Timothy Gowers

MATHEMATICS

A Very Short Introduction

OXFORD
UNIVERSITY PRESS

Contents

Preface

Early in the 20th century, the great mathematician David Hilbert noticed that a number of important mathematical arguments were structurally similar. In fact, he realized that at an appropriate level of generality they could be regarded as the same. This observation, and others like it, gave rise to a new branch of mathematics, and one of its central concepts was named after Hilbert. The notion of a Hilbert space sheds light on so much of modern mathematics, from number theory to quantum mechanics, that if you do not know at least the rudiments of Hilbert space theory then you cannot claim to be a well-educated mathematician.

What, then, is a Hilbert space? In a typical university mathematics course it is defined as a complete inner-product space. Students attending such a course are expected to know, from previous courses, that an inner-product space is a vector space equipped with an inner product, and that a space is complete if every Cauchy sequence in it converges. Of course, for those definitions to make sense, the students also need to know the definitions of vector space, inner product, Cauchy sequence and convergence. To give just one of them (not the longest): a Cauchy sequence is a sequence $x_1, x_2, x_3, \ldots$ such that for every positive number ϵ there exists an integer N such that for any two integers p and q greater than N the distance from x_p to x_q is at most ϵ.

In short, to have any hope of understanding what a Hilbert space is, you must learn and digest a whole hierarchy of lower-level concepts first. Not surprisingly, this takes time and effort. Since the same is true of many of the most important mathematical ideas, there is a severe limit to what can be achieved by any book that attempts to offer an accessible introduction to mathematics, especially if it is to be very short.

Instead of trying to find a clever way round this difficulty, I have focused on a different barrier to mathematical communication. This one, which is more philosophical than technical, separates those who are happy with notions such as infinity, the square root of minus one, the twenty-sixth dimension, and curved space from those who find them disturbingly paradoxical. It is possible to become comfortable with these ideas without immersing oneself in technicalities, and I shall try to show how.

If this book can be said to have a message, it is that one should learn to think abstractly, because by doing so many philosophical difficulties simply disappear. I explain in detail what I mean by the abstract method in Chapter 2. Chapter 1 concerns a more familiar, and related, kind of abstraction: the process of distilling the essential features from a real-world problem, and thereby turning it into a mathematical one. These two chapters, and Chapter 3, in which I discuss what is meant by a rigorous proof, are about mathematics in general.

Thereafter, I discuss more specific topics. The last chapter is more about mathematicians than about mathematics and is therefore somewhat different in character from the others. I recommend reading Chapter 2 before the later ones, but apart from that the book is arranged as unhierarchically as possible: I shall not assume, towards the end of the book, that the reader has understood and remembered everything that comes earlier.

Very little prior knowledge is needed to read this book – a British GCSE course or its equivalent should be enough – but I do presuppose some interest on the part of the reader rather than trying to drum it up myself.

For this reason I have done without anecdotes, cartoons, exclamation marks, jokey chapter titles, or pictures of the Mandelbrot set. I have also avoided topics such as chaos theory and Gödel's theorem, which have a hold on the public imagination out of proportion to their impact on current mathematical research, and which are in any case well treated in many other books. Instead, I have taken more mundane topics and discussed them in detail in order to show how they can be understood in a more sophisticated way. In other words, I have aimed for depth rather than breadth, and have tried to convey the appeal of mainstream mathematics by letting it speak for itself.

I would like to thank the Clay Mathematics Institute and Princeton University for their support and hospitality during part of the writing of the book. I am very grateful to Gilbert Adair, Rebecca Gowers, Emily Gowers, Patrick Gowers, Joshua Katz, and Edmund Thomas for reading earlier drafts. Though they are too intelligent and well informed to count as general readers, it is reassuring to know that what I have written is comprehensible to at least some non-mathematicians. Their comments have resulted in many improvements. To Emily I dedicate this book, in the hope that it will give her a small idea of what it is I do all day.

Chapter 1
Models

How to throw a stone

Suppose that you are standing on level ground on a calm day, and have in your hand a stone which you would like to throw as far as possible. Given how hard you can throw, the most important decision you must make is the angle at which the stone leaves your hand. If this angle is too flat, then although the stone will have a large horizontal speed it will land quite soon and will therefore not have a chance to travel very far. If on the other hand you throw the stone too high, then it will stay in the air for a long time but without covering much ground in the process. Clearly some sort of compromise is needed.

The best compromise, which can be worked out using a combination of Newtonian physics and some elementary calculus, turns out to be as neat as one could hope for under the circumstances: the direction of the stone as it leaves your hand should be upwards at an angle of 45 degrees to the horizontal. The same calculations show that the stone will trace out a parabolic curve as it flies through the air, and they tell you how fast it will be travelling at any given moment after it leaves your hand.

It seems, therefore, that a combination of science and mathematics enables one to predict the entire behaviour of the stone from the

moment it is launched until the moment it lands. However, it does so only if one is prepared to make a number of simplifying assumptions, the main one being that the only force acting on the stone is the earth's gravity and that this force has the same magnitude and direction everywhere. That is not true, though, because it fails to take into account air resistance, the rotation of the earth, a small gravitational influence from the moon, the fact that the earth's gravitational field is weaker the higher you are, and the gradually changing direction of 'vertically downwards' as you move from one part of the earth's surface to another. Even if you accept the calculations, the recommendation of 45 degrees is based on another implicit assumption, namely that the speed of the stone as it leaves your hand does not depend on its direction. Again, this is untrue: one can throw a stone harder when the angle is flatter.

In the light of these objections, some of which are clearly more serious than others, what attitude should one take to the calculations and the predictions that follow from them? One approach would be to take as many of the objections into account as possible. However, a much more sensible policy is the exact opposite: decide what level of accuracy you need, and then try to achieve it as simply as possible. If you know from experience that a simplifying assumption will have only a small effect on the answer, then you should make that assumption.

For example, the effect of air resistance on the stone will be fairly small because the stone is small, hard, and reasonably dense. There is not much point in complicating the calculations by taking air resistance into account when there is likely to be a significant error in the angle at which one ends up throwing the stone anyway. If you want to take it into account, then for most purposes the following rule of thumb is good enough: the greater the air resistance, the flatter you should make your angle to compensate for it.

What is a mathematical model?

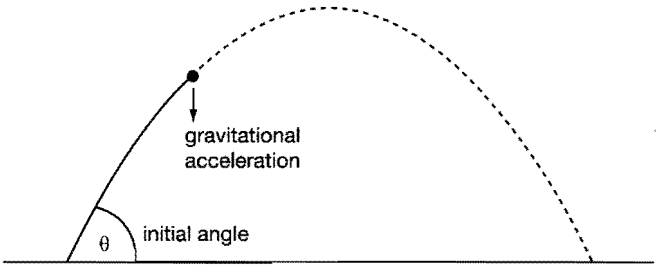
When one examines the solution to a physical problem, it is often, though not always, possible to draw a clear distinction between the contributions made by science and those made by mathematics. Scientists devise a theory, based partly on the results of observations and experiments, and partly on more general considerations such as simplicity and explanatory power. Mathematicians, or scientists doing mathematics, then investigate the purely logical consequences of the theory. Sometimes these are the results of routine calculations that predict exactly the sorts of phenomena the theory was designed to explain, but occasionally the predictions of a theory can be quite unexpected. If these are later confirmed by experiment, then one has impressive evidence in favour of the theory.

The notion of confirming a scientific prediction is, however, somewhat problematic, because of the need for simplifications of the kind I have been discussing. To take another example, Newton's laws of motion and gravity imply that if you drop two objects from the same height then they will hit the ground (if it is level) at the same time. This phenomenon, first pointed out by Galileo, is somewhat counter-intuitive. In fact, it is worse than counter-intuitive: if you try it for yourself, with, say, a golf ball and a table-tennis ball, you will find that the golf ball lands first. So in what sense was Galileo correct?

It is, of course, because of air resistance that we do not regard this little experiment as a refutation of Galileo's theory: experience shows that the theory works well when air resistance is small. If you find it too convenient to let air resistance come to the rescue every time the predictions of Newtonian mechanics are mistaken, then your faith in science, and your admiration for Galileo, will be restored if you get the chance to watch a feather fall in a vacuum – it really does just drop as a stone would.

Nevertheless, because scientific observations are never completely direct and conclusive, we need a better way to describe the relationship between science and mathematics. Mathematicians do not apply scientific theories directly to the world but rather to *models*. A model in this sense can be thought of as an imaginary, simplified version of the part of the world being studied, one in which exact calculations are possible. In the case of the stone, the relationship between the world and the model is something like the relationship between Figures 1 and 2.

1. A ball in flight I



2. A ball in flight II

There are many ways of modelling a given physical situation, and we must use a mixture of experience and further theoretical considerations to decide what a given model is likely to teach us about the world itself. When choosing a model, one priority is to make its behaviour correspond closely to the actual, observed behaviour of the world. However, other factors, such as simplicity and mathematical elegance, can often be more important. Indeed, there are very useful models with almost no resemblance to the world at all, as some of my examples will illustrate.

Rolling a pair of dice

If I roll a pair of dice and want to know how they will behave, then experience tells me that there are certain questions it is unrealistic to ask. For example, nobody could be expected to tell me the outcome of a given roll in advance, even if they had expensive technology at their disposal and the dice were to be rolled by a machine. By contrast, questions of a probabilistic nature, such as, 'How likely is it that the numbers on the dice will add up to seven?' can often be answered, and the answers may be useful if, for example, I am playing backgammon for money. For the second sort of question, one can model the situation very simply by representing a roll of the dice as a random choice of one of the following thirty-six pairs of numbers.

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

The first number in each pair represents the number showing on the first die, and the second the number on the second. Since exactly six of the pairs consist of two numbers that add up to seven, the chances of rolling a seven are six in thirty-six, or one in six.

One might object to this model on the grounds that the dice, when rolled, are obeying Newton's laws, at least to a very high degree of precision, so the way they land is anything but random: indeed, it could in principle be calculated. However, the phrase 'in principle' is being overworked here, since the calculations would be extraordinarily complicated, and would need to be based on more precise information about the shape, composition, initial velocities, and rotations of the dice than could ever be measured in practice. Because of this, there is no advantage whatsoever in using some more complicated deterministic model.

Predicting population growth

The 'softer' sciences, such as biology and economics, are full of mathematical models that are vastly simpler than the phenomena they represent, or even deliberately inaccurate in certain ways, but nevertheless useful and illuminating. To take a biological example of great economic importance, let us imagine that we wish to predict the population of a country in 20 years' time. One very simple model we might use represents the entire country as a pair of numbers $(t, P(t))$. Here, t represents the time and $P(t)$ stands for the size of the population at time t. In addition, we have two numbers, b and d, to represent birth and death rates. These are defined to be the number of births and deaths per year, as a proportion of the population.

Suppose we know that the population at the beginning of the year 2002 is P. According to the model just defined, the number of births and deaths during the year will be bP and dP respectively, so the population at the beginning of 2003 will be $P + bP - dP = (1 + b - d)P$. This argument works for any year, so we have the formula $P(n + 1) = (1 + b - d)P(n)$, meaning that the population at the beginning of year $n + 1$ is $(1 + b - d)$ times the population at the beginning of year n. In other words, each year the population multiplies by $(1 + b - d)$. It follows that in 20 years

it multiplies by $(1 + b - d)^{20}$, which gives an answer to our original question.

Even this basic model is good enough to persuade us that if the birth rate is significantly higher than the death rate, then the population will grow extremely rapidly. However, it is also unrealistic in ways that can make its predictions very inaccurate. For example, the assumption that birth and death rates will remain the same for 20 years is not very plausible, since in the past they have often been affected by social changes and political events such as improvements in medicine, new diseases, increases in the average age at which women start to have children, tax incentives, and occasional large-scale wars. Another reason to expect birth and death rates to vary over time is that the ages of people in the country may be distributed rather unevenly. For example, if there has been a baby boom 15 years earlier, then there is some reason to expect the birth rate to rise in 10 to 15 years' time.

It is therefore tempting to complicate the model by introducing other factors. One could have birth and death rates $b(t)$ and $d(t)$ that varied over time. Instead of a single number $P(t)$ representing the size of the population, one might also like to know how many people there are in various age groups. It would also be helpful to know as much as possible about social attitudes and behaviour in these age groups in order to predict what future birth and death rates are likely to be. Obtaining this sort of statistical information is expensive and difficult, but the information obtained can greatly improve the accuracy of one's predictions. For this reason, no single model stands out as better than all others. As for social and political changes, it is impossible to say with any certainty what they will be. Therefore the most that one can reasonably ask of any model is predictions of a conditional kind: that is, ones that tell us what the effects of social and political changes will be if they happen.

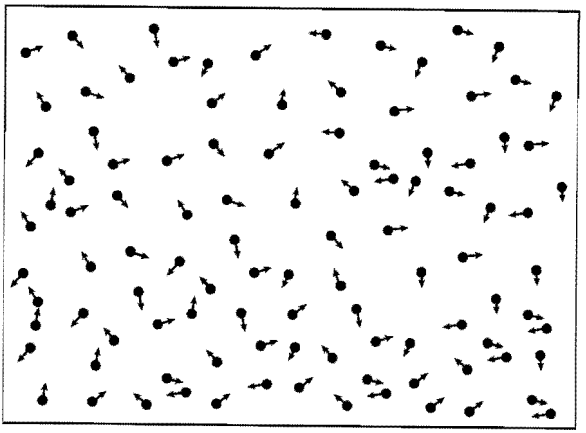
The behaviour of gases

According to the kinetic theory of gases, introduced by Daniel Bernoulli in 1738 and developed by Maxwell, Boltzmann, and others in the second half of the 19th century, a gas is made up of moving molecules, and many of its properties, such as temperature and pressure, are statistical properties of those molecules. Temperature, for example, corresponds to their average speed.

With this idea in mind, let us try to devise a model of a gas contained in a cubical box. The box should of course be represented by a cube (that is, a mathematical rather than physical one), and since the molecules are very small it is natural to represent them by points in the cube. These points are supposed to move, so we must decide on the rules that govern how they move. At this point we have to make some choices.

If there were just one molecule in the box, then there would be an obvious rule: it travels at constant speed, and bounces off the walls of the box when it hits them. The simplest conceivable way to generalize this model is then to take N molecules, where N is some large number, and assume that they all behave this way, with absolutely no interaction between them. In order to get the N-molecule model started, we have to choose initial positions and velocities for the molecules, or rather the points representing them. A good way of doing this is to make the choice randomly, since we would expect that at any given time the molecules in a real gas would be spread out and moving in many directions.

It is not hard to say what is meant by a random point in the cube, or a random direction, but it is less clear how to choose a speed randomly, since speed can take any value from 0 to infinity. To avoid this difficulty, let us make the physically implausible assumption that all the molecules are moving at the same speed, and that it is only the initial positions and directions that are chosen randomly. A



3. A two-dimensional model of a gas

two-dimensional version of the resulting model is illustrated in Figure 3.

The assumption that our N molecules move entirely independently of one another is quite definitely an oversimplification. For example, it means that there is no hope of using this model to understand why a gas becomes a liquid at sufficiently low temperatures: if you slow down the points in the model you get the same model, but running more slowly. Nevertheless, it does explain much of the behaviour of real gases. For example, imagine what would happen if we were gradually to shrink the box. The molecules would continue to move at the same speed, but now, because the box was smaller, they would hit the walls more often and there would be less wall to hit. For these two reasons, the number of collisions per second in any given area of wall would be greater. These collisions account for the pressure that a gas exerts, so we can conclude that if you squeeze a gas into a smaller volume, then its pressure is likely to increase – as is confirmed by observation. A similar argument explains why, if you increase the temperature of a gas without increasing its volume, its pressure also increases.

And it is not too hard to work out what the numerical relationships between pressure, temperature, and volume should be.

The above model is roughly that of Bernoulli. One of Maxwell's achievements was to discover an elegant theoretical argument that solves the problem of how to choose the initial speeds more realistically. To understand this, let us begin by dropping our assumption that the molecules do not interact. Instead, we shall assume that from time to time they collide, like a pair of tiny billiard balls, after which they go off at other speeds and in other directions that are subject to the laws of conservation of energy and momentum but otherwise random. Of course, it is not easy to see how they will do this if they are single points occupying no volume, but this part of the argument is needed only as an informal justification for some sort of randomness in the speeds and directions of the molecules. Maxwell's two very plausible assumptions about the nature of this randomness were that it should not change over time and that it should not distinguish between one direction and another. Roughly speaking, the second of these assumptions means that if d_1 and d_2 are two directions and s is a certain speed, then the chances that a particle is travelling at speed s in direction d_1 are the same as the chances that it is travelling at speed s in direction d_2. Surprisingly, these two assumptions are enough to determine exactly how the velocities should be distributed. That is, they tell us that if we want to choose the velocities randomly, then there is only one natural way to do it. (They should be assigned according to the normal distribution. This is the distribution that produces the famous 'bell curve', which occurs in a large number of different contexts, both mathematical and experimental.)

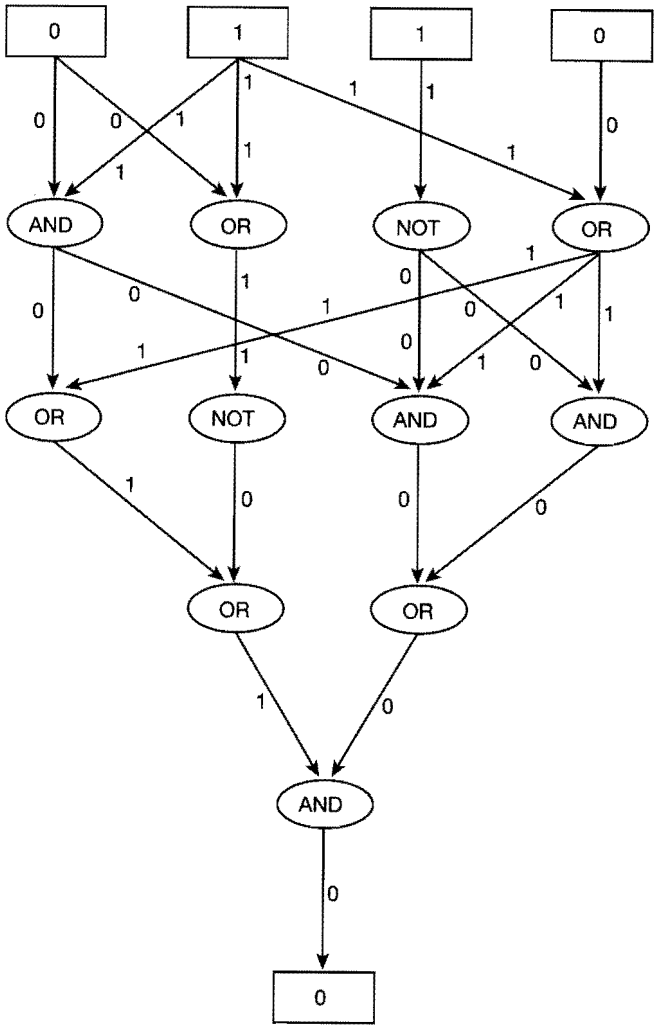
Once we have chosen the velocities, we can again forget all about interactions between the molecules. As a result, this slightly improved model shares many of the defects of the first one. In order to remedy them, there is no choice but to model the interactions

somehow. It turns out that even very simple models of systems of interacting particles behave in a fascinating way and give rise to extremely difficult, indeed mostly unsolved, mathematical problems.

Modelling brains and computers

A computer can also be thought of as a collection of many simple parts that interact with one another, and largely for this reason theoretical computer science is also full of important unsolved problems. A good example of the sort of question one might like to answer is the following. Suppose that somebody chooses two prime numbers p and q, multiplies them together and tells you the answer pq. You can then work out what p and q are by taking every prime number in turn and seeing whether it goes exactly into pq. For example, if you are presented with the number 91, you can quickly establish that it is not a multiple of 2, 3, or 5, and then that it equals 7×13.

If, however, p and q are very large – with 200 digits each, say – then this process of trial and error takes an unimaginably long time, even with the help of a powerful computer. (If you want to get a feel for the difficulty, try finding two prime numbers that multiply to give 6901 and another two that give 280123.) On the other hand, it is not inconceivable that there is a much cleverer way to approach the problem, one that might be used as the basis for a computer program that does not take too long to run. If such a method could be found, it would allow one to break the codes on which most modern security systems are based, on the Internet and elsewhere, since the difficulty of deciphering these codes depends on the difficulty of factorizing large numbers. It would therefore be reassuring if there were some way of showing that a quick, efficient procedure for calculating p and q from their product pq does not exist. Unfortunately, while computers continually surprise us with what they can be used for, almost nothing is known about what they cannot do.



4. A primitive computer program

Before one can begin to think about this problem one must find some way of representing a computer mathematically, and as simply as possible. Figure 4 shows one of the best ways of doing this. It consists of layers of nodes that are linked to one another by lines that are called edges. Into the top layer goes the 'input', which is a sequence of 0s and 1s, and out of the bottom layer comes the 'output', which is another sequence of 0s and 1s. The nodes are of three kinds, called AND, OR, and NOT gates. Each of these gates receives some 0s and 1s from the edges that enter it from above. Depending on what it receives, it then sends out 0s or 1s itself, according to the following simple rules: if an AND gate receives nothing but 1s then it sends out 1s, and otherwise it sends out 0s; if an OR gate receives nothing but 0s then it sends out 0s, and otherwise it sends out 1s; only one edge is allowed to enter a NOT gate from above, and it sends out 1s if it receives a 0 and 0s if it receives a 1.

An array of gates linked by edges is called a *circuit*, and what I have described is the circuit model of computation. The reason 'computation' is an appropriate word is that a circuit can be thought of as taking one sequence of 0s and 1s and transforming it into another, according to some predetermined rules which may, if the circuit is large, be very complicated. This is also what computers do, although they translate these sequences out of and into formats that we can understand, such as high-level programming languages, windows, icons, and so on. There turns out to be a fairly simple way (from a theoretical point of view – it would be a nightmare to do in practice) of converting any computer program into a circuit that transforms 01-sequences according to exactly the same rules. Moreover, important characteristics of computer programs have their counterparts in the resulting circuits.

In particular, the number of nodes in the circuit corresponds to the length of time the computer program takes to run. Therefore, if one can show that a certain way of transforming 01-sequences needs a very large circuit, then one has also shown that it needs a computer

program that runs for a very long time. The advantage of using the circuit model over analysing computers directly is that, from the mathematical point of view, circuits are simpler, more natural, and easier to think about.

A small modification to the circuit model leads to a useful model of the brain. Now, instead of 0s and 1s, one has signals of varying strengths that can be represented as numbers between 0 and 1. The gates, which correspond to neurons, or brain cells, are also different, but they still behave in a very simple way. Each one receives some signals from other gates. If the total strength of these signals – that is, the sum of all the corresponding numbers – is large enough, then the gate sends out its own signals of certain strengths. Otherwise, it does not. This corresponds to the decision of a neuron whether or not to 'fire'.

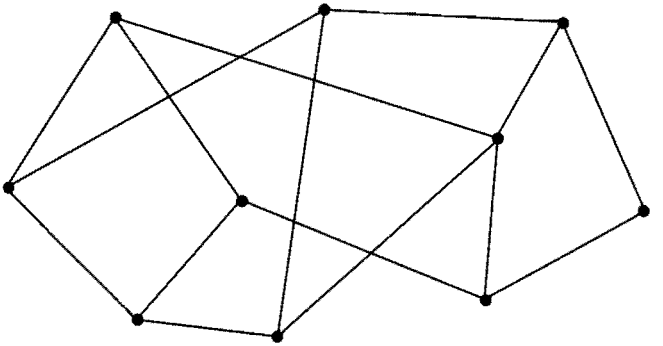
It may seem hard to believe that this model could capture the full complexity of the brain. However, that is partly because I have said nothing about how many gates there should be or how they should be arranged. A typical human brain contains about 100 billion neurons arranged in a very complicated way, and in the present state of knowledge about the brain it is not possible to say all that much more, at least about the fine detail. Nevertheless, the model provides a useful theoretical framework for thinking about how the brain might work, and it has allowed people to simulate certain sorts of brain-like behaviour.

Colouring maps and drawing up timetables

Suppose that you are designing a map that is divided into regions, and you wish to choose colours for the regions. You would like to use as few colours as possible, but do not wish to give two adjacent regions the same colour. Now suppose that you are drawing up the timetable for a university course that is divided into modules. The number of possible times for lectures is limited, so some modules will have to clash with others. You have a list of which students are

taking which modules, and would like to choose the times in such a way that two modules clash only when there is nobody taking both.

These two problems appear to be quite different, but an appropriate choice of model shows that from the mathematical point of view they are the same. In both cases there are some objects (countries, modules) to which something must be assigned (colours, times). Some pairs of objects are incompatible (neighbouring countries, modules that must not clash) in the sense that they are not allowed to receive the same assignment. In neither problem do we really care what the objects are or what is being assigned to them, so we may as well just represent them as points. To show which pairs of points are incompatible we can link them with lines. A collection of points, some of which are joined by lines, is a mathematical structure known as a graph. Figure 5 gives a simple example. It is customary to call the points in a graph vertices, and the lines edges.



5. A graph with 10 vertices and 15 edges

Once we have represented the problems in this way, our task in both cases is to divide the vertices into a small number of groups in such a way that no group contains two vertices linked by an edge. (The graph in Figure 5 can be divided into three such groups, but not

into two.) This illustrates another very good reason for making models as simple as possible: if you are lucky, the same model can be used to study many different phenomena at once.

Various meanings of the word 'abstract'

When devising a model, one tries to ignore as much as possible about the phenomenon under consideration, abstracting from it only those features that are essential to understanding its behaviour. In the examples I have discussed, stones were reduced to single points, the entire population of a country to one number, the brain to a network of gates obeying very simple mathematical rules, and the interactions between molecules to nothing at all. The resulting mathematical structures were abstract representations of the concrete situations being modelled.

These are two senses in which mathematics is an abstract subject: it abstracts the important features from a problem and it deals with objects that are not concrete and tangible. The next chapter will discuss a third, deeper sense of abstraction in mathematics, of which the example of the previous section has already given us some idea. A graph is a very flexible model with many uses. However, when one studies graphs, there is no need to bear these uses in mind: it does not matter whether the points represent regions, lectures, or something quite different again. A graph theorist can leave behind the real world entirely and enter the realm of pure abstraction.