infinitely many elements and we have to work more to get, say, the natural numbers. To construct a structure we need not only a set, but also relations and operations. Recall that the extensionality principle applies to relations and operations as well. Once we accept the extensionality principle for relations the next step is to realize, that relations are just *sets of pairs*. Similarly we can identify binary operations with certain sets of triples and so on. Thus, for example, the relation $\leq$ on the natural numbers is just a set of pairs (a, b) where a is less than or equal to b and $+$ is the set of triples (a, b, c) such that $a + b = c$.

So extensionality easily gives us an explanation of what are relations and operations. What still remains to be defined are pairs, triples, etc. We could assume that these are primitive concepts determined by axioms, such as the concept of a set, but there is a better solution. It turns out that pairs, triples, etc., can easily be constructed from sets. I will describe it in more detail in the next section; for now, let us only observe that an *unordered* pair of elements a and b can be simply identified with the *set* $\{a, b\}$ having the two elements a and b.

Note that it is just a matter of convenience that we reduce the concept of a function to the concept of a set. We could do it otherwise too. Historically people were first interested in the concept of a function and only much later the concept of a set emerged in mathematics. In Newton's time people thought of functions as physical quantities and thus attributed to them properties which are common in physical phenomena such as continuity and the existence of derivatives. When the theory developed, further examples of functions with some bad properties were found. The most striking among them is a continuous function with no derivative in any point. This means that the curve that the function defines does not have a tangent in any point, so to say, any point is a like a sharp edge. The question arose then: what is an arbitrary function? This question is closely related to the question about arbitrary sets, as sets can be defined as the points where a function is zero; on the other hand, a function is, as we understand it today, a set of pairs.

[6] It is consistent to assume the existence of *some* sets x which are equal to $\{x\}$, but usually they are prohibited by other axioms, as they are rather unnatural.

The Natural Numbers

Richard Dedekind
Courtesy of Universität Hamburg

The numbers which count the number of elements in a set are called *cardinal numbers*, or simply *cardinals*. Later on I will also talk on infinite cardinal numbers, but here I will only consider the finite ones. These are the numbers that we denote by $0, 1, 2, 3, \ldots$ and call the *natural numbers*. The nature of numbers was studied by many philosophers and mathematicians. The first rigorous foundations of the natural numbers was given by Richard Dedekind (1831–1916) in the book *What are and what should the numbers be?* published in 1888 [59]. In 1889 Giuseppe Peano (1858–1932) published a book *The principles of arithmetic presented by a new method* [216] where he presented Dedekind's formalization in a more precise form. According to Peano, the natural numbers are defined as a structure with a universe N, a function S and a constant 0 satisfying:

1. for every x, $S(x) \neq 0$,
2. if $x \neq y$, then $S(x) \neq S(y)$,
3. for every set $X \subseteq N$, if $0 \in X$ and $x \in X$ implies $S(x) \in X$, then $X = N$.

The function $S(x)$ is the *successor function*, which is the unary operation of adding one: $x + 1$. The notation with $+$ looks as if we implicitly used $+$ to define it, therefore logicians prefer to use a special symbol for it.

The third axiom is the basic principle of the natural numbers: *mathematical induction*. This principle is usually stated as the following rule:

Mathematical Induction *For a given property of the natural numbers $\varphi(x)$, if φ holds for 0 and $\varphi(x)$ always implies $\varphi(x + 1)$, then all numbers satisfy φ.*

Giuseppe Peano [7]

Note that the only difference between 3. and the statement of Mathematical Induction is that sets are replaced by the informal concept of properties.

This obvious and seemingly trivial principle is used in many proofs, simple and difficult ones as well. In fact, it is a universal principle—since this axiom determines the natural numbers, *all* results in number theory and finite combinatorics can be rewritten so that they only use this principle.[8]

The system based on the three axioms above is called *Dedekind-Peano Arithmetic*. The structures satisfying these

[7] This media file is in the public domain in the United States.

[8] More precisely, we must also use definitions of arithmetical operations and axioms about sets.

axioms are uniquely determined up to isomorphism. But note that it is not an axiomatization in logic since the third axiom speaks about sets. In other words, the natural numbers defined in this way are a second order structure. Thus when using these axioms, we must also use set theory.

It is possible to approximate this system by axiomatizations in logic, but then we can never achieve uniqueness up to isomorphism. The most natural axiomatization based only on logic is traditionally called *Peano Arithmetic* (see page 116).

The German logician Gottlob Frege (1848–1925) studied the question from a more philosophical point of view. His idea is very natural (later also used by Russell and others): a number n is the property shared by all sets with this number of elements. Using set-theoretical terminology, *a number is just the set of all sets of the same cardinality*. This presupposes knowing what it means to be of *the same cardinality*. But this causes no problems; two sets have the same cardinality, if we can find a one-to-one assignment of elements of the first set to the second set so that every element is matched with an element from the other set. In set theory a one-to-one assignment is a function and this in turn is just a set of pairs. So this can be expressed purely in terms of sets. Note that instead of saying that two sets have the same cardinality, we also say that they are *equinumerous*.

When formalizing the natural numbers in set theory we need to represent numbers by sets. Frege's definition of the natural numbers is not suitable for the formal system currently accepted as the standard. In Zermelo-Fraenkel Set Theory the class of all sets of a given cardinality greater than 0 is never a set. We can solve this problem by choosing one representative from each class of equinumerous sets. Another possibility is to use the Dedekind-Peano definition and just say that the natural numbers are one of the structures satisfying the three axioms above. Since in Zermelo-Fraenkel Set Theory we have to state an axiom of infinity anyway, we could just state the axiom saying that such a structure exists. But it is better to use more esthetically pleasing construction. Such a construction is due to the Hungarian mathematician John von Neumann (1903–1957). He defined the number n to be the set of numbers $0, 1, \ldots, n-1$; thus n is identified with the set of numbers smaller than n. Note that this works very well: as there are no natural numbers smaller than 0, 0 *is* the empty set; 1 contains only 0; 2 has two elements 0 and 1 etc. We get the next number by adding it to itself as an element. In set theoretic notation the numbers $0, 1, 2, \ldots, 5$ are: 0 (zero) is $\emptyset$ (the empty set; the two objects are the same), 1 is $\{0\}$, 2 is $\{0, 1\}$, 3 is $\{0, 1, 2\}$, 4 is $\{0, 1, 2, 3\}$, 5 is $\{0, 1, 2, 3, 4\}$. Since zero and the empty set are the same, set-theorists prefer to use 0 instead of $\emptyset$. If we substitute for the numerals their definitions, we can express all numbers only using the symbol 0 for the empty set, braces and commas. Thus $0, \ldots, 5$ become:

$$
\begin{aligned}
&0,\\
&\{0\},\\
&\{0, \{0\}\},\\
&\{0, \{0\}, \{0, \{0\}\}\},\\
&\{0, \{0\}, \{0, \{0\}\}, \{0, \{0\}, \{0, \{0\}\}\}\},\\
&\{0, \{0\}, \{0, \{0\}\}, \{0, \{0\}, \{0, \{0\}\}\}, \{0, \{0\}, \{0, \{0\}\}, \{0, \{0\}, \{0, \{0\}\}\}\}.
\end{aligned}
$$

This notation is, of course, not good for practical purposes.

So far we have only described a few small numbers, but we need a general definition. First we note that the successor is defined very easily by

$$S(x) := x \cup \{x\},$$

which is the key idea of this definition of the natural numbers. We want to say that every number n can be obtained from 0 by applying the successor function a finite number of times. We cannot do it directly because defining a '*finite number of times*' is equivalent to defining the natural numbers. So such a definition would be circular. Thus instead we use Dedekind's idea and say that

> *n must be contained in all sets which contain 0 and which are closed under the successor function.*

For a set a to be closed under S means that whenever it contains m it also contains $S(m)$. Note that the condition stated above is a property of elements n, hence, by the comprehension principle, there exists a set consisting of such elements. This set is the smallest set that contains 0 and is closed under the successor function. So it is natural to think of it as the set of numbers that can be obtained from 0 by applying the successor function. We define the set of natural numbers $\mathbb{N}$ to be this set.

To prove that the principle of mathematical induction holds for $\mathbb{N}$ defined in this way, we argue as follows. Suppose φ is a formula such that φ holds for 0 and $\varphi(x)$ implies $\varphi(S(x))$. Let N' be the set of numbers satisfying φ. By the assumptions of induction, N' contains 0 and it is closed under S. Hence, by definition, $\mathbb{N} \subseteq N'$, which means that all n satisfy φ.

Once we have defined the set of all natural numbers, the universe of the structure, and the successor function, it remains to define the ordering relation and the operations. The ordering is defined very simply: $a \leq b$ if and only if a is a subset of b. Here we see the advantage of having sets as elements: the structure of the elements enables us to define some relations very easily.

We define addition by saying that $n + m$ is the number whose cardinality is equal to the union of two disjoint sets A and B, where A is equinumerous to n and B is equinumerous to m. Unfortunately the sets representing the two numbers are not disjoint (unless one of them is 0), but it is a trivial task to construct such A and B. As regards the multiplication we are luckier, we can define $n \cdot m$ as the number equinumerous to the Cartesian product $n \times m$ of the sets n and m.

These constructions use special properties of the two operations. There is a much more general way of defining arithmetical function called *definition by recursion*. Let us consider a recursive definition of addition. Addition is determined by the following equations

$$\begin{aligned} x + 0 &= x, \\ x + S(y) &= S(x + y). \end{aligned} \tag{1.1}$$

Here we define what it means to add 0 and then we define how to add a number bigger than 0 using the successor function and the addition for a smaller number.

So these equations uniquely determine the operation. Having addition, we can give a recursive definition of multiplication:

$$x \cdot 0 = 0,$$
$$x \cdot S(y) = x \cdot y + x.$$

We can go on and define x^y and other functions. (See also the general form of recursion on page 142.)

The Real Numbers

I will skip the constructions of the integers and the rational numbers from the natural numbers because they are easy and well-known. A more interesting problem is to construct the real numbers.

In the 18th century, mathematicians were aware of the fact that calculus, the theory of real functions, limits, integrals, infinite series, etc., needs some axioms of continuity.

Example Let f be a continuous function defined on the closed interval $[0, 1]$. If $f(0) < 0$ and $f(1) > 0$, then there exists a real number a, $0 < a < 1$ such that $f(a) = 0$.

No one doubts that principles such the one above are true. But in order to develop the theory, we either have to state them as axioms, or we need a definition of the real numbers from which they follow. When using set theory as the foundations, we do not want to add axioms that are not about sets. We would like to derive everything from the basic axioms about sets, so only the second option is of concern to us. To this end we must define a mathematical structure representing real numbers, in a similar way as we defined a mathematical structure representing the natural numbers.

The classical approach is based on *Cauchy sequences*, named after the French mathematician Augustin-Luis Cauchy (1789–1857).The starting point is, as in all constructions of the real numbers, the rational numbers. An infinite sequence of rational numbers $r_0, r_1, r_2, \ldots$ is called a *Cauchy sequence* if the elements of the sequence get closer and closer as n increases. This is a rather subtle concept that needs a more precise explanation. It does not suffice that the distance between consecutive elements decreases. What we need is that if n is large, then the distance between r_n and *all* r_m, for $m > n$, is small. Once we know that all Cauchy sequences converge, then all the properties of real numbers follow. So the idea is to ensure that a Cauchy sequence converges by choosing an object representing its limit. Clearly, we have to choose the same object for all Cauchy sequences that should converge to the same limit. Thus the whole construction boils down to the definition of what it means that two Cauchy sequences converge to the same real number, which must be stated without mentioning the real number itself. The formal definitions of these concepts are in Notes.

For formalization in set theory, Cauchy sequences are as good as any other formalization. However, the option preferred by set-theorists is based on Dedekind's cuts. It is also a more acceptable answer to the philosophical question '*what are the real numbers?*' Dedekind used the fact that we only need to add irrational numbers to the rationals. Then an irrational number is defined as a '*hole*' in the line of rational numbers. Set theory enables us to easily define what a hole means: it is a partition of the rational numbers into two parts, one below the hole, the other above. Arithmetic operations with holes are done by shifting these partitions appropriately.

Example $\sqrt{2}$ is thus identified with the pair of sets (X, Y), where X is the set of all rational numbers less than $\sqrt{2}$ and Y is the set of all rational numbers bigger than $\sqrt{2}$. However, it would be a circular definition if we used this to define $\sqrt{2}$. Therefore we must say that Y is the set of all positive rational numbers y such that $2 < y^2$ and X is the complement of Y.

If we only want to show that the real numbers can be formalized in set theory, we can ignore tradition and philosophy and use some simple straightforward construction, such as decimal representation. In this representation a real number is an infinite sequence of numbers $0, \ldots, 9$ with a period and a sign. In order to get uniqueness, we disallow sequences that end with a tail of 9s.

Interestingly, the formalization of structures in set theory is a similar task as the formalization of structures for computers. Programming languages seldom use sets, they rather use lists and arrays, in which elements are given in some order, but this is not essential. The only essential difference between representing objects in set theory and in computers is that in computers we do not have infinite structures.

Notes

1. *Urelements.*[9] It is possible to develop set theory using true elements which are not sets. Such elements are called *urelements*. Another possibility is to use sets which have themselves as the only one element, sets that satisfy $x = \{x\}$. Thus we can mimic urelements while preserving extensionality for all objects, which we cannot do in the first case. The standard approach is, however, to use neither of the two kinds of urelements since we do not need them for practical purposes and the theory is simpler without them.
2. *Pairs and sequences.* The pair (a, b) is defined, following Kuratowski, by $\{\{a\}, \{a, b\}\}$. If $a = b$, then (a, b) contains one element that contains one element a. If $a \neq b$, then it contains two elements; one element is a one element set containing a, this determines a; the other is a two element set containing both elements, this determines b as the element that is not in the one element set.

[9] *Ur-*, originally a German prefix now also used in English, means *primitive, original.*

To represent a finite sequence with n elements $a_1, \ldots, a_n$ we take the set of pairs $\{(1, a_1), \ldots, (n, a_n)\}$. This is, in fact, a function defined on the set $\{1, \ldots, n\}$. Other indexed structures (matrices, infinite sequences, etc.) are done in similar way.

3. *Recursive definitions in set theory.* It is not difficult to prove in set theory that functions defined by recursion exist. We can reduce it to induction, which we already have. For example, we prove by induction that, for all n, there exists a unique partial operation defined on the interval $[0, n]$ satisfying the equations (1.1). Then the operation of addition defined on all natural numbers is the union of these partial operations.
4. *Cauchy sequences.* A sequence $r_0, r_1, r_2, \ldots$ is Cauchy, if for every $\varepsilon > 0$, there exists n such that for all $k, m > n$, the inequality $|r_k - r_m| < \varepsilon$ is satisfied.

   Two Cauchy sequences $r_0, r_1, r_2, \ldots$ and $s_0, s_1, s_2, \ldots$ converge to the same real number if for every $\varepsilon > 0$, there exists n such that for all $m > n$, $|r_m - s_m| < \varepsilon$. Note that we are able to define it without knowing the number to which they converge. Thus we can use the above condition to define an equivalence relation on Cauchy sequences. Then we define the real numbers as equivalence classes.

   The advantage of this construction is that it works for all metric spaces. Thus one can prove that every metric space can be extended to a complete metric space.
5. *Dedekind's real numbers.* Let $\mathbb{Q}$ denote the set of rational numbers. Dedekind's definition can be simplified by considering only one set of rational numbers for every real number. Thus we define a real number to be a nonempty proper subset of $\mathbb{Q}$ which is closed downwards (with any rational number it contains all smaller ones) and which does not have the largest element. For two real numbers r, s, we say that r is less than or equal to s, if r is a subset of s; $r + s$ is defined as the set of rational numbers which are less than or equal to $a + b$ for some $a \in r$ and $b \in s$; multiplication is defined in a similar way. The rational numbers $\mathbb{Q}$ are not a subset of $\mathbb{R}$ constructed in this way, but they are embedded in $\mathbb{R}$ by the assignment $a \mapsto (-\infty, a)$.
6. *Other structures.* As regards a small finite structure there is no problem to construct it now. We take, say, an initial segment of the natural numbers as the universe and to define a subset, relation or function, we simply list the elements. In the case of infinite structures, we have to find a particular construction in each case. This may depend on the axioms of set theory that we use! We can talk freely about classes of structures satisfying some properties, but to prove that there exists at least one such structure we need a construction.

   So far we are only using naive set theory, which is inconsistent, if taken strictly logically. We will have to restrict the general principles to get consistency and then add new axioms to retain the necessary strength. For instance, the existence of the power set $\mathcal{P}(X)$ for every set X is a consequence of the Principle of Comprehension, but it will be postulated as an axiom later. In order to prove that $\mathbb{N}$ is a set, we will also need an axiom—the Axiom of Infinity.

1.3 Antinomies of Set Theory

The decadent mood of the end of the 19th century influenced also the views on the future of science and technology. People thought that all important inventions had been discovered and there were no substantial discoveries going to happen in physics. Mathematics has always been different because it has had famous open problems. They will never be exhausted, as new problems arise at least as fast as old problems are solved. The foundations of mathematics are, however, a different thing. In foundations there is a clear convergence to more complete and more precise systems. From this point of view the state of the affairs in mathematics was similar to physics. During the 19th century all concepts of mathematics were reduced to natural numbers. This process, called arithmetization, started with Descartes's introduction of analytic geometry, continued with the formal definitions of convergence, derivations and integrals, and ended with the introduction of sets. Set theory was able to reduce even the remaining natural numbers to the abstract concept of a set.

For mathematicians this was a positive thing. Except for a few, they are interested in doing research on real mathematical problems. The problems on foundations are seldom clear cut and often it is more philosophy than science. Having firm foundations meant that they could discard those pseudoproblems forever. But even before set theory became generally accepted, it received a serious blow. This was because a contradiction was derived from basic principles.

Before considering the contradictions, I will briefly digress to explain why a contradiction is fatal for any theory. A contradiction is a pair of statements such that one is the negation of the other. When we derive such statements we can derive also their conjunction (also called a contradiction) which is logically false. It follows from the rules of logic that any statement is a consequence of a false statement. In Latin this rule is referred to as '*ex falso sequitur quodlibet*'.[10] This is also used in natural language. When we want to stress that something is blatantly false, we say that if that is true then something ridiculous is also true. However, the natural human interpretation of implication is that the parts of the implication, the *antecedent* and the *consequent* share some content. Therefore it is not easy to accept that a single statement can imply everything. The best way to see that a contradiction implies everything is to use a proof by contradiction. In such proofs we assume that the statement that we want to prove is false and derive a contradiction. Then we argue that therefore it is not possible that the statement is false, hence it is true. Now, if we are able to derive a contradiction without any assumptions (except for the basic principles), then it is formally derivable from any assumption. Thus any assumption can be rejected, hence everything can be proved.

Once we know that we can prove everything, there is no point in actually proving anything. Such a system gives us no information and certainly does not describe any real phenomenon, as in the real world a statement cannot be true and false at the same time.

[10] From falsehood, it follows anything you like.

Here we are, of course, concerned with mathematical truth. In our life it is quite different. We get a lot of contradictory information. One source of contradiction is unreliable information, another one is the use of generalization based on partial data. We are always ready to reject such statements and recompute our model of the world.

Contradictions in set theory are often called *paradoxes* or *antinomies* because they contradict our intuition.[11] The simplest and the most important one is *Russell's Paradox* discovered by the philosopher and logician Bertrand Russell (1872–1970) in 1901. He showed that already one particular instance of the Comprehension Principle is contradictory. Namely, he applied this principle to the property '*of not containing itself as an element*'. The principle asserts that there is a set, let us call it R, whose elements are just the sets with this property. For example, the empty set belongs to R since it does not contain any set as an element. On the other hand, the set of all sets (suppose we proved that it existed) contains any set, hence also itself, thus it does not belong to R. We obtain a contradiction if we consider the question, whether R is an element of R. For suppose R is an element of R, then R does not satisfy the defining property of R, hence it does not belong to R. This contradiction shows that R cannot belong to itself. But if it does not, then it does satisfy the defining property of R so it must belong to itself. Thus we get a contradiction in any case.

Russell was probably not the first to discover this paradox. Logicians in Hilbert's circle knew this paradox and attributed it to Ernst Zermelo (1871–1951). Zermelo did not publish the paradox, but according to his recollections, he thought about it around 1900. He used it to prove that the largest cardinality does not exist. But Cantor had been aware of the problems with certain sets already before Zermelo. He said that they "*cannot be conceived as determinate, well-defined, finished sets*". He also called them "*absolutely infinite sets*".[12] However, there is an essential difference between the approaches of Russell on the one hand, and Cantor and Zermelo on the other. While Cantor and Zermelo studied sets as mathematical entities, Russell's focus was on the principles of logic. Cantor and Zermelo viewed the paradoxes as proofs that "very large sets" do not exist. In contrast, Russell presented his paradox as a proof that the principle of comprehension is not a universally valid logical principle. If we want to have a consistent system, we must restrict the class of properties to which it is applied. Presenting the paradox in this way had a decisive impact on the further development of set theory.

In fact, Russell arrived at his paradox analyzing an earlier paradox found by Cantor. Cantor proved that for every set, the set of all subsets of it is strictly larger. The problem then is with the set of all sets. This set exists by the Comprehension Principle, where one uses as the defining property any property which is generally

[11] Strictly speaking, we should distinguish between paradoxes—apparent contradictions, and antinomies—actual contradictions, but when using informal reasoning it is difficult to make this distinction. Therefore, these words are used interchangeably.

[12] Letters to Hilbert, September 26 and October 2, 1897. See [65], page 42.

true (say, the property of being equal to itself). This set is, by definition, the largest set, so it contradicts to Cantor's theorem.

Apparently most mathematicians were not very impressed by the antinomies. They felt that what they were doing was sufficiently well tested and they used mathematical objects that were in some sense more real than sets. In any case, the historical experience suggested that even if a part of the present mathematics would have to be abandoned because of its contradictory character, it would be only a small part. Still, it was rather disturbing that the contradiction was derived from what seemed an intuitively obvious principle.

At this point several other paradoxes were discussed. One of them, whose roots go back to the ancient Greeks, is the well-known *liar's* or *Epimenides* paradox.' The story says that Epimenides was a Cretan who said: "*All Cretans are liars.*" Was he a liar?[13] A modern version of this paradox is *the paradox of the barber: "There is a man in a village who shaves all men in the village who do not shave themselves, and only those. Does he shave himself?*"

Another, known as *Berry's paradox*, goes as follows. We know that any nonempty subset of the natural numbers has the first element. (This is just an equivalent form of the induction principle.) Also it is clear that there are only finitely many English sentences with at most 100 letters, hence there are natural numbers which cannot be defined by such sentences. Thus we can define a number n to be *the first number that cannot be defined by an English sentence with at most one hundred letters.* This is a contradiction, as we have just defined n by such a sentence!

At first it may seem that the problem with the paradox may have something to do with infinity. After all, we have no idea how large the largest number definable in this way is. But in fact we can easily give an upper bound on the numbers that have to be considered. With 26 letters used in English and one more character for the space between words (or at the end of the sentence) we can estimate the number of English sentences with at most 100 letters by 27^{100}. Thus the alleged number should be amongst the numbers $0, 1, 2, \ldots, 27^{100}$ since at least one of these numbers cannot be defined using 100 letters.

The number 27^{100} is, unfortunately, too big even for a computer. Furthermore, English is rather complex, so it would be difficult even to generate all syntactically admissible sentences. But you can design, or at least imagine, your own special purpose language with a simple and precise syntax and such that one can state the paradox using a sufficiently small number instead of 100.

Another version of this paradox is based on the assumption that our universe is finite. Under this assumption we do not have to give an explicit estimate of the length of the definition.

These two paradoxes belong to a class of paradoxes, called *semantic paradoxes*, that are based on natural language and use words such as '*true*' and '*defined*', which are not precisely defined. However, it is possible to formalize these concepts when we have a formal language. Then, the paradoxes are resolved by strictly distinguish-

[13] This is the traditional version of the paradox which assumes that a liar is *always* lying.

ing between the object language and the language that we use to define these concepts. These paradoxes have never been perceived as a real threat. After all, such paradoxes have been known for thousands of years and they never interfered with mathematics.

Paradoxes in Mathematics

There are paradoxical results in mathematics that do not present inconsistencies. They are exact theorems, except that they are counterintuitive. A classical example of a paradoxical object is the function, constructed by Bolzano and Weierstrass which is continuous but not differentiable at any point, which I already mentioned above. Another classical example is a curve constructed by Peano in 1890 that completely fills a square.

Paradoxical results are present in many fields of mathematics, the more the field is connected with our a priori intuition the more likely we can find some. For instance, human understanding of the geometry of three dimensions, which is to a large extent inborn, is quite good. When thinking about higher dimensions we try to use our three dimensional intuition, but it often fails badly. It is an easy exercise to construct two circles C_1 and C_2 in four dimensions such that the distances between all pairs of points one on C_1 and the other on C_2 are the same. We can do it using analytical geometry, but we are not able to visualize it because in three dimensions it is impossible.

Since a lot of our everyday decisions are based on estimating probabilities of various events, one would expect that our intuition about probability is fairly good. But there are examples of the failure of our intuition also in this field. Perhaps the most popular is the well-known *Birthday Paradox* of Richard von Mises. It seems very counterintuitive that with a probability greater than $1/2$ among 23 randomly chosen people there are two with the same birthday. One would expect to need essentially more to get this probability, but the above fact can easily be shown by a simple calculation.

A more recent and more tricky one is the following nice puzzle, so nice that it made it into the pages of the New York Times as *the Hat Problem*. There are n people each having a blue or red hat. Each person can see the color of everybody else's hat, but not his or her own. According to the rules of the game they play, at some point they are asked to guess the colors of their own hats. They have to answer at once and independently of each other, but anybody can abstain. If everybody abstains or one of them guesses incorrectly they loose. If at least one does not abstain and everybody who answers gives the correct answer, they win. The players can agree on a suitable strategy beforehand, but once the game starts they are not allowed to communicate. It is clear that every strategy may fail since either everybody abstains, which is a failure, or at least one player always answers, but the player surely may give the wrong answer, as the players do not know their colors. The question is, what is the best strategy when one wants to get the highest chance

3. *Hilbert's Paradox.* Consider two operations: the union of a set X (denoted by $\bigcup X$) and the set of all mappings on a set X (denoted by X^X). Start with a non-empty set and let U be the set obtained by applying these operations in all possible ways. Then U^U must be a subset of U, by definition. But U^U has a larger cardinality than U (as one can show by the same argument as in the theorem above). Hence we cannot consistently assume the existence of such a set.

   Hilbert considered this paradox to be more serious than others because the set U is apparently constructed from below only using basic set-theoretical operations. But this is only what it looks like if we do not use a precise definition. As soon as one tries to define U precisely, one sees that it is not possible to avoid referring to "large entities". For example, the standard way to define such a U is to use transfinite recursion. To this end, however, one has to use *all* ordinals. Ordinals do not form a set, hence also U will not be a set. In set theories with classes, ordinals form a proper class (a class which is not a set) and so will also be U.
4. *Burali-Forti's Antinomy.* For the sake of completeness we mention also Burali-Forti's antinomy. It was published by Burali-Forti, but had been known to Cantor before. After all, it is not so much different from Cantor's paradox. By the theory of ordinal numbers, which we will consider later, an initial segment of ordinal numbers has an ordinal number which is bigger than any element of the segment. Thus the existence of the set of all ordinal numbers leads to a contradiction.
5. *The Hat Problem.* The reason why the intuitive argument is wrong is the following. While it is surely true that when a player answers, he gives the right answer with probability $1/2$ and the wrong one with $1/2$, this is only the conditional probability with the condition being that the player answers. If we take into account that a player sometimes abstains, we have the probability ε of the correct answer, ε of the wrong answer and $1-2\varepsilon$ that he abstains, for some $0 \leq \varepsilon \leq 1/2$. If ε is small, then the player gives wrong answer with small probability. Now the trick is that with more players it may be possible to arrange it so that the bad cases overlap, so the probability of failure remains small, but the good cases are distinct, so the probabilities add up. Indeed, in the optimal solution that exists for n of the form $2^k - 1$,

   a. either all players answer incorrectly, and this happens with probability $\frac{1}{n+1}$,
   b. or exactly one player answers correctly while others abstain; each of the players does so with probability $\frac{1}{n+1}$, hence they win with probability $\frac{n}{n+1}$

   The solution is based on Hamming codes, which is a hint for the reader who wants to solve it.
6. *Paradoxes in computational complexity theory.*

   a. Consider computations of Boolean functions $f : \{0,1\}^n \to \{0,1\}^n$ by Boolean circuits. This model of computation will be introduced later. For now, think of it as a piece of hardware consisting of electronic gates that works as follows. If you fix the input values on input wires, the circuit computes for a while and when the values on all gates stabilize, you get the output value of the function f that the circuit computes. Suppose, for some function f, the minimal size

of a Boolean circuit computing f is S. Now suppose that you want to compute f in parallel on two independent inputs. This means that you want to compute the function F that from $2n$ bits x, y produces $2n$ bits $f(x), f(y)$. Intuition tells us that the minimal size of a circuit computing F should be $2S$. The following is an intuitive reason that it cannot be less. Given a circuit for F we can think of it as two overlapping circuits, one computing $f(x)$ the other $f(y)$. The overlap consists of the gates that depend on both inputs x and y. But if a gate may have an arbitrary value depending on y, then it should be useless for computing $f(x)$ and symmetrically with x replaced by y. Thus the gates from the overlap should be useless, hence the best we can do should be to take two disjoint circuits. Yet, one can show that, for some functions f, one needs only a tiny fraction more than S to compute F (namely $(1+\varepsilon_n)S$, where $\varepsilon_n \to 0$ as $n \to \infty$). (See [295].)

b. Consider the following three player communication game. Player 1 gets a bit string x of length n, $x = (x_0, \ldots, x_{n-1})$ and a number i; Player 2 gets the same string x and a number j; Player 3 gets i and j. Their information is private, so, for example, Player 1 does not know j. Then Player 1 and Player 2 send independently of each other messages to Player 3. They have agreed beforehand on what messages they will send in all possible situations and they have done so in such a way that Player 3 is always able to say correctly what is the value of x_k for $k \equiv i + j \bmod n$. The question is what is the total length of the messages they would have to send to Player 3 in the worst case. Clearly, a possible protocol on which they may agree is that they would send all the bits of x to Player 3, which is n bits. Intuitively this seems the best possible thing they can do. The argument is as follows. For Player 1, the information about i is totally irrelevant, as for a given i the $k \equiv i + j \bmod n$ may be completely arbitrary. Hence the only relevant information Player 1 can send concerns x. Similarly for Player 2. So they will send some information about x, independently on the indices i, j. But then they have to send at least n bits, as the information on x cannot be compressed. Yet, the minimal number of bits that the players have to exchange is bounded by a function $f(n)$ such that $f(n)/n \to 0$ as $n \to \infty$. (See [230].)

1.4 The Axiomatic Method

At the age of eleven, I began Euclid, with my brother as my tutor. This was one of the great events in my life, as dazzling as first love. I had not imagined that there was anything so delicious in the world.

Bertrand Russell, *The Autobiography of Bertrand Russell*[14]

The oldest mathematical texts contain examples of mathematical problems with solutions. They served as guides of how to solve equations, how to construct geometric figures etc. The first proofs of mathematical theorems appeared in ancient

[14] [254], Vol. 1, page 36.

Greece probably in the 6th century BCE. They are attributed to Thales and members of the Pythagorean School (for example, the proofs of Thales' Theorem and the Pythagorean Theorem). Convincing evidence that mathematical proofs had been used in the ancient Greece in the 5th century BCE is the discovery of the incommensurability of the side and diagonal of a square. (This is essentially the fact that $\sqrt{2}$ is not a rational number, see page 257.) This is a kind of statement that requires a proof; you cannot claim that it is *impossible* to write $\sqrt{2}$ as a fraction of two whole numbers, unless you can prove it.

This was not only the time when first proofs appeared, but also the time when western philosophy emerged. According to tradition, it was Pythagoras who coined the term *philosopher*. The emergence of philosophy meant that science ceased to be considered to be a tool serving to efficiently accomplish practical tasks, but rather an environment for intellectual activity, disregarding any possible applications. Once people started to ask, not only '*how?*', but also '*why?*', they could not have been satisfied with mere statements of mathematical facts. They needed *proofs*.

Aristotle (384–322), the greatest philosopher of Antiquity, studied logic and the scientific method in general. He determined a set of logical rules, which he called *syllogisms* and described logical deductions as successive applications of these rules starting from some basic assumptions. By this, he described what we now call the *axiomatic method*.

Aristotle distinguished between two types of basic assumptions: *postulates* and *axioms*. Postulates are those that are common to all sciences, whereas axioms are special for a particular field. In the modern terminology of mathematical logic we do not use the word '*postulate*'; however, we do distinguish between *logical axioms* and *mathematical axioms*.

A prime example of an application of the axiomatic method are *Elements* written by Euclid of Alexandria around 330 BCE. Euclid starts by explaining the basic concepts such as '*A point is what does not have a part.*' Part of these statements are not definitions in the modern mathematical sense; they relate the abstract mathematical concepts to reality. We would rather call them *intended interpretation*. Then he presents two lists of statements. The first one can be interpreted as geometrical axioms, the second as logical and arithmetical axioms. The results are presented as theorems, constructions and algorithms.

We know about some gaps in proofs and that the postulates in *Elements* are not sufficient to derive all theorems. Yet, it is an impressive work, whose style is surprisingly close to present-day mathematical monographs. Many mathematicians used *Elements* as a prototype for their treatment of geometry. In fact, this book is among the most influential ones of Western civilization. Finally, a modern axiomatization of geometry was given by the great German mathematician David Hilbert (1862–1943) in his *Foundations of Geometry*[15] in 1899.

The axiomatic method is a way to reduce assumptions used in a theory to a few basic principles. But this does not only concern assumptions; at the same time,

[15] Grundlagen der Geometrie, [124].

we are also reducing concepts to simpler ones. Thus the reduction goes on in two parallel lines: on the one line we are reducing the assumptions, on the other we are reducing the concepts.

axioms	theorems
primitive concepts	defined concepts

Reducing the assumptions means that we show that they are derivable from others; reducing concepts means that they are definable from others. Eventually no further reduction is possible and then we talk about *axioms* and *primitive concepts*.[16] The primitive concepts are those which are not defined. The main reason is that they cannot be further reduced, but we usually also assume that they are clear and do not need further explanation. Similarly axioms are statements that we are not able to reduce to more primitive ones.

In principle, we could develop theory only using primitive concepts, but it would be very cumbersome. Definitions enable us to use short terms to express more complicated concepts and thus we can express ideas more efficiently.

An ideal mathematical text starts with axioms, followed by definitions, theorems and proofs of theorems. Definitions do not have to be all at the beginning. Furthermore, proofs may use auxiliary theorems, which are called lemmas. Proofs may also use auxiliary concepts that are not used in the axioms and the statements of the theorems. Although we use a special word 'lemma' for auxiliary theorems, we do not have words distinguishing auxiliary terms and their definitions from the genuine concepts and their definitions. However, mathematical articles and monographs do not only consist of definitions and theorems. Reading a completely formal mathematical text would be difficult and readers need to know the motivation for the theorems, how the results relate to those in other articles etc. It also helps to give informal descriptions of difficult proofs.

Example In elementary plane geometry the primitive concepts are points, lines and the incidence relation between points and lines. Thus we have two kind of objects, *points* and *lines*, and the relation 'a point *lies* on a line'. The basic axioms of plane geometry are:

1. *for every two different points, there is a unique line incident with them*;
2. *every line has at least two points*;
3. *any two different lines have at most one point in common*;
4. *there are three points which do not lie on one line.*

Using these basic concepts one can define other objects, such as triangles, quadrilaterals, etc., but also relations such as two lines being parallel ($\Leftrightarrow$ no point lies on both lines). These axioms are only a part of the list that Euclid needed, but already

[16] Sometimes it is useful to keep some redundancy; sometimes we are not able to prove that further reduction is impossible, but it is.

using these axioms one can prove many theorems. Also the concepts available in this system are rather simple and we have to add more primitive ones and more axioms to get interesting theorems. In particular we need the relation of congruence in order to be able to say that two line segments have the same length.

The main reason for using the axiomatic method is that we want to understand the subject that we study, we want to know what is essential—we need a *theory*. By a theory we usually understand a collection of statements which explain certain phenomena. It is very difficult to define what it means *to explain*. There are, however, some attributes that are quite clear: simplicity and universality. Thus a good theory must be based on a small number of general statements. The simplest theories may consist of a single postulate. The law of free fall asserts that the speed of falling objects is proportional to the square of the elapsed time. The universal nature of this theory is in its applicability to any object. A more general theory is Newton's theory of gravity. It explains much more than just the attraction of bodies to the Earth. It can also be given by a single equation asserting that the attraction of bodies is proportional to the product of their masses and to the square of the distance. Maxwell's theory unifies electrostatic and magnetic forces using a few differential equations. The ultimate goal of theoretical physics is a unification of all physical theories, dubbed the *Theory of Everything*; presented more modestly, it should be one theory for all forces in nature.

This is just to name a few examples from physics. Theories are present in all scientific disciplines. They are not always called theories; sometimes they are called *models* (when there are alternative theories), sometimes they are called *laws*. Formally, they are all just axiomatic systems.

Ancient Greeks not only discovered that one can axiomatize mathematics, but also the striking fact that one needs only a small number of very basic principles to do that. This also concerns some other fields of science. If nature were evil, we would need to get more experimental data every time we wanted to get more knowledge. That would mean accepting more and more axioms, which eventually would make the axiomatic method almost useless. But on the contrary, especially in physics, we are witnessing a reduction to fewer and fewer basic principles, one needs fewer and fewer absolute constants, etc. Already the present physical theories are able to reduce all chemistry to a few physical laws. In principle, it is possible to compute the chemical properties of all atoms and molecules only using quantum electrodynamics. We can go on and reduce molecular biology to chemistry etc. These are, of course, only theoretical reductions. In practice, the computational problems involved are so difficult that it is unlikely that one will ever be able to do without experiments.

In the foundations of mathematics the axiomatic method plays an extremely important role. Russell's paradox taught us a lesson: set theory cannot be based only on intuitive principles. In particular, it is necessary to restrict the use of the Principle of Comprehension. In this situation, it is reasonable to present the modification as precisely as possible. Although stating axioms of set theory explicitly does not guarantee the consistency of the resulting theory, it gives us at least something that

containing a metavariable for formulas. We obtain an instance of the schema, a concrete axiom, by substituting a formula for the metavariable. The two most important theories axiomatized by schemata are Peano Arithmetic and Zermelo-Fraenkel Set Theory, which I will describe in the following chapters. One can relax the condition on the set of axioms to the mere requirement that there is an algorithm for deciding whether or not a given formula is an axiom or not. But this is as far as we can go; if the set of axioms is algorithmically undecidable, we cannot consider it to be a formal system. In a formal system, we should be able to decide whether or not a given text is a proof; if we are not able to decide if a sentence is an axiom, then this is impossible.

In this book, I will only consider theories that are axiomatized by an algorithmically decidable set of axioms. To stress the latter fact, I will sometimes use the term '*formal theory*' or '*formal system*'. The latter one has a little broader meaning—the system does not have to be based on logic. I will also use '*axiomatic system*', '*axiomatization*', etc. with the same meaning as '*formal theory*'. The reader not familiar with the concept of decidability can simply imagine a formal theory as a theory axiomatized by a finite set of axioms and schemas since for a large class of theories, axiomatizability by a decidable set of axioms is equivalent to axiomatizability by a schema (according to a result of R.L. Vaught [298]).

The assumption that the set of axioms must be algorithmically decidable has profound consequences. It implies that certain structures cannot be axiomatized. This concerns, in particular, the structure of the natural numbers, as well as all structures that contain the natural numbers. This fact is the essence of the Gödel Incompleteness Theorem, which I will explain in Chap. 2 and then in more detail in Chap. 4. An important consequence is that nonelementary theories that use the natural numbers as primitive concepts cannot be fully formalized. In particular, none of the currently used physical theories can be fully formalized.

Properties of Theories

1. The most important property of an axiom system is its *consistency*. This means that the system is free of contradiction. In an inconsistent system one can derive any sentence, hence such a system is useless, as we noted in the section on antinomies in set theory. Actually, axiomatization of set theory was historically the first case where the question of consistency became important. Before people axiomatized concrete structures. Assuming that a particular structure exists, we get the consistency of any set of sentences that are satisfied in the structure. In particular, we believe that the natural numbers exist, therefore the axioms about them are consistent. For sets there is no such "canonical" structure. The only place where they occur is our natural language, which is imprecise and inconsistent in many ways. There is nothing to which we could reduce the consistency of set theory.

   Upon closer inspection, we realize that the situation is not much better even if we have a canonical structure for the theory. For example, we may firmly believe

that the natural numbers are a real object and as such they must be consistent. But how can we test that a sentence that talks about all numbers is true in the structure? We cannot test all infinitely many numbers. So our argument that the axioms about natural numbers are consistent is based on the *belief* that the axioms are satisfied in this structure. What we, however, can do completely formally is to reduce the consistency of one theory to another one. Thus, for example, we can reduce the consistency of an axiomatic system for the natural numbers to the consistency of an axiomatic system for set theory.

Consistency is the key concept in the foundations, so we will learn more about it later; it will occupy us essentially for the rest of this book.

2. The second most important property of axiomatic systems is the *completeness*. A system is complete, if we can derive all sentences that are true in the structure that we are axiomatizing. In the case the system should describe a class of structures, we require that any sentence which is true in all structures of a given class is derivable in the system. For some simple structures, it is possible to find a complete axiomatization, for more complex ones, it is impossible. Note that completeness depends on the language that we consider. Thus, for example, it is possible to give a complete axiomatization of elementary geometry of the plane in the style of Euclid and Hilbert. However, if we want to study deeper problems, say differential geometry, the task becomes impossible. Another example is the natural numbers with addition as the only operation, which we denote by $(\mathbb{N}; +)$. This structure is axiomatizable, whereas if we also include multiplication, that is, if take the structure $(\mathbb{N}; +, \cdot)$, it is not.

   In the case of classes of structures defined by axioms we get completeness automatically. For example, groups are precisely those structures (with one binary operation, one unary operation and a constant) that satisfy the three axioms on page 18. Thus the three equations form a complete set of axioms. This looks terrific, as if we could just let a computer generate all the theorems about groups from these axioms. Unfortunately there is again the problem of the language that one considers. If we only use the elementary language of group theory $\{1, \cdot, x^{-1}\}$ we get only trivial theorems. In order to express interesting concepts, for example, to define a simple group, we need either to use a higher order language, or work in set theory. In both cases a complete axiomatization is elusive.

   A more technical remark concerns *relative completeness*. I touched on this subject already above when talking on real vector spaces. The set axioms of real vector spaces is complete relative to the structure of the real numbers $(\mathbb{R}; +, \cdot)$, which means that we can derive all true sentences about real vector spaces using the axioms and sentences true in $(\mathbb{R}; +, \cdot)$. Incidentally, there is a complete axiomatization of $(\mathbb{R}; +, \cdot)$, which implies that we can also completely axiomatize real vector spaces. But again, interesting problems concern *sets* of vectors.

3. We say that a collection of axioms is *independent*, if no axiom can be derived from the others. Put otherwise, axioms are dependent, if they can be further reduced to a smaller set. So it is important to know, if a given set is independent.

The famous case of the fifth postulate of Euclid concerns this property. The original statement of this axiom was that *two lines a, b intersecting a line c so that at one side of c the sum of inner angles is less than* $180°$ *("two right angles") must intersect at that side of c*. This is equivalent, using the other axioms, to: *for a line a and a point B not on the line, there is a unique line b through B which does not intersect a*. A lot of people tried to derive this axiom from the others. It took a long time for people to accept the possibility that it cannot be done. A positive outcome of this were new structures, the *non-Euclidean geometries*. We will come back to this topic later and I will explain how is possible to show independence. For now, let us just say that one needs to construct a structure which satisfies all axioms except the one that we want to show to be independent.

Independence is not as important as consistency and completeness. If we want to axiomatize a structure or a class of structures, we are satisfied with any consistent and complete set of axioms. We are interested in the dependence of axioms only because we want to fully understand the concept and, possibly, find its generalizations.

Notes

1. *First-order logic.* In this chapter I have been using the term '*logic*' for what is more precisely called '*first-order logic*'. The name stems from the fact that the logic uses *first-order language*, the language for first-order structures. I will explain this connection and the key role of first-order logic among other logics in the next chapter.
2. *The axioms of Euclidean geometry on a plane.* Above I have stated only the most basic axioms, the axioms about the incidence relation between points and lines. To develop elementary geometry one needs axioms about two more relations:

   a. "*point A is between points B and C*";
   b. "*segment AB has the same length as segment CD*"; we say that AB is *congruent* to CD.

   There are two groups of axioms one for each of the two relations. These are a few cleverly chosen statements that rather surprisingly suffice to derive all that one needs. What they say can be informally described as follows.

   a. The axioms about the relation 'between' say that on every line, once we fix a direction by taking two points, we can define a linear ordering that is dense and does not have the largest or the smallest elements.
   b. The axioms about the congruence relation say, roughly speaking, that we can drag a segment on a line and to any line and that all distances in congruent triangles are preserved.

   Once we have congruence on segments, we can define congruence on angles.

   A large part of elementary geometry can be developed using these axioms and only using logic. In particular, although we do not have the circle as a primitive

concept, we can emulate it by a point C that determines the center and a segment CD whose length determines the diameter. Essentially the same can be done with the ellipse and other quadrics. However when using more complex objects one has to resort to set theory. For example, one cannot express in logic concepts such as *polygon* and *connected*, and cannot define curves that are not determined by algebraic equations. (Below I will show that connectedness of graphs is not expressible in logic.) Hence, in order to get a completely formal system in which we can develop more advanced parts of geometry, we have to accept some axioms of set theory on top of the Euclidean axioms.

This set of axioms is not complete. To make it complete one has to add axioms about the topology; it suffices to do it on lines. This is usually done by talking about sets of points. Adding axioms on sets results in a system that cannot be completed, due to Gödel theorems, but if we restrict ourselves to the primitive concepts of these axioms and only use order logic, one can get a complete theory. The idea is to replace the axiom on sets by an infinite schema that states it for every formula. For example, one can take the following set of axioms for every two formulas ϕ and ψ.

> Let a line be given and an ordering on the line be fixed. Suppose that on the line every point that satisfies ϕ is before every point B that satisfies ψ, then there exists a point A on the line that is between the points that satisfy ϕ and ψ (A may satisfy one of the two formulas).

Note that this is very much related to the axiomatization of the structure $(\mathbb{R}; +, \cdot)$.

3. *Gaps in Elements.* One kind of important missing axioms are instances of the continuity principle. In particular, the axioms telling when a circle and a line intersect and when two circles intersect. This axiom is needed already in the first theorem that proves the existence of an equilateral triangle with a given side AB. Euclid relied on the intuitively clear fact that if we want to connect a point inside of a circle with a point outside using a line, we have to intersect the circle. This is correct, but it does not follow from his axioms.
4. *Connected graphs is a nonelementary class.* We will use the class of connected graphs to illustrate some limitations of the axiomatic method.

   A graph is connected if every two different vertices are connected by a path. This is a clear and natural definition, but there is a problem: we need the concept of a path. We can define connected graphs equivalently by saying that *a graph is* not *connected, if there is a partition of the vertices into two nonempty disjoint sets such that there are no edges between the two blocks*. In this definition we need the concept of a set. Without using such concepts we cannot define connected graphs.

   Suppose that connected graphs can be defined by a first-order sentence Φ. Consider an infinite sequence of symbols $v, u_1, u_2, u_3, \ldots$, which will be interpreted as vertices of a graph. Furthermore, consider the following infinite set of axioms:

   a. $u_i \neq u_j$, (for all $i \neq j$);
   b. $u_i \neq v$, (for all i);

they see. Thus the problem of axiomatizing Relativity has attracted a lot of researchers. Most axiomatic systems proposed so far only formalize *Special Relativity*. Axiomatizations help us understand what are basic principles and what are their consequences. Then one can clearly see that specific mathematical concepts, such as the Lorenz transformation and the Minkowski norm, follow from the assumption that the speed of light is the same for all inertial observers and a few other basic principles.

General Relativity is a much more difficult theory. In this theory space-time is described by *Einstein's Field Equations*, which are nonlinear partial differential equations. One can use some axiomatizations of Special Relativity and extend it by adding Einstein's Field Equations to obtain an axiomatization of General Relativity.[17] It would be more interesting to have a theory in which Einstein's Field Equations would logically follow from basic principles.

1.5 The Necessity of Using Abstract Concepts

Building a good theory is the main goal in any field of science. Having a theory we can give explanations of a variety of phenomena and make predictions. Making predictions means that we are able to compute what happens more precisely and more efficiently. A characteristic feature of theories is that they use more abstract concepts than those that we can observe immediately. Philosophers argue whether or not one should use concepts that do not correspond to things that we can observe. The *Occam's Razor*, also called *the law of parsimony*, tells us that we should avoid any use of concepts that are not inevitable for describing the situations that we study. Logical positivism was based on a similar axiom, the aim being to avoid meaningless 'metaphysical' considerations. In mathematics essentially all concepts are abstract, so these problems may seem irrelevant, but it is not true. What should be called '*abstract*' and what should not is difficult to decide and mathematicians do not care anyway. What is however undeniable is that there is a hierarchy of mathematical concepts. The words '*more abstract*' and '*higher order*' correspond to our feelings about the concepts higher in the hierarchy. Furthermore, mathematics, being the most precise of all fields of science, gives us the possibility to study the role of abstract concepts systematically. More than that, we can even prove that abstract concepts help in several ways. In fact, the field of logical foundations is all about it.

A Tough Nut for Computers

It's high time now to be less abstract and give some concrete examples. I will start with a very elementary example, which is a well-known problem from recre-

[17] This is not quite precise. One has to first generalize the theory and only then it is possible to add Einstein's Field Equations. The generalizations without the Field Equations are also interesting theories and can describe nontrivial phenomena.

that the generalized problem for boards $n \times n$, with n an even number of the order of thousands, any proof in this proof system is so large that it cannot be practically performed. For showing that the argument using coloring cannot be done in the proof system, this is enough because if one could use this argument, the proof would be still relatively short even for large boards.

Transcendental Numbers

Let us consider something more serious. Most examples of using abstract concepts for solving problems that are stated in elementary terms come from number theory. One of the popular problems in number theory is proving that a number is not a solution of an algebraic equation with integer coefficients. Numbers that are solutions of such equations are called *algebraic*; those that are not are called *transcendental*. For example, $\sqrt{2}$ is a solution of the equation

$$x^2 - 2 = 0,$$

so $\sqrt{2}$ is algebraic. On the other hand π is not a solution of any such equation, hence it is a transcendental number. Proving that a particular number is transcendental is usually hard. The first proof that a number is transcendental was given as late as in the 19th century. Later, when Cantor discovered set theory, he showed that the existence of such numbers can be proved very easily using set theoretical concepts. He proved that the cardinality of the set of all real numbers is not countable, whereas the cardinality of the set of algebraic numbers is countable. Therefore, there are transcendental numbers.

Notice the similarity with the previous problem. Again the main idea is counting. Such counting proofs are often very simple, but we have to pay for it: such proofs do not give us explicit examples of the objects claimed to exist. We will encounter proofs that prove the existence without giving explicit examples again later.

Diophantine Equations

There are many problems about natural numbers that can be stated only using the basic arithmetical operations. Problems of this type were studied by Diophantus of Alexandria, who lived in the 3rd century. The problems he solved can be presented in modern terms as follows. Given an equation with integer coefficients, find a solution that is also an integer (or several integers, if the equation contains more variables). A classical problem, solved already in antiquity, is to give all such solutions to the Pythagorean equation

$$x^2 + y^2 = z^2.$$

There are infinitely many such triples (3, 4, 5 is the smallest one) and they have a simple characterization. The proof is completely elementary. One may be tempted

8. *Random generic sets.* When explaining generic sets, I stressed the fact that they do not have properties typical for random sets. This is not quite precise and it concerns only the original constructions of Cohen. There are generic extensions by sets that look random. To this end we do not have to develop a new kind of forcing; we only have to take a suitable set of forcing conditions, or, which is equivalent, to take a suitable complete Boolean algebra.

   Let us demonstrate it by the problem of adding a non-constructible subset of natural numbers. To this end Cohen used forcing conditions that lead to the Boolean algebra B_2 defined above. In order to obtain a randomly looking generic subset of natural numbers, we take a different Boolean algebra, which we will denote by B_3. To define B_3 we start with the same set as we did with B_2, the set of countably infinite sequences of zeros and ones $\{0, 1\}^\omega$, but instead of using topology, we will use measure. The measure that we need is the natural measure such that the whole space has measure 1, the set of sequences starting with 0 (respectively 1) has measure $1/2$, etc., (that is, the measure of the set of infinite sequences that extend a fixed sequence of length n has measure $1/2^n$). The details of how this is precisely defined are not important, but let me stress the fact that measure is defined only for some subsets of $\{0, 1\}^\omega$, which are called *measurable sets*. To define the elements of B_3 we identify measurable sets whose difference is a set of measure 0. For example, all countable sets have measure 0 (but not only those), hence, in particular, two sets that differ in a countable number of points will be identified. Formally, the elements of B_3 are classes of measurable sets that differ by sets of measure 0. The operations are defined by taking representatives from the classes, applying the corresponding Boolean operation, and taking the class containing the result.

   Thus Boolean algebra B_3 produces a generic extension $M[F]$ in which F is not constructible. The disadvantage of this construction is that whereas B_2 has a succinct description by a countable set of forcing conditions (finite strings of zeros and ones), B_3 does not have such a representation: B_3 is not generated by a countable set of forcing conditions.

   Let us now compare $M[F]$ with a generic extension $M[G]$ produced by Cohen's forcing conditions. We know that F is different from G (for example, the average number of zeros in initial segments of F converges to $1/2$), but this is not enough to prove that the two models are different. We know that such extensions contain a lot of other subsets of natural numbers that are not present in M. Thus G could be among the sets generated from F. To prove that it is not so, we have to find some property that distinguishes these two generic extensions. An interesting property that does it is the following. In $M[F]$ every function on natural numbers is bounded from above by a function from M, while in $M[G]$ there are functions that grow faster than any function in M. Though models constructible by such '*random forcing*' are different from those produced by '*Cohen forcing*', many results, including the unprovability of the Continuum Hypothesis, can be reproved using random forcing.
9. *Martin's Axiom.* Researchers in set theory prefer to assume the negation of the Continuum Hypothesis, since the universe of sets satisfying this axiom is richer.

Chapter 5
The Complexity of Computations

Hiding in the alternating patterns of digits, deep inside the transcendental number, was a perfect circle, its form traced out by unities in a field of naughts.

Carl Sagan, *Contact*

COMPLEXITY is a notion about which we do not learn in schools, but which is very familiar to us. Our generation has witnessed a tremendous increase of complexity in various parts of our life. It is not only the complexity of industrial products that we use. The world economy is a much more complex system now than it used to be; the same is true about transportation, laws and so on. Computers help us to cope with it, but they also enhance the process of making our lives more complex. The progress in science reveals more and more about the complexity of nature. This concerns not only biology and physics, but also mathematics. In spite of the great role that it plays in our lives, complexity has become an object of mathematical research only recently. More precisely, the word complexity had not been used until about the 1960s, but many parameters introduced long before can be thought of as some sort of complexity measures. Already the words used for these parameters suggest that they are used to classify concepts according to their complexity: *degree, rank, dimension*, etc. The most important instantiation of the notion of complexity is in computability theory, which is the subject of this chapter.

Originally the motivation for studying computational complexity was to understand which algorithms can be used in practice. It had been known that some problems, although algorithmically solvable, require so large a number of steps that they never can be used. It was, therefore, necessary to develop a theory for classifying problems according to their feasibility. When theoretical studies began, it turned out that there are fundamental problems concerning computational complexity. Moreover, some of these problems appeared to be very difficult. We now appreciate their difficulty because only a few of them have been solved after many years.

These problems concern the relationship of the basic resources used by algorithms: time, space, nondeterminism and randomness. Our inability to make any substantial progress in solving them suggests that there may be fundamental obstacles that prevent us from solving them. It is conceivable that these problems not only need new methods, but may need new axioms. This seems to be a rather bold

P. Pudlák, *Logical Foundations of Mathematics and Computational Complexity*, Springer Monographs in Mathematics, DOI 10.1007/978-3-319-00119-7_5,

conjecture, but recall the history of Diophantine equations. The problem appeared to be just a difficult number-theoretical problem and Hilbert even assumed that it was algorithmically solvable. Now we know that this is not the case: there is no theory that would suffice to prove the unsolvability of every unsolvable equation. History may repeat itself in computational complexity and we may need mathematical logic to solve the fundamental problems of computational complexity theory.

In the next chapter, we will see slightly more explicit connections of computational complexity with logic and the foundations of mathematics, mediated by proof complexity.

5.1 What Is Complexity?

From our daily experience we know that there are easy tasks and there are difficult ones. Everybody knows that it is more difficult to multiply two numbers than to add them. Those who use computers more extensively also know that they are able to solve certain problems fast, while some other problems require a long time. But we also know that some people are faster than others, that we can solve a task more easily if we know more about it and that some programs are slow for a given problem, but sometimes a sophisticated program can solve the same problem very efficiently. Thus it is not clear whether there is a particular property of problems that prevents us (and computers) from solving some problems quickly, or if it is just the question of knowing how to solve a particular problem fast.

Therefore the first thing to learn is that, indeed, there is a quantity associated with every problem, which we call the complexity of the problem, that determines how efficiently the problem can be solved. This quantity is represented by a *natural number*. When studying computational complexity, we always consider only algorithmically solvable problems, problems solvable using a finite amount of computational resources. Since algorithms make discrete steps, also the resources can be measured in discrete units. The amount of computational resources needed to solve a particular instance of a problem is this number. In fact there is not only one, but several such quantities corresponding to the type of resources that we study. Furthermore, each one depends on the particular model of computation that we use.

Let us start with the most important type of complexity, which is called *time*. If we use the classical model of computation, Turing machines, then the time complexity of a problem is the minimal number of steps that a Turing machine needs to solve the problem. However, the time complexity of computations cannot be defined for a single input. Recall that when we considered the concept of decidability, it was important to have an infinite *set of instances of a given problem.* Typically, we asked if a property of natural numbers was decidable. For a finite set, there always exists an algorithm—a look up table. So the same is true about complexity; it only makes sense, if we have an infinite, or at least very large set of inputs.

Suppose, for example, that the problem is to decide if a given number N has an even number of prime divisors. The problem is, clearly, decidable: we can enumerate all primes less than N and try to divide N by each of them. This is certainly not the

Name Index

P. Pudlák, *Logical Foundations of Mathematics and Computational Complexity*, Springer Monographs in Mathematics, DOI 10.1007/978-3-319-00119-7,

Subject Index

P. Pudlák, *Logical Foundations of Mathematics and Computational Complexity*, Springer Monographs in Mathematics, DOI 10.1007/978-3-319-00119-7,