OXFORD

DISCRETE MATHEMATICS

second edition

...it is a wonderful book. Biggs' expository style is of the highest quality.
Professor James Reid, University of Mississippi

...a well written book by a world-renowned expert. The explanations go to the heart of the matter and the proofs given are elegant.
Professor Mohan Shrikhande, University of Michigan

...the material is well written in a clear and lucid style...
Dr Jim Renshaw, University of Southampton

...the new chapters are elegantly written.
Dr Peter Robinson, University of Cambridge

NORMAN L. BIGGS

www.oup.com/mathematics/discretemath

Now with fully worked solutions on the web!

4

Exercises 4.4

1 Use the principle of induction to prove that

$$1^2 + 2^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$$

for all natural numbers n.

2 Use the method of induction to show that the following statement is true for all natural numbers n:

$$\sum_{r=1}^{n} r(r+2)(r+4) = \frac{1}{4}n(n+1)(n+4)(n+5).$$

3 Guess a formula for the sum

$$\sum_{r=1}^{n} (br + c),$$

where b and c are given numbers, and prove it by using the principle of induction.

4 Using Exercises 1 and 3, write down the formula for

$$\sum_{r=1}^{n} (ar^2 + br + c),$$

where a, b, c are any given numbers.

4.5 Recursive definitions

Suppose that after one year of gambling on the National Lottery you have made a net loss of two pounds, and thereafter your losses double each year. Then your losses for the first few years could be calculated as follows:

$$s_1 = 2, \quad s_2 = 2 \times s_1 = 4, \quad s_3 = 2 \times s_2 = 8, \quad s_4 = 2 \times s_3 = 16,$$
$$s_5 = 2 \times s_4 = 32, \quad s_6 = 2 \times s_5 = 64, \quad s_7 = 2 \times s_6 = 128.$$

The successive numbers are given by the rule $s_{n+1} = 2s_n$. This kind of a definition is known as a *recursive definition.*

We could also say that your Lottery losses are given by the formula $s_n = 2^n$. Although we usually think of 2^n as meaning 'the product of n twos', this is actually just a rather sloppy version of the correct definition, which is recursive:

$$2^1 = 2, \qquad 2^{n+1} = 2 \times 2^n \quad (n \geq 1).$$

The following examples show that this method of definition is very common.

Example 1 The *factorial* of n, written as $n!$ is obtained by multiplying together all the natural numbers from 1 to n. It is easy to see what this means when n has a specific value, such as $n = 6$:

$$6! = 1 \times 2 \times 3 \times 4 \times 5 \times 6 = 720.$$

When we try to be explicit about what it means for a general value of n, we end up by writing something such as $n! = 1 \times 2 \times \ldots \times n$. The ... notation is very useful, and we shall use it frequently. But it is important to realize that its meaning is not self-evident; in fact the 'three dots' are a shorthand for a recursive definition. The correct way to define $n!$ is

$$1! = 1, \qquad (n+1)! = (n+1) \times n! \quad (n \geq 1).$$

□

Example 2 Another example of a recursive definition is the *sigma notation* introduced in the previous section. We *define* the sum

$$\sum_{r=1}^{n} a_r \quad \text{or equivalently} \quad a_1 + a_2 + \cdots + a_n$$

as follows:

$$\sum_{r=1}^{1} a_r = a_1, \qquad \sum_{r=1}^{n+1} a_r = a_{n+1} + \sum_{r=1}^{n} a_r \quad (n \geq 2).$$

□

Example 3 Suppose we define

$$f_1 = 1, \quad f_2 = 1, \quad f_{n+1} = f_n + f_{n-1} \quad (n \geq 2).$$

To calculate the first few values of f_n we proceed as follows:

$$f_3 = f_2 + f_1 = 2 + 1 = 3,$$
$$f_4 = f_3 + f_2 = 3 + 2 = 5,$$
$$f_5 = f_4 + f_3 = 5 + 3 = 8,$$

and so on. These numbers are very famous: they are called the *Fibonacci numbers*, and they have many interesting properties (see Ex. 4.6.4, for example). The first fifteen Fibonacci numbers are:

1 1 2 3 5 8 13 21 34 55 89 144 233 377 610. □

Exercises 4.5

1 Calculate the values of u_1, u_2, u_3, u_4, and u_5 given by the recursive definition

$$u_1 = 1, \qquad u_2 = 1, \qquad u_{n+1} = u_n + 2u_{n-1} \quad (n \geq 2).$$

2 Write down formulae for the numbers u_n defined by the following equations.

(i) $u_1 = 1, \qquad u_{n+1} = u_n + 3 \qquad (n \geq 1).$

(ii) $u_1 = 1, \qquad u_{n+1} = (n+1)^2 u_n \qquad (n \geq 1).$

3 Natural numbers q_n are defined by the rule:

$$q_1 = 2, \quad q_{n+1} = q_n^2 - q_n + 1.$$

Find a counter-example to the statement that q_n is a prime for all $n \geq 1$.

4 Suppose that t_n is defined by the equations

$$t_1 = 2, \quad t_{n+1} = 2^{t_n} \quad (n \geq 1).$$

What is the greatest value of n for which you can evaluate t_n using your calculator?

4.6 Other forms of the principle of induction

Consider the statement $n^2 > 7n + 1$. Clearly, it is false when $n = 1$, but true for a value such as $n = 10$. Calculating more systematically:

n	1	2	3	4	5	6	7	8	9	10
n^2	1	4	9	16	25	36	49	64	81	100
$7n+1$	8	15	22	29	36	43	50	57	64	71.

10

Exercises 10.2

1 In Dr Cynthia Angst's Calculus class, 32 of the students are boys. Each boy knows five of the girls in the class and each girl knows eight of the boys. How many girls are in the class?

2 Suppose we have a number of different subsets of $\mathbb{N}_8$, with the property that each one has four members, and each member of $\mathbb{N}_8$ belongs to exactly three of the subsets. How many subsets are there? Write down a collection of subsets which satisfies the conditions.

3 Is it possible to find a collection of subsets of $\mathbb{N}_8$ such that each one has three members and each member of $\mathbb{N}_8$ belongs to exactly five of the subsets?

4 If $X_1, X_2, \ldots, X_n$ are sets, the **product set**

$$X_1 \times X_2 \times \cdots \times X_n$$

is defined to be the set of all ordered n-tuples $(x_1, x_2, \ldots, x_n)$, with $x_i \in X_i (1 \leq i \leq n)$. Use the principle of induction to prove that

$$|X_1 \times X_2 \times \cdots \times X_n| = |X_1| \times |X_2| \times \cdots \times |X_n|.$$

5 In a simple language there are the usual 26 letters and all words have four letters. Any arrangement of the letters, including repetitions, is allowed. How many words are there? How many of them do not contain the letter b?

10.3 Euler's function

In this section we shall prove an important and useful theorem using only the most basic counting principles.

The theorem is concerned with the divisibility properties of integers. Recall that two integers x and y are *coprime* if $\gcd(x, y) = 1$; and for each $n \geq 1$ let $\phi(n)$ denote the number of integers x in the range $1 \leq x \leq n$ such that x and n are coprime. We can calculate the first few values of $\phi(n)$ by making a table (Table 10.3.1).

Table 10.3.1

n	Coprime to n	$\phi(n)$
1	1	1
2	1	1
3	1, 2	2
4	1, 3	2
5	1, 2, 3, 4	4
6	1, 5	2
7	1, 2, 3, 4, 5, 6	6
8	1, 3, 5, 7	4

The function ϕ is called **Euler's function** after Leonhard Euler (1707–1783). When n is a prime, say $n = p$, each one of the integers $1, 2, \ldots, p-1$ is coprime to p, so we have

$$\phi(p) = p - 1 \quad (p \text{ prime}).$$

An explicit general formula for $\phi(n)$ will be given in Theorem 11.5.1.

Our present task is to prove a result concerning the sum of the values $\phi(d)$ taken over all divisors d of a given positive integer n.

For example, when $n = 12$ the divisors d are 1, 2, 3, 4, 6, and 12, and we find that

$$\begin{aligned} &\phi(1) + \phi(2) + \phi(3) + \phi(4) + \phi(6) + \phi(12) \\ &= 1 + 1 + 2 + 2 + 2 + 4 \\ &= 12. \end{aligned}$$

We shall show that the sum is always equal to the given integer n.

Theorem 10.3 For any positive integer n,

$$\sum_{d|n} \phi(d) = n.$$

Proof Let S denote the set of pairs of integers (d, f) satisfying

$$d \mid n, \qquad 1 \leq f \leq d, \qquad \gcd(f, d) = 1.$$

Table 10.3.2 illustrates S when $n = 12$; the 'mark' indicating that (d, f) is in S is a number whose significance will appear in a moment. In general, the number of 'marks' in row d is just the number of f in the range $1 \leq f \leq d$ satisfying $\gcd(f, d) = 1$; that is, $\phi(d)$. Hence, by counting S by the row method we obtain

$$|S| = \sum_{d|n} \phi(d).$$

Table 10.3.2

	f												
d	1	2	3	4	5	6	7	8	9	10	11	12	$\phi(d)$
1	12												1
2	6												1
3	4	8											2
4	3		9										2
6	2				10								2
12	1				5		7				11		4
													12

In order to show that $|S| = n$ we shall construct a bijection β from S to $\mathbb{N}_n$. Given a pair (d, f) in S, we define

$$\beta(d, f) = fn/d.$$

In the table, $\beta(d, f)$ is the 'mark' in row d and column f. Since $d \mid n$ the value of β is an integer, and since $1 \leq f \leq d$ it lies in $\mathbb{N}_n$.

To show that β is an injection we remark that

$$\beta(d, f) = \beta(d', f') \quad \Rightarrow \quad fn/d = f'n/d' \quad \Rightarrow \quad fd' = f'd.$$

But f and d are coprime, as are f' and d', so we can conclude that $d = d'$ and $f = f'$.

To show that β is a surjection, suppose we are given x in $\mathbb{N}_n$. Let g_x denote the gcd of x and n, and let

$$d_x = n/g_x, \qquad f_x = x/g_x.$$

Since g_x is a divisor of x and n both d_x and f_x are integers, and since it is the gcd, d_x and f_x are coprime. Now

$$\beta(d_x, f_x) = f_x n/d_x = x,$$

and so β is a surjection.

Thus β is a bijection and $|S| = n$, as required. □

Exercises 10.3

1 Find the values of $\phi(19)$, $\phi(20)$, $\phi(21)$.

2 Show that if x and n are coprime, so are $n - x$ and n. Deduce that $\phi(n)$ is even for all $n \geq 3$.

3 Show that, if p is a prime and m is a positive integer, then an integer x in the range $1 \leq x \leq p^m$ is *not* coprime to p^m if and only if it is a multiple of p. Deduce that $\phi(p^m) = p^m - p^{m-1}$.

4 Find a counter-example which disproves the conjecture that $\phi(a)\phi(b) = \phi(ab)$ for any positive integers a and b. Try to modify the conjecture so that it cannot be disproved.

10.4 Functions, words, and selections

We shall consider functions (not necessarily bijections) defined on a set of positive integers $\mathbb{N}_m$, and with values in a given set Y. The values of such a function f determine an m-tuple

$$(f(1), f(2), \ldots, f(m))$$

of elements of Y. According to the general definition of a product set (Ex. 10.2.4) this m-tuple belongs to the set $Y \times Y \times \cdots \times Y$ (m factors), which we shall denote by Y^m. Each element of Y^m is an m-tuple $(y_1, y_2, \ldots, y_m)$ and corresponds to a function f from $\mathbb{N}_m$ to Y defined by the equations

$$f(1) = y_1, \qquad f(2) = y_2, \ldots, f(m) = y_m.$$

These remarks lead us to the conclusion that a function from $\mathbb{N}_m$ to Y is logically the same thing as an element of the product set Y^m.

There is another way of looking at this relationship, which is very useful in practice. If we think of the members of Y as the letters of an alphabet, then the sequence $f(1), f(2), \ldots, f(m)$ can be regarded as the m letters of a word. For

selected, and so $i(m)$ must be one of the remaining $n - (m - 1)$ objects. Hence the total number is as stated. □

For example, if we have a pool of 16 players the number of ways of selecting a batting order for a baseball team of nine is

$$16 \times 15 \times 14 \times 13 \times 12 \times 11 \times 10 \times 9 \times 8 = 4\,151\,347\,200.$$

Exercises 10.5

1 In how many ways can we select a batting order of 11 from a pool of 14 cricketers?

2 How many four-letter words can be made from an alphabet of 10 symbols if there are no restrictions on spelling except that no letter can be used more than once?

3 Explain briefly how you would make a systematic list of all the ordered selections, without repetition, of three things from the set $\{a, b, c, d, e, f\}$.

4 Let $(n)_m = n(n-1)\cdots(n-m+1)$. By interpreting the result in terms of ordered selections, show that

$$(n)_m \times (n-m)_{r-m} = (n)_r$$

for any positive integers satisfying $n > r > m$.

10.6 Permutations

A **permutation** of a non-empty finite set X is a bijection from X to X. Frequently we take X to be $\mathbb{N}_n = \{1, 2, \ldots, n\}$. For example, a typical permutation of $\mathbb{N}_5$ is the function α defined by the equations

$$\alpha(1) = 2, \quad \alpha(2) = 4, \quad \alpha(3) = 5, \quad \alpha(4) = 1, \quad \alpha(5) = 3.$$

A bijection from a finite set to itself is necessarily an injection, and conversely any such injection is a bijection (Ex. 6.7.2). Thus the number of permutations of an n-set is the same as the number of injections from $\mathbb{N}_n$ to itself, and by Theorem 10.5 this number is

$$n \times (n-1) \times \cdots \times 1 = n!.$$

We shall denote the set of all permutations of $\mathbb{N}_n$ by S_n. For example, S_3 contains the following $3! = 6$ permutations:

$$\begin{array}{ccc} 1\ 2\ 3 & 1\ 2\ 3 & 1\ 2\ 3 \\ \downarrow\ \downarrow\ \downarrow & \downarrow\ \downarrow\ \downarrow & \downarrow\ \downarrow\ \downarrow \\ 1\ 2\ 3 & 1\ 3\ 2 & 2\ 1\ 3 \\ \\ 1\ 2\ 3 & 1\ 2\ 3 & 1\ 2\ 3 \\ \downarrow\ \downarrow\ \downarrow & \downarrow\ \downarrow\ \downarrow & \downarrow\ \downarrow\ \downarrow \\ 2\ 3\ 1 & 3\ 1\ 2 & 3\ 2\ 1. \end{array}$$

In practice, we usually assign some concrete interpretation to an element of S_n. As in the previous section we can use the interpretation as an 'ordered selection without repetition' where, in this case, we select the members of $\{1, 2, \ldots, n\}$ in some order until there are none left. A related interpretation is that a permutation

effects a rearrangement of $\{1, 2, \ldots, n\}$; for example, the permutation α given above effects the rearrangement of 12345 into 24513 as follows:

$$\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 4 & 5 & 1 & 3. \end{array}$$

In some circumstances it is convenient to regard a permutation and the corresponding rearrangement as the same thing, but this can lead to difficulties when we have to consider successive rearrangements. By and large, it is advisable to remember that

a permutation is a kind of a function.

When permutations are treated as functions it is clear how they should be combined. Let us take α to be the permutation of $\mathbb{N}_5$ specified above, and suppose β is the permutation of $\mathbb{N}_5$ given by

$$\beta(1) = 3, \qquad \beta(2) = 5, \qquad \beta(3) = 1, \qquad \beta(4) = 4, \qquad \beta(5) = 2.$$

The composite function $\beta\alpha$ is the permutation defined by $\beta\alpha(i) = \beta(\alpha(i))$ $(1 \leq i \leq 5)$, that is

$$\beta\alpha(1) = 5, \quad \beta\alpha(2) = 4, \quad \beta\alpha(3) = 2, \quad \beta\alpha(4) = 3, \quad \beta\alpha(5) = 1.$$

(Recall that, as always, $\beta\alpha$ means 'first α, then β'.) In terms of rearrangements we have

$$\begin{array}{cccccc} & 1 & 2 & 3 & 4 & 5 \\ \alpha & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ & 2 & 4 & 5 & 1 & 3 \\ \beta & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ & 5 & 4 & 2 & 3 & 1. \end{array}$$

There are four features of the composition of permutations which are of paramount importance, and they are listed in the next theorem. In Part IV of the course we shall review these properties in a more general context.

Theorem 10.6 The following properties hold in the set S_n of all permutations of $\{1, 2, \ldots, n\}$.

(i) If π and σ are in S_n, so is $\pi\sigma$.

(ii) For any permutations π, σ, τ in S_n,

$$(\pi\sigma)\tau = \pi(\sigma\tau).$$

(iii) The identity function, denoted by id and defined by $\mathrm{id}(r) = r$ for all r in $\mathbb{N}_n$, is a permutation and for any σ in S_n we have

$$\mathrm{id}\,\sigma = \sigma\,\mathrm{id} = \sigma.$$

(iv) For every permutation π in S_n there is an inverse permutation π^{-1} in S_n such that

$$\pi\pi^{-1} = \pi^{-1}\pi = \mathrm{id}.$$

Proof Statement (i) follows immediately from the fact that the composite of two bijections is a bijection (Theorem 5.3), and statement (ii) is a standard property of composition (Ex. 5.3.3). Statement (iii) is plainly true, and statement (iv) follows from the fact that every bijection has an inverse (Theorem 5.4). □

It is convenient to have a more compact notation for permutations. Consider once again the permutation α of $\{1, 2, 3, 4, 5\}$, and note in particular that

$$\alpha(1) = 2, \quad \alpha(2) = 4, \quad \alpha(4) = 1.$$

Thus α takes 1 to 2, 2 to 4, and 4 back to 1, and for this reason we say that the symbols 1, 2, 4 form a *cycle* (of length 3). Similarly, the symbols 3 and 5 form a cycle of length 2, and we write

$$\alpha = (1\,2\,4)(3\,5).$$

This is the *cycle notation* for α. Any permutation π can be written in cycle notation in the following manner:

begin with any symbol (say 1) and trace the effect of π on it and its successors until we reach 1 again, so that we have a cycle;

choose a symbol not already dealt with and construct the cycle derived from it;

repeat the procedure until every symbol has been dealt with.

For example, the permutation β defined above has the cycle notation

$$\beta = (13)(25)(4),$$

where we note especially that the symbol 4 forms a 'degenerate' cycle on its own, since $\beta(4) = 4$. In some circumstances we can omit such cycles of length 1 when writing a permutation in cycle notation, since they correspond to symbols which are not affected by the permutation. However, it is usually helpful *not* to adopt this convention until one is quite familiar with the notation.

Although the representation of a permutation in cycle notation is essentially unique, there are two obvious ways in which we can change the notation without altering the permutation. First, each cycle can begin with any one of its symbols—for example (7 8 2 1 3) and (1 3 7 8 2) describe the same cycle. Secondly, the order of the cycles is unimportant—for example (124)(35) and (35)(124) denote the same permutation. But the important features are the numbers and lengths of the cycles, and the disposition of the symbols within the cycles, and these are uniquely determined. Consequently, the cycle notation tells us many useful things about a permutation.

Example Cards numbered 1 to 12 are laid out in the manner shown towards the left-hand side below. They are picked up in row order and re-dealt in the same array, but by columns rather than rows, so that they appear as seen on the right-hand side.

$$\begin{array}{ccc} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{array} \qquad \begin{array}{ccc} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{array}$$

How many times must this procedure be carried out before the cards reappear in their original positions?

Solution Let π be the permutation which effects the rearrangement; that is $\pi(i) = j$ if card j appears in the position previously occupied by card i. On working out the cycle notation for π we find

$$\pi = (1)(2\ 5\ 6\ 10\ 4)(3\ 9\ 11\ 8\ 7)(12).$$

The degenerate cycles (1) and (12) indicate that cards 1 and 12 never change their positions. The other cycles have length 5, so after the procedure has been carried out five times all the cards will reappear in their original positions. (Try it.) Another way of expressing the result is to say that $\pi^5 = \mathrm{id}$, where π^5 signifies the five-fold repetition of π. □

Exercises 10.6

1 Write down the cycle notation for the permutation which effects the rearrangement

$$\begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & 5 & 7 & 8 & 4 & 6 & 1 & 2 & 9. \end{array}$$

2 Let σ and τ be the permutations of $\{1, 2, \ldots, 8\}$ whose representations in cycle notation are

$$\sigma = (1\ 2\ 3)(4\ 5\ 6)(7\ 8), \qquad \tau = (1\ 3\ 5\ 7)(2\ 6)(4)(8).$$

Write down the cycle notations for $\sigma\tau$, $\tau\sigma$, σ^2, σ^{-1}, τ^{-1}.

3 Solve the problem posed in the *Example* when there are 20 cards arranged in five rows of four.

4 Show that there are just three members of S_4 which have two cycles of length 2 when written in cycle notation.

5 Let K denote the subset of S_4 which contains the identity permutation id and the three permutations α_1, α_2, α_3 described in the previous exercise. Write out the 'multiplication table' for K, when multiplication is interpreted as composition of permutations.

10.7 Miscellaneous Exercises

1 A committee of nine people must elect a chairman, secretary and treasurer. In how many ways can this be done? (Explain carefully the assumptions you make in your solution.)

2 In the usual set of dominos each domino may be represented by the symbol $[x \mid y]$, where x and y are members of the set $\{0, 1, 2, 3, 4, 5, 6\}$. The numbers x and y may be equal. Explain as to why the total number of dominos is 28 rather than 49.

3 In how many ways can we select a black square and a white square on a chessboard in such a way that the two squares are not in the same rank or the same file?

4 In how many ways can we place eight Rooks on a chessboard in such a way that no two of them are on the same rank or file?

5 Suppose there are m girls and n boys in a class. What is the number of ways of arranging them in a line so that all the girls are together?

6 If we have nine different subsets of $\mathbb{N}_{12}$, each of which has eight members, and each member of $\mathbb{N}_{12}$ occurs in the same number r of subsets, what is the value of r? Is it possible to find nine different subsets of $\mathbb{N}_{12}$ each of which has seven members,

11

wrong envelope? (This is often called the *derangement problem*: there are several other picturesque formulations.)

Solution We can regard each letter and its correct envelope as being labelled with an integer i in the range $1 \leq i \leq n$. The act of putting letters into envelopes is described by a permutation π of $\mathbb{N}_n$: $\pi(i) = j$ if letter i goes into envelope j. We require the number of **derangements**, that is, permutations π such that $\pi(i) \neq i$ for all i in $\mathbb{N}_n$.

Let A_i $(1 \leq i \leq n)$ denote the subset of S_n (the set of all permutations of $\mathbb{N}_n$) containing those π for which $\pi(i) = i$. We say that the members of A_i *fix* i. By the sieve principle, the required number of derangements is

$$d_n = n! - \alpha_1 + \alpha_2 + \cdots + (-1)^n \alpha_n,$$

where α_r is the sum of the cardinalities of the intersections of the A_i taken r at a time. In other words, α_r is the number of permutations which fix r given symbols, taken over all ways of choosing the r symbols. Now there are $\binom{n}{r}$ ways of choosing r symbols, and the number of permutations which fix them is just the number of permutations of the remaining $n - r$ symbols, that is, $(n - r)!$ Hence

$$\alpha_r = \binom{n}{r} \times (n-r)! = \frac{n!}{r!}, \quad d_n = n!\left(1 - \frac{1}{1!} + \frac{1}{2!} - \cdots + (-1)^n \frac{1}{n!}\right). \qquad \square$$

Exercises 11.4

1 In a class of 67 mathematics students, 47 can read French, 35 can read German and 23 can read both languages. How many can read neither language? If, furthermore, 20 can read Russian, of whom 12 also read French, 11 read German also and 5 read all three languages, how many cannot read any of the three languages?

2 Find the number of ways of arranging the letters A, E, M, O, U, Y in a sequence in such a way that the words ME and YOU do not occur.

3 Calculate the number d_4 of derangements of $\{1, 2, 3, 4\}$ and write down the relevant permutations in cycle notation.

4 Use the principle of induction to prove that the formula for d_n satisfies the recursion

$$d_1 = 0, \quad d_2 = 1, \quad d_n = (n-1)(d_{n-1} + d_{n-2}) \quad (n \geq 3).$$

5 Show that the number of derangements of $\{1, 2, \ldots, n\}$ in which a given object (say 1) is in a 2-cycle is $(n-1)d_{n-2}$. Hence construct a direct proof of the recursion formula given in the previous exercise.

11.5 Some arithmetical applications

For many hundreds of years mathematicians have studied problems about prime numbers and the factorization of integers. The discussion of these matters in earlier chapters should have convinced the reader that such problems are difficult, because the primes themselves are irregularly distributed, and because there is no straightforward way to find the prime factorization of a given integer.

However, if the prime factorization of an integer is given to us, then it is relatively easy to answer some questions about its arithmetical properties. Suppose,

for instance, that we wish to list all the divisors of an integer n, and we know that the prime factorization of n is

$$n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}.$$

Then an integer d is divisor of n if and only if it has no prime divisors different from those of n, and no prime divides it more often than it divides n. Thus the divisors are precisely the integers which can be written in the form

$$d = p_1^{f_1} p_2^{f_2} \cdots p_r^{f_r},$$

where each f_i $(1 \leq i \leq r)$ satisfies $0 \leq f_i \leq e_i$. For example, given that $60 = 2^2 \times 3 \times 5$ we can quickly list all the divisors of 60: a good way of arranging them is illustrated in Fig. 11.2. (Technically, Fig. 11.2 is a diagram of the *lattice* of divisors of 60, but we shall not need the precise mathematical description of this term.)

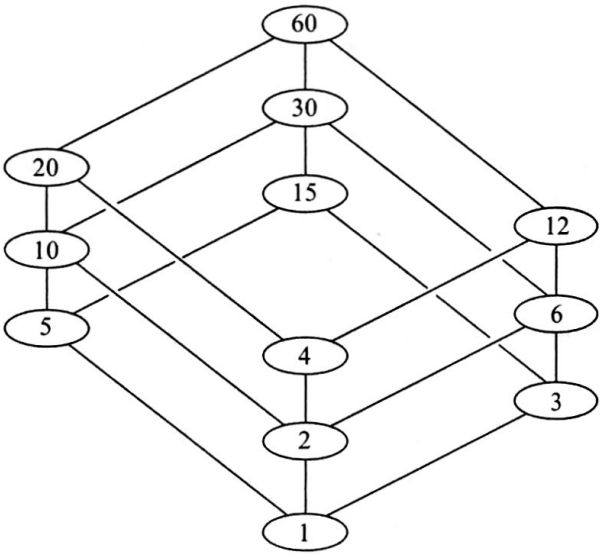


Fig. 11.2
The divisors of 60.

A similar problem is to find the number of integers x in the range $1 \leq x \leq n$ which are coprime to n. In Section 10.3 we denoted this number by $\phi(n)$, the value of Euler's function ϕ at n. We now show that if the prime factorization of n is known then $\phi(n)$ can be calculated by the sieve principle.

Example What is the value of $\phi(60)$? In other words, how many integers x in the range $1 \leq x \leq 60$ satisfy gcd $(x, 60) = 1$?

Solution We know that $60 \times 2^2 \times 3 \times 5$, so we have to count the number of integers x in the range $1 \leq x \leq 60$ which are *not* divisible by 2, 3 or 5. Let $A(2)$ denote the subset of $\mathbb{N}_{60}$ containing those integers which *are* divisible by 2, $A(2, 3)$ those which *are* divisible by 2 and 3, and so on. Then we have

$$\begin{aligned}\phi(60) &= 60 - |A(2) \cup A(3) \cup A(5)| \\ &= 60 - (|A(2)| + |A(3)| + |A(5)|) \\ &\quad +(|A(2, 3)| + |A(2, 5)| + |A(3, 5)|) - |A(2, 3, 5)|,\end{aligned}$$

by the sieve principle. Now $|A(2)|$ is the number of multiples of 2 in $\mathbb{N}_{60}$, which is $60/2 = 30$. Similarly, $|A(2, 3)|$ is the number of multiples of 2×3, which is $60/(2 \times 3) = 10$, and so on. Hence

$$\phi(60) = 60 - (30 + 20 + 12) + (10 + 6 + 4) - 2 = 16. \qquad \square$$

The same method can be used to give an explicit formula for $\phi(n)$ in the general case.

Theorem 11.5.1 Let $n \geq 2$ be an integer whose prime factorization is $n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$. Then

$$\phi(n) = n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right).$$

Proof Let A_j denote the subset of $\mathbb{N}_n$ containing the multiples of p_j $(1 \leq j \leq r)$. Then

$$\begin{aligned} \phi(n) &= n - |A_1 \cup A_2 \cup \cdots \cup A_r| \\ &= n - \alpha_1 + \alpha_2 - \cdots + (-1)^r \alpha_r, \end{aligned}$$

where α_i is the sum of the cardinalities of the intersections taken i at a time. Now a typical intersection such as

$$A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_i}$$

contains the multiples of $P = p_{j_1} \times p_{j_2} \times \cdots \times p_{j_i}$ in $\mathbb{N}_n$, and these are just the integers

$$P, 2P, 3P, \ldots, \left(\frac{n}{P}\right)P.$$

Hence the cardinality of the typical intersection is n/P, and α_i is the sum of all terms of the form

$$\frac{n}{P} = n\left(\frac{1}{p_{j_1}}\right)\left(\frac{1}{p_{j_2}}\right) \cdots \left(\frac{1}{p_{j_i}}\right).$$

It follows that

$$\begin{aligned} \phi(n) &= n - n\left(\frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_r}\right) + n\left(\frac{1}{p_1 p_2} + \frac{1}{p_1 p_3} + \cdots\right) + \cdots \\ &\quad \cdots + (-1)^r n\left(\frac{1}{p_1 p_2 \cdots p_r}\right) \\ &= n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right). \end{aligned} \qquad \square$$

It is easy to use the formula—provided, of course, that we know the prime factorization of n. For example,

$$\phi(60) = 60(1 - \tfrac{1}{2})(1 - \tfrac{1}{3})(1 - \tfrac{1}{5}) = 60 \times \tfrac{1}{2} \times \tfrac{2}{3} \times \tfrac{4}{5} = 16.$$

The sieve principle and the resulting formula for $\phi(n)$ also have some important theoretical consequences, which we shall now discuss.

Suppose we reverse the last step of the proof, and multiply out the factors in the formula for $\phi(n)$. In the case $n = 60$, we obtain

$$\phi(60) = \frac{60}{1} - \left(\frac{60}{2} + \frac{60}{3} + \frac{60}{5}\right) + \left(\frac{60}{6} + \frac{60}{10} + \frac{60}{15}\right) - \frac{60}{30}.$$

There is a term $60/d$ for each divisor d of 60 which is the product of distinct primes, and its coefficient is $+1$ or -1 according to whether the number of primes is even or odd. The divisors 4, 12, and 20, which have repeated prime factors (2^2 in this case) do not contribute, but for the sake of uniformity we may say that they contribute a term with coefficient 0. In general, we can express $\phi(n)$ as a sum of terms $\mu(d) \times (n/d)$, one for each divisor d of n; that is

$$\phi(n) = \sum_{d|n} \mu(d)\frac{n}{d},$$

where the coefficients $\mu(d)$ are given by

$$\mu(d) = \begin{cases} 1 & \text{if } d = 1, \\ (-1)^k & \text{if } d \text{ is a product of } k \text{ distinct primes} \\ 0 & \text{if } d \text{ has a repeated prime factor.} \end{cases}$$

The function μ is known as the **Möbius function**, after A. F. Möbius (1790–1868); it will play a vital part in some of the algebraic theorems in Part IV of this book. For the moment we shall show that it has some rather unexpected properties.

We begin by showing that for any integer $n \geq 2$ the sum of $\mu(d)$ taken over all divisors d of n is zero; that is

$$\sum_{d|n} \mu(d) = 0.$$

To prove this, suppose $n = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$. Each divisor d has the form $p_1^{f_1} p_2^{f_2} \dots p_r^{f_r}$ with $0 \leq f_i \leq e_i$, and $\mu(d)$ is zero unless each f_i is 0 or 1. Thus each divisor d with $\mu(d) \neq 0$ corresponds to the subset of $\{p_1, p_2, \dots, p_r\}$ containing those p_i with $f_i = 1$. The number of such subsets of size k is $\binom{r}{k}$, and $\mu(d)$ is $(-1)^k$, so we have

$$\sum_{d|n} \mu(d) = 1 - \binom{r}{1} + \binom{r}{2} - \dots + (-1)^r \binom{r}{r} = 0.$$

Once again, we have used the fundamental fact about the alternating sum of binomial numbers (*Example 2*, Section 11.1).

Now we are ready to establish the characteristic property of the Möbius function; it is usually known as the **Möbius inversion formula**.

Theorem 11.5.2 Let g be a function defined on $\mathbb{N}$ and suppose that f is the function obtained from g by the rule

$$f(n) = \sum_{d|n} g(d).$$

Then g can be obtained from f by the rule

$$g(n) = \sum_{d|n} \mu(d) f\left(\frac{n}{d}\right).$$

Proof Substituting for $f(n/d)$ in the right-hand side of the second equation we obtain

$$\begin{aligned}\sum_{d|n} \mu(d) f\left(\frac{n}{d}\right) &= \sum_{d|n} \mu(d) \sum_{c|n/d} g(c) \\ &= \sum_{(c,\, d)\in S} \sum \mu(d) g(c).\end{aligned}$$

The double sum is taken over the set S of all pairs (c, d) for which $d \mid n$ and $c \mid n/d$. But this is the same as the set of pairs (c, d) for which $c \mid n$ and $d \mid n/c$ and so we may rearrange the sum as follows:

$$\sum_{c|n} g(c) \left(\sum_{d|n/c} \mu(d) \right).$$

The sum in brackets is zero when $n/c \geq 2$, by the result obtained above. Hence the only term remaining is the one for which $n = c$, which reduces to

$$g(n) \sum_{d|1} \mu(d) = g(n)\mu(1) = g(n),$$

as required. □

Exercises 11.5

1 Calculate $\phi(n)$ and $\mu(n)$ for each n in the range $95 \leq n \leq 100$.

2 Show that if the prime factorization of n is $p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}$ then the number of divisors of n is

$$(e_1 + 1)(e_2 + 1) \dots (e_r + 1).$$

3 Use Theorem 11.5.1 to show that if $\gcd(m, n) = 1$ then $\phi(mn) = \phi(m)\phi(n)$.

4 Show that if $1 \leq x \leq n$ then

$$\gcd(x, n) = \gcd(n - x, n).$$

Hence prove that the sum of all the integers x which satisfy $1 \leq x \leq n$ and $\gcd(x, n) = 1$ is $\frac{1}{2} n\phi(n)$.

5 Suppose that the function f is obtained from g by the rule

$$f(n) = \sum_{d|n} \mu(d) g\left(\frac{n}{d}\right).$$

Show that g can be obtained from f by the rule

$$g(n) = \sum_{d|n} f(d).$$

(This is the converse of the Möbius inversion formula.) Hence use the formula

$$\phi(n) = \sum_{d|n} \mu(n) \frac{n}{d}$$

to give another proof of Theorem 10.3; that is,

$$\sum_{d|n} \phi(d) = n.$$

12

where the sum is taken over all k-tuples of non-negative integers $(n_1, n_2, \ldots, n_k)$ such that $n_1 + n_2 + \cdots + n_k = n$.

Proof When the n factors $x_1 + x_2 + \cdots + x_k$ are multiplied, a typical product term $x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}$ arises by choosing the x_1 term from n_1 of the factors, the x_2 term from n_2 of the factors, and so on. In other words, the typical term corresponds to a function from the set of n factors to the set $\{x_1, x_2, \ldots, x_k\}$, with the property that n_1 of the factors go to x_1, n_2 of them go to x_2, and so on. By the definition of the multinomial numbers, there are

$$\binom{n}{n_1, n_2, \ldots, n_k}$$

functions of this kind, and so this is the number of terms $x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}$ in the product. □

Because of their occurrence as coefficients of the terms in the expansion of $(x_1 + x_2 + \cdots + x_k)^n$, the multinomial numbers are often referred to as *multinomial coefficients*. However, the proof of the multinomial theorem shows that they occur in this way because they represent the number of functions of a certain kind, and for this reason we prefer to use a name which emphasizes their definition as basic counting numbers.

Exercises 12.3

1 Formulate and solve the following problem as a question about surjections of an 11-set onto a 4-set: 'How many 11-letter words can be made from the letters of the word MISSISSIPPI?'

2 Evaluate the multinomial numbers

$$\binom{10}{4, 3, 2, 1} \quad \text{and} \quad \binom{9}{5, 2, 2}.$$

3 Show that the number of different positions possible after four moves in a game of noughts-and-crosses (tic-tac-toe) is 756.

4 Show that if $a + b + c = n$, then

$$\binom{n}{a, b, c} = \binom{n-1}{a-1, b, c} + \binom{n-1}{a, b-1, c} + \binom{n-1}{a, b, c-1}.$$

Write down the analogous formula for a general multinomial number.

5 Calculate the coefficient of

(i) $x^5 y^3 z^2$ in $(x + y + z)^{10}$,

(ii) $x^3 y z^4 t$ in $(x + y + z + t)^9$.

6 Let p be a prime. Show that the multinomial number

$$\binom{p}{n_1, n_2, \ldots, n_k}$$

is divisible by p unless one of the n_i $(1 \leq i \leq k)$ is equal to p.

12.4 Partitions of a positive integer

If we are given a partition of an n-set X, such as

$$X = X_1 \cup X_2 \cup \cdots \cup X_k,$$

then there is a corresponding equation

$$n = n_1 + n_2 + \cdots + n_k,$$

where n_i is the size of X_i $(1 \leq i \leq k)$. We refer to this equation as a **partition of the integer n** into k parts. It must be stressed that the parts are non-zero (since the sets X_i are non-empty) and that the order of the parts is unimportant. Thus the partitions of the integer 6 are

$6,$	$5+1,$	$4+2,$
$4+1+1,$	$3+3,$	$3+2+1,$
$3+1+1+1,$	$2+2+2,$	$2+2+1+1.$
$2+1+1+1+1,$	$1+1+1+1+1+1.$	

The standard notation for partitions of a positive integer n involves counting the number of parts of each size. If there are α_i parts of size i, then the partition is written as

$$[1^{\alpha_1}2^{\alpha_2}\cdots n^{\alpha_n}].$$

With this notation, and making some trivial abbreviations, the partitions of 6 are

$[6],$	$[3^2],$
$[15],$	$[2^3],$
$[1^24],$	$[24]$
$[1^33],$	$[123]$
$[1^42],$	$[1^22^2].$
$[1^6],$	

It is rather unfortunate that this conventional notation for partitions uses a multiplicative symbol for an additive decomposition. It is worth keeping in mind that when we say, for example, that $[2^3]$ is a partition of 6, we mean that 6 is the *sum* of three twos.

The problem of counting partitions of n is a very interesting one, but it requires techniques which we have not yet dealt with in this book. We shall return to the problem in Chapter 26.

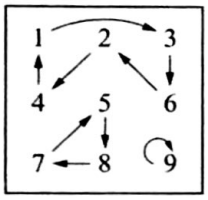
Exercises 12.4

1 Write down the partitions of 7 in standard notation.

2 Let $p_k(n)$ denote the number of partitions of n into k parts. Prove that

$$p_k(n) = p_k(n-k) + p_{k-1}(n-k) + \cdots + p_1(n-k).$$

3 Use the formula given in Ex. 2 to construct a table of the numbers $p_k(n)$ for $1 \leq k \leq n \leq 7$, and hence check your answer to Ex. 1.



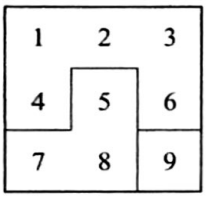


Fig. 12.4
A permutation of $\mathbb{N}_9$ and the corresponding partition.

12.5 Classification of permutations

In Section 10.6 we showed how any permutation can be written in terms of disjoint cycles. For example, the permutation of $\mathbb{N}_9$ indicated in Fig. 12.4 is written as (13624)(587)(9); clearly its cycles are the parts of a partition of $\mathbb{N}_9$, as shown on the right. This observation may be used to provide a formal justification for the cycle notation, by means of equivalence relations. The details are indicated in Ex. 12.7.18.

In this section we shall study the classification of permutations according to their cycle structure. Recall that S_n denotes the set of all permutations of $\mathbb{N}_n$. Associated with each permutation π in S_n is the partition of $\mathbb{N}_n$ whose parts are the cycles of π, and this in turn yields a corresponding partition of the integer n. We shall refer to the latter as the **type** of π. In other words, if π has α_i cycles of length i $(1 \leq i \leq n)$, then the type of π is the partition $[1^{\alpha_1}2^{\alpha_2}\cdots n^{\alpha_n}]$ of n. The permutation depicted in Fig. 12.4 has type [135].

It is fairly easy to count the number of permutations of a given type, provided we remember the conventions of the cycle notation. Suppose, for instance, we wish to count the number of members of S_{14} which have type $[2^2 3^2 4]$; we have to put the symbols $1, 2, \ldots, 14$ into the cycle pattern

$$(\cdot\,\cdot)(\cdot\,\cdot)(\cdot\,\cdot\,\cdot)(\cdot\,\cdot\,\cdot)(\cdot\,\cdot\,\cdot\,\cdot)\,,$$

and there are 14! ways of doing this. However, a given permutation π arises from this procedure in many different ways. With regard to each cycle, any member of that cycle can be put in the first position and the order of the rest is then determined by π. So there are two ways of getting each 2-cycle, three ways of getting each 3-cycle, and four ways of getting the 4-cycle. Hence the internal arrangement of the cycles can be made in $2^2 \times 3^2 \times 4$ ways for each given π. In general, for a permutation of type $[1^{\alpha_1}2^{\alpha_2}\cdots n^{\alpha_n}]$ there are

$$1^{\alpha_1}2^{\alpha_2}\cdots n^{\alpha_n}$$

ways of making the internal arrangements.

Also, the order of cycles of equal length is arbitrary. In the example, there are 2! ways of ordering the two 2-cycles and 2! ways of ordering the two 3-cycles. In general, the relevant number is

$$\alpha_1!\,\alpha_2!\cdots\alpha_n!.$$

Thus the number of permutations of type $[2^2 3^2 4]$ is

$$\frac{14!}{2^2 \times 3^2 \times 4 \times 2! \times 2!};$$

and the number of type $[1^{\alpha_1}2^{\alpha_2}\cdots n^{\alpha_n}]$ is

$$\frac{n!}{1^{\alpha_1}2^{\alpha_2}\cdots n^{\alpha_n}\alpha_1!\,\alpha_2!\cdots\alpha_n!}.$$

In simple cases it is often easier to use commonsense methods rather than this cumbersome formula. For example, the numbers in the classification of S_5 shown in Table 12.5.1 can be obtained in a variety of simple ways.

Table 12.5.1

Type	Example	Number
$[1^5]$	id	1
$[1^3 2]$	(12)(3)(4)(5)	10
$[1^2 3]$	(123)(4)(5)	20
$[12^2]$	(12)(34)(5)	15
[14]	(1234)(5)	30
[23]	(123)(45)	20
[5]	(12345)	24
		120

The classification of permutations by type is doubly useful, because there is an alternative description of the classes which has useful consequences in the algebraic theory of permutations (Chapters 21 and 27). Let α and β be permutations in S_n. If there is a permutation σ in S_n such that

$$\sigma\alpha\sigma^{-1} = \beta,$$

then we say that α and β are **conjugate**.

Theorem 12.5 Two permutations are conjugate if and only if they have the same type.

Proof Suppose α and β conjugate, so that $\sigma\alpha\sigma^{-1} = \beta$. If $\alpha(x_1) = x_2$, put $y_1 = \sigma(x_1)$, $y_2 = \sigma(x_2)$; if $\alpha(x_2) = x_3$ put $y_3 = \sigma(x_3)$; and so on (Fig. 12.5). Then we have

$$\beta(y_1) = \sigma\alpha\,\sigma^{-1}(\sigma(x_1)) = \sigma\alpha(x_1) = \sigma(x_2) = y_2.$$

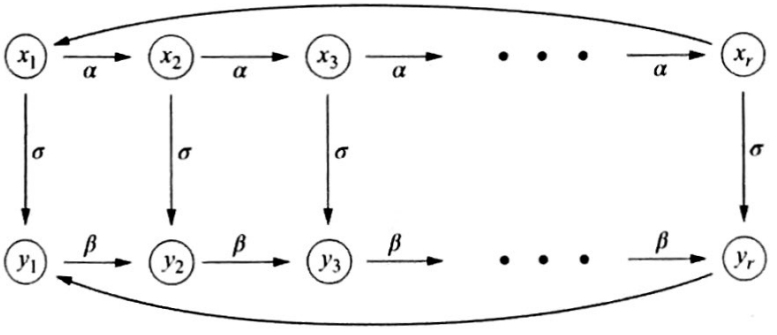


Fig. 12.5
Conjugate permutations.

Similarly, $\beta(y_2) = y_3$, $\beta(y_3) = y_4$, and so on. Consequently, for each cycle $(x_1 x_2 \cdots x_r)$ of α there is a corresponding cycle $(y_1 y_2 \ldots y_r)$ of β, and it follows that α and β have the same type.

Conversely, suppose α and β have the same type. Since they have the same number of cycles of each length, we can set up a bijective correspondence between their cycles, in which a typical cycle $(x_1 x_2 \ldots x_r)$ of α will correspond to a cycle $(z_1 z_2 \ldots z_r)$ of β. Let us define σ by the rule $\sigma(x_i) = z_i$ $(1 \leq i \leq r)$, and use similar rules for the other cycles. Then $\sigma\alpha\sigma^{-1} = \beta$, since

$$\sigma\alpha\sigma^{-1}(z_1) = \sigma\alpha(x_1) = \sigma(x_2) = z_2 = \beta(z_1).$$

and so on. Hence α and β are conjugate. □

Exercises 12.5

1 Write out the classification of the elements of S_6 in the same form as the table for S_5 given on p. 135.
2 Let $\alpha = (13624)(587)(9)$, $\beta = (15862)(394)(7)$. Write down a permutation σ in S_9 such that $\sigma\alpha\sigma^{-1} = \beta$.
3 Prove that if π and τ are any members of S_n then $\pi\tau$ and $\tau\pi$ have the same type. [Hint: Use Theorem 12.5.]
4 Prove directly from the definition (without using Theorem 12.5) that conjugacy is an equivalence relation on S_n.
5 Use the classification of S_6 obtained in Ex. 1 to find the number of *derangements* in S_6 (as defined in Section 11.4).
6 Find the number of permutations σ which have the property specified in Ex. 2.

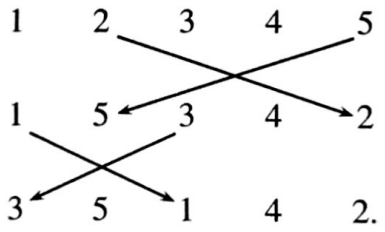
12.6 Even and odd permutations

The classification of S_5 obtained in the previous section has a remarkable feature, which should be evident from the following tabulation of the classes.

Type	Number	Type	Number
$[1^5]$	1	$[1^3 2]$	10
$[1^2 3]$	20	[14]	30
$[1\,2^2]$	15	[23]	20
[5]	24		
	60		60

Clearly, we have a partition of S_5 into two parts of equal size with 60 permutations in each. In this section we shall prove that S_n can be split into two equal parts for every integer $n \geq 2$, and we shall see that there is a very simple way of deciding which part contains a given permutation.

The key observation is that every rearrangement can be achieved by successively switching certain pairs of objects. For example, in order to obtain 35142 from 12345 we need just two switches

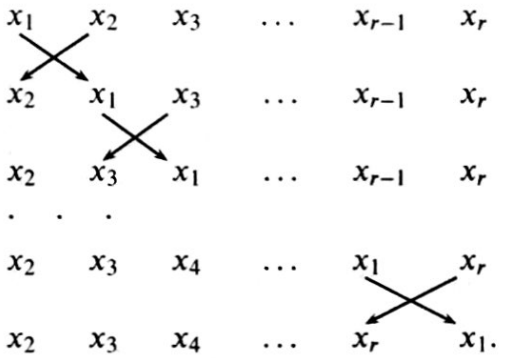


In terms of permutations, the permutation which effects the transformation of 12345 into 35142 is (13) (25) (4), and clearly the two 2-cycles correspond exactly to the switches carried out above. More generally it may not be so obvious that a given permutation can be expressed in terms of 2-cycles in this way, and our first task is to show that this is indeed so.

The technical name for a permutation which interchanges two objects and leaves the rest unaltered is a **transposition**. Thus an element of S_n is a transposition if it has type $[1^{n-2}2]$: it has $n-2$ 1-cycles and just one 2-cycle. Now any cycle, such as $(x_1x_2\ldots x_{r-1}x_r)$, effects the rearrangement taking

$$x_1x_2\cdots x_{r-1}x_r \quad \text{to} \quad x_2x_3\cdots x_rx_1,$$

and this can be achieved by successive transpositions in the following manner:



So we can write

$$(x_1x_2\ldots x_{r-1}x_r) = (x_1x_r)\ldots(x_1x_3)(x_1x_2).$$

Here each transposition is represented by the corresponding 2-cycle, and the 1-cycles are suppressed since they do not have any effect. Also, it must be kept in mind that the transposition written last is the first one to be carried out, in accordance with the rule for combining permutations. Since any permutation can be decomposed into cycles it can thereby be decomposed into transpositions, using the rule given above. For example

$$(136)(2457) = (16)(13)(27)(25)(24).$$

Two things must be stressed about the decomposition into transpositions. First, the transpositions are not disjoint, so that one object may be moved several times. Secondly, the decomposition is far from unique, for not only can the order of the transpositions be altered, but also we can actually use an entirely different set of transpositions, such as

$$(136)(2457) = (15)(35)(36)(57)(14)(27)(12).$$

Our aim is to show that, although there are many different ways of decomposing a given permutation, there is one feature common to all of them.

Let $c(\pi)$ denote the total number of cycles of a permutation π, so that if π has type $[1^{\alpha_1}2^{\alpha_2}\ldots n^{\alpha_n}]$ then $c(\pi) = \alpha_1 + \alpha_2 + \cdots + \alpha_n$. Suppose we combine π with a transposition τ, giving a new permutation $\tau\pi$. What is the relation between $c(\tau\pi)$ and $c(\pi)$?

Suppose τ switches a and b, so that $\tau(a) = b$, $\tau(b) = a$, and $\tau(k) = k$ when $k \neq a, b$. If a and b are in the same cycle of π we can write

$$\pi = (ax\ldots yb\ldots z)\ldots \text{ and some other cycles.}$$

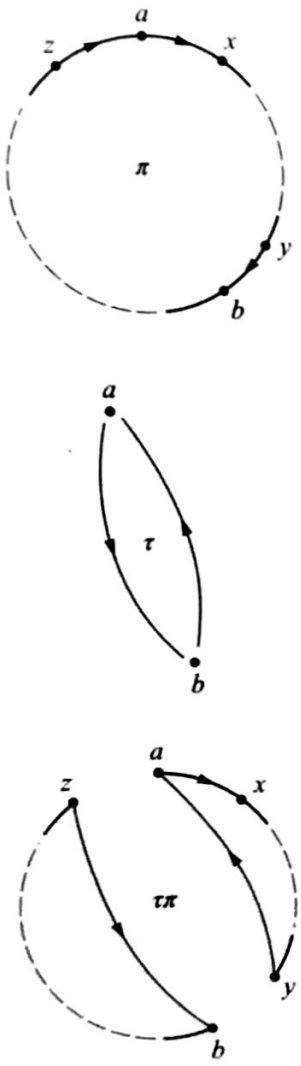


Fig. 12.6
Combining τ with a cycle of π (first case).

The composite permutation $\tau\pi$ (first π, then τ) can be calculated quite simply; as illustrated in Fig. 12.6 it is

$$\tau\pi = (ax \ldots y)(b \ldots z) \ldots \text{ and the same other cycles.}$$

In this case $c(\tau\pi) = c(\pi) + 1$. On the other hand, if a and b are in different cycles of π, so that

$$\pi = (ax \ldots y)(b \ldots z) \ldots \text{ and some other cycles,}$$

then a similar calculation shows that

$$\tau\pi = (ax \ldots yb \ldots z) \ldots \text{ and the same other cycles.}$$

In this case $c(\tau\pi) = c(\pi) - 1$. In both cases the result of following π with a transposition is to alter the number of cycles by one, and this simple fact leads to the next theorem.

Theorem 12.6.1 Suppose the permutation π is S_n can be written as the composite of r transpositions, and also as the composite of r' transpositions. Then either r and r' are both even or r and r' are both odd.

Proof Let $\pi = \tau_r \tau_{r-1} \cdots \tau_2 \tau_1$, where τ_i $(1 \leq i \leq r)$ is a transposition. Since τ_1 has one 2-cycle and $n-2$ 1-cycles we have

$$c(\tau_1) = 1 + (n-2) = n - 1.$$

When $\tau_2, \tau_3, \ldots, \tau_r$ are combined in turn with τ_1 the final result is π. At each stage, the number of cycles is altered by one: suppose it increases by one g times and decreases by one h times. Then the final number of cycles is

$$(n-1) + g - h = c(\pi).$$

But $g + h$ is the total number of stages, $r - 1$. So we have

$$\begin{aligned} r &= 1 + g + h = 1 + g + (n - 1 + g - c(\pi)) \\ &= n - c(\pi) + 2g. \end{aligned}$$

By the same argument, if π is the composite of r' transpositions, there is an integer g' such that $r' = n - c(\pi) + 2g'$. Hence

$$r - r' = 2(g - g'),$$

and since the right-hand side is even, we have the result. □

As a consequence of the theorem we can say that a permutation is **even** or **odd**, according to whether the number of transpositions in any decomposition of it is even or odd. We can also define the **sign** of a permutation π, written $\operatorname{sgn} \pi$, to be $+1$ if π is even, and -1 if π is odd. Thus

$$\operatorname{sgn} \pi = (-1)^r,$$

where r is the number of transpositions in a decomposition of π. In particular $\operatorname{sgn} \mathrm{id} = (-1)^0 = +1$. If π can be decomposed into r transpositions and σ can be

decomposed into s transpositions, then clearly the composite $\pi\sigma$ can be decomposed into $r + s$ transpositions and so

$$\operatorname{sgn} \pi\sigma = (-1)^{r+s} = (-1)^r(-1)^s = \operatorname{sgn} \pi \operatorname{sgn} \sigma.$$

This implies, for example, that $\operatorname{sgn} \pi^{-1} = \operatorname{sgn} \pi$, since $\pi^{-1}\pi$ is the identity and

$$\operatorname{sgn} \pi^{-1} \operatorname{sgn} \pi = \operatorname{sgn} \pi^{-1}\pi = \operatorname{sgn} \mathrm{id} = +1.$$

Now we can establish the general form of the partition of S_n observed at the beginning of the Section in the case $n = 5$.

Theorem 12.6.2 For any integer $n \geq 2$ exactly half of the permutations in S_n are even and exactly half of them are odd.

Proof Let $\pi_1, \pi_2, \ldots, \pi_k$ be a list of the even permutations in S_n. (Certainly there are such permutations, since id is one of them.) Let τ be any transposition in S_n, say $\tau = (12)(3)(4)\ldots(n)$.

The permutations $\tau\pi_1, \tau\pi_2, \ldots, \tau\pi_k$ are all distinct, since if $\tau\pi_i = \tau\pi_j$ we should have

$$\pi_i = (\tau^{-1}\tau)\pi_i = \tau^{-1}(\tau\pi_i) = \tau^{-1}(\tau\pi_j) = (\tau^{-1}\tau)\pi_j = \pi_j,$$

using the fundamental rules given in Theorem 10.6. Furthermore, these permutations are all odd, since

$$\operatorname{sgn} \tau\pi_i = \operatorname{sgn} \tau \operatorname{sgn} \pi_i = (-1) \times (+1) = -1.$$

Finally, we show that *any* odd permutation ρ is one of the $\tau\pi_i$ $(1 \leq i \leq n)$. Since

$$\operatorname{sgn} \tau^{-1}\rho = \operatorname{sgn} \tau^{-1} \operatorname{sgn} \rho = (-1) \times (-1) = +1,$$

it follows that $\tau^{-1}\rho$ is one of the even permutations π_i. Thus

$$\rho = (\tau\tau^{-1})\rho = \tau(\tau^{-1}\rho) = \tau\pi_i,$$

as claimed. Hence there are just as many odd permutations as even permutations, and the result follows. □

The partition of S_n into two equal parts has many interesting consequences. The following is an example from 'recreational' mathematics.

Example Eight blocks labelled A, E, I, O, U, Y, R, T, are placed in a square frame as shown in the first diagram (Fig. 12.7). A legal move consists of sliding one block into the currently empty space. Prove that it is impossible to obtain the arrangement shown in the second diagram by a sequence of legal moves.

Solution Let us denote the space by □ so that the initial arrangement is AEIOUYRT□ and the final arrangement is YOUAREIT□. Moving a letter X into the space corresponds to the 2-cycle (X□). If we are given any final arrangement with □ in its original place, then in order to achieve it □ must have been moved

A	E	I
O	U	Y
R	T	

Y	O	U
A	R	E
I	T	

Fig. 12.7
Can it be done?

upwards the same number of times as downwards, and leftwards the same number of times as rightwards. Consequently, the total number of moves required is even and, since each move is a transposition, the final arrangement must be effected by an *even* permutation. However, the permutation which effects the rearrangement of

AEIOUYRT□

into

YOUAREIT□

is (AYEO)(IUR)(T)(□). This is an *odd* permutation, since the 4-cycle is equivalent to three transpositions and the 3-cycle is equivalent to two transpositions. It follows that it is impossible to achieve the required result by a sequence of legal moves. □

Exercises 12.6

1 Express the following members of S_8 in terms of 2-cycles and find the sign of each one.

$$\alpha = (1357)(2468),$$
$$\beta = (127)(356)(48),$$
$$\gamma = (135)(678)(2)(4).$$

2 Without using Theorem 12.5 prove that

$$\operatorname{sgn} \pi\sigma\pi^{-1} = \operatorname{sgn} \sigma \qquad (\sigma, \pi \in S_n).$$

3 Show that if π has type $[1^{\alpha_1}2^{\alpha_2}\dots n^{\alpha_n}]$ then

$$\operatorname{sgn} \pi = (-1)^{\alpha_2+\alpha_4+\alpha_6+\cdots}.$$

4 Which of the following positions can be obtained by legal moves starting from the initial arrangement specified in the *Example*?

(i) A R E (ii) Y E A
Y O U T O
I T U R I

5 Show that any cycle of odd length can be written as the composite of (not necessarily disjoint) 3-cycles. Deduce that a permutation is even if and only if it can be expressed as a composite of 3-cycles.

12.7 Miscellaneous Exercises

1 How many 14-letter words can be made from the letters of the word CLASSIFICATION?

2 Calculate the coefficient of

(i) $x^3y^2z^4$ in $(x+y+z)^9$,
(ii) xy^3zt^2u in $(x+y+z+t+u)^8$.

3 Calculate $p(8)$, the total number of partitions of 8, and verify that the number which have distinct parts is equal to the number whose parts are all odd. Can you explain this equality?

4 Let α, β be the elements of S_8 represented in cycle notation as

$$\alpha = (123)(456)(78), \quad \beta = (1357)(26)(4)(8).$$

Find sgn α and sgn β, and express α and β in terms of transpositions, using the least number of transpositions possible in each case.

5 Show that

$$S(n, 3) = \tfrac{1}{2}(3^{n-1} + 1) - 2^{n-1}.$$

6 Let $\sim$ denote the relation on $\mathbb{Z}$ defined by

$$a \sim b \Leftrightarrow a - b \text{ is divisible by } 11.$$

Show that $\sim$ is an equivalence relation on $\mathbb{Z}$. What is the number of equivalence classes?

7 Find the number of ten-digit positive integers (in base 10 notation) which have the property that the number of distinct odd digits is half the number of distinct even digits.

8 Prove that

$$\sum_{k=1}^{m} S(m, k)n(n-1)\cdots(n-k+1) = n^m.$$

13

13
Modular arithmetic

Contents

13.1 Congruences 142
13.2 $\mathbb{Z}_m$ and its arithmetic 144
13.3 Invertible elements of $\mathbb{Z}_m$ 147
13.4 Cyclic constructions for designs 149
13.5 Latin squares 153
13.6 Miscellaneous Exercises 156

13.1 Congruences

One of the most familiar partitions is the partition of $\mathbb{Z}$ into even and odd integers. This partition corresponds to an equivalence relation on $\mathbb{Z}$ which we may define by saying that x_1 is related to x_2 whenever $x_1 - x_2$ is divisible by 2. It is customary to use the notation

$$x_1 \equiv x_2 \,(\mathrm{mod}\, 2)$$

for this relation, and to say that x_1 is *congruent* to x_2 *modulo* 2. Thus, x_1 and x_2 are in the same part of the partition if and only if x_1 is congruent to x_2 modulo 2.

Clearly any positive integer m can be used instead of 2.

Definition Let x_1 and x_2 be integers, and m a positive integer. We say that x_1 is **congruent** to x_2 **modulo** m, and write

$$x_1 \equiv x_2 \,(\mathrm{mod}\, m)$$

whenever $x_1 - x_2$ is divisible by m.

It is easy to verify that congruence modulo m is an equivalence relation. It is reflexive, since $x - x$ is zero and is divisible by m for any x. It is symmetric, since if $x_1 - x_2 = km$, then $x_2 - x_1 = (-k)m$. It is transitive, since if $x_1 - x_2 = km$ and $x_2 - x_3 = lm$, then $x_1 - x_3 = (k + l)m$.

The usefulness of congruence relations stems mainly from the fact that they are compatible with the arithmetical operations. Specifically, we have the following theorem.

Theorem 13.1 Let m be a positive integer and x_1, x_2, y_1, y_2 integers such that

$$x_1 \equiv x_2 \,(\mathrm{mod}\, m), \qquad y_1 \equiv y_2 \,(\mathrm{mod}\, m).$$

Then

$$\text{(i)}\quad x_1 + y_1 \equiv x_2 + y_2 \,(\mathrm{mod}\, m), \qquad \text{(ii)}\quad x_1 y_1 \equiv x_2 y_2 \,(\mathrm{mod}\, m).$$

Proof (i) We are given that $x_1 - x_2 = mx$, $y_1 - y_2 = my$, for some x and y in $\mathbb{Z}$. It follows that

$$\begin{aligned}(x_1 + y_1) - (x_2 + y_2) &= (x_1 - x_2) + (y_1 - y_2)\\ &= mx + my\\ &= m(x + y),\end{aligned}$$

and so the left-hand side is divisible by m, as required.

(ii) Here we have

$$\begin{aligned}x_1y_1 - x_2y_2 &= (x_1 - x_2)y_1 + x_2(y_1 - y_2)\\ &= mxy_1 + x_2my\\ &= m(xy_1 + x_2y),\end{aligned}$$

and again the left-hand side is divisible by m, as required. □

Example Let $(x_nx_{n-1}\dots x_0)_{10}$ be the representation of the positive integer x in base 10. Show that

$$x \equiv x_0 + x_1 + \cdots + x_n \pmod 9$$

and use this result to check the calculation

$$54\,321 \times 98\,765 = 5\,363\,013\,565.$$

Solution Using the definition of the base 10 representation, we have

$$\begin{aligned}x - (x_0 + x_1 + \cdots + x_n) &= (x_0 + 10x_1 + \cdots 10^nx_n) - (x_0 + x_1 + \cdots + x_n)\\ &= (10^1 - 1)x_1 + \cdots + (10^n - 1)x_n.\end{aligned}$$

Now, for each $r \geq 1$,

$$\begin{aligned}10^r - 1 &= (99\dots9)_{10} \qquad (r \text{ nines})\\ &= 9 \times (11\dots1)_{10}.\end{aligned}$$

Thus 9 divides $x - (x_0 + x_1 + \cdots + x_n)$, as required.

For convenience, let us write $\theta(x)$ instead of $x_0 + x_1 + \cdots + x_n$. We have shown that $\theta(x) \equiv x \pmod 9$. By part (ii) of Theorem 13.1

$$\theta(x)\theta(y) \equiv xy \pmod 9,$$

and so if $xy = z$ we must have $\theta(x)\theta(y) \equiv \theta(z) \pmod 9$. In the given calculation

$$\theta(54\,321) = 15, \quad \theta(98\,765) = 35, \quad \theta(5\,363\,013\,565) = 37,$$

and

$$\theta(15) = 6, \quad \theta(35) = 8, \quad \theta(37) = 10.$$

Since 6×8 is not congruent to 10 (mod 9), it follows that 15×35 is not congruent to 37 (mod 9) and $54\,321 \times 98\,765$ is not congruent to 5 363 013 565 (mod 9). If

the calculation were correct, the expressions would be equal, and consequently congruent mod 9. Hence the calculation is wrong.

This procedure is known as 'casting out nines'. □

Exercises 13.1

1 Without carrying out any 'long multiplication' show that

(i) $1\,234\,567 \times 90\,123 \equiv 1 \pmod{10}$,

(ii) $2468 \times 13\,579 \equiv -3 \pmod{25}$.

2 Use the method of casting out nines to show that two of the following equations are false. What can be said about the other equation?

(i) $5783 \times 40\,162 = 233\,256\,846$,

(ii) $9787 \times 1258 = 12\,342\,046$,

(iii) $8901 \times 5743 = 52\,018\,443$.

3 Suppose we are given $m \geq 2$ and an integer x. The remainder r when x is divided by m satisfies

$$x \equiv r \pmod{m}, \quad 0 \leq r \leq m-1,$$

and is sometimes called the *least non-negative residue* of $x \pmod{m}$. Find the least non-negative residue of $3^{15} \pmod{17}$ and $15^{81} \pmod{13}$.

4 Let $(x_n x_{n-1} \ldots x_0)_{10}$ be the base 10 representation of the positive integer x. Show that

$$x \equiv x_0 - x_1 + x_2 - \cdots + (-1)^n x_n \pmod{11},$$

and use this result to test whether 1 213 141 516 171 819 is divisible by 11.

13.2 $\mathbb{Z}_m$ and its arithmetic

In this section we shall introduce a more compact method of dealing with congruence properties of integers.

For any integer x and positive integer m we shall use the notation $[x]_m$ for the equivalence class of x with respect to congruence modulo m. In other words, $[x]_m$ consists of all the integers x' for which $x' - x$ is a multiple of m. For example,

$$[5]_3 = \{\ldots, -4, -1, 2, 5, 8, 11, \ldots\},$$
$$[6]_7 = \{\ldots, -8, -1, 6, 13, 20, \ldots\}.$$

As usual, each equivalence class has many aliases, depending on which representative is used. For example,

$$\ldots = [-8]_7 = [-1]_7 = [6]_7 = [13]_7 = [20]_7 = \ldots.$$

The general theory of equivalence relations ensures that, for each m, the set $\mathbb{Z}$ is partitioned into disjoint equivalence classes by the relation of congruence modulo m. When $m = 3$ we have

$$\mathbb{Z} = X_0 \cup X_1 \cup X_2,$$

where

$$X_0 = [0]_3 = \{\ldots, -3, 0, 3, 6, \ldots\},$$
$$X_1 = [1]_3 = \{\ldots, -2, 1, 4, 7, \ldots\},$$
$$X_2 = [2]_3 = \{\ldots, -1, 2, 5, 8, \ldots\}.$$

For any given m, there are m distinct equivalence classes $[0]_m, [1]_m, \ldots, [m-1]_m$. This follows from the fact that any x in $\mathbb{Z}$ can be expressed uniquely in the form $qm + r$ with $0 \leq r \leq m - 1$ (Theorem 8.2), so that x is in $[r]_m$ for just one such r.

Definition The set of **integers modulo m**, written as $\mathbb{Z}_m$, is the set of distinct equivalence classes under the relation of congruence modulo m in $\mathbb{Z}$.

So $\mathbb{Z}_m$ is the set $\{[0]_m, [1]_m, \ldots, [m-1]_m\}$. It must be emphasized that the members of $\mathbb{Z}_m$ are defined as *subsets* of $\mathbb{Z}$. However, it is often convenient to think of them as the integers $0, 1, 2, \ldots, m-1$ with a modified arithmetical structure, and we can make this idea precise in the following way.

Let us define new operations of 'addition' and 'multiplication' for the members of $\mathbb{Z}_m$, written $\oplus$ and $\otimes$, by the rules

$$[x]_m \oplus [y]_m = [x + y]_m, \quad [x]_m \otimes [y]_m = [xy]_m.$$

Since x and y are integers, the expressions $x + y$ and xy on the right-hand side are defined, and they have all the familiar properties. The new operations inherit their properties from those properties. But before we study them we must deal with a difficulty concerning the definitions.

The difficulty arises because each equivalence class has many names. Suppose, for example, that $[x]_m$ and $[x']_m$ denote the same class, and $[y]_m$ and $[y']_m$ denote the same class. Then, in order that the definition of $\oplus$ be reasonable, we must ensure that $[x]_m \oplus [y]_m$ and $[x']_m \oplus [y']_m$ denote the same class. The fact that this is so is a simple consequence of Theorem 13.1. For we are given that $x \equiv x' \pmod{m}$ and $y \equiv y' \pmod{m}$, and so $x + x' \equiv y + y' \pmod{m}$; consequently $[x + x']_m = [y + y']_m$ as required. A similar proof holds for multiplication.

We can now list the arithmetical properties of $\mathbb{Z}_m$.

Theorem 13.2 The operations $\oplus$ and $\otimes$ satisfy the following rules, where a, b, c denote any members of $\mathbb{Z}_m$, and $0 = [0]_m$, $1 = [1]_m$.

M1. $a \oplus b$ and $a \otimes b$ are in $\mathbb{Z}_m$.
M2. $a \oplus b = b \oplus a$, $a \otimes b = b \otimes a$.
M3. $(a \oplus b) \oplus c = a \oplus (b \oplus c)$, $(a \otimes b) \otimes c = a \otimes (b \otimes c)$.
M4. $a \oplus 0 = a$, $a \otimes 1 = a$.
M5. $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$.
M6. For each a in $\mathbb{Z}_m$ there is a unique element $-a$ in $\mathbb{Z}_m$ such that $a \oplus (-a) = 0$.

Proof **M1** is a direct consequence of the definitions of $\oplus$ and $\otimes$. For the first part of **M2**, suppose $a = [x]_m$ and $b = [y]_m$; then

$$\begin{aligned} a \oplus b = [x]_m \oplus [y]_m &= [x + y]_m && \text{(definition of } \oplus) \\ &= [y + x]_m && \text{(property of } \mathbb{Z}) \\ &= [y]_m \oplus [x]_m && \text{(definition of } \oplus) \\ &= b \oplus a. \end{aligned}$$

Similar proofs hold for the second part of **M2** and for **M3, M4, M5**. For **M6**, if $a = [x]_m$ we put $-a = [-x]_m$, and check

$$\begin{aligned} a \oplus (-a) &= [x]_m \oplus [-x]_m \\ &= [x + (-x)]_m \\ &= [0]_m \\ &= 0, \end{aligned}$$

as required. □

In practice, we usually dispense with the cumbersome $[x]_m$ notation for the members of $\mathbb{Z}_m$, and use the integers $0, 1, \ldots, m-1$ to denote the classes $[0]_m, [1]_m, \ldots, [m-1]_m$. The specific value of m under consideration must either be clear from the context, or stated explicitly. Also, we use the usual notations for addition and multiplication, rather than $\oplus$ and $\otimes$. Thus, we write

$$7 + 5 = 3 \text{ (in } \mathbb{Z}_9) \quad \text{instead of} \quad [7]_9 \oplus [5]_9 = [3]_9,$$

and with precisely the same meaning as $7 + 5 \equiv 3 \pmod 9$. The properties stated in Theorem 13.2 justify most of the usual arithmetical manipulations in $\mathbb{Z}_m$, just as in $\mathbb{Z}$.

However, there are some important differences between $\mathbb{Z}_m$ and $\mathbb{Z}$. In $\mathbb{Z}$ it holds (Theorem 7.5.2) that if $ab = ac$ and $a \neq 0$ then $b = c$. This does *not* hold in $\mathbb{Z}_m$: for example, in $\mathbb{Z}_6$ we have

$$3 \times 1 = 3 \times 5 \text{ and } 3 \neq 0, \text{ but } 1 \neq 5.$$

So we must be careful about 'cancellation' in $\mathbb{Z}_m$, a topic we shall discuss in more detail in the next section.

Finally, we remark that there is no order relation in $\mathbb{Z}_m$ resembling the relation $\leq$ in $\mathbb{Z}$. Our intuitive picture of $\mathbb{Z}$ as a regularly spaced set of points on a line, extending indefinitely in either direction, represents the properties of that relation. In $\mathbb{Z}_m$ we have instead a kind of *cyclic* order, represented by a regularly spaced set of points on a circle (Fig. 13.1). For this reason, arithmetic in $\mathbb{Z}_m$, or as we shall say, *modular arithmetic*, is often taught in schools as 'clock arithmetic'.

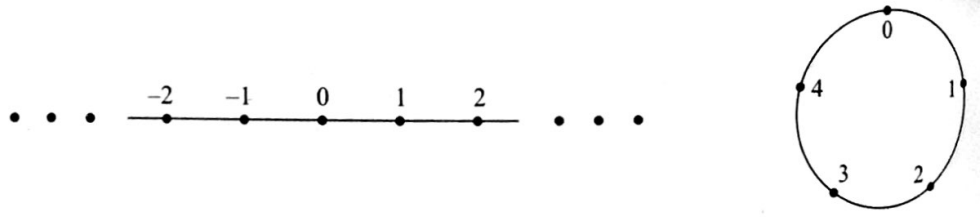


Fig. 13.1
Pictures of $\mathbb{Z}$ and $\mathbb{Z}_5$.

Exercises 13.2

1 Complete Tables 13.2.1(a, b), the addition and multiplication tables for $\mathbb{Z}_6$.

Table 13.2.1(a)

$\oplus$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2						
3						
4						
5						

Table 13.2.1(b)

$\otimes$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2						
3						
4						
5						

2 Deduce from Theorem 7.5.2 that if x and y are integers such that $xy = 0$ and $x \neq 0$, then $y = 0$. Show by counter-examples that this axiom does not hold in $\mathbb{Z}_6$, $\mathbb{Z}_8$, and $\mathbb{Z}_{15}$. Is there a counter-example in $\mathbb{Z}_7$?

3 Solve the simultaneous equations

$$x + 2y = 4, \qquad 4x + 3y = 4$$

in $\mathbb{Z}_7$. Is there a solution in $\mathbb{Z}_5$?

4 Solve the quadratic equation

$$x^2 + 3x + 4 = 0$$

in $\mathbb{Z}_{11}$.

13.3 Invertible elements of $\mathbb{Z}_m$

In Chapter 1 we stressed the fact that the symbol s/r need not represent an integer even though r and s are integers. In other words, if we are given integers r and s then it may be impossible to solve the equation $rx = s$ with x in $\mathbb{Z}$. In this section we shall investigate the corresponding problem in $\mathbb{Z}_m$.

Definition An element r in $\mathbb{Z}_m$ is said to be **invertible** if there is some x in $\mathbb{Z}_m$ such that $rx = 1$ in $\mathbb{Z}_m$. In that case, x is called the **inverse** of r, and we write $x = r^{-1}$.

Since $rx = xr$ in $\mathbb{Z}_m$ we have $xr = 1$ also, and $r = x^{-1}$.

Exercises 13.3

1 Find the invertible elements of $\mathbb{Z}_6$, $\mathbb{Z}_7$, and $\mathbb{Z}_8$.

2 Show that 0 is not invertible in any $\mathbb{Z}_m$, but 1 is always invertible.

3 Show that if x and y are invertible in $\mathbb{Z}_m$ then xy and x^{-1} are invertible in $\mathbb{Z}_m$.

Theorem 13.3.1 The element r in $\mathbb{Z}_m$ is invertible if and only if r and m are coprime in $\mathbb{Z}$. In particular, when p is a prime every element of $\mathbb{Z}_p$ except 0 is invertible.

Proof Suppose r is invertible, so that $rx = 1$ in $\mathbb{Z}_m$. It follows that, in $\mathbb{Z}$, we have $rx - 1 = km$ for some integer k, or

$$rx - km = 1.$$

Now any common divisor of r and m must divide $rx - km$, which is 1, so gcd $(r, m) = 1$.

Conversely, suppose gcd $(r, m) = 1$. Then by Theorem 8.4 there are integers x and y such that $rx + my = 1$. Rearranging this equation we obtain $rx \equiv 1 \pmod{m}$, or $rx = 1$ (in $\mathbb{Z}_m$), as required. □

Recall that the value $\phi(m)$ of Euler's function (Section 10.3) is the number of integers in the range $1 \leq r \leq m$ which are coprime to m. It follows from Theorem 13.3.1 that the number of invertible elements of $\mathbb{Z}_m$ is $\phi(m)$.

The next theorem is one of the classics of elementary number theory, and it has many useful consequences. We prepare for it with a simple observation concerning the set U_m of invertible elements of $\mathbb{Z}_m$. If y is in U_m define yU_m to be the set obtained when we multiply each member of U_m by y; that is

$$yU_m = \{z \in \mathbb{Z}_m \mid z = yx \text{ for some } x \text{ in } U_m\}.$$

We shall show that $yU_m = U_m$. For example, taking $m = 9$ and $y = 5$ we obtain

$$U_9 = \{1, 2, 4, 5, 7, 8\}, \quad 5U_9 = \{5, 1, 2, 7, 8, 4\}.$$

First, we have $yU_m \subseteq U_m$ since if $z = yx$ and both y and x are in U_m, then z is also in U_m (Ex. 13.3.3). On the other hand, $U_m \subseteq yU_m$, since given x in U_m we can write

$$x = y(y^{-1}x),$$

which is clearly an element of yU_m (again by Ex. 13.3.3). Thus $yU_m = U_m$, as claimed.

Theorem 13.3.2 If y is invertible in $\mathbb{Z}_m$ then

$$y^{\phi(m)} = 1 \quad \text{in } \mathbb{Z}_m.$$

Proof Let u denote the product of all the members of U_m; say $u = x_1x_2\cdots x_k$, where (as noted above) $k = \phi(m)$. Since $yU_m = U_m$ the elements $yx_1, yx_2, \ldots, yx_k$ are just a rearrangement of $x_1, x_2, \ldots, x_k$. If follows that

$$\begin{aligned} u = x_1x_2\cdots x_k &= (yx_1)(yx_2)\cdots(yx_k) \\ &= y^k u. \end{aligned}$$

But u itself is invertible (its inverse is $x_k^{-1}\cdots x_2^{-1}x_1^{-1}$), so multiplying by u^{-1} we obtain $y^k = 1$ as required. □

Theorem 13.3.2 can also be stated as a theorem about integers, in the following manner:

$$\text{if} \quad \gcd(y, m) = 1 \quad \text{then} \quad y^{\phi(m)} \equiv 1 \,(\text{mod } m).$$

This is known as **Euler's Theorem**. The particular case when m is replaced by a prime p is **Fermat's Theorem**:

$$\text{if} \quad p \nmid y \quad \text{then} \quad y^{p-1} \equiv 1 \,(\text{mod } p).$$

Example Show that for any positive integer n and prime p

$$n^p \equiv n \,(\text{mod } p).$$

Deduce that, in base 10, the last digits of n and n^5 are always the same.

Solution Suppose $p \nmid n$; then by Fermat's theorem $n^{p-1} \equiv 1 \,(\text{mod } p)$ and so $n^p \equiv n \,(\text{mod } p)$. On the other hand, if $p \mid n$ then both n and n^p are congruent to 0 modulo p.

Using this result with $p = 5$ we have $n^5 - n \equiv 0 \,(\text{mod } 5)$. Also, $n^5 - n = n(n-1)(n^3+n^2+n+1)$ and, since one of the first two factors is even, $n^5 - n \equiv 0 \,(\text{mod } 2)$. Thus $n^5 - n$ is divisible by 5 and by 2, and so it is divisible by 10, which is equivalent to the result stated. □

Exercises 13.3 (continued)

4 Find the inverses of

(i) 2 in $\mathbb{Z}_{11}$, (ii) 7 in $\mathbb{Z}_{15}$,

(iii) 7 in $\mathbb{Z}_{16}$, (iv) 5 in $\mathbb{Z}_{13}$.

5 Use Fermat's Theorem to calculate the remainder when 3^{47} is divided by 23.

6 Suppose that a and b are integers and p is a prime. Use Fermat's Theorem to show that

$$(a+b)^p \equiv a^p + b^p \,(\text{mod } p).$$

7 Show that the equation $x = x^{-1}$ in $\mathbb{Z}_p$ implies that $x^2 - 1 = 0$, and deduce that 1 and -1 are the only elements of $\mathbb{Z}_p$ which are equal to their own inverse.

8 By considering the product of all the non-zero elements of $\mathbb{Z}_p$ show that

$$(p-1)! \equiv -1 \,(\text{mod } p).$$

13.4 Cyclic constructions for designs

In this section and the next one we shall study some constructions based on the cyclic properties of modular arithmetic.

If S is a subset of $\mathbb{Z}_m$ and i is any member of $\mathbb{Z}_m$, then we denote by $S + i$ the subset of $\mathbb{Z}_m$ obtained by adding i to each member of S. For example, when $m = 12$ and $S = \{0, 1, 3, 11\}$, we have

$$S + 1 = \{1, 2, 4, 0\}, \qquad S + 2 = \{2, 3, 5, 1\},$$

and so on. We shall investigate the possibility of using the subsets $S + i$ ($i \in \mathbb{Z}_m$) as the blocks of a design.

15

15
Graphs

Contents

15.1 Graphs and their representation 178
15.2 Isomorphism of graphs 179
15.3 Degree 181
15.4 Paths and cycles 183
15.5 Trees 185
15.6 Colouring the vertices of a graph 187
15.7 The greedy algorithm for vertex-colouring 188
15.8 Miscellaneous Exercises 191

15.1 Graphs and their representation

The objects which we shall call *graphs* are very useful throughout Discrete Mathematics. Their name is derived from the fact that they can be regarded as a form of graphical (or pictorial) notation, and in this respect alone they resemble the familiar graphs of functions which are studied in elementary mathematics. But our *graphs* are quite different from graphs of functions, and correspond more closely with the objects which, in everyday language that we call 'networks'.

Definition A **graph** G consists of a finite set V, whose members are called **vertices**, and a set E of 2-subsets of V, whose members are called **edges**. We usually write $G = (V, E)$ and say that V is the **vertex set** and E is the **edge set**.

The restriction to finite sets is not essential, but it is convenient for us because we shall not consider infinite 'graphs' in this book.

A typical example of a graph $G = (V, E)$ is given by the sets

$$V = \{a, b, c, d, z\}, \qquad E = \{\{a, b\}, \{a, d\}, \{b, z\}, \{c, d\}, \{d, z\}\}.$$

This example and the definition itself are not very illuminating, and it is only when we turn to the *pictorial representation* of a graph that the light dawns. We represent the vertices as points, and join two points by a line whenever the corresponding pair of vertices is an edge. Thus Fig. 15.1 is a pictorial representation of the graph given in the example above. This kind of a representation is extremely convenient for working 'by hand' with relatively small graphs. Furthermore, its intuitive familiarity is of great assistance in formulating and understanding abstract arguments. Here is an example; in fact it is the same problem that was discussed in Section 6.3.

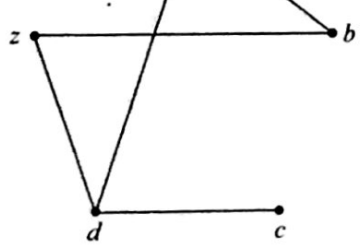


Fig. 15.1
A pictorial representation of a graph.

Example Professor McBrain and his wife April give a party at which there are four other married couples. Some pairs of people shake hands when they meet, but naturally no couple shake hands with each other. At the end of the party the Professor asks everyone else how many people they have shaken hands with, and he receives nine different answers. How many people shook hands with April?

Solution We construct a graph whose vertices are the people at the party, and there is an edge $\{x, y\}$ whenever x and y shook hands. Since there are nine people apart from Professor McBrain, and the maximum number of handshakes in which

any one person can be involved is eight, it follows that the nine different answers received by the Professor must be 0, 1, 2, 3, 4, 5, 6, 7, 8. We denote the vertices by these numbers and use M for McBrain himself. So we have a pictorial representation as in Fig. 15.2.

Now, vertex 8 is joined to all the other vertices except one, which must therefore represent the spouse of 8. This vertex must be 0, since it is certainly not joined to 8 (or any other vertex, for that matter). Thus 8 and 0 are a married couple, and 8 is joined to 1, 2, ..., 7 and M. In particular 1 is joined to 8 and this is the only edge from 1. Hence vertex 7 is not joined to 0 and 1 (only), and the spouse of 7 must be 1, since 0 is married to 8. Continuing in the same way, we see that 6 and 2, and 5 and 3 are married couples. It follows that M and 4 are married, so vertex 4 represents April, who shook hands with four people. □

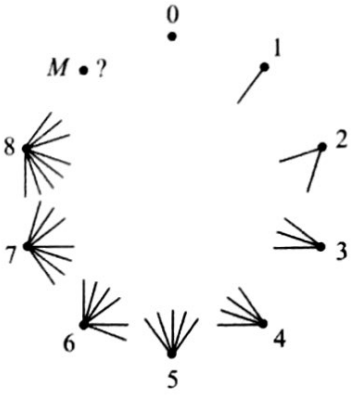


Fig. 15.2
April's party.

Although the pictorial representation of graphs is intuitively appealing to human beings, it is clearly useless when we wish to communicate with a computer. For that purpose we must represent a graph by some kind of list or table. Let us say that two vertices x and y of a graph are **adjacent** whenever $\{x, y\}$ is an edge. (We also say that x and y are **neighbours**.) Then we can represent a graph $G = (V, E)$ by its **adjacency list**, wherein each vertex v heads a list of those vertices adjacent to v. The graph in Fig. 15.1 has the adjacency list

a	b	c	d	z
b	a	d	a	b
d	z		c	d
			z	

.

Exercises 15.1

1 Three houses A, B, C each has to be connected to the gas, water, and electricity supplies: G, W, E. Write down the adjacency list for the graph which represents this problem, and construct a pictorial representation of it. Can you find a picture in which the lines representing the edges do not cross?

2 The pathways in a formal garden are to be laid out in the form of a **wheel graph** W_n, whose vertex set is $V = \{0, 1, 2, \ldots, n\}$ and whose edges are

$$\{0, 1\}, \quad \{0, 2\}, \ldots, \{0, n\},$$
$$\{1, 2\}, \quad \{2, 3\}, \ldots, \{n-1, n\}, \quad \{n, 1\}.$$

Describe a route around the pathways which starts and ends at vertex 0 and visits every vertex once only.

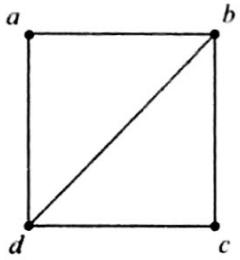
3 For each positive integer n we define the **complete graph** K_n to be the graph with n vertices in which each pair of vertices is adjacent. How many edges has K_n? For which values of n can you find a pictorial representation of K_n with the property that the lines representing the edges do not cross?

4 A 3-*cycle* in a graph is a set of three mutually adjacent vertices. Construct a graph with five vertices and six edges which contains no 3-cycles.

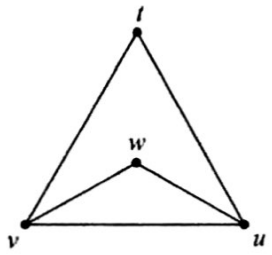
15.2 Isomorphism of graphs

At this point it must be emphasized that a graph is defined as an abstract mathematical entity. It is in this light that we shall discuss the important question of what we mean by saying that two graphs are 'the same'.

Clearly, the important thing about a graph is not the names of the vertices, nor is it their representation pictorially or in any other concrete way. The characteristic property of a graph is the way in which the vertices are linked by its edges. This motivates the following definition.



G_1



G_2

Fig. 15.3
G_1 and G_2 are isomorphic.

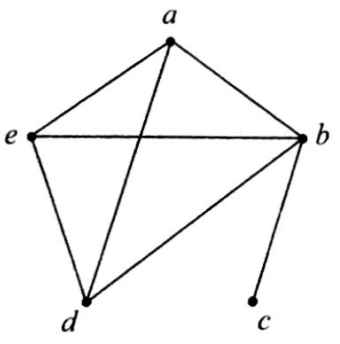
Definition Two graphs G_1 and G_2 are said to be **isomorphic** when there is a bijection α from the vertex set of G_1 to the vertex set of G_2 such that $\{\alpha(x), \alpha(y)\}$ is an edge of G_2 if and only if $\{x, y\}$ is an edge of G_1. The bijection α is said to be an **isomorphism**.

For example, consider the two graphs depicted in Fig. 15.3. In this case there is a bijection from the vertex set of G_1 to the vertex set of G_2 which has the required property; it is given by

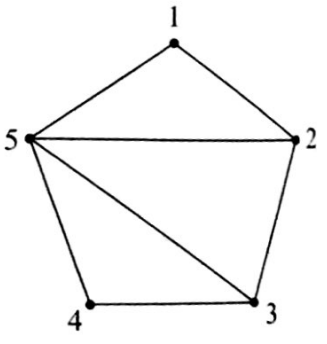
$$\alpha(a) = t, \qquad \alpha(b) = v, \qquad \alpha(c) = w, \qquad \alpha(d) = u.$$

We can verify that each edge of G_1 corresponds uniquely to an edge of G_2, and conversely. For instance, the edge bc of G_1 corresponds to the edge vw of G_2, and so on. (We shall customarily use the abbreviation xy for an edge $\{x, y\}$, keeping in mind that an edge is an unordered pair, so that xy and yx mean the same thing.)

When, as in Fig. 15.3, two graphs G_1 and G_2 are isomorphic we usually regard them as being 'the same' graph. In order to show that two graphs are *not* isomorphic, we must demonstrate that there can be no bijection from the vertex set of one to the vertex set of the other which takes edges to edges. If the two graphs have different numbers of vertices, then no bijection is possible, and the graphs cannot be isomorphic. If the graphs have the same number of vertices, but different numbers of edges, then there are bijections but none of them can be an isomorphism (Ex. 15.8.10). Even if the graphs have the same numbers of vertices and edges, they need not be isomorphic. For example, the two graphs in Fig. 15.4 both have five vertices and seven edges, but they are not isomorphic. One way to show this is to remark that the vertices a, b, d, e form a *complete* subgraph of G_1 (each pair of them is linked by an edge). Any isomorphism must take these vertices to four vertices of G_2 which have the same property, and since there is no such set of vertices in G_2 there can be no isomorphism.



G_1



G_2

Fig. 15.4
G_1 and G_2 are not isomorphic.

Exercises 15.2

1 Prove that the graphs shown in Fig. 15.5 are not isomorphic.

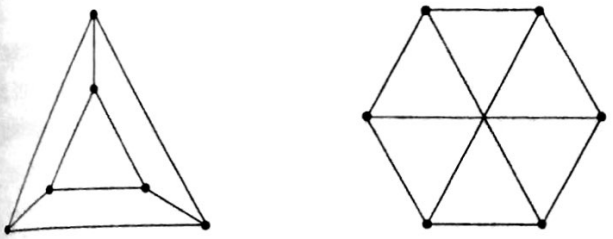


Fig. 15.5
Show that these graphs are not isomorphic.

2 Find an isomorphism between the graphs defined by the following adjacency lists. (Both lists specify versions of a famous graph, known as **Petersen's graph**. See also Ex. 15.8.3.)

a	b	c	d	e	f	g	h	i	j	0	1	2	3	4	5	6	7	8	9
b	a	b	c	d	a	b	c	d	e	1	2	3	4	5	0	1	0	2	6
e	c	d	e	a	h	i	j	f	g	5	0	1	2	3	4	4	3	5	7
f	g	h	i	j	i	j	f	g	h	7	6	8	7	6	8	9	9	9	8

3 Let $G = (V, E)$ be the graph defined as follows. The vertex set V is the set of all words of length 3 in the alphabet $\{0, 1\}$, and the edge set E contains those pairs of words which differ in exactly one position. Show that G is isomorphic to the graph formed by the corners and edges of an ordinary cube.

15.3 Degree

The **degree** of a vertex v in a graph $G = (V, E)$ is the number of edges of G which contain v. We shall use the notation $\delta(v)$ for the degree of v, so formally

$$\delta(v) = |D_v|, \quad \text{where} \quad D_v = \{e \in E \mid v \in e\}.$$

The graph depicted in Fig. 15.1 has $\delta(a) = 2$, $\delta(b) = 2$, $\delta(c) = 1$, $\delta(d) = 3$, $\delta(z) = 2$. The first theorem of graph theory tells us that the sum of these numbers is twice the number of edges; it is a simple application of the method for counting sets of pairs given in Section 10.2.

Theorem 15.3 The sum of the values of the degree $\delta(v)$, taken over all the vertices v of a graph $G = (V, E)$, is equal to twice the number of edges:

$$\sum_{v \in V} \delta(v) = 2\,|E|.$$

Proof Let S denote the subset of $V \times E$ consisting of those pairs (v, e) for which v belongs to e. For each v in V the 'row total' $r_v(S)$ is the number of edges containing v, and so it is equal to $\delta(v)$. For each e in E the 'column total' $c_e(S)$ is the number of vertices in e, which is 2. Hence, by Theorem 10.2

$$\sum_{v \in V} \delta(v) = 2 + 2 + \cdots + 2 = 2\,|E|,$$

as required. □

There is a useful corollary of this result. Let us say that a vertex of G is **odd** if its degree is odd, and **even** if its degree is even. Denote by V_o and V_e the sets

of odd and even vertices, respectively, so that $V = V_o \cup V_e$ is a partition of V. By Theorem 15.3, we have

$$\sum_{v \in V_o} \delta(v) + \sum_{v \in V_e} \delta(v) = 2|E|.$$

Now each term in the second sum is even, and so this sum is an even number. Since the right-hand side is an even number, the first sum must also be even. But a sum of odd numbers can only be even if there is an even number of them. In other words

the number of odd vertices is even.

This result is sometimes known as the handshaking lemma, in view of its interpretation in terms of people and handshakes: given any set of people, the number of people who have shaken hands with an odd number of other members of the set is even.

A graph in which all vertices have the same degree r is said to be **regular** with degree r. In this case, the result of Theorem 15.3 becomes

$$r|V| = 2|E|.$$

Many of the graphs which occur in applications are regular. We have already met the complete graphs K_n (Ex. 15.1.3); they are regular, with degree $n-1$. In elementary geometry we discuss the n-sided polygons, which correspond in graph theory to **cycle graphs** C_n. Formally, we can say that the vertex set of C_n is $\mathbb{Z}_n$, the vertices i and j being joined by an edge whenever $j = i+1$ or $j = i-1$ in $\mathbb{Z}_n$. Clearly, C_n is a regular graph with degree 2, provided $n \geq 3$.

An important application of the notion of degree is to the problem of testing whether or not two graphs are isomorphic. If $\alpha : V_1 \to V_2$ is an isomorphism between G_1 and G_2, and $\alpha(v) = w$, then each edge containing v is transformed by α into an edge containing w. Consequently, $\delta(v) = \delta(w)$. On the other hand, if G_1 has a vertex x, with $\delta(x) = \delta_0$ say, and G_2 has no vertex with degree δ_0, then G_1 and G_2 cannot be isomorphic. This gives us another way of distinguishing between the two graphs in Fig. 15.4, since the first graph has a vertex of degree 1 and the second graph has no such vertex.

A further extension of this idea is given in Ex. 15.3.4.

Exercises 15.3

1 Is it possible that the following lists are the degrees of all the vertices of a graph? If so, give a pictorial representation of such a graph. (Remember that there is at most one edge joining each pair of vertices.)

(i) 2, 2, 2, 3. (ii) 1, 2, 2, 3, 4.

(iii) 2, 2, 4, 4, 4. (iv) 1, 2, 3, 4.

2 If $G = (V, E)$ is a graph, the **complement** $\bar{G}$ of G is the graph whose vertex set is V and whose edges join those pairs of vertices which are not joined in G. If G has n vertices and their degrees are $d_1, d_2, \ldots, d_n$, what are the degrees of the vertices of $\bar{G}$?

3 Find as many different (non-isomorphic) regular graphs with degree 4 and seven vertices as you can. [Hint: consider the complement of such a graph.]

4 Suppose G_1 and G_2 are isomorphic graphs. For each $k \geq 0$ let $n_i(k)$ be the number of vertices of G_i which have degree k ($i = 1, 2$). Show that $n_1(k) = n_2(k)$.

5 Show that if G is a graph with at least two vertices then G has two vertices with the same degree.

15.4 Paths and cycles

Frequently, we use graphs as models of practical situations involving routes: the vertices represent towns or junctions, and each edge represents a road or some other form of communication link. The definitions in this section are best conceived with that kind of a picture in mind.

Definition A **walk** in a graph G is a sequence of vertices

$$v_1, v_2, \ldots, v_k,$$

such that v_i and v_{i+1} are adjacent ($1 \leq i \leq k-1$). If all its vertices are distinct, a walk is called a **path**.

Thus a walk specifies a route in G which proceeds from a vertex to an adjacent one, and so on. A walk may visit any vertex several times, and in particular, it may reverse its direction by going from x to y and immediately back to x again. In a path, each vertex is visited at most once.

Let us write $x \sim y$ whenever vertices x and y of G can be joined by a path in G: strictly speaking, this means that there is a path $v_1, v_2, \ldots, v_k$ in G with $x = v_1$ and $y = v_k$. It is a simple matter to verify that $\sim$ is an equivalence relation on the vertex set V of G, and so V is partitioned into disjoint equivalence classes. Two vertices are in the same class if they can be joined by a path, and in different classes if there is no such path.

Definition Suppose $G = (V, E)$ is a graph and the partition of V corresponding to the equivalence relation $\sim$ is

$$V = V_1 \cup V_2 \cup \cdots \cup V_r.$$

Let E_i ($1 \leq i \leq r$) denote the subset of E comprising those edges whose ends are both in V_i. Then the graphs $G_i = (V_i, E_i)$ are called the **components** of G. If G has just one component, it is said to be **connected**.

The terminology is almost self-explanatory. The graph shown in Fig. 15.6 has two components, and is therefore not connected. The decomposition of a graph into components is very useful, since many properties of graphs can be established by considering each component separately. For this reason, theorems about graphs are often proved only for the class of connected graphs.

When a fairly small graph is given by a pictorial representation it is quite easy to spot as to whether it is connected or not. However, when a graph is given by an adjacency list we shall need an efficient algorithm to decide if it is connected. This problem will be studied in the next chapter.

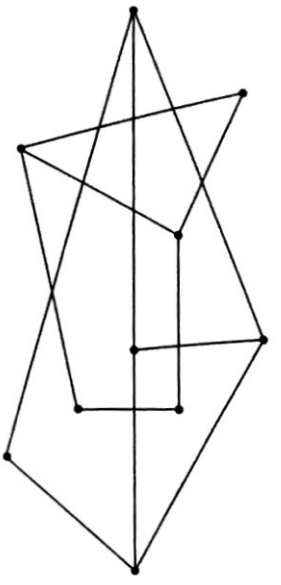


Fig. 15.6
A graph with two components.

A walk $v_1, v_2, \ldots, v_{r+1}$ whose vertices are all distinct except that $v_1 = v_{r+1}$ is called a **cycle**. It has r distinct vertices and r edges, and we often speak of an **r-cycle**, or cycle of **length** r.

Example Two senior members of the Mathematics Department at the University of Folornia plan to spend their vacation on the island of Wanda. Figure 15.7 represents the interesting places on the island and the roads linking them. Dr Elsie Chunner is a tourist by nature, and wishes to visit each place once and return to her starting point. Dr Bob Dodder is an explorer, and wishes to traverse every road just once, in either direction; he is prepared to start and finish in different places. Can suitable routes for Drs Chunner and Dodder be found?

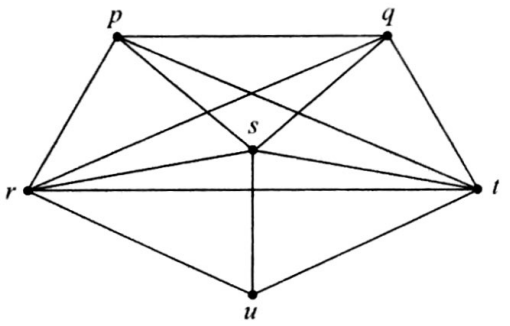


Fig. 15.7
The grand tour.

Solution Dr C can use several routes: one possibility is the cycle p, q, t, s, u, r, p.

However, Dr D has a problem. Let us label his starting vertex x and his finishing vertex y, and suppose for the moment that $x \neq y$. Then he uses one edge at x when he starts and each time he returns to x he must arrive and depart by new edges. In this way, he uses an odd number of edges at x, and so x must be an odd vertex. Similarly, y must be an odd vertex also, since he uses two edges each time he passes through y, and one edge to finish at y. All the remaining vertices must be even, since every time he arrives at an intermediate vertex he also departs, and thereby uses two edges.

In summary, a route for Dr D starting and finishing at distinct vertices x and y is possible only if x and y are odd vertices, and the rest are even. But in the given graph the degrees are as follows:

v:	p	q	r	s	t	u
$\delta(v)$:	4	4	5	5	5	3.

So there are too many odd vertices, and consequently no route for Dr D. If we allow the possibility that $x = y$, the situation is worse still, for then all the vertices would have to be even. □

In general, Dr C's route is a cycle which contains all vertices of the given graph. Such cycles were studied by the Irish mathematician W. R. Hamilton (1805–65), and consequently a cycle with this property is known as a **Hamiltonian cycle**. In our example, it was very easy to find a Hamiltonian cycle, but this was a misleading special case. For a specific graph, it may be a difficult problem to decide whether or not a Hamiltonian cycle exists.

On the other hand, Dr D's problem is easily settled. A walk which uses each edge of a graph exactly once is called a **Eulerian walk**, because Euler was the first to study such walks. He found that if $x \neq y$ a necessary condition for a Eulerian walk starting at x and finishing at y is that x and y are odd vertices and the rest are even, while if $x = y$ the condition is that all vertices are even. Thus a necessary condition for the existence of a Eulerian walk in a graph G is that G has at most two odd vertices. Furthermore, it can be shown that this condition is also sufficient. Since it is easy to compute the degrees of vertices in any graph, it is correspondingly easy to decide if a given graph has a Eulerian walk.

Exercises 15.4

1 Find the number of components of the graph whose adjacency list is

a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
f	*c*	*b*	*h*	*c*	*a*	*b*	*d*	*a*	*a*
i	*g*	*e*		*g*	*i*	*c*		*f*	*f*
j		*g*			*j*	*e*			

2 How many components are there in the graph of April's party (Section 15.1)?

3 Find a Hamiltonian cycle in the graph formed by the vertices and edges of an ordinary cube.

4 Next year Dr Chunner and Dr Dodder intend to visit the island of Meanda, where the interesting places and the roads joining them are represented by the graph whose adjacency list is

0	1	2	3	4	5	6	7	8
1	0	1	0	3	0	1	0	1
3	2	3	2	5	4	5	2	3
5	6	7	4		6	7	6	5
7	8		8		8		8	7.

Is it possible to find routes for them which satisfy the requirements set out in the *Example* on p.184?

5 A mouse intends to eat a $3 \times 3 \times 3$ cube of cheese. Being tidy-minded, it begins at a corner and eats the whole of a $1 \times 1 \times 1$ cube before going on to an adjacent one. Can the mouse end in the centre?

15.5 Trees

Definition We say that a graph T is a **tree** if it has two properties:

(T1) T is connected;

(T2) there are no cycles in T.

Some typical trees are depicted in Fig. 15.8. Because of their special structure and properties, trees occur in many different applications of mathematics, especially in operations research and computer science. We begin our study of them by establishing some simple properties.

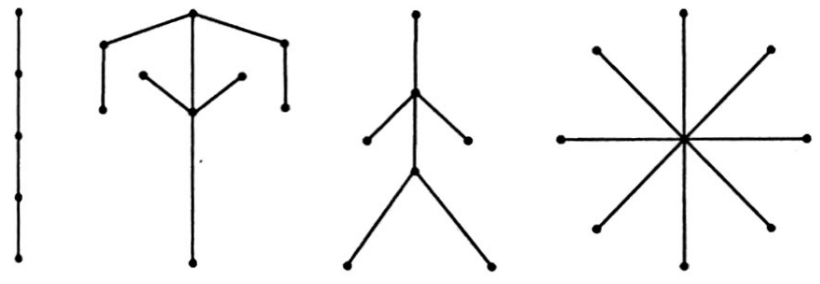


Fig. 15.8
Some trees.

Theorem 15.5 If $T = (V, E)$ is a tree with at least two vertices, then:

(T3) for each pair x, y of vertices there is a unique path in T from x to y;

(T4) the graph obtained from T by removing any edge has two components, each of which is a tree;

(T5) $|E| = |V| - 1$.

Proof (T3) Since T is connected, there is a path from x to y, say

$$x = v_0, v_1, \ldots, v_r = y.$$

If there is a different path, say

$$x = u_0, u_1, \ldots, u_s = y,$$

then let i be the smallest subscript for which $u_{i+1} \neq v_{i+1}$ (Fig. 15.9).

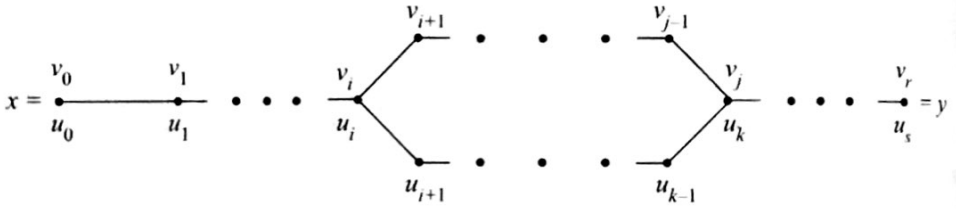


Fig. 15.9
Two distinct paths determine a cycle.

Since both paths finish at y they will meet again, and we can define j to be the smallest subscript such that

$$j > i \quad \text{and} \quad v_j = u_k \text{ for some } k.$$

Then $v_i, v_{i+1}, \ldots, v_j, u_{k-1}, u_{k-2}, \ldots, u_{i+1}, v_i$ is a cycle in T, contrary to the hypothesis. Hence there is just one path in T from x to y.

(T4) Suppose uv is an edge of T, and let $S = (V, E')$ be the graph with the same vertex set as T and edge set $E' = E \setminus uv$. Let V_1 be the set of vertices x of T for which the unique path from x to v in T passes through u. Clearly the relevant path must end with the edge uv, otherwise T would have a cycle. Let V_2 be the complement of V_1 in V.

Each vertex in V_1 is joined by a path in S to u, and each vertex in V_2 is joined in S to v, but there is no path from u to v in S. It follows that V_1 and V_2 are the vertex sets of the two components of S. Each component is (by definition) connected, and it contains no cycles, since there are no cycles in T. Hence the two components are trees.

(T5) The result is true when $|V| = 1$, since the only possible tree has no edges in that case.

Suppose it is true for all trees with k or fewer vertices. Let T be a tree with $|V| = k + 1$, and let uv be any edge of T. If $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ are the trees obtained by removing uv from T, we have

$$|V_1| + |V_2| = |V|, \qquad |E_1| + |E_2| = |E| - 1.$$

Applying the induction hypothesis of T_1 and T_2 we have

$$|E| = |E_1| + |E_2| + 1 = |V_1| - 1 + |V_2| - 1 + 1 = |V| - 1,$$

as required. Hence the result is true for all positive integers. □

The properties (T1)–(T5) provide several alternative ways of defining a tree. For example, property (T3) alone can be taken as the defining property, instead of (T1) and (T2). We have already shown that (T3) is a consequence of (T1) and (T2), so it only remains to show that the converse holds (Ex. 15.5.3).

Exercises 15.5

1 There are six different (that is, mutually non-isomorphic) trees with six vertices: draw them.

2 Let $T = (V, E)$ be a tree with $|V| \geq 2$. Using property (T5) and Theorem 15.3 show that T has at least two vertices with degree 1.

3 Show that property (T3) implies (T1) and (T2).

4 A **forest** is a graph satisfying (T2) but not necessarily (T1). Prove that if $F = (V, E)$ is a forest with c components then

$$|E| = |V| - c.$$

15.6 Colouring the vertices of a graph

A problem which occurs frequently in modern life is that of timetabling a set of events in such a way as to avoid clashes. We shall consider a very simple case as an example of how the theory of graphs can help us to study this problem.

Suppose we wish to schedule six one-hour lectures, $v_1, v_2, v_3, v_4, v_5, v_6$. Among the potential audience there are people who wish to hear both v_1 and v_2, v_1 and v_4, v_3 and v_5, v_2 and v_6, v_4 and v_5, v_5 and v_6, and v_1 and v_6. How many hours are necessary in order that the lectures can be given without clashes?

We can represent the situation by a graph (Fig. 15.10). The vertices correspond to the six lectures, and the edges signify the potential clashes. A timetable which achieves the object of avoiding clashes is as follows:

Hour 1	Hour 2	Hour 3	Hour 4
v_1 and v_3	v_2 and v_4	v_5	v_6

In mathematical terms, we have a partition of the vertex set into four parts, with the property that no part contains a pair of adjacent vertices of the graph. A more graphic description utilizes the function

$$c\colon \{v_1, v_2, v_3, v_4, v_5, v_6\} \rightarrow \{1, 2, 3, 4\}$$

which assigns to each vertex (lecture) the hour scheduled for it. Usually, we speak of colours assigned to the vertices, rather than hours, but clearly the exact nature of the objects 1, 2, 3, 4 is unimportant. We can use the names of actual colours, red, green, blue, yellow, or we can speak of colour 1, colour 2, and so on. The important point is that vertices which are adjacent in the graph must be given different colours.

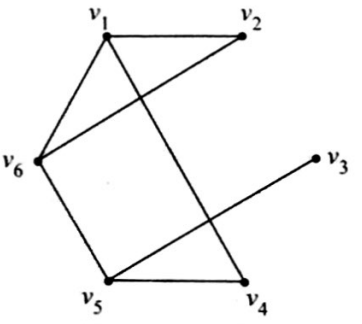


Fig. 15.10
The graph for a timetable problem.

Definition A **vertex colouring** of a graph $G = (V, E)$ is a function $c: V \to \mathbb{N}$ with the property that

$$c(x) \neq c(y) \quad \text{whenever} \quad \{x, y\} \in E.$$

The **chromatic number** of G, written $\chi(G)$, is defined to be the least integer k for which there is a vertex colouring of G using k colours. In other words, $\chi(G) = k$ if and only if there is a vertex colouring c which is a function from V to $\mathbb{N}_k$, and k is the least integer with this property.

Returning to Fig. 15.10, we see that our first attempt at a timetable is equivalent to a vertex colouring using four colours. The smallest number of hours needed is the chromatic number of the graph, and we have now to ask if this number is less than 4. A quick trial with three colours will lead us to a solution

Colour 1	Colour 2	Colour 3
v_1	v_2 and v_5	v_3, v_4, and v_6.

Furthermore, at least three colours are needed, since v_1, v_2, and v_6 are mutually adjacent and must have different colours. So we conclude that the chromatic number of this graph is 3.

In general, in order to show that the chromatic number of a given graph is k, we have to do two things:

(i) find a vertex colouring using k colours;

(ii) show that no vertex colouring uses fewer than k colours.

Exercises 15.6

1 Find the chromatic numbers of the following graphs:

(i) a complete graph K_n;

(ii) a cycle graph C_{2r} with an even number of vertices;

(iii) a cycle graph C_{2r+1} with an odd number of vertices.

2 Determine the chromatic numbers of the graphs depicted in Fig. 15.11.

3 Describe all graphs G for which $\chi(G) = 1$.

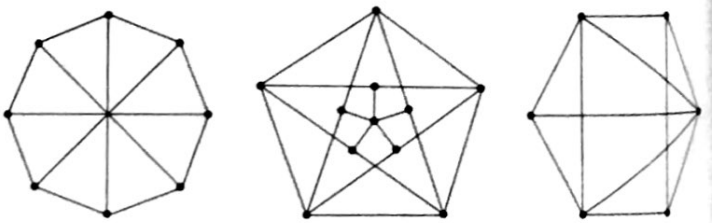


Fig. 15.11
Find the chromatic numbers.

15.7 The greedy algorithm for vertex colouring

The problem of finding the chromatic number of a given graph is a difficult one. Indeed, there is no known algorithm for the problem which works in 'polynomial-time', and most people believe that no such algorithm exists. However, there is a simple method of constructing a vertex colouring using a 'reasonable' number of colours.

The method is to assign colours to the vertices in order, in such a way that each vertex receives the first colour which has not already been assigned to one of its neighbours. In this algorithm we insist on making the best choice that we can at

each step, without looking ahead to see if that choice will create problems later on. An algorithm of this kind is usually referred to as a **greedy algorithm**.

The greedy algorithm for vertex colouring is easy to program. Suppose we have arranged the vertices in some order $v_1, v_2, \ldots, v_n$. We assign colour 1 to v_1; for each v_i $(2 \leq i \leq n)$ we form the set S of colours assigned to vertices v_j $(1 \leq j < i)$ which are adjacent to v_i; and we give v_i the first colour not in S. (In practice, more sophisticated methods of handling the data may be used.)

Greedy vertex colouring algorithm

```
assign colour 1 to v1;
for i:=2 to n do
 begin
 let S be the empty set of colours;
 for j:=1 to i-1 do
  if vj is adjacent to vi
    then add the colour of vj to S;
 k:=1;
 while colour k is in S do k:=k+1;
 assign colour k to vi
 end
```

Because the greedy strategy is a shortsighted one, the number of colours which it uses will normally be greater than the minimum possible. For example, the greedy algorithm applied to the graph in Fig. 15.10 gives precisely the vertex colouring with four colours which we first proposed as our 'timetable', whereas we later found a vertex colouring using only three colours. Of course, everything depends upon the order initially given to the vertices. It is quite easy to see that if we hit on the right order, then the greedy algorithm will give us a best possible colouring (Ex. 15.7.2). But there are $n!$ orders altogether, and if we have to check each one of them, the algorithm will require 'exponential-time'.

Despite its wastefulness, the greedy algorithm is useful both in theory and in practice. We shall prove two theorems by using the greedy strategy.

Theorem 15.7.1 If G is a graph with maximum degree k, then

(i) $\chi(G) \leq k + 1$,

(ii) if G is connected and not regular, $\chi(G) \leq k$.

Proof (i) Let $v_1, v_2, \ldots, v_n$ be any ordering of the vertices of G. Each vertex v_i has at most k neighbours, and so the set S of colours assigned by the greedy algorithm to vertices v_j which are adjacent to v_i $(1 \leq j < i)$ has cardinality k at most. Hence at least one of the colours $1, 2, \ldots, k + 1$ is not in S, and the greedy algorithm will assign the first of these to v_i. In this way the greedy algorithm produces a vertex colouring of G using at most $k+1$ colours, and so $\chi(G) \leq k+1$.

(ii) For this part, we arrange the vertices in a special order, starting with v_n and working backwards. Since G has maximum degree k and is not regular, there is at least one vertex of G whose degree is less than k: call it v_n. List the neighbours of

v_n as $v_{n-1}, v_{n-2}, \ldots, v_{n-r}$; there are at most $k-1$ of them. Next, list the neighbours of v_{n-1} (except for v_n), remarking that since the degree of v_{n-1} is at most k, there are at most $k-1$ such vertices. Next list all the neighbours of v_{n-2} which have not already been listed, and so on. Since G is connected, the list will eventually contain all vertices of G. Furthermore, the method of construction ensures that every vertex is adjacent to at most $k-1$ of its predecessors in the order $v_1, v_2, \ldots, v_n$.

It follows from the same argument as used in part (i) that (for this ordering) the greedy algorithm will require at most k colours. Hence $\chi(G) \leq k$. □

Part (ii) of the theorem is false if we allow G to be regular. The reader who has correctly answered Ex. 15.6.1 will be able to supply two examples of this fact: the complete graphs, and the odd cycle graphs, both of which require $k+1$ colours. However, it can be shown that these are the only counter-examples.

Another useful consequence of the greedy algorithm concerns graphs G for which $\chi(G) = 2$. For such a graph, the sets V_1 and V_2 of vertices assigned the colours 1 and 2, respectively, form a partition of V, with the property that every edge of G has one vertex in V_1 and the other in V_2. For this reason, when $\chi(G) = 2$ we say that G is **bipartite**. A vertex colouring of the cube with two colours is illustrated in Fig. 15.12, together with an alternative picture which emphasizes the bipartite nature of the graph. We frequently use the latter kind of illustration in dealing with bipartite graphs.

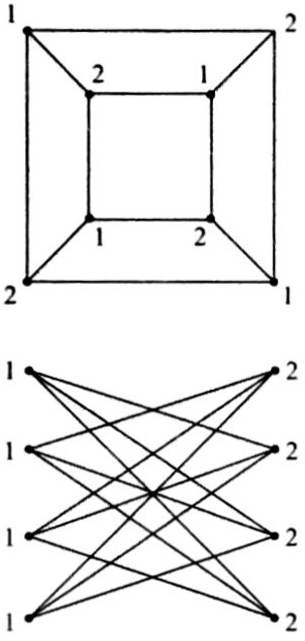


Fig. 15.12
The cube as a bipartite graph.

Theorem 15.7.2 A graph is bipartite if and only if it contains no cycles with odd length.

Proof If there is a cycle with an odd number of vertices then three colours are required for a vertex colouring of that cycle alone, and the chromatic number of the graph is at least 3. Hence if the graph is bipartite it can have no odd cycles.

Conversely, suppose G is a graph with no odd cycles. We shall construct an ordering of G for which the greedy algorithm produces a vertex colouring with two colours. Choose any vertex and call it v_1; we shall say that v_1 is at *level 0*. Next, list the neighbours of v_1, calling them $v_2, v_3, \ldots, v_r$; we shall say that these vertices are at *level 1*. Next, list the neighbours of the level 1 vertices (except v_1); we shall say that these vertices are at *level 2*. Continue in this way, listing at *level* l all those vertices adjacent to the *level* $l-1$ vertices, except for those previously listed at *level* $l-2$. When no new vertices can be added in this way, we have a component G_0 of G (if G is connected, $G_0 = G$).

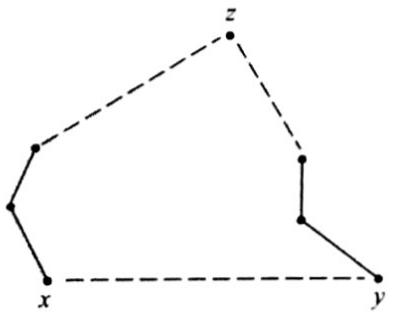


Fig. 15.13
Adjacent vertices at the same level yield an odd cycle.

The crucial feature of this ordering is that a vertex at level l can be adjacent only to vertices at levels $l-1$ and $l+1$, not to vertices at the same level. For suppose x and y are vertices at the same level; then they are joined by paths of equal length m to some vertex z at a previous level, and the paths can be chosen so that z is the only common vertex (Fig. 15.13). If x and y were adjacent, there would be a cycle of odd length $2m+1$ in G_0, contrary to the hypothesis.

It follows that the greedy algorithm assigns colour 1 to vertices at levels 0, 2, 4, ..., and colour 2 to vertices at levels 1, 3, 5, ... Hence $\chi(G_0) = 2$. Repeating the same argument for each component of G, we obtain the required result. □

Exercises 15.7

1 Find orderings of the vertices of the cube graph (Fig. 15.12) for which the greedy algorithm requires 2, 3, and 4 colours, respectively.

2 Show that for any graph G there is an ordering of the vertices for which the greedy algorithm requires $\chi(G)$ colours. [Hint: use a vertex colouring with $\chi(G)$ colours to define the required ordering.]

3 Let $e_i(G)$ denote the number of vertices of a graph G whose degree is strictly greater than i. Use the greedy algorithm to show that if $e_i(G) \leq i + 1$ for some i, then $\chi(G) \leq i + 1$.

4 The graph M_r ($r \geq 2$) is obtained from the cycle graph C_{2r} by adding extra edges joining each pair of opposite vertices. Show that

(i) M_r is bipartite when r is odd,

(ii) $\chi(M_r) = 3$ when r is even and $r \neq 2$,

(iii) $\chi(M_2) = 4$.

15.8 Miscellaneous Exercises

1 For which values of n is it true that the complete graph K_n has a Eulerian walk?

2 Use the principle of induction to show that if $G = (V, E)$ is a graph with $|V| = 2m$, and G has no 3-cycles, then $|E| \leq m^2$.

3 Let $X = \{1, 2, 3, 4, 5\}$ and let V denote the set of all 2-subsets of X. Let E denote the set of pairs of members of V which are disjoint (as subsets of X). Show that the graph $G = (V, E)$ is isomorphic with the graph depicted in Fig. 15.14. Also show that this is just another version of Petersen's graph, introduced in Ex. 15.2.2.

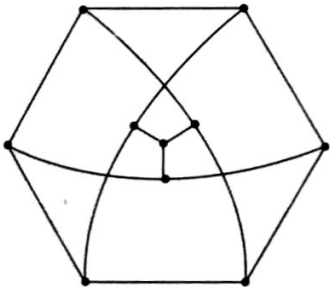


Fig. 15.14 Petersen's graph.

4 Let G be a bipartite graph with an odd number of vertices. Show that G cannot have a Hamiltonian cycle.

5 The **k-cube** Q_k is the graph whose vertices are the words of length k in the alphabet $\{0, 1\}$ and whose edges join words which differ in exactly one position. Show that

(i) Q_k is a regular graph with degree k,

(ii) Q_k is bipartite.

6 Prove that the graph Q_k defined in Ex. 15.8.5 has a Hamiltonian cycle.

7 Show that Petersen's graph does not have a Hamiltonian cycle.

8 In a game of dominos (Ex. 10.7.2) the rules require that the dominos be placed in a line so that adjacent dominos have matching numbers: $[x|y]$ is next to $[y|z]$, and so on. By regarding the dominos for which $x \neq y$ as the edges of the complete graph K_7, show that it is possible to have a game in which all the dominos are used.

9 Calculate the number of Eulerian walks in K_7 and the number of complete games of dominos.

10 Show that if $\alpha: V_1 \to V_2$ is an isomorphism of the graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ then the function $\beta: E_1 \to E_2$ defined by

$$\beta\{x, y\} = \{\alpha(x), \alpha(y)\} \quad (\{x, y\} \in E_1)$$

is a bijection.

11 If G is a regular graph with degree k and n vertices show that

$$\chi(G) \geq \frac{n}{n-k}.$$

12 Construct five mutually non-isomorphic connected regular graphs with degree 3 and eight vertices.

13 Show that the complete graph K_{2n+1} is the union of n Hamiltonian cycles, no two of which have a common edge.

14 Is it possible for a knight to visit all the squares of a chessboard exactly once and return to its starting square? Interpret your answer in terms of Hamiltonian cycles in a certain graph.

15 The **odd graph** O_k is defined as follows (when $k \geq 2$): the vertices are the $(k-1)$-subsets of a $(2k-1)$-set, and the edges join disjoint subsets. (Thus O_3 is Petersen's graph.) Show that $\chi(O_k) = 3$ for all $k \geq 2$.

17

17
Bipartite graphs and matching problems

Contents

17.1 Relations and bipartite graphs 210
17.2 Edge colourings of graphs 212
17.3 Application of edge colouring to latin squares 213
17.4 Matchings 216
17.5 Maximum matchings 219
17.6 Transversals for families of finite sets 221
17.7 Miscellaneous Exercises 223

17.1 Relations and bipartite graphs

In Section 10.2 we introduced the wide class of problems which can be expressed in terms of counting a given subset of a product set $X \times Y$. One way of describing such a subset is to say that a member x of X and a member y of Y are 'related' whenever the pair (x, y) belongs to the given subset. For example, if X is a set of students and Y is a set of courses, we could say that x and y are related whenever x is a student who is taking course y.

These remarks lead to the conclusion that a **relation** R between two sets X and Y is simply a subset of $X \times Y$, and so the statements

$$x \text{ and } y \text{ are related (by } R),$$

$$\text{the pair } (x, y) \text{ is in } R,$$

mean precisely the same thing. It is possible that $X = Y$; but in this chapter we shall study relations which are defined when X and Y are *disjoint* sets. We shall base our discussion on a representation of such a relation by a bipartite graph, which we shall now describe.

When R is a relation between disjoint sets X and Y (that is, when R is a subset of $X \times Y$), we define a bipartite graph G representing R as follows. The vertex set of G is the union of X and Y, and the edge set E contains those edges xy for which (x, y) is in R. Since every edge has one vertex in X and one vertex in Y it is clear that G is bipartite, and it is convenient to think of G in pictorial terms, as in Fig. 17.1 for example. In order to emphasize that G is bipartite, we shall use the notation $G = (X \cup Y, E)$ for a graph of this kind.

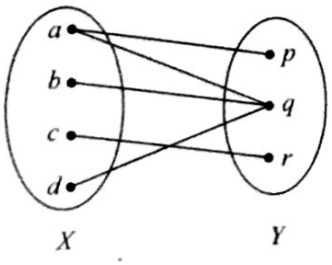


Fig. 17.1
The bipartite graph representing the relation $R = \{(a, p), (a, q), (b, q), (c, r), (d, q)\}$.

The basic theorem on counting sets of pairs has a simple interpretation in graphical terms.

Theorem 17.1 Let $G = (X \cup Y, E)$ be a bipartite graph and let $\delta(v)$ denote the degree of a vertex v in G. Then

$$\sum_{x \in X} \delta(x) = \sum_{y \in Y} \delta(y) = |E|.$$

Proof Since each edge has exactly one vertex in X, the total number of edges is the sum of the degrees of the vertices in X. Similarly, the total number of edges is equal to the sum of the degrees of the vertices in Y. Hence we have the required result. □

Throughout the rest of this chapter we shall frame our results in terms of bipartite graphs, rather than relations. But all the results can be interpreted as results about relations, if we so wish.

Example Suppose we are given a set of people and a set of jobs, such that each person is qualified to do exactly k of the jobs and there are exactly k people qualified to do each job. Show that

(i) the number of people is equal to the number of jobs;

(ii) given any n-subset A of the people there are at least n jobs for which some member of A is qualified.

Solution Let X denote the set of people and Y the set of jobs. Define x and y to be related whenever person x is qualified to do job y. The given condition says that the bipartite graph $G = (X \cup Y, E)$ representing this relation is a regular graph with degree k. Hence, by Theorem 17.1, we have

$$|X|\, k = |Y|\, k = |E|,$$

so that $|X| = |Y|$, as claimed in part (i).

For part (ii), let A be any n-subset of X, and define $J(A)$ to be the set of jobs for which at least one member of A is qualified; that is

$$J(A) = \{y \in Y \mid xy \in E \text{ for some } x \in A\}.$$

Since each vertex belongs to exactly k edges of G, the set E_A of edges which has one vertex in A has cardinality $k\,|A| = kn$. By the definition of $J(A)$, each of these edges has one vertex in $J(A)$, and the given condition says that the total number of edges with one vertex in $J(A)$ is $k\,|J(A)|$. Hence

$$|E_A| = kn \leq k\,|J(A)|,$$

and it follows that $|J(A)| \geq n$, as claimed in part (ii). □

Exercises 17.1

1 Let $X = \{2, 3, 5, 7, 11\}$, $Y = \{99, 100, 101, 102, 103\}$, and define x and y to be related whenever x is a divisor of y. Draw the bipartite graph representing this relation, and verify that Theorem 17.1 is satisfied.

2 The **complete bipartite graph** $K_{r,s}$ is the bipartite graph $(X \cup Y, E)$ in which $|X| = r$, $|Y| = s$, and every pair xy with $x \in X$ and $y \in Y$ is an edge.

(i) What is the degree of each vertex in X?

(ii) What is the degree of each vertex in Y?

(iii) How many edges are there in $K_{r,s}$?

(iv) Describe in ordinary language the relation which $K_{r,s}$ represents.

(v) Show that for any $s \geq 1$ the graph $K_{1,s}$ is a tree.

(vi) Show that $K_{r,s}$ is not a tree whenever $r \geq s \geq 2$.

3 Is the graph illustrated in Fig. 17.2 bipartite?

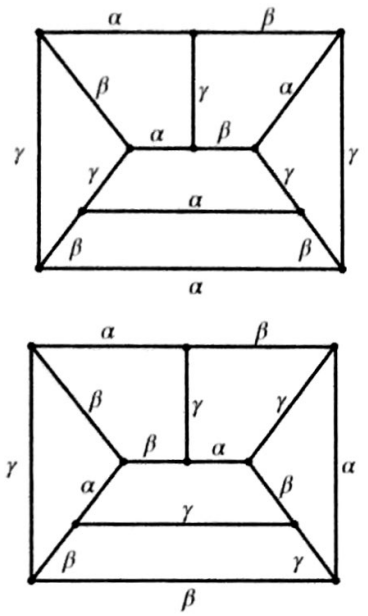


Fig. 17.2
Which one is an edge colouring?

17.2 Edge colourings of graphs

There are many problems which can be interpreted in terms of a partition of the edge set E of a graph, that is, a decomposition of the form

$$E = E_1 \cup E_2 \cup \cdots \cup E_r$$

where $E_1, E_2, \ldots, E_r$ are disjoint, non-empty sets. Intuitively, it is helpful to describe such a partition by means of a 'colouring' of the edges: the edges in E_1 are given a certain colour, those in E_2 a different colour, and so on. We shall use small Greek letters $\alpha, \beta, \gamma, \ldots$, for the names of the colours, and usually we shall require that the colouring satisfies a condition analogous to the one we imposed for a vertex colouring.

Definition Let G be a graph with edge set E. A colouring of E is said to be an **edge colouring** of G if any two edges containing the same vertex have different colours.

The diagram (Fig. 17.2) illustrates two colourings of the edges of the same graph. One is an edge colouring but the other is not.

Exercises 17.2

1 What is the least number of colours required for an edge colouring of

(i) the complete graph K_4;
(ii) the complete graph K_5;
(iii) the cube graph (Fig. 15.12).

2 Suppose there is an edge colouring of Petersen's graph (Fig. 15.14) using only three colours. Show that each colour must be used twice on the edges of the 'outer hexagon' in the Figure, and that there are only two essentially different ways of doing this. Deduce that no such edge colouring of the whole graph is possible.

3 Prove that for any positive integer n the complete bipartite graph $K_{n,n}$ has an edge colouring with n colours.

If v is a vertex with degree k, then there are k edges containing v. In order that these edges shall receive different colours, it is clear that at least k colours must be available. Hence, if K is the maximum degree of G, at least K colours are needed for an edge colouring of G. In general, K colours will not be enough (as in Ex. 17.2.2); however, we can prove that when G is *bipartite* K colours are always sufficient.

Theorem 17.2 If $G = (X \cup Y, E)$ is a bipartite graph, then the minimum number of colours needed for an edge colouring of G is equal to the maximum degree of G.

Proof We shall proceed by induction on m, the number of edges. If $m = 1$, then G has maximum degree 1 and clearly one colour is sufficient to colour the one edge.

Suppose the result is true for any bipartite graph with m edges, and suppose G has $m+1$ edges and maximum degree K. Remove any edge xy from G to obtain a bipartite graph G' with m edges. Since the maximum degree of G' is K or $K-1$, it

follows from the induction hypothesis that there is an edge colouring of G' using at most K colours $\alpha, \beta, \ldots$, and so on.

The degree of x in G' is at most $K - 1$ (since xy has been removed) and so there must be a colour, say α, not used on the edges at x. Similarly, there must be colour β not used at y. If we can choose α and β to be the same then we can give this colour to xy and thereby obtain an edge colouring of G as required. We shall call this the *easy case*. Thus we need only consider the situation where $\alpha \neq \beta$, and we shall show how, in this situation, we can modify the given colouring of G' so that the easy case applies.

Suppose then that $\alpha \neq \beta$. Define a path $x, y_1, x_1, y_2, x_2, \ldots$ as follows:

(1) xy_1 is the edge at x coloured β;
(2) if there is an edge at y_1 coloured α, call it y_1x_1, otherwise stop;
(3) if there is an edge at x_1 coloured β, call it x_1y_2, otherwise stop;
(4) continue with edges coloured α and β alternately until forced to stop.

The path is illustrated in Fig. 17.3. It must stop eventually, since the graph is finite. Also, the path does not contain y, since it arrives at each vertex in Y by an edge coloured β, and β is defined to be a colour not used at y.

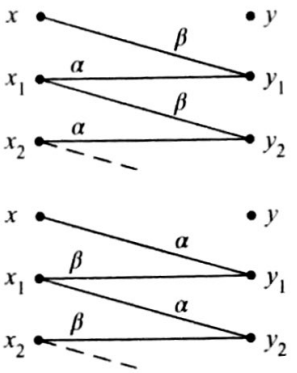


Fig. 17.3
An alternating path and its recolouring.

Now we alter the edge colouring of G' by interchanging the colours α and β on the path, leaving the colours on other edges unchanged (Fig. 17.3). The result is an edge colouring of G' in the easy case: no edge at x is coloured β. From this colouring we obtain the required edge colouring of G by assigning xy the colour β.

By the principle of induction the result holds for all bipartite graphs. □

The crucial part of the proof is the construction of the 'alternating path' $x, y_1, x_1, y_2, \ldots$, which depends critically on the bipartite character of G. We shall make more use of this construction in later sections.

Exercises 17.2 (continued)

4 Show that the graph illustrated in Fig. 17.4 is bipartite and construct an edge colouring of it using only three colours.
5 According to Ex. 15.2.3, the graph Q_3 formed by the corners and edges of a cube can be represented in the following manner: the vertices are the words of length 3 in the alphabet $\{0, 1\}$, and edges join words which differ in just one letter. Use this representation to show that Q_3 is bipartite and to construct an edge colouring of Q_3 using only three colours.
6 Generalize the results of Ex. 5 to the graph Q_k obtained by replacing 3 by k throughout.

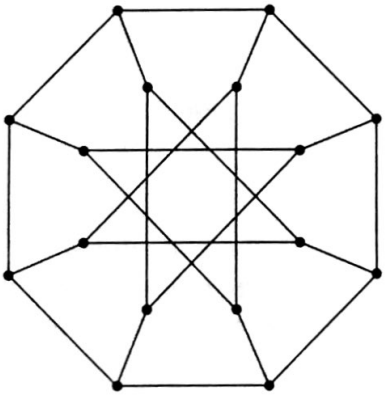


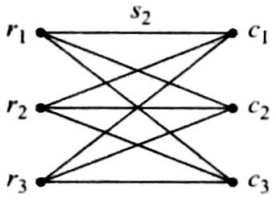
Fig. 17.4
A bipartite graph.

17.3 Application of edge colouring to latin squares

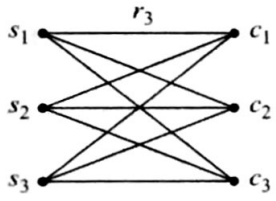
There is a simple relationship between latin squares and edge colourings of bipartite graphs. We may describe an $n \times n$ latin square in terms of the rows $r_1, r_2, \ldots, r_n$,

	c_1	c_2	c_3
r_1	s_2	s_3	s_1
r_2	s_3	s_1	s_2
r_3	s_1	s_2	s_3

(a)



(b)



(c)

Fig. 17.5
A 3 × 3 latin square and two ways of edge colouring $K_{3,3}$.

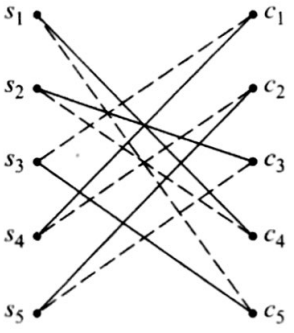
the columns $c_1, c_2, \ldots, c_n$, and the symbols $s_1, s_2, \ldots, s_n$, arranged so that each symbol occurs just once in each row and column. In practice, of course, we often use the same labels for the names of the rows and columns, and for the symbols, but the amplified notation will make the ensuing discussion clearer.

If we are given a latin square, for example the one in Fig. 17.5a, then we can use it to assign colours to the edges of a complete bipartite graph, as in Fig. 17.5b. The vertices of the graph are regarded as the rows and columns of the latin square and the edge $r_i c_j$ is assigned the 'colour' s_k, where s_k is the symbol in row r_i and column c_j of the square. The defining property of the latin square ensures that this assignment of colours actually is an edge colouring. So any latin square of order n defines an edge colouring of $K_{n,n}$.

In mathematics, it sometimes pays to take a twisted view, and that is true here. Instead of constructing an edge colouring of $K_{n,n}$ by the obvious method described above, we shall use the latin square in a different way. We take the vertices to be the symbols $s_1, s_2, \ldots, s_n$ and the columns $c_1, c_2, \ldots, c_n$, and assign the edge $s_i c_j$ the 'colour' r_k, where r_k is the row such that s_i appears in row r_k and column c_j of the square (Fig. 17.5c). Here again, the defining property of the latin square ensures that we have an edge colouring of $K_{n,n}$.

We shall adopt a twisted view to investigate problems about the construction of latin squares. Suppose we set out to construct an $n \times n$ latin square by filling in one row at a time. Naturally, we make sure that each row contains every symbol once only, and that no symbol occurs more than once in a column. The result of filling in m rows in this way is an $m \times n$ **latin rectangle** ($1 \leq m < n$). An example of a 3×5 latin rectangle is given in Fig. 17.6.

	c_1	c_2	c_3	c_4	c_5
r_1	s_1	s_2	s_3	s_4	s_5
r_2	s_5	s_1	s_4	s_3	s_2
r_3	s_2	s_3	s_1	s_5	s_4



	c_1	c_2	c_3	c_4	c_5
r_1					
r_2					
r_3					
r_4	s_3	s_4	s_5	s_2	s_1
r_5	s_4	s_5	s_2	s_1	s_3

Fig. 17.6
A latin rectangle and how to complete it.

Given an $m \times n$ latin rectangle, can we fill in the remaining $n - m$ rows to obtain an $n \times n$ latin square? Rather surprisingly, the answer is yes, without any extra conditions on the latin rectangle. In other words, if we construct a latin square one row at a time then, provided we observe the obvious constraints at each stage, we shall never get stuck.

Theorem 17.3.1 Any $m \times n$ latin rectangle with $1 \leq m < n$ can be completed to form an $n \times n$ latin square.

Proof The m rows of the latin rectangle enable us to colour some of the edges of $K_{n,n}$, using the twisted rule given above. Let E denote the set of edges which remain as yet uncoloured; that is

$$E = \{s_i c_j \mid \text{symbol } s_i \text{ does } \textit{not} \text{ appear in column } c_j\}.$$

The graph formed by the uncoloured edges is $G = (S \cup C, E)$, where S and C denote the sets of symbols and columns, respectively. The graph G is regular, with degree $n - m$, and so (by Theorem 17.2) it has an edge colouring using $n - m$ colours. Call these colours $r_{m+1}, r_{m+2}, \ldots, r_n$.

Now we can fill in the remaining rows of the square, by placing s_i in row r_k and column c_j whenever the edge $s_i c_j$ is coloured r_k $(m + 1 \leq k \leq n)$. □

The procedure for a 3×5 latin rectangle is illustrated in Fig. 17.6. The graph formed by 'uncoloured' edges has degree $5 - 3 = 2$, and can be edge-coloured with two colours as shown. The last two rows of the square can then be filled in.

We now investigate another problem concerning the step-by-step construction of latin squares. Suppose that we wish to construct an $n \times n$ latin square using the symbols $s_1, s_2, \ldots, s_n$, and we have filled in a rectangle of size $m \times p$, where m and p are strictly less than n. In this case, both rows and columns are incomplete, and even if we have filled in the rectangle so that no symbol occurs more than once in any row or column, it may yet be impossible to complete the latin square. For example, the 3×4 partial latin rectangle

$$\begin{array}{cccc} A & C & D & E \\ C & E & A & B \\ E & A & C & D \end{array}$$

cannot be completed to give a 5×5 latin square. To observe this we need only remark that B occurs just once in the rectangle, and it can occur only three times more in the square (once in the last column and twice in the last two rows); hence there are only four possible occurrences of B, whereas five are needed. The following theorem shows that the condition that each symbol occurs often enough in the rectangle is both necessary and sufficient for a completion.

Theorem 17.3.2 Let R be a partial $m \times p$ latin rectangle in which the symbols $\{s_1, s_2, \ldots, s_n\}$ are used, and let $n_R(s_i)$ denote the number of times s_i occurs in R $(1 \leq i \leq n)$. Then R can be completed to an $n \times n$ latin square if and only if

$$n_R(s_i) \geq m + p - n \qquad (1 \leq i \leq n).$$

Proof Suppose that R can be completed. Since there are $n - m$ rows and $n - p$ columns to be filled in, there are at most $(n - m) + (n - p)$ further occurrences of any symbol. Hence

$$n_R(s_i) + (n - m) + (n - p) \geq n,$$

which leads to the stated condition.

Conversely, suppose the condition holds. Construct the bipartite graph whose vertices are the rows $\{r_1, \ldots, r_m\}$ and the symbols $\{s_1, \ldots, s_n\}$, and whose edge set is

$$E = \{r_i s_j \mid s_j \text{ does } \textit{not} \text{ occur in row } r_j\}.$$

Since each row of R contains p different symbols, the degree of each r_i vertex is

$$\delta(r_i) = n - p \qquad (1 \leq i \leq n).$$

Since each symbol s_i occurs at least $m + p - n$ times in R, and each occurrence is in a different row, s_j does not occur in at most $m - (m + p - n)$ rows. That is,

$$\delta(s_j) \leq n - p.$$

Since the maximum degree of the bipartite graph is $n - p$, it follows from Theorem 17.2 that it has an edge colouring using $n - p$ colours, which we shall call $c_{p+1}, c_{p+2}, \ldots, c_n$.

Using this colouring we can complete the m rows of R, by placing s_j in row r_i and column c_k when the edge $r_i s_j$ is coloured c_k. We now have an $m \times n$ latin rectangle with complete rows and, by Theorem 17.3.1, this can be completed to form an $n \times n$ latin square. □

Exercises 17.3

1 Use the edge-colouring method to extend the following latin rectangle to a 5×5 latin square.

$$\begin{array}{ccccc} A & B & C & D & E \\ C & D & B & E & A \\ B & C & E & A & D \end{array}$$

2 Find all values of Q for which the rectangle R_1 can be extended to a 6×6 latin square. Show that R_2 cannot be so extended, whatever value Q has.

$$R_1: \begin{array}{cccc} A & B & C & D \\ F & E & A & B \\ C & D & F & A \\ D & A & B & Q \end{array} \qquad R_2: \begin{array}{cccc} A & B & C & D \\ F & E & A & B \\ B & D & F & A \\ D & A & B & Q \end{array}$$

3 Show that any $n \times n$ latin square can be used as the 'top left quarter' of a $2n \times 2n$ latin square.

17.4 Matchings

In Section 17.1 (*Example*) we discussed a special case of the situation where we have a set X of people and a set Y of jobs, and each person is qualified to do some of the jobs. A question with obvious practical implications is the following. How shall we assign people to jobs, so that the maximum number of people get jobs for which they are qualified?

We shall translate the question into the language of bipartite graphs. The relation of 'being qualified' enables us to set up a bipartite graph $G = (X \cup Y, E)$ in the usual way: xy is an edge if and only if x is qualified to do job y. An assignment of people to jobs for which they are qualified corresponds to a 'matching' in G, in the technical sense defined below.

Definition A **matching** in a bipartite graph $G = (X \cup Y, E)$ is a subset M of E with the property that no two edges in M have a common vertex.

In Fig. 17.7 two matchings M_1 and M_2 in the same graph are illustrated; the edges which belong to the matchings are indicated by heavy lines. Using the people-and-jobs terminology, M_1 yields an assignment of jobs for the two people x_1 and x_4 and M_2 for the three people x_1, x_3, x_4. In fact, M_2 cannot be bettered, since it is impossible for all four people to get jobs for which they are qualified. To observe this, we simply remark that the three people $\{x_1, x_2, x_3\}$ are collectively qualified only for the two jobs y_2 and y_3, and so one of these people must be disappointed however the jobs are filled.

We shall say that a matching M is a **maximum matching** for $G = (X \cup Y, E)$ if no other matching has a greater cardinality. If $|M| = |X|$ (all the people get jobs), then we say that M is a **complete matching**. In the example, M_2 is a maximum matching, but not a complete matching.

The first step in the study of matchings is to decide when a complete matching is possible. A necessary condition has already been mentioned above, where we remarked that three people are collectively qualified for only two jobs. More generally, if $G = (X \cup Y, E)$ and A is a subset of X, let

$$J(A) = \{y \in Y \mid xy \in E \text{ for some } x \in A\},$$

so that $J(A)$ is the set of jobs for which the people in A are collectively qualified. Our remark amounts to saying that, if $|J(A)| < |A|$, then someone in A is bound to be disappointed. So, if there is a complete matching, it must be true that $|J(A)| \geq |A|$ for all $A \subseteq X$. This is known as **Hall's condition**, after the mathematician Philip Hall, who studied a similar problem in 1935. (See Section 17.6.)

The fundamental theorem about complete matchings says that Hall's condition is both necessary and sufficient.

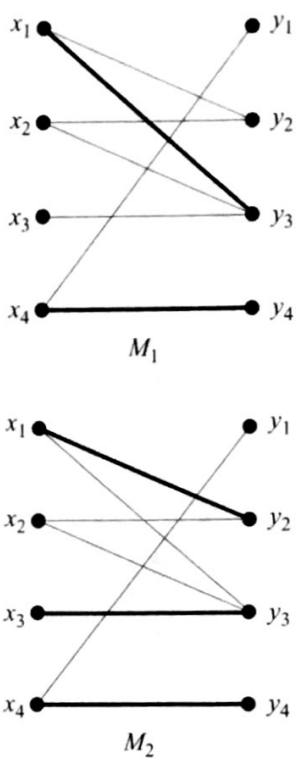


Fig. 17.7
A matching M_1 and a maximum matching M_2.

Theorem 17.4 The bipartite graph $G = (X \cup Y, E)$ has a complete matching if and only if Hall's condition is satisfied, that is

$$|J(A)| \geq |A| \quad \text{for all } A \subseteq X.$$

Proof Suppose there is a complete matching. For any $A \subseteq X$ the vertices in Y matched with those in A form a subset of $J(A)$ with size $|A|$. Hence $|J(A)| \geq |A|$.

Conversely, suppose Hall's condition holds. Given any matching M with $|M| < |X|$ we shall show how to construct a matching M' with $|M'| = |M| + 1$.

Let x_0 be any vertex not matched by M. Since $|J\{x_0\}| \geq |\{x_0\}| = 1$, there is at least one edge x_0y_1. If y_1 is unmatched then we can add x_0y_1 to M and obtain the required M'.

If y_1 is matched, with x_1 say, then

$$|J\{x_0, x_1\}| \geq |\{x_0, x_1\}| = 2,$$

and so there is another vertex y_2 apart from y_1 adjacent to x_0 or x_1. If y_2 is unmatched, stop. If y_2 is matched, to x_2 say, repeat the argument and select a new vertex y_3 adjacent to at least one of x_0, x_1, x_2. Continuing in this way, we must eventually stop at an unmatched vertex y_r, since G is finite.

Each vertex y_i $(1 \leq i \leq r)$ is adjacent to at least one of $x_0, x_1, \ldots, x_{i-1}$. Thus, on retracing our steps, we have a path

$$y_r, x_s, y_s, x_t, y_t, \ldots, y_w, x_0$$

in which the edges $x_i y_i$ are in M, and the alternate edges are not in M. We construct the new matching M' such that the edges $x_i y_i$ of the path are not in M', but the alternate edges are in M'. Since the terminal edges $y_r x_s$ and $y_w x_0$ are both in M', we have $|M'| = |M| + 1$, as required. □

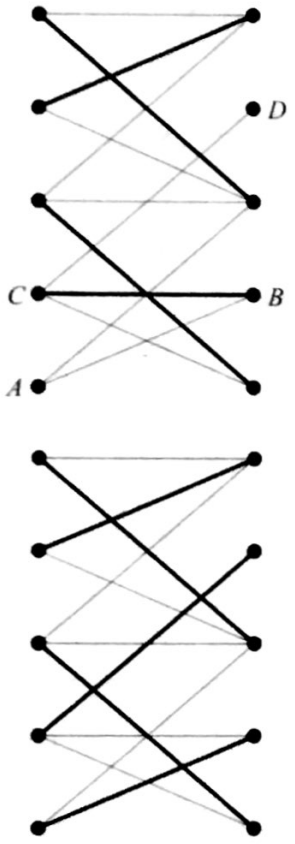


Fig. 17.8
An alternating path $ABCD$ and the result of switching.

The key idea in the proof is the construction of a path whose edges are alternately in M and not in M. In general, suppose $G = (X \cup Y, E)$ is a bipartite graph and M is a matching in G. We say that the path

$$x_0, y_1, x_1, y_2, x_2, \ldots, x_{k-1}, y_k$$

is an **alternating path** (for M) if the edges $y_i x_i$ are in M, the edges $x_{i-1} y_i$ are not in M, and x_0 and y_k do not belong to any edge of M. Note that the first and last edges are not in M, so that the path has one edge fewer in M than it has not in M. The proof of the theorem shows that if Hall's condition is satisfied, and M is an incomplete matching, then an alternating path for M exists. Switching the status of the edges on this path yields a new matching M' with one more edge.

The idea is not only the basis of the proof, it is also a practical device for the construction of matchings. In Fig. 17.8 the path $ABCD$ is an alternating path, and switching the status of the edges on this path yields the complete matching shown.

In Section 17.5 we shall explain how this approach leads to an algorithm for finding a maximum matching in any bipartite graph.

Exercises 17.4

1 Use Hall's condition to show that the graph in Fig. 17.9 has no complete matching.

2 Let M be the matching denoted by heavy lines in Fig. 17.9.

(i) Find an alternating path for M beginning at x_2.
(ii) Use it to construct a matching M' with $|M'| = 4$.
(iii) Check that there is no alternating path for M'.
(iv) Is M' a maximum matching?

3 Suppose that each member of a set of people has a list of k books which he or she wishes to borrow from a library. Suppose also that each book appears on exactly k lists. Show that it is possible for everyone to borrow one of the books on his or her list at the same time. [Hint: use the result established in the *Example*, Section 17.1.]

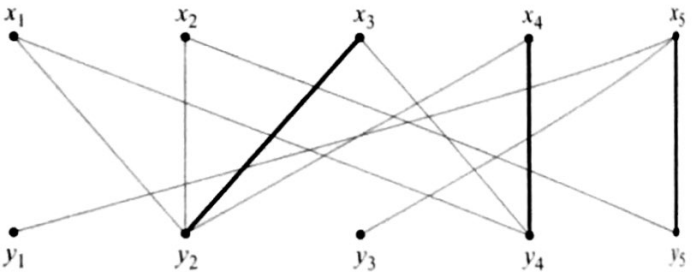


Fig. 17.9
Illustration for Ex. 17.4.2.

17.5 Maximum matchings

In general, a bipartite graph will not have a complete matching. This remark leads us back to our original question of finding the maximum size of a matching. In the people-and-jobs terminology, we are trying to find the assignment which results in the largest possible number of people getting suitable jobs. The solution to this problem is, in fact, a fairly straightforward deduction from the result on complete matchings, Theorem 17.4.

The crucial point is the remark that, if a set of people A are collectively qualified for the set of jobs $J(A)$, and $|A| > |J(A)|$, then some of the people will be disappointed. Indeed, at least $|A| - |J(A)|$ people will be disappointed.

Definition The **deficiency** d of a bipartite graph $G = (X \cup Y, E)$ is defined to be

$$d = \max_{A \subseteq X} \{|A| - |J(A)|\}.$$

We remark that the empty set $\varnothing$ is a subset of X and $|\varnothing| = |J(\varnothing)| = 0$, so that $d \geq 0$ in all cases. Theorem 17.4 asserts that G has a complete matching if and only if $d = 0$, and the next theorem deals with the size of a maximum matching in the general case.

Theorem 17.5.1 The size of a maximum matching M in a bipartite graph $G = (X \cup Y, E)$ is

$$|M| = |X| - d$$

where d is the deficiency of G.

Proof By the definition of d, there is a set $A_0 \subseteq X$ for which $|A_0| - |J(A_0)| = d$. In any matching, at least d members of A_0 remain unmatched, and so $|M| \leq |X| - d$. We have to show that there is a matching with size $|X| - d$.

Let D be a new set of size d, and let G^* be the graph $(X^* \cup Y^*, E^*)$ given by

$$X^* = X, \qquad Y^* = Y \cup D, \qquad E^* = E \cup K,$$

where K is the set of all possible edges linking X and D. The set of vertices $J^*(A)$ linked to a subset A of X in G^* is $D \cup J(A)$, and so

$$|J^*(A)| - |A| = d + |J(A)| - |A| \geq 0,$$

by the definition of d. Hence G^* satisfies Hall's condition, and it has a complete matching M^*. Removing from M^* the d edges which have a vertex in D, we obtain the required matching in G. □

Theorem 17.5.1 does not tell us how to find a maximum matching. Indeed, it is not even the basis of a good practical method for finding the *size* of a maximum matching, since in order to calculate d we must examine all the $2^{|X|}$ subsets of X.

A more practical approach is based on the fact that if we have an alternating path for a matching M then we can construct a better matching M'. In order to make this idea work, we need the following result.

Theorem 17.5.2 If the matching M in a bipartite graph G is not a maximum matching, then G contains an alternating path for M.

Proof Let M^* be a maximum matching, and let F denote the set of edges which are in one of M or M^*, but not both. (F is the 'symmetric difference' of M and M^*.) The edges in F and the vertices they contain form a graph in which every vertex has degree 1 or 2, so the components of this graph are paths and cycles. In each path or cycle the edges in M alternate with edges not in M, and so in any cycle the number of edges in M is equal to the number not in M. Since $|M^*| > |M|$, there must be at least one component which is a path with an odd number of edges, and this is an alternating path for M. □

On the basis of Theorem 17.5.2, we can outline a strategy for finding a maximum matching.

(1) Begin with any matching M (one edge alone will do).
(2) Search for an alternating path for M.
(3) If an alternating path is found, construct a better matching M' in the usual way, and return to (2) with M' replacing M.
(4) If no alternating path can be found, stop: M is a maximum matching.

The search for an alternating path can be carried out by a modified BFS procedure. We choose an unmatched vertex x_0 and construct a tree of 'partial' alternating paths starting from x_0 as follows.

(1) At level 1 insert all the vertices $y_1, y_2, \ldots, y_k$ adjacent to x_0. If any one of these vertices y_i is unmatched, stop: x_0y_i is an alternating path.
(2) If all level 1 vertices are matched, insert the vertices $x_1, x_2, \ldots, x_k$ with which they are matched at level 2.
(3) At level 3, insert all new vertices adjacent to the level 2 vertices. If any one of them is unmatched, stop: the path leading from this vertex to x_0 is an alternating path.
(4) If all level 3 vertices are matched, insert the vertices with which they are matched at level 4.

... and so on.

It remains only to remark that the construction may be halted because there are no new vertices to insert at an odd-numbered level. When this happens there is no alternating path beginning at the chosen vertex x_0. We must, however, repeat the procedure for every unmatched vertex in X before we can be sure that no alternating path whatsoever can be found in G.

Example Let $G = (X \cup Y, E)$ be the bipartite graph with $X = \{x_1, x_2, x_3, x_4, x_5\}$, $Y = \{y_1, y_2, y_3, y_4, y_5\}$, and E specified by Table 17.5.1. Let M denote the

matching $\{x_1y_3, x_2y_1, x_3y_5, x_4y_4\}$. Construct the tree of partial alternating paths rooted at x_5 and use it to find a complete matching in G.

Table 17.5.1

x_1	x_2	x_3	x_4	x_5
y_1	y_1	y_1	y_2	y_3
y_3	y_3	y_3	y_4	y_4
		y_5	y_5	

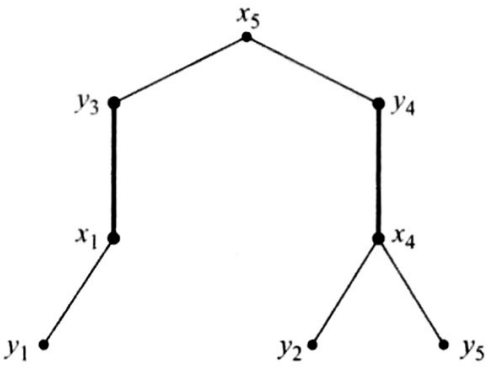


Fig. 17.10
The tree of partial alternating paths rooted at x_5.

Solution The tree is illustrated in Fig. 17.10. At level 3 we see that y_2 is unmatched, so that $x_5y_4x_4y_2$ is an alternating path. Switching the status of the edges on this path gives the complete matching $\{x_1y_3, x_2y_1, x_3y_5, x_4y_2, x_5y_4\}$. □

Exercises 17.5

1 Let $G = (X \cup Y, E)$ be the bipartite graph with $X = \{a, b, c, d, e\}$, $Y = \{v, w, x, y, z\}$, and $E = \{av, ax, bv, bz, cw, cy, cz, dy, dz, ez\}$. Use the algorithmic method to find a complete matching in G, starting from the matching $M = \{av, bz, cy\}$.

2 Let $G = (X \cup Y, E)$ be the graph depicted in Fig. 17.7. For which 3-subsets of X is it possible to find a maximum matching in G such that the three given vertices are matched?

3 Suppose $G = (X \cup Y, E)$ is a bipartite graph with $|X| = |Y| = n$. Show that if δ is the minimum degree of G then

$$|A| - |J(A)| \leq n - \delta \quad \text{for all } A \subseteq X.$$

Deduce that if $|E| > (m-1)n$ then G has a matching with at least m edges.

17.6 Transversals for families of finite sets

At the University of Folornia, the Mathematics Department is run by committees. There are only six members of the department (Professor McBrain, Dr Angst, Dr Blott, Dr Chunner, Dr Dodder, and Dr Elder), and they have organized themselves into four committees:

Teaching: {McBrain, Angst},

Administration: {McBrain, Blott},

Research: {McBrain, Angst, Blott},

Car Parking: {Chunner, Dodder, Elder}.

It has been decided that each committee must select a representative to serve on the department's new Committee for Committees. No one is allowed to represent more than one committee. Can this be done?

Given the stated membership of the committees, there are several ways to select distinct representatives: one way is to select Angst, Blott, McBrain, and Chunner to represent the respective committees in the order listed above. However, if the Car Parking committee contained only Angst and Blott, then the selection would be impossible. (Why?)

The general form of this problem is best expressed by using the notion of a family of sets, as introduced in Section 12.1. We are given a family

$$\mathscr{S} = \{S_i \mid i \in I\}$$

of sets, not necessarily distinct, and we wish to choose representatives s_i $(i \in I)$ such that

$$s_i \in S_i \qquad \text{and} \qquad i \neq j \Rightarrow s_i \neq s_j.$$

Such a set of distinct representatives is usually called a **transversal** for $\mathscr{S}$. The basic problem is to find conditions which will ensure that a given family $\mathscr{S}$ has a transversal.

In fact, this problem is merely a disguised form of the problem of finding a sufficient condition for the existence of a complete matching in a bipartite graph. To observe this, we construct the bipartite graph whose two parts correspond to the names of the sets and the members of the sets, respectively, and whose edges signify which sets contain which members. (The situation at the University of Folornia is illustrated in Fig. 17.11.)

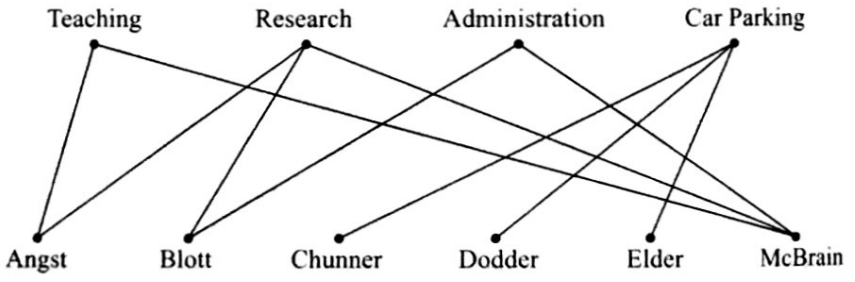


Fig. 17.11
Democracy at the University of Folornia.

Generally, we define $G = (X \cup Y, E)$ as follows:

$$X = I \qquad \text{(the names of the sets)},$$
$$Y = \bigcup_{i \in I} S_i \quad \text{(the union of the sets)},$$

and place the edge iy in E whenever y is a member of the set S_i. Then a transversal for $\mathscr{S}$ is simply a complete matching for G; if y_i represents S_i then the edge joining i to y_i is in the matching. Now, Hall's condition is easily expressed in the transversal terminology. A subset H of I is simply a subfamily of sets in $\mathscr{S}$, and $J(H)$ is the total membership of those sets, that is

$$\bigcup_{i \in H} S_i.$$

Translating Hall's condition into this language leads to the following version of Theorem 17.4.

Theorem 17.6 The finite family of finite sets

$$\mathscr{S} = \{S_i \mid i \in I\}$$

has a transversal if and only if

$$\left|\bigcup_{i\in H} S_i\right| \geq |H| \quad \text{for all } H \subseteq I. \qquad \square$$

A useful way of expressing the condition is to say that any k of the sets must have at least k members collectively ($k \geq 1$).

Exercises 17.6

1 Let $\mathscr{S}$ be the family of sets $\{a, b, l, e\}$, $\{l, e, s, t\}$, $\{s, t, a, b\}$, $\{s, a, l, e\}$, $\{t, a, l, e\}$, $\{s, a, l, t\}$. Find a transversal for $\mathscr{S}$.

2 Show that there is no transversal for the family $\mathscr{S}$ given in Ex. 1 in which the first three sets are represented by e, l, s, respectively.

3 Prove that the family of sets $\{a, m\}$, $\{a, r, e\}$, $\{m, a, r, e\}$, $\{m, a, s, t, e, r\}$, $\{m, e\}$, $\{r, a, m\}$ has no transversal, by showing explicitly that Hall's condition does not hold.

4 Let $\{X_1, X_2, \ldots, X_n\}$ be a family of sets and let X denote the union of the sets. Show that if the family has a transversal then, given any x in X, there is a transversal containing x.

17.7 Miscellaneous Exercises

1 Show that if a regular graph with degree 3 has a Hamiltonian cycle, then it has an edge colouring with three colours.

2 There are n married couples at a party. Conversations only take place between two people of the opposite sex who are not married. Represent this situation by means of a bipartite graph and show explicitly that your graph has an edge colouring using $n - 1$ colours.

3 Let $S = \{a, d, i, m, o, r, s, t\}$ and let $\mathscr{S}$ be the family of subsets $\{r, o, a, d\}$, $\{r, i, o, t\}$, $\{r, i, d, s\}$, $\{s, t, a, r\}$, $\{m, o, a, t\}$, $\{d, a, m, s\}$, $\{m, i, s, t\}$. Show that any 7-subset of S is a transversal for $\mathscr{S}$.

4 Let X denote the union of the family of sets $\{X_1, X_2, \ldots, X_n\}$, and suppose x and y are members of X. Show by an example that the family may have a transversal, but not a transversal which contains both x and y.

5 Let the elements of $\mathbb{Z}_{14}$ represent the vertices of the cycle graph C_{14}, and let G be the graph obtained from C_{14} by adding the edges $\{i, i + 5\}$ ($i = 0, 2, 4, 6, 8, 10, 12$). ($G$ is known as **Heawood's graph**.) Show that G is bipartite and construct an edge colouring of G which uses the smallest possible number of colours.

6 Suppose there are five committees: $C_1 = \{a, c, e\}$, $C_2 = \{b, c\}$, $C_3 = \{a, b, d\}$, $C_4 = \{d, e, f\}$, $C_5 = \{e, f\}$. Each committee must send a different representative to the Annual Congress of Committees, and C_1 wishes to nominate e, C_2 wishes to nominate b, C_3 wishes to nominate a, and C_4 wishes to nominate f.

(i) Show that it is not possible to respect the wishes of C_1, C_2, C_3, and C_4.

(ii) Use the alternating path method and the associated graph to find a complete system of distinct representatives.

(iii) Is it possible to find a complete system of distinct representatives if committee C_1 refuses to change its nomination?

7 A bipartite graph $G = (V \cup W, E)$ may be represented by an $m \times n$ matrix $B = (b_{ij})$ where $m = |V|$, $n = |W|$ and

$$b_{ij} = \begin{cases} 1 & \text{if } \{v_i, w_j\} \in E; \\ 0 & \text{if not.} \end{cases}$$

Describe the alternating path algorithm for finding a maximum matching in G, in terms of operations on B.

19

will deduce from the formula that d_n is approximately equal to $n!/e$. In other words, the proportion of the $n!$ permutations which are derangements is about $e^{-1} = 0.367\ldots$.

Exercises 19.1

1 Show that the formula for d_n can be written as follows:

$$d_n = 3 \times 4 \times \cdots \times (n-1) \times n$$
$$-4 \times \cdots \times (n-1) \times n + \cdots$$
$$+(-1)^{n-1} n + (-1)^n.$$

Show that the number of multiplications required to compute d_n by this formula is $O(n^2)$. What is the number of multiplications required if the recursion is used?

2 (i) Write a program to calculate values of d_n by means of the recursion.

(ii) Modify your program so that it finds the smallest value of n for which $d_n > 10^{10}$.

3 Show that the derangement numbers d_n also satisfy the recursion

$$d_1 = 0, \qquad d_n = nd_{n-1} + (-1)^n \quad (n \geq 2).$$

Is there any advantage in using this recursion (rather than the usual one) for the calculation of d_n?

19.2 Linear recursion

A simple kind of recursion is the following:

$$u_0 = c_0, \qquad u_1 = c_1, \ldots, \qquad u_{k-1} = c_{k-1},$$
$$u_{n+k} + a_1 u_{n+k-1} + a_2 u_{n+k-2} + \cdots + a_k u_n = 0 \quad (n \geq 0),$$

where $c_0, c_1, \ldots, c_{k-1}$ and $a_1, a_2, \ldots, a_k$ are constants. This is called a **linear recursion** of degree k. We shall find that there is always an explicit formula for the terms of a sequence given by a linear recursion, although it may not always be practicable (or sensible) to use it. Here we shall restrict our attention to the case $k = 2$; the general case will be discussed in Chapter 25.

Theorem 19.2 Let (u_n) be a sequence satisfying the linear recursion

$$u_0 = c_0, \quad u_1 = c_1,$$
$$u_{n+2} + a_1 u_{n+1} + a_2 u_n = 0 \quad (n \geq 0),$$

and let α and β be the roots of the **auxiliary equation**

$$t^2 + a_1 t + a_2 = 0.$$

If $\alpha \neq \beta$ then there are constants A and B such that

$$u_n = A\alpha^n + B\beta^n \quad (n \geq 0),$$

while if $\alpha = \beta$ there are constants C and D such that

$$u_n = (Cn + D)\alpha^n \quad (n \geq 0).$$

The constants A and B (or C and D) are determined by the values of c_0 and c_1.

Proof If $\alpha \neq \beta$ the equations

$$A + B = c_0, \quad A\alpha + B\beta = c_1,$$

determine A and B as follows:

$$A = \frac{c_1 - c_0\beta}{\beta - \alpha}, \quad B = \frac{c_1 - c_0\alpha}{\alpha - \beta}.$$

Consequently, when A and B are assigned these values the result holds for u_0 and u_1, and we have the basis for a proof by means of the strong principle of induction.

For the induction hypothesis, suppose that the result holds for all u_r with $0 \leq r \leq n+1$ (where $n \geq 0$). Then using the recursion equation for u_{n+2} and the induction hypothesis, we have

$$\begin{aligned} u_{n+2} &= -(a_1u_{n+1} + a_2u_n) \\ &= -[a_1(A\alpha^{n+1} + B\beta^{n+1}) + a_2(A\alpha^n + B\beta^n)] \\ &= -A\alpha^n(a_1\alpha + a_2) - B\beta^n(a_1\beta + a_2) \\ &= A\alpha^{n+2} + B\beta^{n+2}. \end{aligned}$$

At the last step we use the fact that α and β are roots of the auxiliary equation. Hence the result holds for u_{n+2}, and by the strong principle of induction it holds for all u_n with $n \geq 0$.

When $\alpha = \beta$ we apply the same method using the alternative formula. □

The theorem provides a simple method of finding a formula for the terms of a sequence (u_n) which satisfies a linear recursion with $k = 2$. We have only to solve a quadratic equation and determine the constants A and B (or C and D) so that we obtain the specified values of u_0 and u_1.

Example Find an explicit formula for u_n when

$$u_0 = 0, \quad u_1 = 1, \quad u_{n+2} - 5u_{n+1} + 6u_n = 0 \quad (n \geq 0).$$

Solution The auxiliary equation is

$$t^2 - 5t + 6 = 0.$$

Its roots are $\alpha = 2$ and $\beta = 3$, so the formula for u_n takes the form $A(2^n) + B(3^n)$. The specified values of u_0 and u_1 yield the equations

$$A + B = 0, \quad 2A + 3B = 1,$$

whence it follows that $A = -1$ and $B = 1$. So the formula for u_n is

$$u_n = 3^n - 2^n.$$

□

It should be noted that the roots α and β may be real or complex numbers rather than integers. When the initial values c_0 and c_1, and the coefficients a_1 and a_2, are integers, then it is clear from the form of the recursion itself that every u_n must be an integer. However, the *formula* for u_n may involve powers of non-integral numbers. This is the case in Ex. 19.2.2, for example, where the roots

α and β are $\frac{1}{2}(1+\sqrt{5})$ and $\frac{1}{2}(1-\sqrt{5})$. When α and β belong to the set of complex numbers $\mathbb{C}$, we can often use De Moivre's theorem to manipulate the solution into a recognizable form. For example, suppose we are given that

$$u_0 = 2, \quad u_1 = 0, \quad u_{n+2} + u_n = 0 \quad (n \geq 0).$$

The auxiliary equation is $t^2 + 1 = 0$, and its roots are i and $-$i. From the given values of u_0 and u_1 we obtain the formula

$$u_n = \mathrm{i}^n + (-\mathrm{i})^n,$$

which can be expressed in the form

$$u_n = 2\cos\tfrac{1}{2}n\pi.$$

However, this is just a rather complicated way of saying that the terms of the sequences are $2, 0, -2, 0, 2, 0, -2, 0, \ldots$, which is an obvious direct consequence of the recursion equation. So here again is an illustration of the maxim that in many circumstances the equation itself is more useful than a formula.

Exercises 19.2

1 Find an explicit formula for u_n when

(i) $u_0 = 1, \quad u_1 = 1, \quad u_{n+2} - 3u_{n+1} - 4u_n = 0 \quad (n \geq 0)$;

(ii) $u_0 = -2, \quad u_1 = 1, \quad u_{n+2} - 2u_{n+1} + u_n = 0 \quad (n \geq 0)$.

2 The *Fibonacci numbers* f_n are defined by the recursion

$$f_1 = 1, \quad f_2 = 1, \quad f_{n+1} = f_n + f_{n-1} \quad (n \geq 2).$$

Show that

$$f_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right].$$

3 Let q_n denote the number of words of length n in the alphabet $\{0, 1\}$ which have the property that no two consecutive terms are 0. Show that

$$q_1 = 2, \quad q_2 = 3, \quad q_{n+2} = q_{n+1} + q_n \quad (n \geq 0).$$

4 What is the relationship between the Fibonacci numbers and the numbers q_n defined in the previous exercise? Use this relationship to give an explicit formula for q_n.

5 Without using the formula for the Fibonacci numbers f_n, show that

(i) $f_{n+2} = f_n + f_{n-1} + \cdots + f_1 + 1$,

(ii) $f_n f_{n+2} = f_{n+1}^2 + (-1)^n$.

19.3 Recursive bisection

In the discussion of insertion sorting (Section 14.8) we mentioned the method of repeated bisection. A similar method is used in a number of other algorithms, and it leads us to study recursions of the general form

$$u_{2n} = Pu_n + Q(n).$$

where P is a constant and Q is a function of n. Of course, such a recursion does not determine u_n for all values of n, but if we are given u_2 (for example) we can calculate u_4, u_8, u_{16} and so on. The behaviour of these terms may be sufficient to indicate the behaviour of the sequence (u_n). This method is sometimes known as 'divide and conquer'.

20

20
Groups

20.1 The axioms for a group

In order to make a serious study of Discrete Mathematics it is essential to be familiar with modern algebraic techniques. The theories of permutations, designs, and latin squares (for example) are inextricably linked with certain aspects of the algebraic theories of groups, rings, and fields. In this book we shall study these algebraic structures from a utilitarian point of view, and hope that the reader will thereby be inspired to proceed to the study of algebra for its own sake.

The basic idea underlying the definition of an algebraic structure is that of a set with a 'binary operation'. Suppose we have a set X of objects, with the property that any pair of them, x and y, can be combined in some way to form an object z. This rule of combination can be expressed by the equation

$$x * y = z,$$

where the symbol $*$ indicates a **binary operation**, the word 'binary' signifying here that two objects are involved. The most familiar examples are the arithmetical operations such as $+$ and $\times$, defined on the set of integers $\mathbb{Z}$. Another important example is the rule of composition defined on the set S_n of permutations of $\{1, 2, \ldots, n\}$.

Any given binary operation has certain algebraic properties. In Chapter 4 we listed the 'laws of algebra' for $\mathbb{N}$. In Theorem 10.6 we obtained four properties of the composition operation in S_n, and remarked that the full significance of those properties would emerge later. That time is nigh, for we shall see that the four properties are shared by many other systems, and that they have far-reaching consequences. For these reasons, mathematicians use the special name *group* to describe such a system.

Definition A **group** consists of a set G, together with a binary operation $*$ defined on G which satisfies the following axioms.

G1 (*Closure*). For all x and y in G

$$x * y \text{ is in } G.$$

G2 (*Associativity*). For all x, y, and z in G

$$(x * y) * z = x * (y * z).$$

Contents

20.1 The axioms for a group 259
20.2 Examples of groups 260
20.3 Basic algebra in groups 263
20.4 The order of a group element 265
20.5 Isomorphism of groups 266
20.6 Cyclic groups 268
20.7 Subgroups 270
20.8 Cosets and Lagrange's theorem 273
20.9 Characterization of cyclic groups 277
20.10 Miscellaneous Exercises 279

G3 (*Identity*). There is an element e in G such that

$$e * x = x * e = x$$

for all x in G.

G4 (*Inverse*). For all x in G there is an x' in G such that

$$x * x' = x' * x = e.$$

The axioms are known by the names indicated. Also, an element e with the property stated in **G3** is said to be an **identity** element for the $*$ operation in G, and an element x' in **G4** is said to be an **inverse** element for x.

If G is a group and $|G|$ is finite, then $|G|$ is known as the **order** of G; a group with infinitely many elements is said to have **infinite order**.

Exercises 20.1

1 Let $G = \mathbb{Z}$. Complete the following table, where $+$, $-$, and $\times$ represent the usual operations of arithmetic.

*	Closure	Associativity	Identity	Inverse
+	√			
−	√			
×	√			

2 Repeat Ex. 1 taking $G = \mathbb{N}$, with the same operations.

20.2 Examples of groups

We have already come across two examples of groups. The set of permutations of $\{1, 2, \ldots, n\}$ is a group with respect to the composition operation. It is known as the **symmetric group** (which explains the notation S_n) and its order is $n!$. The set $\mathbb{Z}$ of integers is a group with respect to the addition operation, and it has infinite order. If these were the only examples, the study of groups would be useful for that reason alone. But in fact the concept is so widely applicable that it pervades the whole of higher mathematics. We give two more examples here, typical of the many that can be cited in support of this claim.

Our first example is geometrical. Let $\triangle$ be an equilateral triangle—it may help to think of $\triangle$ as a flat piece of card with its corners labelled ABC. There are six different transformations of $\triangle$ which have the property that $\triangle$ occupies the same position in space before and after the transformation. These transformations are known as **symmetries** of $\triangle$, and they are indicated in Fig. 20.1 by their effect on the initial position of $\triangle$ in which A is at the top and A, B, C are in clockwise order.

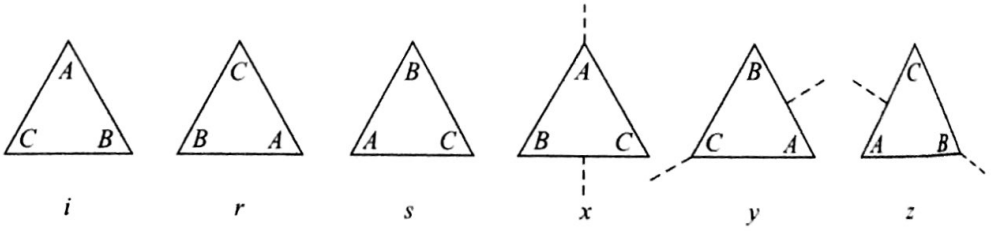


Fig. 20.1
Symmetries of an equilateral triangle.

In the figure, i is the trivial symmetry in which nothing happens, r and s are rotations through 120° and 240° about the centroid of $\triangle$, and x, y, z are reflections in the axes indicated. Alternatively, x, y, z can be regarded as the operations of turning $\triangle$ over about the relevant axis. (Note that the axes are fixed in space, and do not move when $\triangle$ is transformed.)

Example 1 Show that $G_\triangle = \{i, r, s, x, y, z\}$ is a group, with respect to the operation $*$ representing the successive implementation of symmetry transformations.

Solution It follows from the geometrical interpretation that the successive implementation of two symmetries is another symmetry. Thus, if we consider y and s, for example, the result of working out s first and then y is denoted by $y * s$ (note the order), and to find the value of $y * s$ we can trace the effect of the transformations as in Fig. 20.2. On comparing the final effect of $y * s$ with Fig. 20.1 we see that $y * s = z$.

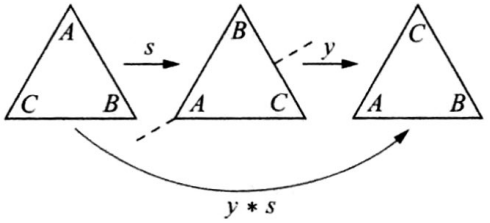


Fig. 20.2
$y * s$ has the same effect as z.

The results of all such calculations can be conveniently arranged in the **group table** for $G_\triangle$ (Table 20.2.1).

Table 20.2.1

	i	r	s	x	y	z
i	i	r	s	x	y	z
r	r	s	i	y	z	x
s	s	i	r	z	x	y
x	x	z	y	i	s	r
y	y	x	z	r	i	s
z	z	y	x	s	r	i

In this table the symbol in row y and column s is z, corresponding to the fact that $y * s = z$. The other entries in the table are obtained similarly.

Now it is easy to verify that all the group axioms are satisfied. The closure and associativity properties are immediate consequences of the nature of symmetry transformations. The identity for $G_\triangle$ is clearly the trivial symmetry i, and the inverse of each element can be found from Table 20.2.1, as follows.

Element:	i	r	s	x	y	z
Inverse:	i	s	r	x	y	z

Hence $G_\triangle$ is a group. □

Example 2 Let G_M denote the set of 2×2 matrices of the form

$$\begin{bmatrix} \alpha & \beta \\ 0 & 1 \end{bmatrix},$$

where α and β are elements of $\mathbb{Z}_3$ and $\alpha \neq 0$. Show that G_M is a group with respect to the usual rule for multiplying matrices.

Solution We shall verify the axioms **G1** – **G4** in turn.
(**G1**) The product of two matrices in G_M is

$$\begin{bmatrix} \alpha & \beta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma & \delta \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \alpha\gamma & \alpha\delta + \beta \\ 0 & 1 \end{bmatrix},$$

and since the right-hand side has the correct form it is an element of G_M. (Note that $\alpha\gamma$ cannot be zero since $\alpha \neq 0$ and $\gamma \neq 0$.)
(**G2**) The multiplication of matrices is always associative—in the present example this could be verified explicitly if necessary.
(**G3**) Taking $\alpha = 1$ and $\beta = 0$ we obtain the identity matrix, which therefore is the identity element of G_M.
(**G4**) In order to find the inverse of a typical element of G_M we note that

$$\begin{bmatrix} \alpha & \beta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma & \delta \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

if and only if

$$\alpha\gamma = 1 \quad \text{and} \quad \alpha\delta + \beta = 0.$$

When α and β are given we can solve these equations for γ and δ (remembering that $\alpha \neq 0$ so α^{-1} exists in $\mathbb{Z}_3$). Explicitly

$$\gamma = \alpha^{-1} \quad \text{and} \quad \delta = -(\alpha^{-1}\beta).$$

Thus each member of G_M has an inverse determined by these equations. □

Exercises 20.2

1 Write down explicitly the matrices belonging to the group in *Example 2*. (There are just six of them.) Find the inverse of each one.

2 There are eight symmetry transformations of a square. List them, and draw up the group table as in *Example 1*.

3 For which values of $n \geq 2$ is it true that the following are groups?

(i) $\mathbb{Z}_n$ with the + operation.

(ii) $\mathbb{Z}_n$ with the × operation.

(iii) $\mathbb{Z}_n \backslash \{0\}$ with the × operation.

4 Draw up the group table for the set of complex numbers $\{1, -1, i, -i\}$ with respect to multiplication, verifying that it is indeed a group.

20.3 Basic algebra in groups

In this section we shall being to study the consequences of the four axioms for a group. It will be necessary to avoid making any assumptions about the symbols being manipulated, except for the fact that they satisfy the group axioms. Students sometimes find this difficult because they have had many years of training in algebraic manipulation, and have learned many rules, only some of which apply here. For this reason it is helpful to remember that the present study is part of *abstract algebra*, where the symbols have no properties beyond what is stated in the axioms.

From the outset we shall drop the cumbersome $*$ notation for the group operation and use what is usually called the *multiplicative* notation, in which

$$\begin{array}{lll} x * y & \text{becomes} & xy, \\ e & \text{becomes} & 1, \\ x' & \text{becomes} & x^{-1}. \end{array}$$

This notation has the advantage of economy, but its disadvantage is that it can be confused with the ordinary multiplication of numbers. It must be stressed again that the only properties which can be assumed are the group properties, that is

$$x(yz) = (xy)z, \qquad x1 = 1x = x, \qquad xx^{-1} = x^{-1}x = 1.$$

In particular, it must *not* be assumed that $xy = yx$. If a given pair of elements do satisfy $xy = yx$ then we say that x and y **commute**; a group in which every pair commutes is said to be **commutative**. (An alternative term is **abelian**, commemorating the mathematician N. H. Abel (1802–1829).)

Theorem 20.3.1 (i) Let x, y, z, a, b, be any elements of a group G. Then

$$xy = xz \quad \Rightarrow \quad y = z \qquad \text{(left cancellation)},$$

and

$$ax = bx \quad \Rightarrow \quad a = b \qquad \text{(right cancellation)}.$$

Proof Since G is a group, x has an inverse x^{-1}. Multiplying the equation $xy = xz$ on the left by x^{-1} and using the axioms as indicated, we obtain

$$\begin{array}{ll} x^{-1}(xy) = x^{-1}(xz) & \\ (x^{-1}x)y = (x^{-1}x)z & \text{(by } \mathbf{G2}\text{)} \\ 1y = 1z & \text{(by } \mathbf{G4}\text{)} \\ y = z. & \text{(by } \mathbf{G3}\text{)}. \end{array}$$

The right cancellation rule is proved similarly. □

Let $\{g_1, g_2, \ldots, g_n\}$ be the elements of a finite group. The row of the group table corresponding to an element g_i contains the entries

$$g_ig_1, g_ig_2, \ldots, g_ig_n.$$

These are all different, since if $g_i g_r = g_i g_s$ then the left cancellation rule implies that $g_r = g_s$. Similarly, the right cancellation rule implies that the entries in every column of the group table are all different. In other words

a group table is a latin square.

The simplest example of a latin square of order m is the group table of the group $\mathbb{Z}_m$ with respect to the + operation. We now know that *any* finite group will yield a latin square, but as we might expect, not every latin square is the group table for some group (see Ex. 20.3.5).

Theorem 20.3.2 If a and b are any elements of a group G, then the equation

$$ax = b$$

has a unique solution in G.

Proof If x and $\bar{x}$ are both solutions of the equation, then

$$ax = a\bar{x} \quad (= b),$$

and so, by the left cancellation rule, $x = \bar{x}$. Hence there is at most one solution. Clearly $x = a^{-1}b$ is a solution, since

$$a(a^{-1}b) = (aa^{-1})b = 1b = b,$$

and so there is exactly one solution. □

In the case $a = b$ the theorem implies that there is a unique solution of the equation $ax = a$. Since any identity element of G satisfies this equation we conclude that

G has a unique identity element.

Similarly, in the case $b = 1$ the theorem implies that there is a unique solution of the equation $ax = 1$. Since any inverse of a satisfies this equation, we conclude that

each element of G has a unique inverse.

As a consequence of these observations we are justified in speaking of *the* identity element of G and *the* inverse of a in G.

Exercises 20.3

1 Show that the inverse of ab is $b^{-1}a^{-1}$.

2 Establish the following implications, where x and y are any elements of a group:

(i) $xy = 1 \quad \Rightarrow yx = 1$;

(ii) $(xy)^2 = x^2y^2 \Rightarrow xy = yx$.

(In part (ii) x^2 stands for xx.)

3 Suppose that G is a group with the property that $g^2 = 1$ for all g in G. Prove that G is a commutative group.

4 (i) Let $G = \{1, a, b, c\}$ be a group, where 1 is the identity and $a^2 = b^2 = c^2 = 1$. Using the latin square property, write out the complete group table for G.

(ii) Give a reason why the latin square obtained from the group table for $\mathbb{Z}_4$ (with respect to $+$) is essentially different from the one obtained in part (i).

5 Show that the following latin square of order 5 is not a group table.

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

20.4 The order of a group element

Given any element x of a group G we may define the positive and negative powers of x recursively, as follows:

$$x^1 = x, \qquad x^r = xx^{r-1} \qquad (r \geq 2),$$
$$x^{-1} = x^{-1}, \quad x^{-s} = x^{-1}x^{-(s-1)} \quad (s \geq 2).$$

If we make the convention that $x^0 = 1$ (the identity) then x^n is defined for all integers n, and the familiar rules for manipulating powers hold. That is

$$x^m x^n = x^{m+n}, \quad (x^m)^n = x^{mn} \quad (m, n \in \mathbb{Z}).$$

Since G is a group, x^n is a member of G for each n, but this does not mean that all the powers x^n represent different elements of G. Indeed, when G is a *finite* group some of the powers must be equal, since there are infinitely many of them and only finitely many elements of G. Suppose that $x^a = x^b$, where $a > b$. Multiplying both sides of the equation by x^{-b} we obtain $x^{a-b} = 1$, where $a - b > 0$. So we can be sure that there is some positive integer n for which $x^n = 1$ and, by Theorem 4.7, there is a least integer with this property.

Definition If x is an element of a finite group G then the least positive integer m such that $x^m = 1$ is called the **order** of x in G. When G is an infinite group the order of x is defined in the same way, provided that m exists; otherwise x is said to have **infinite order**.

In practice, we compute the order of a group element by calculating its positive powers until the identity is obtained. For example, in the group of the triangle $G_\triangle$, the powers of the element r are

$$r^1 = r, \qquad r^2 = s, \qquad r^3 = rs = i.$$

Since i is the identity in $G_\triangle$ it follows that the order of r is 3.

The most useful fact about the order of a group element is contained in the next theorem.

Theorem 20.4 Let x be an element of order m in a finite group G. Then

$$x^s = 1 \quad \text{in} \quad G$$

if and only if s is a multiple of m.

Proof If s is a multiple of m, say $s = mk$, then

$$x^s = x^{mk} = (x^m)^k = (1)^k = 1,$$

where we use the rules for manipulating powers and the fact that $x^m = 1$.

Conversely, suppose $x^s = 1$. By Theorem 8.2 we can write

$$s = mq + r, \qquad 0 \leq r < m.$$

Then

$$1 = x^s = x^{mq+r} = (x^m)^q x^r = x^r$$

since $x^m = 1$. Now if $r > 0$ the equation $x^r = 1$ contradicts the definition of m as the least positive integer for which $x^m = 1$. Consequently, we must have $r = 0$ and $s = mq$, so s is a multiple of m, as required. □

Exercises 20.4

1 Let α and β denote the permutations of $\mathbb{N}_7$ whose representations in cycle notation are

$$\alpha = (15)(27436), \qquad \beta = (1372)(46)(5).$$

Calculate the orders of α and β, considered as elements of the symmetric group S_7. What are the orders of $\alpha\beta$ and $\beta\alpha$?

2 Let x and y be elements of a finite group G. Show that the orders of x and yxy^{-1} are the same.

3 Let M denote the set of matrices of the form

$$\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix}$$

where the entries are members of $\mathbb{Z}_7$ and $a \neq 0$. Show that M is a group with respect to matrix multiplication, and find the orders of the elements

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}.$$

4 Let G be the group defined as in Ex. 3, except that the entries of the matrices are now real numbers rather than elements of $\mathbb{Z}_7$. Show that G contains infinitely many elements of order 2.

5 Let u and v be elements of a commutative group, and suppose that their orders are r and s, respectively. Show that if r and s are distinct primes then the order of uv is rs.

20.5 Isomorphism of groups

Recall the two examples considered in Section 20.2. In *Example 1* we discussed the group $G_\triangle$ whose elements are the six symmetries of an equilateral triangle, $G_\triangle = \{i, r, s, x, y, z\}$. In *Example 2* we discussed another group G_M which turns out to have six elements; they are the matrices which we shall denote by

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad S = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix},$$

$$X = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}, \quad Z = \begin{bmatrix} 2 & 2 \\ 0 & 1 \end{bmatrix},$$

where the symbols 0, 1, and 2 are elements of $\mathbb{Z}_3$. For the moment let us forget the definitions of $G_\triangle$ and G_M and concentrate on their group tables (Tables 20.5.1a,b).

Table 20.5.1(a)

	i	r	s	x	y	z
i	i	r	s	x	y	z
r	r	s	i	y	z	x
s	s	i	r	z	x	y
x	x	z	y	i	s	r
y	y	x	z	r	i	s
z	z	y	x	s	r	i

Table 20.5.1(b)

	I	R	S	X	Y	Z
I	I	R	S	X	Y	Z
R	R	S	I	Y	Z	X
S	S	I	R	Z	X	Y
X	X	Z	Y	I	S	R
Y	Y	X	Z	R	I	S
Z	Z	Y	X	S	R	I

Clearly, the group tables are essentially the same—the notation has been carefully chosen to emphasize this. Thus, insofar as group properties are concerned, $G_\triangle$ and G_M are identical; they differ only in the names given to their elements. More formally, we have a bijection β: $G_\triangle \to G_M$ which takes i to I, r to R, and so on, and which preserves the group operation. For example, $rx = y$ in $G_\triangle$ and $RX = Y$ in G_M, which implies that

$$\beta(rx) = \beta(y) = Y = RX = \beta(r)\beta(x).$$

Definition If G_1 and G_2 are groups (both written in the multiplicative notation), then a bijection β: $G_1 \to G_2$ is said to be an **isomorphism** if, for all g and g' in G_1

$$\beta(gg') = \beta(g)\beta(g').$$

When there is such an isomorphism, G_1 and G_2 are said to be **isomorphic**, and we write $G_1 \approx G_2$.

Exercises 20.5

1 Describe the four symmetries of a rectangle, and construct the group table. By writing down a suitable bijection show that your group is isomorphic to the one whose group table is Table 20.5.2.

Table 20.5.2

	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

2 By analysing the possible group tables show that, if isomorphic groups are regarded as the same, then

(i) there is just one group of order 2;
(ii) there is just one group of order 3;
(iii) there are just two groups of order 4.

3 Show that the isomorphism relation $\approx$ is an equivalence relation.

4 Suppose that G_1 and G_2 are finite groups and β: $G_1 \to G_2$ is an isomorphism. If $x_2 = \beta(x_1)$ for a given element x_1 in G_1, prove that x_1 and x_2 have the same order.

20.6 Cyclic groups

From an abstract point of view, two isomorphic groups are the same. So the notion of isomorphism allows us to classify groups in an obvious way. In practice, if we encounter some group G in a particular context, we usually set out to show that it is isomorphic to a 'standard example' H, whose properties are already worked out. Then the group-theoretical properties of G are precisely those of H, and we do not have to study them afresh.

In order to carry out the program outlined above we must, naturally, have a stock of standard examples. We now begin to establish this stock.

Definition A group G is said to be **cyclic** if it contains an element x such that every member of G is a power of x. The element x is said to **generate** G, and we write $G = \langle x \rangle$.

If x generates G, and all the powers of x are *distinct*, then

$$G = \{\ldots, x^{-3}, x^{-2}, x^{-1}, 1, x, x^2, x^3, \ldots\}.$$

In this case we say that G is an **infinite cyclic group**, and we use the symbol C_∞ for a typical group of this kind. Groups isomorphic to C_∞ are very common and they occur in many forms.

Example 1 Show that the set $\mathbb{Z}$ of integers, with the ordinary addition operation, is an infinite cyclic group.

Solution We have to construct a bijection β from $\mathbb{Z}$ to C_∞ such that $\beta(n_1 + n_2) = \beta(n_1)\beta(n_2)$. Note that the $+$ sign occurs on the left-hand side because addition is the specified operation in $\mathbb{Z}$. Taking x to be a generator of C_∞ define β by the rule

$$\beta(n) = x^n \quad (n \in \mathbb{Z}).$$

There is a power x^n in C_∞ for each n in $\mathbb{Z}$, and these powers are all distinct, so β is a bijection. Furthermore,

$$\beta(n_1 + n_2) = x^{n_1+n_2} = x^{n_1}x^{n_2} = \beta(n_1)\beta(n_2),$$

so that β is an isomorphism, as required. □

We now turn to the case of a cyclic group G with generator x such that the powers of x are not all distinct. In this case x is an element of finite order m and we can show that

$$G = \{1, x, x^2, \ldots, x^{m-1}\}.$$

For if k is any integer, we can write $k = mq + r$ with $0 \leq r < m$, and so

$$x^k = x^{mq+r} = (x^m)^q x^r = 1^q x^r = x^r.$$

Consequently, any power of x is equal to one of the m elements listed above. Furthermore, these elements are all different, since if $x^i = x^j$ $(0 \leq i < j \leq m-1)$

then $x^{j-i} = 1$, where $0 < j - i < m$, contradicting the definition of m. In this case G is said to be a **cyclic group of order m**, denoted by C_m. The most familiar concrete instance of a group isomorphic to C_m is the set $\mathbb{Z}_m$ of integers modulo m with respect to the + operation. As in the infinite case it is easy to show that the function β from $\mathbb{Z}_m$ to C_m defined by $\beta(n) = x^n$ is an isomorphism.

One way of extending the list of standard examples of groups is to combine known groups in some way. For example, if we are given two groups A and B (both written in the multiplicative notation), we can form their **direct product** $A \times B$ as follows. The elements of $A \times B$ are the ordered pairs

$$(a, b) \quad (a \in A, b \in B),$$

and these elements are combined under the group operation defined by

$$(a_1, b_1)(a_2, b_2) = (a_1 a_2, b_1 b_2).$$

Note that on the right-hand side $a_1 a_2$ denotes the result of combining a_1 and a_2 under the group operation in A, and $b_1 b_2$ denotes the result of combining b_1 and b_2 under the group operation in B.

It is easy to verify that $A \times B$ is indeed a group under the operation defined above (Ex. 20.6.3).

Example 2 List the elements of $C_2 \times C_3$ and $C_2 \times C_4$. Show that $C_2 \times C_3$ is isomorphic to C_6, but $C_2 \times C_4$ is not isomorphic to C_8.

Solution Suppose C_2 and C_3 are generated by x and y, respectively, so that

$$C_2 = \langle x \rangle = \{1, x\}, \qquad C_3 = \langle y \rangle = \{1, y, y^2\}.$$

According to the definition of the direct product, the elements of $C_2 \times C_3$ are

$$(1, 1), (1, y), (1, y^2), (x, 1), (x, y), (x, y^2).$$

If we let $z = (x, y)$, then calculating according to the rule given above we find that

$$z^2 = (1, y^2), \qquad z^3 = (x, 1), \qquad z^4 = (1, y),$$
$$z^5 = (x, y^2), \qquad z^6 = (1, 1).$$

Since $(1, 1)$ is the identity in $C_2 \times C_3$, we have shown that the elements of $C_2 \times C_3$ can be written as $1, z, z^2, z^3, z^4, z^5$, and consequently the group is a cyclic group of order 6.

Suppose that C_4 is generated by u, so that

$$C_4 = \langle u \rangle = \{1, u, u^2, u^3\}.$$

The elements of $C_2 \times C_4$ are

$$(1, 1), (1, u), (1, u^2), (1, u^3), (x, 1), (x, u), (x, u^2), (x, u^3),$$

and their orders are, respectively,

$$1, \quad 4, \quad 2, \quad 4, \quad 2, \quad 4, \quad 2, \quad 4.$$

Thus there is no element of order 8 in $C_2 \times C_4$ and consequently, $C_2 \times C_4$ cannot be isomorphic to C_8. □

The fact that $C_2 \times C_3 \approx C_6$ is a special case of the following theorem.

Theorem 20.6 If m and n are coprime positive integers then

$$C_m \times C_n \approx C_{mn}.$$

Proof Suppose C_m, C_n are generated by x, y, respectively, and let z denote the element (x, y) of $C_m \times C_n$. Let r be the order of z in $C_m \times C_n$. We shall show that $r = mn$.

Since $z^r = (x^r, y^r)$, and the identity of $C_m \times C_n$ is $(1, 1)$, the equation $z^r = 1$ implies that $x^r = 1$ in C_m and $y^r = 1$ in C_n. Consequently, by Theorem 20.4, r is a multiple of m and a multiple of n. Since r is the least positive integer for which $z^r = 1$, it must be the *least* common multiple of m and n. Hence, by Ex. 8.7.8,

$$r = \operatorname{lcm}(m, n) = \frac{mn}{\gcd(m, n)} = mn,$$

since m and n are coprime.

Since $C_m \times C_n$ has mn elements, and it contains an element z of order mn, it must be a cyclic group. □

Exercises 20.6

1 Let U be the subset of $\mathbb{Z}_7$ which contains all the elements of $\mathbb{Z}_7$ except 0. Show that multiplication in $\mathbb{Z}_7$ defines a group operation for U, and that $U \approx C_6$.

2 Let M denote the set of 2×2 matrices of the form

$$\begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix},$$

where the entries are integers. Show that M, with respect to matrix multiplication, is an infinite cyclic group.

3 Write out the details of the proof that the direct product $A \times B$ of two groups is a group.

4 Show that the two distinct groups of order 4 which you obtained in Ex. 20.5.2 are isomorphic to C_4 and $C_2 \times C_2$.

20.7 Subgroups

A subset H of a group G is said to be a **subgroup** of G if the members of H form a group with respect to the group operation in G.

Example 1 Let $G_\triangle = \{i, r, s, x, y, z\}$ be the group of symmetries of the triangle, as in *Example 1*, Section 20.2. Which of the following subsets of $G_\triangle$ are subgroups?

$$H_1 = \{r, s, z\}, \quad H_2 = \{i, r, s\}, \quad H_3 = \{i, r, s, x\}.$$

Solution H_1 is not a subgroup, for several reasons. It is not closed, because although r and z are H_1 the product rz is equal to x, which is not in H_1. Also, H_1 has no identity element.

On the other hand, H_2 *is* a subgroup of $G_\triangle$. On constructing the 'multiplication table' for the elements of H_2 we obtain

	i	r	s
i	i	r	s
r	r	s	i
s	s	i	r

which shows that H_2 is closed. The associativity property holds in H_2 since it holds in $G_\triangle$, and clearly i is the identity in H_2, as in $G_\triangle$. Finally $i^{-1} = i$, $r^{-1} = s$, and $s^{-1} = r$.

The subset H_3 is not a subgroup since it is not closed; for example, $rx = y$ and y is not in H_3. □

Although we can always verify the subgroup property by working out the group table, this is not often a very practical method. Indeed, it might be said that the main purpose of group theory is to study groups without recourse to group tables. For this reason we shall formulate a theorem giving sufficient conditions for a subset to be a subgroup.

Theorem 20.7 Let G be a group, and suppose that H is a non-empty subset of G satisfying the conditions

$$\textbf{S1}.\quad x, y \in H \Rightarrow xy \in H;$$
$$\textbf{S2}.\quad x \in H \Rightarrow x^{-1} \in H.$$

Then H is a subgroup of G. If G is finite, then **S1** alone is sufficient to ensure that H is a subgroup.

Proof The stated conditions assert that H is closed, and that every element of H has an inverse in H. The associativity property for elements of H follows from the corresponding property in G, since the elements of H all belong to G. Finally, to show that the identity 1 of G is in H (and consequently that it is the identity in H), we may argue as follows. If x is any member of H then x^{-1} is in H (by **S2**), where xx^{-1} is in H (by **S1**). But $xx^{-1} = 1$, so 1 is in H, as required.

It remains to show that, in the finite case, **S1** implies **S2**. If H contains only the element 1 it is clearly a subgroup. Otherwise, let x be any element of H except 1, and let m be the order of x. On multiplying the equation $x^m = 1$ by x^{-1} we obtain $x^{m-1} = x^{-1}$ and, since $m > 1$, x^{-1} is equal to a positive power of x. But on using **S1** repeatedly we see that any positive power of x is in H, and so x^{-1} is in H. Thus **S2** is a consequence of **S1**. □

Example 2 Given any group G, let $Z(G)$ denote the subset consisting of those elements which commute with every element of G, that is

$$Z(G) = \{z \in G \mid zg = gz \text{ for all } g \text{ in } G\}.$$

Show that $Z(G)$ is a subgroup of G (it is called the **centre** of G).

Solution We verify the conditions **S1** and **S2**. Suppose x and y are in $Z(G)$, so that $xg = gx$ and $yg = gy$ for all g in G. We have

$$(xy)g = x(yg) = x(gy) = (xg)y = (gx)y = g(xy),$$

so that xy is in $Z(G)$, and **S1** is verified. Also, on multiplying the equation $xg = gx$ by x^{-1} on both sides we obtain

$$x^{-1}(xg)x^{-1} = x^{-1}(gx)x^{-1},$$

and using the associative law we obtain $gx^{-1} = x^{-1}g$. Thus x^{-1} is in $Z(G)$ and **S2** is verified. □

If x is an element of order m in a group G then the elements

$$1, x, x^2, \ldots, x^{m-1}$$

of G form the **cyclic subgroup** $\langle x \rangle$ generated by x. It is very easy to verify that $\langle x \rangle$ is indeed a subgroup (Ex. 20.7.6), and clearly

the order of x is equal to the order of the subgroup $\langle x \rangle$.

For example, let U be the group of non-zero elements of $\mathbb{Z}_7$, with respect to multiplication (Ex. 20.6.1). On computing the powers of 2 in U we find

$$2^1 = 2, \quad 2^2 = 4, \quad 2^3 = 1,$$

so 2 has order 3 and the cyclic subgroup $\langle 2 \rangle$ contains the three elements 1, 2, 4. On the other hand, the cyclic subgroup $\langle 3 \rangle$ is the whole of U since

$$3^1 = 3, \quad 3^2 = 2, \quad 3^3 = 6, \quad 3^4 = 4, \quad 3^5 = 5, \quad 3^6 = 1.$$

Exercises 20.7

1 Which of the following are subgroups of G_Δ?

$$K_1 = \{i, x\}, \qquad K_2 = \{i, x, y\}, \qquad K_3 = \{i, r, s, x, y\}.$$

2 Use the group G_Δ to provide an example of the fact that if H and K are subgroups then $H \cup K$ need not be a subgroup.

3 Show that if H and K are subgroups of G then so is $H \cap K$.

4 Let g be a given element of a group G and let $C(g)$ denote the set of elements of G which commute with g, that is

$$C(g) = \{x \in G \mid xg = gx\}.$$

Show that $C(g)$ is a subgroup of G. What is the relationship between these subgroups and the centre $Z(G)$?

5 If G is the group of the square (Ex. 20.2.2), find $C(g)$ for each g in G, and then find $Z(G)$.

6 Use Theorem 20.7 to verify that if x is an element of order m in a group G then $\langle x \rangle = \{1, x, \ldots, x^{m-1}\}$ is a subgroup of G.

20.8 Cosets and Lagrange's theorem

In this section we shall prove the first really interesting theorem about groups. Anyone who understands the significance of this theorem will have no difficulty in believing that the theory of groups is a subject rich in elegant and fascinating results.

The theorem asserts that if H is a subgroup of a finite group G then $|H|$ is a divisor of $|G|$. So, for example, a group of order 20 can have subgroups only of the orders 1, 2, 4, 5, 10, 20. The idea of the proof is to partition G in such a way that each of the parts has the same size as H. If there are k such parts, then we must have $|G| = k\,|H|$, and the result follows. The parts are known as cosets.

Definition Let H be a subgroup of a (not necessarily finite) group G. The **left coset** gH of H with respect to an element g in G is defined to be the set obtained by multiplying each element of H on the left by g, that is

$$gH = \{x \in G \mid x = gh \text{ for some } h \in H\}.$$

The **right coset** of H with respect to g is defined as

$$Hg = \{x \in G \mid x = hg \text{ for some } h \in H\}.$$

If H is a finite subgroup, say $H = \{h_1, h_2, \ldots, h_m\}$, then the elements which belong to the left coset gH are

$$gh_1, gh_2, \ldots, gh_m.$$

It is clear that they are all distinct, since if $gh_i = gh_j$ the left cancellation rule implies that $h_i = h_j$. Thus we have a fundamental property of cosets:

$$|gH| = |H| \quad (g \in G).$$

For example, let $G_\triangle = \{i, r, s, x, y, z\}$ be the group of the triangle, and let H be the subgroup $\{i, x\}$. The left cosets of H in $G_\triangle$ are

$$iH = \{ii, ix\} = \{i, x\}, \qquad rH = \{ri, rx\} = \{r, y\},$$
$$sH = \{si, sx\} = \{s, z\}, \qquad xH = \{xi, xx\} = \{x, i\},$$
$$yH = \{yi, yx\} = \{y, r\}, \qquad zH = \{zi, zx\} = \{z, s\}.$$

We remark that only three distinct subsets of $G_\triangle$ occur as left cosets H, and they are disjoint. Consequently, we have the partition

$$G_\triangle = \{i, x\} \cup \{r, y\} \cup \{s, z\},$$

as displayed in Fig. 20.3, where each part is a left coset of H.

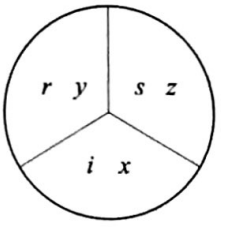


Fig. 20.3
The left cosets of $\{i, x\}$ in $G_\triangle$.

The point which often confuses beginners is that the parts have alternative names when they are regarded as cosets, so that $\{r, y\}$ (for example) may be denoted by rH or yH. Consequently, there are several different ways of *writing* $G_\triangle$ as a disjoint union of left cosets, such as

$$G_\triangle = iH \cup rH \cup sH \quad \text{or} \quad G_\triangle = xH \cup yH \cup zH,$$

although the parts themselves are the same in each case.

The next theorem establishes the general property observed in the example.

Theorem 20.8.1 Let H be a subgroup of a group G. If g_1 and g_2 are any elements of G, the left cosets g_1H and g_2H are either identical or they have no elements in common.

Proof We shall show that if g_1H and g_2H have one common element, then they are identical. Suppose that x belongs to g_1H and g_2H, that is

$$x = g_1h_1 \text{ for some } h_1 \in H, \qquad x = g_2h_2 \text{ for some } h_2 \in H.$$

In order to show that $g_1H \subseteq g_2H$, let y be any element of g_1H, so that $y = g_1h$ for some $h \in H$. Then

$$\begin{aligned} y = g_1h &= (xh_1^{-1})h = x(h_1^{-1}h) \\ &= (g_2h_2)h_1^{-1}h \\ &= g_2(h_2h_1^{-1}h). \end{aligned}$$

Since H is a subgroup, $h_2h_1^{-1}h$ is in H, and it follows that y is in g_2H. Thus $g_1H \subseteq g_2H$. An analogous argument with g_1 and g_2 reversed shows that $g_2H \subseteq g_1H$, so g_1H and g_2H are identical, as claimed. □

Another way of proving Theorem 20.8.1 uses the theory of equivalence relations. If we define a relation $\sim$ on G by saying that

$$x \sim y \quad \text{means} \quad x^{-1}y \in H$$

then $\sim$ is an equivalence relation and the equivalence classes are the left cosets (Ex. 20.8.1). If follows from the basic property of equivalence relations (Theorem 12.2) that the distinct left cosets form a partition of G. This fundamental observation leads immediately to our main theorem, which is known as **Lagrange's theorem**, after J. L. Lagrange (1736–1813).

Theorem 20.8.2 If G is a finite group of order n and H is a subgroup of order m, then m is a divisor of n.

Proof We have observed that each left coset has the same cardinality m as H, and the distinct left cosets form a partition of G. So if there are k distinct left cosets we must have $n = km$. □

The number of distinct left cosets is called the **index** of H in G, and is written as $|G:H|$; that is

$$|G : H| = |G|/|H|.$$

Of course, we could use right cosets instead of left cosets, and obtain the same results. However, it should be noted that although the numbers of left and right cosets are the same, it is not true that the left cosets and the right cosets form the same partition of G. (See Ex. 20.8.2.)

Many useful theorems about groups flow directly from Lagrange's theorem. We give two examples here.

Theorem 20.8.3 Let g be an element of a finite group G and suppose $|G| = n$. Then

(i) the order of g divides n, (ii) $g^n = 1$.

Proof (i) The order of g is the same as the order of the cyclic subgroup $\langle g \rangle$, and by Lagrange's theorem this is a divisor d of n.

(ii) If $dk = n$, then, since $x^d = 1$, we have

$$x^n = (x^d)^k = 1^k = 1.$$

□

Theorem 20.8.4 If G is a group whose order is a prime p, then G is isomorphic to the cyclic group C_p.

Proof Since $p > 1$, G has an element $x \neq 1$. The order of the cyclic subgroup $\langle x \rangle$ is greater than 1, and by Lagrange's theorem it is a divisor of p. Since p is prime, the order of $\langle x \rangle$ is p and so $\langle x \rangle$ is the whole of G. Consequently, G is a cyclic group of order p. □

In practice, Lagrange's theorem is important because it restricts the order of subgroups (and of group elements) so drastically. However, it does not provide any information about the number of subgroups. In the next section we shall show that when G is cyclic there is exactly one subgroup corresponding to each divisor d of $|G|$. The following *Example* displays some of the more complicated features which the set of subgroups can have in the general case.

Example Show that the set A_4 of all even permutations of $\{1, 2, 3, 4\}$ is a group of order 12, and find all its subgroups.

Solution In Section 12.6 we showed that the composite of two even permutations is even and so, by Theorem 20.7, A_4 is a group (actually, a subgroup of the symmetric group S_4). According to Theorem 12.6.2 the order of A_4 is $\frac{1}{2}(4!) = 12$, and the list of permutations which belong to A_4 is as follows.

The identity permutation:	id.
Three permutations of order 2:	(12)(34), (13)(24), (14)(23).
Eight permutations of order 3:	(123), (132), (124), (142), (134), (143), (234), (243).

By Lagrange's theorem, the possible orders of subgroups of A_4 are 1, 2, 3, 4, 6, and 12. We examine the possibilities in turn.

Order 1 The trivial subgroup {id} is the only possibility.

Order 2 By Theorem 20.8.4 (or the simple-minded method of Ex. 20.5.2) any subgroup of order 2 must be cyclic. Since A_4 has just three elements of order 2 there are three subgroups of order 2, such as {id, (12)(34)}.

Order 3 By similar arguments, there are four (why not eight?) subgroups of order 3, such as $\{\mathrm{id}, (123), (132)\}$.

Order 4 A subgroup of order 4 is isomorphic to C_4 or $C_2 \times C_2$ (Ex. 20.6.4). Since there are no elements of order 4, C_4 is impossible. The three elements of order 2, namely (12)(34), (13)(24), (14)(23), together with the identity, do form a subgroup isomorphic to $C_2 \times C_2$; and this is the only possibility, since the elements of order 3 cannot belong to a subgroup of order 4 (why?).

Order 6 Suppose K is a subgroup of order 6. Since there are only four elements of A_4 which do not have order 3, K must contain at least one element of order 3, say $x = (123)$, and its inverse $x^{-1} = (132)$. The elements of order 3 in K occur in pairs, such as x and x^{-1}, and so there are an even number of them. But K also contains the identity, and so in order to make the total number of elements even, K must contain at least one of the elements of order 2. Now if one such element y is in K, so are xyx^{-1} and $x^{-1}yx$ (by the subgroup property). These elements are conjugate to y and so they have the same type as y, and they are distinct. Hence the elements y, xyx^{-1}, and $x^{-1}yx$ are the three elements of order 2 (that is, (12)(34), (13)(24), and (14)(23)) is some order.

But we showed that these three elements, together with the identity, form a subgroup of order 4, which must be a subgroup of K. By Lagrange's theorem, the order of K is a multiple of 4, contradicting the assumption that $|K| = 6$. Hence there are no subgroups of order six.

Order 12 Clearly, the only possibility is A_4 itself. □

It is convenient to arrange the subgroups in a *lattice* (Fig. 20.4). In general, the lattice of subgroups can be extremely complicated, but in the next section we shall see that in the case of a cyclic group the lattice can be described in simple arithmetical terms.

Exercises 20.8

1 Let H be a subgroup of G, and define a relation $\sim$ on G by the rule that $x \sim y$ means $x^{-1}y \in H$. Show that $\sim$ is an equivalence relation and its equivalence classes are the left cosets of H.

2 Describe explicitly the partition of the triangle group by the *right* cosets of the subgroup $H = \{i, x\}$. Check that the partition is not the same as that given by the left cosets of H.

3 The symmetry group of a regular pentagon is a group of order 10. Show that it has subgroups of each of the orders allowed by Lagrange's theorem, and sketch the lattice of subgroups.

4 Suppose that a finite group G and a prime number p are given, and G has exactly m subgroups of order p. Show that the number of elements of order p in G is $m(p-1)$.

5 Use Ex. 4 to show that a non-cyclic group of order 55 has at least one subgroup of order 5 and at least one subgroup of order 11.

6 Sketch the lattice of subgroups of the symmetric group S_4. [Hint: you will need a large sheet of paper.]

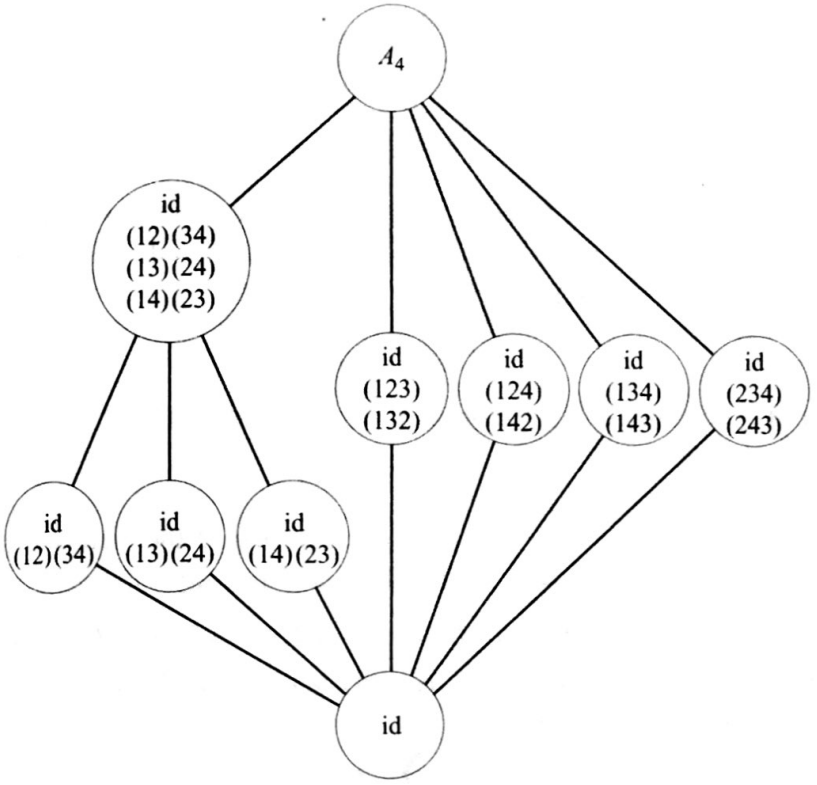


Fig. 20.4
The lattice of subgroups of A_4.

20.9 Characterization of cyclic groups

In Lagrange's theorem we begin with an algebraic hypothesis (H is a subgroup of G) and end with an arithmetical conclusion ($|H|$ divides $|G|$). Results of this kind are extremely useful because they reduce hard algebraic problems to simpler arithmetical ones. In this section we shall discuss the arithmetical properties of cyclic groups, and show that such groups can be characterized by numerical properties. We shall use some of the counting techniques developed in Chapters 10 and 11, in particular the method of Möbius inversion and its relation with Euler's function ϕ.

Theorem 20.9 If G is a finite group of order $n \geq 2$, the following statements are equivalent.

(i) G is a cyclic group.

(ii) For each divisor d of n the number of elements x in G which satisfy $x^d = 1$ is d.

(iii) For each divisor d of n the number of elements x in G which have order d is $\phi(d)$.

Proof We shall show that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ x(i), from which it follows that the statements are equivalent.

(i) $\Rightarrow$ (ii) Suppose G is a cyclic group of order n generated by g. Given any divisor d of n, let $dk = n$. The elements

$$1, g^k, g^{2k}, \ldots, g^{(d-1)k}$$

are distinct, and each of them satisfies the equation $x^d = 1$ since

$$(g^{ik})^d = (g^{kd})^i = (g^n)^i = 1^i = 1.$$

Thus we have d elements of G satisfying $x^d = 1$. We must show that there are no other solutions. Let y be any element of G such that $y^d = 1$; since G is generated by g we have $y = g^e$ for some $e \geq 0$, and consequently

$$g^{ed} = (g^e)^d = y^d = 1.$$

The order of g is n and so, by Theorem 13.4, ed is a multiple of n, say ln. Thus

$$ed = ln = l(dk),$$

so $e = lk$ and $y = g^e = g^{lk}$, which is equal to one of the elements g^{ik} already known to be solutions of $x^d = 1$. Hence these are the only solutions.

(ii) $\Rightarrow$ (iii) Suppose statement (ii) holds. By Theorem 20.4 an element x satisfies $x^d = 1$ if and only if the order of x is a divisor c of d. Hence if there are $\alpha(c)$ elements of order c we must have

$$d = \sum_{c|d} \alpha(c).$$

By the Möbius inversion formula,

$$\alpha(d) = \sum_{c|d} \mu(c)\frac{d}{c}.$$

But the relation between μ and ϕ (page 118) shows that the right-hand side is equal to $\phi(d)$. Hence $\alpha(d) = \phi(d)$ as claimed.

(iii) $\Rightarrow$ (i) If (iii) holds we know, in particular, that the number of elements of order n is $\phi(n)$. Now $\phi(n) \geq 1$, since 1 is always coprime to n, and hence G contains at least one element of order n. This element generates the whole of G (since $|G| = n$), and so G is a cyclic group. □

In Chapter 23 we shall use this numerical characterization to show that an important class of groups, arising in a wider algebraic context, are cyclic groups.

For the moment, it is instructive to use the theorem to determine all the subgroups H of a cyclic group G of order n. By Lagrange's theorem, we know that $|H| = d$, where $d \mid n$, and by Theorem 20.8.3 (ii), each of the d elements of H satisfies $x^d = 1$. But we have shown that G contains *exactly* d elements satisfying $x^d = 1$: specifically, the elements $1, g^k, \ldots, g^{(d-1)k}$, where g generates G and $dk = n$. Thus H must contain precisely those elements. In summary, we have shown that

a cyclic group of order n has just one subgroup of each order d dividing n, and these subgroups are cyclic.

For example, consider the cyclic group C_{12} generated by an element z. Each of the elements $1, z^1, \ldots, z^{11}$ of C_{12} generates a cyclic subgroup of C_{12}, and we

now know that these are the only subgroups. Furthermore, any two subgroups having the same order are identical. By simple calculations we can verify these facts explicitly, as in Table 20.9.1. Another way of illustrating the result is to say that the lattice of subgroups of C_{12} is the same as the lattice of divisors of 12 (Fig. 20.5).

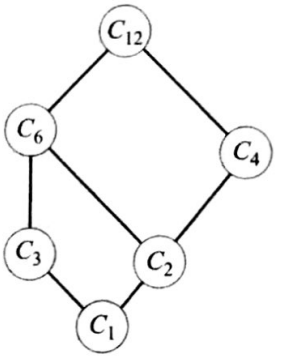


Fig. 20.5
The lattice of subgroups of C_{12}.

Table 20.9.1

Subgroup	Elements	Isomorphism class
$\langle 1\rangle$	1	C_1
$\langle z^6\rangle$	$1, z^6$	C_2
$\langle z^4\rangle$, $\langle z^8\rangle$	$1, z^4, z^8$	C_3
$\langle z^3\rangle$, $\langle z^9\rangle$	$1, z^3, z^6, z^9$	C_4
$\langle z^2\rangle$, $\langle z^{10}\rangle$	$1, z^2, z^4, z^6, z^8, z^{10}$	C_6
$\langle z\rangle$, $\langle z^5\rangle$, $\langle z^7\rangle$, $\langle z^{11}\rangle$	$1, z, z^2, \ldots, z^{11}$	C_{12}

Exercises 20.9

1 Sketch the lattice of subgroups of the cyclic group C_{24}. If z is a generator of C_{24}, identify the subgroups generated by z^7, z^8, and z^9.

2 How many elements of C_{60} generate the whole group?

3 Suppose that r and s are divisors of n, and the cyclic group C_n is generated by x. Write down generators for the cyclic subgroups C_r and C_s of C_n. Find the order of $C_r \cap C_s$, and write down a generator for this subgroup.

4 Explain how Theorem 20.8.4 can be deduced from Theorem 20.9.

20.10 Miscellaneous Exercises

1 Suppose x, y, and z are elements of a group G. Simplify the following expressions in G.

(i) $(x^{-1}z^{-1})(y^{-1}x)^{-1}y$, (ii) $(xzy)^{-1}x(xyz^{-1})^{-1}$.

2 When x and y are elements of a group G the **conjugate** of x with respect to y is yxy^{-1}, which is written x^y, and the **commutator** of x and y is $xyx^{-1}y^{-1}$, which is written $[x, y]$. Show that, for any elements a, b, c in G,

(i) $[ab, c] = [b, c]^a[a, c]$, (ii) $[a, bc] = [a, b][a, c]^b$.

3 Let $t_{a,b}$ denote the function from the set $\mathbb{R}$ of real numbers to itself defined by

$$t_{a,b}(x) = ax + b,$$

where a and b are real numbers and $a \neq 0$. Show that the set of all such functions forms a group under the operation of composition of functions.

4 Show that the group defined in the previous exercise contains infinitely many elements of order 2.

21

21
Groups of permutations

21.1 Definitions and examples

When we began our study of permutations we remarked that the rule of composition in the symmetric group S_n has four fundamental properties. In the previous chapter we used those properties as the axioms for a group. Now we shall return to the study of permutations, using the group-theoretical terminology to guide our investigations.

Let G be a set of permutations of a finite set X. If G is a group (with respect to the rule for combining permutations) then we say that G is a **group of permutations of X**.

If we take $X = \{1, 2, \ldots, n\}$ then a group of permutations of X is simply a subgroup of S_n. For example, here is a list of all the subgroups of S_3, each of which is also a group of permutations of $\{1, 2, 3\}$.

$$H_1 = \{\mathrm{id}\}, \qquad H_2 = \{\mathrm{id}, (12)\}, \qquad H_3 = \{\mathrm{id}, (13)\},$$
$$H_4 = \{\mathrm{id}, (23)\}, \qquad H_5 = \{\mathrm{id}, (123), (132)\}, \qquad H_6 = S_3.$$

In order to verify whether or not a given subset of S_n is a subgroup it is convenient to use Theorem 20.7, which tells us that (since S_n is finite) we need only verify the closure property.

Contents

21.1 Definitions and examples 281
21.2 Orbits and stabilizers 283
21.3 The size of an orbit 285
21.4 The number of orbits 288
21.5 Representation of groups by permutations 290
21.6 Applications to group theory 292
21.7 Miscellaneous Exercises 295

Exercises 21.1

1 Which of the following are groups of permutations of the set $\{1, 2, 3, 4, 5\}$, that is, which of them are subgroups of S_5?

(i) $\{(12345), (124)(35)\}$.

(ii) $\{\mathrm{id}, (12345), (13524), (14253), (15432)\}$.

(iii) $\{\mathrm{id}, (12)(34), (13)(24), (14)(23)\}$.

(iv) $\{\mathrm{id}, (12)(345), (135)(24), (15324), (12)(45), (134)(25), (143)(25)\}$.

An important subgroup of S_n is the subgroup consisting of all the even permutations, which is known as the **alternating group** A_n. The results obtained in Section 12.6 imply that the composite of two even permutations is even (so that A_n is indeed a subgroup of S_n) and that the order of A_n is $\frac{1}{2}n!$

Many examples of groups of permutations arise as the symmetry groups of geometrical objects. For example, if we label the corners of a square in clockwise order 1, 2, 3, 4, then each symmetry induces a permutation of the set $\{1, 2, 3, 4\}$, and we obtain the eight permutations listed in Table 21.1.1.

Table 21.1.1

Identity	id
Clockwise rotation through 90°	(1234)
Clockwise rotation through 180°	(13)(24)
Clockwise rotation through 270°	(1432)
Reflection in diagonal 13	(24)
Reflection in diagonal 24	(13)
Reflection in perpendicular bisector of 12	(12)(34)
Reflection in perpendicular bisector of 14	(14)(23).

It follows from the geometrical interpretation that these eight permutations form a group; more specifically, it is a subgroup of S_4.

A similar situation arises when we study graphs rather than geometric objects. In this case the 'symmetries' are the permutations of the vertices which transform edges into edges. Such a permutation is called an **automorphism** of the graph. The permutation (15)(24) is an automorphism of the graph depicted in Fig. 21.1, whereas (12345) is not an automorphism since the edge $\{2, 4\}$ is transformed into $\{3, 5\}$, which is not an edge. Clearly, the set of all automorphisms of any graph forms a group, known as the **automorphism group** of the graph.

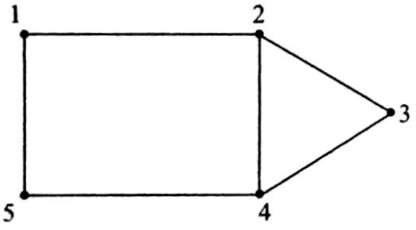


Fig. 21.1
A graph with two automorphisms.

The most important fact about automorphisms is that if v and w are vertices of a graph Γ, and there is an automorphism α such that $\alpha(v) = w$, then v and w have the same properties with respect to Γ. For instance, every edge containing v is transformed by α into an edge containing w, and so the degree of w is the same as that of v. (See Fig. 21.2.) Similarly, every cycle through v is transformed into a cycle of the same length through w. In simple cases we can use such facts to determine the automorphism group of G completely.

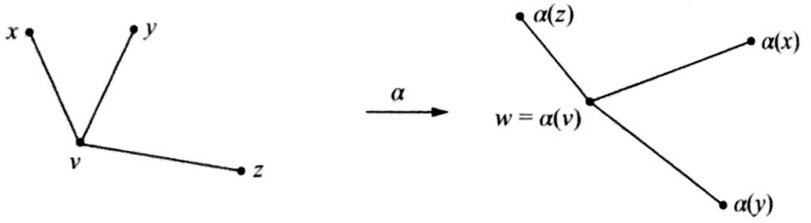


Fig. 21.2
Automorphisms preserve degree.

Example Find the automorphism group of the graph depicted in Fig. 21.3.

Solution We notice first that the vertices fall naturally into two sets: the set $\{1, 3, 5\}$ having degree four and the set $\{2, 4, 6\}$ having degree two. For the reasons given above, no automorphism can transform a member of the first set to a member of the second set. On the other hand, it is easy to see that we can take *any* permutation of $\{1, 3, 5\}$ and extend it to an automorphism of the graph.

For example, if the permutation (135) is part of an automorphism α, then α must transform 2 to 4, since 2 is the only vertex adjacent to both 1 and 3, and 4 is only vertex adjacent to both 3 and 5. Similarly, α must transform 4 to 6 and 6 to 2, and so we must have $\alpha = (135)(246)$. In the same way, each of the six permutations of $\{1, 3, 5\}$ can be extended in a unique manner to give an automorphism of the graph

id	extends to	id,	(13)	extends to	(13)(46),
(135)	extends to	(135)(246),	(15)	extends to	(15)(24),
(153)	extends to	(153)(264),	(35)	extends to	(35)(26).

It follows that there are just six automorphisms of the graph, and they are the six extended permutations listed above. □

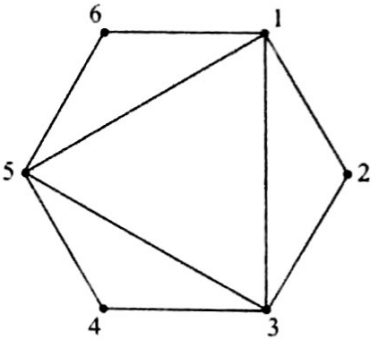


Fig. 21.3
How many automorphisms?

Exercises 21.1 (continued)

2 Find the orders of the following permutations, considered as elements of the symmetric group S_8:

(i) (1235)(48)(67); (ii) (12)(35)(48)(67);

(iii) (13672)(458).

3 According to Theorem 20.8.3 the order of any element of S_8 is a divisor of $|S_8| = 8!$ By considering the cycle structure of the permutations in S_8 write down the orders of elements which actually occur in S_8, and give an example of a divisor of 8! which is not the order of an element of S_8.

4 List all the symmetries of a regular pentagon, regarded as permutations of the corners 1, 2, 3, 4, 5, labelled in cyclic order.

5 Find the group of automorphisms of the graph given by the following adjacency list. (A picture will help.)

1	2	3	4	5	6	7	8
2	1	1	1	2	3	4	4
3	3	2	7	7	7	5	5
4	5	6	8	8	8	6	6

21.2 Orbits and stabilizers

Suppose that G is a group of permutations of a set X. We shall show that the group structure of G leads naturally to a partition of X.

Define a relation $\sim$ on X by the rule

$$x \sim y \quad \Leftrightarrow \quad g(x) = y \text{ for some } g \in G.$$

We can check that $\sim$ is an equivalence relation in the usual way.

(Reflexivity) Since id belongs to any group, and $\mathrm{id}(x) = x$ for all $x \in X$, we have $x \sim x$.

(Symmetry) Suppose $x \sim y$, so that $g(x) = y$ for some $g \in G$. Since G is a group, g^{-1} belongs to G, and since $g^{-1}(y) = x$ we have $y \sim x$, as required.

(Transitivity) If $x \sim y$ and $y \sim z$ we must have $y = g_1(x)$ and $z = g_2(y)$ for some g_1 and g_2 in G. Since G is a group, $g_2 g_1$ is in G and since $g_2 g_1(x) = z$ we have $x \sim z$, as required.

It follows that the distinct equivalence classes of $\sim$ form a partition of X; x and y are in the same part if and only if there is a permutation in G which transforms x

to y. These parts (equivalence classes) are known as **orbits** of G on X. The orbit of x contains all the members of X which are of the form $g(x)$ for some $g \in G$, and it is usually written as Gx. Explicitly,

$$Gx = \{y \in X \mid y = g(x) \text{ for some } g \in G\}.$$

Intuitively, the orbit Gx contains all the objects which are indistinguishable from x under the action of G. For example, when G is the group of automorphisms of the graph shown in Fig. 21.3 the vertex set is partitioned into two orbits, $\{1, 3, 5\}$ and $\{2, 4, 6\}$, and we have

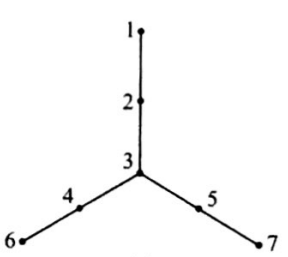
$$G1 = G3 = G5 = \{1, 3, 5\}, \qquad G2 = G4 = G6 = \{2, 4, 6\}.$$

Exercises 21.2

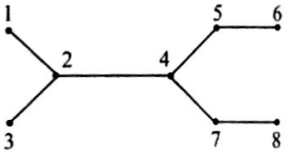
1 Write down all the automorphisms of the graph shown in Fig. 21.1. (There are only two of them!) Show that the group of automorphisms induces a partition of the vertex set into three orbits.

2 Let G be the group of automorphisms of the tree shown in Fig. 21.4a, acting on the set X of vertices. Determine the orbits of G on X.

3 Repeat Ex. 2 for the tree shown in Fig. 21.4b.



(a)



(b)

Fig. 21.4
Find the orbits.

There are two obvious numerical problems concerning orbits: we need to know how to find the size of each orbit, and how to find the number of orbits. The solution of these problems involves the combination of group-theoretical ideas on the one hand and elementary counting techniques on the other.

When G is a group of permutations of a set X we shall denote by $G(x \to y)$ the set of elements of G which take x to y; that is

$$G(x \to y) = \{g \in G \mid g(x) = y\}.$$

In particular, when $x = y$ the set $G(x \to x)$ contains all the permutations γ in G which **fix** x: that is, the permutations such that $\gamma(x) = x$. The set $G(x \to x)$ is called the **stabilizer** of x, and is written G_x. We note that if γ_1 and γ_2 are in G_x then

$$\gamma_1\gamma_2(x) = \gamma_1(x) = x,$$

and so $\gamma_1\gamma_2$ also belongs to G_x. Thus G_x is actually a *subgroup* of G.

Theorem 21.2 Let G be a group of permutations of X and suppose that h belongs to $G(x \to y)$. Then

$$G(x \to y) = hG_x,$$

the left coset of G_x with respect to h.

Proof If α is in the left coset hG_x we have $\alpha = h\beta$ for some β in G_x. Thus

$$\alpha(x) = h\beta(x) = h(x) = y,$$

and so α belongs to the set $G(x \to y)$. Conversely, if γ is in $G(x \to y)$ then

$$h^{-1}\gamma(x) = h^{-1}(y) = x$$

so that $h^{-1}\gamma = \delta$ where δ is in the stabilizer G_x. Thus $\gamma = h\delta$ is in hG_x. We have proved that every member of hG_x is also in $G(x \to y)$ and conversely, so it follows that the two sets are the same. □

Theorem 21.2 is best regarded as a rule for finding the size of the set $G(x \to y)$. We recall that any coset of a given subgroup has the same size as the subgroup itself, and so $|hG_x| = |G_x|$. Thus whenever there is an element h of G which takes x to y (in other words, whenever y is in the orbit Gx) we have the equation

$$|G(x \to y)| = |G_x| \quad (y \in Gx).$$

This holds for any y in Gx. On the other hand, if y is not in the orbit Gx then, by definition, there are no permutations taking x to y, and so

$$|G(x \to y)| = 0 \quad (y \notin Gx).$$

Exercises 21.2 (continued)

4 In the *Example* given in Section 21.1 verify explicitly that (i) $|G(2 \to 6)| = |G_2|$, (ii) $|G(3 \to 1)| = |G_3|$.

5 Let G be a group of permutations of a set X and let k be an element of $G(x \to y)$. Prove that $G(x \to y)$ is equal to the *right* coset $G_y k$, and deduce that if u and v are any two elements in the same orbit of G then $|G_u| = |G_v|$.

6 Let $X = \mathbb{Z}_5$ and suppose that G is the cyclic group of permutations of X generated by the permutation π defined by the rule $\pi(x) = 2x$. Write down the elements of G in cycle notation and determine the orbits of G on X.

21.3 The size of an orbit

In this section we shall establish a fundamental relationship between the size of an orbit Gx and the size of the stabilizer G_x. We shall need the results obtained in the previous section, together with the techniques for counting sets of pairs developed in Section 10.2.

Let G be a group of permutations of a set X, and let x be a chosen element of X. The set S of pairs (g, y) such that $g(x) = y$ can be described by means of a table as in Section 10.2.

	$\cdots \quad y \quad \cdots$	
$\vdots$		
g	$\surd$ means that (g, y) is in S	$r_g(S)$
$\vdots$		
	$c_y(S)$	

The two methods of counting S, using the row totals $r_g(S)$ and the column totals $c_y(S)$, form the basis for the proof of our main theorem.

Theorem 21.3 Let G be a group of permutations of a set X, and let x be any chosen element of X. Then we have the equation

$$|Gx| \times |G_x| = |G|.$$

Proof Let S denote the set of pairs illustrated in the table above, that is

$$S = \{(g, y) \mid g(x) = y\}.$$

Since g is a permutation there is just one y such that $g(x) = y$, for each g. In other words, each row total $r_g(S)$ is equal to 1.

The column total $c_y(S)$ is the number of g such that $g(x) = y$, that is $|G(x \to y)|$. So if y is in the orbit Gx we have

$$c_y(S) = |G(x \to y)| = |G_x|.$$

On the other hand, if y is not in Gx there are no permutations in G which take x to y, and so $c_y(S) = 0$.

The two methods for counting S give the equation

$$\sum_{y \in X} c_y(S) = \sum_{g \in G} r_g(S).$$

On the left-hand side there are $|Gx|$ terms equal to $|G_x|$ and the rest are zero, while on the right-hand side there are $|G|$ terms equal to 1. Hence we have the result. □

For example, it is easy to verify the result when G is the group of symmetries of a square, regarded as permutations of the corners (as in Section 21.1). To determine the orbit of the corner 1 (say), we note that G contains permutations which transform

1 to 1: id	1 to 2: (1234),
1 to 3: (13)(24),	1 to 4: (1432).

Thus the orbit $G1$ is the whole set and $|G1| = 4$. The stabilizer of 1 is

$$G_1 = \{\text{id}, (24)\},$$

and so

$$|G1| \times |G_1| = 4 \times 2 = 8,$$

as expected, since there are eight symmetries in all.

The result can also be used to compute the order of a group of permutations, provided that we can calculate the size of an orbit and the corresponding stabilizer.

Example Let T be a regular tetrahedron in three-dimensional space (Fig. 21.5). Find the order of the group of rotational symmetries of T.

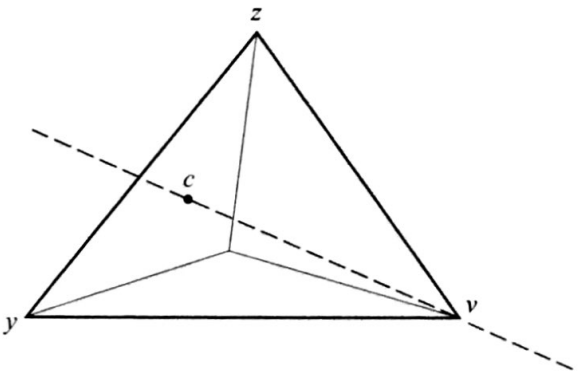


Fig. 21.5
A regular tetrahedron.

Solution Let G be the group of permutations of the corners which correspond to rotational symmetries, and let z be any corner of T. Given any other corner y there is an edge yz of T and there are two faces of T which are bounded by yz. Let c be the centroid of one of these faces and let v be the opposite corner of T (Fig. 21.5). Then a rotation of 120° about the axis cv (in the appropriate sense) takes z to y. Hence the orbit Gz contains all four vertices, and we have $|Gz| = 4$.

The only rotational symmetries which fix z are the rotations through 0°, 120°, and 240° about the altitude passing through z, and so the stabilizer G_z has order 3. Hence

$$|G| = |Gz| \times |G_z| = 4 \times 3 = 12. \qquad \square$$

Exercises 21.3

1 Label the corners of a regular tetrahedron T as 1, 2, 3, 4. Write down the permutations corresponding to the twelve rotational symmetries of T and verify that the group obtained is the alternating group A_4.

2 Let X denote the set of corners of a cube and let G denote the group of permutations of X which correspond to rotations of the cube. Show that:

(i) G has just one orbit on X;

(ii) if z is any corner, then $|G_z| = 3$;

(iii) $|G| = 24$.

3 Let V be the vertex-set of the graph Γ shown in Fig. 21.6, and let G be the automorphism group of Γ. Determine the orbits of G on V, and compute the orders of G_a, G_b, and G.

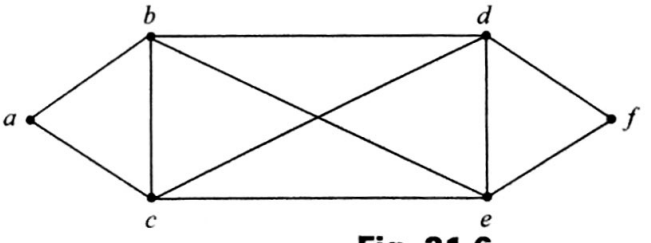


Fig. 21.6
Illustrating Ex. 21.3.3.

21.4 The number of orbits

We turn now to the problem of counting the number of orbits when we are given a group G of permutations of a set X. Each orbit is a subset of X whose members are indistinguishable under the action of G, and so the number of orbits tells us the number of distinguishably different types of object in X.

For example, suppose that it is proposed to manufacture identity cards from plastic squares, marked with a 3×3 grid on both sides, and punched with two holes, as in Fig. 21.7. Since there are 9 positions and 2 holes, the number of ways of punching the holes is $\binom{9}{2} = 36$. We shall refer to these as *configurations*. Now, not all the configurations are distinguishable, since the cards may be rotated and overturned. The first two configurations shown in Fig. 21.7 are indistinguishable, but the third one is essentially different from them.

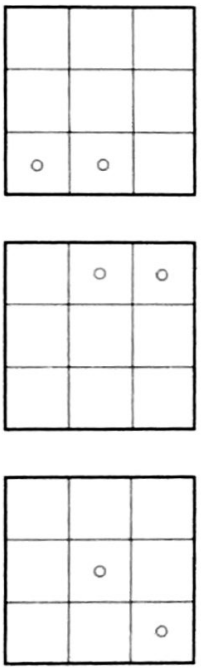


Fig. 21.7
Some identity cards.

Clearly, the group G which acts here is the familiar group of eight symmetries of a square. But we must consider its action on the set X of 36 configurations, rather than on the four corners as hitherto. Then the number of orbits of G on X is just the number of distinguishably different identity cards.

We could find the number of orbits by labelling the 36 members of X in some way, and writing down the eight permutations of these 36 configurations explicitly. However, this would be rather laborious, and fortunately there is a better way. Given any group G of permutations of a set X we define, for each g in G, a set

$$F(g) = \{x \in X \mid g(x) = x\}.$$

Thus $F(g)$ is the set of objects *fixed* by g. The next theorem says that the number of orbits is equal to the average size of the sets $F(g)$.

Theorem 21.4 The number of orbits of G on X is

$$\frac{1}{|G|} \sum_{g \in G} |F(g)|.$$

Proof Once again, we use the method of counting pairs. Let

$$E = \{(g, x) \mid g(x) = x\}.$$

Then the row total $r_g(E)$ is equal to the number of x fixed by g, that is $|F(g)|$. Also, the column total $c_x(E)$ is equal to the number of g which fix x, that is $|G_x|$. Hence the two methods for counting E lead to the equation

$$\sum_{g \in G} |F(g)| = \sum_{x \in X} |G_x|.$$

Suppose there are t orbits, and let z be a chosen element of X. According to Ex. 21.2.5 if x belongs to the orbit Gz then $|G_x| = |G_z|$. Hence on the right-hand side of the equation above there are $|Gz|$ terms equal to $|G_z|$, one for each x in Gz. The total contribution from these terms is

$$|Gz| \times |G_z| = |G|,$$

by Theorem 21.3. In other words, the total contribution from the members of each orbit is $|G|$. Since there are t orbits altogether the right-hand side is equal to $t\,|G|$, and on rearranging the equation we obtain

$$t = \frac{1}{|G|} \sum_{g \in G} |F(g)|. \qquad \square$$

Now we can solve the identity card problem. For each of the eight symmetries g we need only compute $|F(g)|$, the number of configurations fixed by g. For example, when g is the rotation through 180°, there are four fixed configurations, as depicted in Fig. 21.8. In the same way we can verify that the number of configurations fixed by each of the eight symmetries is as listed in Table 21.4.1. Hence the number of orbits is

$$\tfrac{1}{8}(36 + 0 + 4 + 0 + 6 + 6 + 6 + 6) = 8.$$

We conclude that just eight different identity cards can be produced in this way.

Table 21.4.1

Identity	36
Clockwise rotation through 90°	0
Clockwise rotation through 180°	4
Clockwise rotation through 270°	0
Reflection in diagonal 13	6
Reflection in diagonal 24	6
Reflection in perpendicular bisector of 12	6
Reflection of perpendicular bisector of 14	6

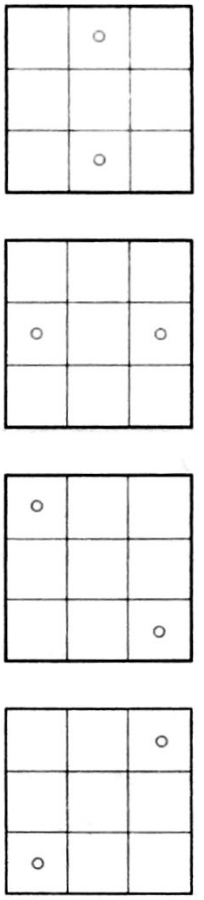


Fig. 21.8
Configurations fixed by rotation through 180°.

Of course, in this particular case it would be quite easy to list the eight cards by trial and error. However, the result of Theorem 21.4 is applicable in much greater generality, and it is useful in the solution of many other problems involving symmetry.

Example Necklaces are manufactured by arranging 13 white beads and three black beads on a loop of string. How many different necklaces can be produced in this way? (The position of the fastening may be ignored.)

Solution We can think of the 16 beads as being placed at the corners of a regular polygon with 16 sides. Each configuration is specified by the choice of the three corners which are occupied by black beads, and so there are $\binom{16}{3} = 560$ configurations in all. Two configurations give the same necklace if one can be obtained from the other by a symmetry transformation of the polygon: either a rotation or a reflection, the latter being equivalent to overturning the polygon. There are 32 symmetries in all, as listed below.

(a) The identity fixes all 560 configurations.

(b) There are 15 rotations through angles $2\pi n/16$ ($n = 1, 2, \ldots, 15$), and each of them has no fixed configurations. (Why?)

(c) There are eight reflections in axes joining the mid-points of opposite sides, and each of them has no fixed configurations.

(d) There are eight reflections in axes passing through opposite corners. The positions of the three black beads are unchanged (as a threesome) by such a reflection only if one of the beads occupies one of the two corners lying on the axis, and the other pair occupies one of the seven pairs of corners symmetrically placed with respect to this axis. Hence there are $2 \times 7 = 14$ fixed configurations for each reflection of this kind.

It follows that the number of different necklaces is

$$\tfrac{1}{32}(560 + (8 \times 14)) = 21. \qquad \square$$

Exercises 21.4

1 Show that there are just five different necklaces which can be constructed from five white beads and three black beads. Sketch them.

2 Suppose identity cards are manufactured from square cards ruled with a 4×4 grid, with two holes punched. How many different cards can be produced in this way?

3 Let V be the vertex set of the binary tree shown in Fig. 21.9, and let G be the group of automorphisms of the tree. Write down the elements of G (as permutations of V), and verify that Theorem 21.4 holds in this case.

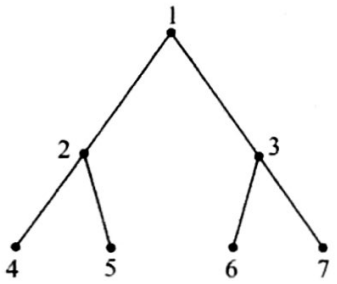


Fig. 21.9
Illustrating Ex. 21.4.3 and Ex. 21.4.4.

4 Let X denote the set of 'coloured trees' which result when each vertex of the tree in Fig. 21.9 is assigned one of the colours red or blue. How many different coloured trees of this kind are there?

5 Let G be a group of permutations of X, and let $I(x)$ be an expression which is constant on each orbit of G, so that

$$I(g(x)) = I(x) \quad \text{for all} \quad g \in G, \quad x \in X.$$

Let D be a set of representatives, one from each orbit, and let $E = \{(g, x) \mid g(x) = x\}$, as in the proof of Theorem 21.4. By evaluating the sum

$$\sum_{(g,x)\in E} I(x)$$

in two different ways, show that

$$\sum_{x\in D} I(x) = \frac{1}{|G|} \sum_{g\in G} \sum_{x\in F(g)} I(x).$$

(This is a 'weighted' version of Theorem 21.4, and reduces to it when $I(x) = 1$ for all $x \in X$. It will be used in Chapter 27.)

21.5 Representation of groups by permutations

Suppose we are given a group G, not specifically a group of permutations, and a set X. A **representation** of G by permutations of X assigns to each element g of G a permutation $\hat{g}$ of X, in such a way that the composition of group elements is compatible with the composition of the corresponding permutations. In other words,

$$\widehat{g_1 g_2} = \hat{g}_1 \hat{g}_2 \quad \text{for all } g_1 \text{ and } g_2 \text{ in } G.$$

22

22
Rings, fields, and polynomials

Contents

22.1 Rings 296
22.2 Invertible elements of a ring 297
22.3 Fields 299
22.4 Polynomials 301
22.5 The division algorithm for polynomials 304
22.6 The Euclidean algorithm for polynomials 306
22.7 Factorization of polynomials in theory 309
22.8 Factorization of polynomials in practice 310
22.9 Miscellaneous Exercises 313

22.1 Rings

Although the concept of a group is a very useful one, a group is a rather restricted algebraic object, since it has only one operation. We are accustomed to dealing with structures in which there are two basic operations, like the addition and multiplication in $\mathbb{Z}$.

The most basic object of this kind is known in mathematics as a *ring*. We shall present the axioms for a ring in a compact form, using the group concept to amalgamate several axioms into a single one.

Definition A **ring** is a set R on which there are defined two binary operations $+$ and $\times$, satisfying the following axioms.

R1. R with the operation $+$ is a commutative group.
R2. The operation $\times$ has the closure, associativity, and identity properties.
R3. (*The distributive laws.*) For all a, b, and c in R,

$$\begin{aligned} a \times (b + c) &= (a \times b) + (a \times c) \\ (a + b) \times c &= (a \times c) + (b \times c). \end{aligned}$$

Almost always we shall suppress the $\times$ sign, and write ab instead of $a \times b$.

Let us review the implications of the definition in detail. According to **R1**, the $+$ operation has the properties

$$\begin{aligned} &(a + b) + c = a + (b + c), \\ &a + 0 = 0 + a = a, \\ &a + (-a) = (-a) + a = 0, \\ &a + b = b + a, \end{aligned}$$

where the existence of the element 0 and the additive inverse $-a$ is a part of the definition. Similarly the $\times$ operation satisfies

$$\begin{aligned} &(ab)c = a(bc), \\ &a1 = 1a = a, \end{aligned}$$

where the existence of the element 1 is also guaranteed. However, it must be stressed that the existence of a multiplicative inverse a^{-1} for each element a is *not* assumed. Also, it is *not* assumed that the $\times$ operation is commutative. Finally, the axiom **R3** gives the rules for dealing with 'brackets'.

Clearly the prototype of a ring is the set $\mathbb{Z}$ of integers with the usual addition and multiplication. The integers also have two extra algebraic properties, not included in the general definition of a ring: the multiplication is commutative and we can cancel a non-zero integer from both sides of an equation (Theorem 7.5.2).

Another familiar example of a ring is the set $\mathbb{Z}_m$ of integers modulo m, with the operations defined in Section 13.2. The ring $\mathbb{Z}_m$ has a commutative multiplication, but the cancellation rule does not hold, since in $\mathbb{Z}_6$ (for example) we have $3 \times 1 = 3 \times 5$ but we cannot conclude that $1 = 5$.

As an example of a ring in which the commutative law for multiplication does not hold we can take the set of 2×2 matrices with integer entries, and the usual operations of matrix addition and multiplication. It is easy to verify that axioms **R1**, **R2**, and **R3** hold, but multiplication is not commutative. For example, if

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix},$$

then

$$AB = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}, \qquad BA = \begin{bmatrix} 2 & 1 \\ 4 & 4 \end{bmatrix}.$$

Exercises 22.1

1 If R is a ring then the set $M_2(R)$ of 2×2 matrices over R is also a ring. (You need not verify this.)

(i) What is the additive identity in $M_2(R)$?
(ii) What is the additive inverse $-A$ of a matrix A in $M_2(R)$?
(iii) What is the multiplicative identity in $M_2(R)$?
(iv) What is the cardinality of $M_2(\mathbb{Z}_m)$?
(v) Show that multiplication is not commutative in $M_2(\mathbb{Z}_2)$.
(vi) Which elements of $M_2(\mathbb{Z}_2)$ have a multiplicative inverse?

2 By using Theorem 7.5.2, prove that if x and y are integers such that $xy = 0$, then $x = 0$ or $y = 0$. Show by examples that the corresponding result does not hold when $\mathbb{Z}$ is replaced by $\mathbb{Z}_6$ or by $M_2(\mathbb{Z})$.

3 Show that if x and y are members of a ring R then $(-x)y = -xy$ and $(-x)(-y) = xy$. (At each stage of the proof, explain which property of R you are using.)

22.2 Invertible elements of a ring

An element x of a ring R is said to be **invertible** if x has a multiplicative inverse in R, that is, if there is an element u in R such that

$$ux = xu = 1.$$

A simple argument (Ex. 22.2.2) shows that if x is invertible then the element u is unique, and so we can use symbol x^{-1} for u without ambiguity. The element x^{-1} is **the inverse** of x, and the set of invertible elements of R is denoted by $U(R)$.

The only invertible elements of the ring $\mathbb{Z}$ are 1 and -1, each of which is its own inverse. In Section 13.3 we studied the invertible elements of $\mathbb{Z}_m$, and according to the results obtained there we have $U(\mathbb{Z}_8) = \{1, 3, 5, 7\}$, for example.

Theorem 22.2 The set $U(R)$ of invertible elements of a ring R is a group with respect to the multiplication in R.

Proof If x and y are invertible, let x^{-1} and y^{-1} be their inverses. Then

$$(xy)(y^{-1}x^{-1}) = (y^{-1}x^{-1})(xy) = 1$$

so that $y^{-1}x^{-1}$ is the inverse of xy. Thus $U(R)$ is closed under multiplication in R. Furthermore, multiplication is associative, and 1 is its own inverse so it certainly belongs to $U(R)$. Finally, if x is in $U(R)$ then its inverse x^{-1} is invertible (its inverse is x) and so x^{-1} is in $U(R)$. □

In Section 13.3 we proved that the element r is invertible in $\mathbb{Z}_m$ if and only if $\gcd(r, m) = 1$. It follows that $U(\mathbb{Z}_m)$ is a group of order $\phi(m)$. For example, $U(\mathbb{Z}_8) = \{1, 3, 5, 7\}$ is a group of order $\phi(8) = 4$; its group table is as given in Table 22.2.1.

Table 22.2.1

	1	3	5	7
1	1	3	5	7
3	3	1	7	5
5	5	7	1	3
7	7	5	3	1

In this case the group is the non-cyclic group $C_2 \times C_2$ of order 4. On the other hand,

$$U(\mathbb{Z}_7) = \{1, 2, 3, 4, 5, 6\}$$

is a cyclic group C_6, with generator 3, since

$$3^1 = 3, \quad 3^2 = 2, \quad 3^3 = 6, \quad 3^4 = 4, \quad 3^5 = 5, \quad 3^6 = 1.$$

In the next chapter we shall prove that $U(\mathbb{Z}_p)$ is a cyclic group of order $\phi(p) = p - 1$ whenever p is a prime.

Exercises 22.2

1 Find the orders of the groups $U(\mathbb{Z}_{10})$, $U(\mathbb{Z}_{11})$, and $U(\mathbb{Z}_{12})$, and describe their structure.

2 Show that if x is an element of a ring R and u, v are elements of R such that

$$ux = xu = 1, \qquad vx = xv = 1,$$

then $u = v$.

3 A complex number of the form $m + ni$, where m and n are integers, is known as a **Gaussian integer**. Verify that the set Γ of Gaussian integers is a ring with respect to the usual addition and multiplication of complex numbers. (You need not verify explicitly the standard properties of complex numbers.) Find the invertible elements of Γ and describe the group structure of $U(\Gamma)$.

22.3 Fields

A **field** is a ring in which the multiplication is commutative and every element except 0 has a multiplicative inverse. Thus, in a field F we have

$$U(F) = F\backslash\{0\}.$$

In order to avoid difficulties about trivial cases, we insist that a field has at least two elements.

We can reorganize the definition, and make it more explicit, by saying that a set F is a field with respect to the operations $+$ and $\times$ if

(i) F is a commutative group with respect to $+$,
(ii) $F\backslash\{0\}$ is a commutative group with respect to $\times$,
(iii) the distributive laws (**R3**) hold.

The groups in (i) and (ii) are usually referred to as the **additive group** and the **multiplicative group** of the field, respectively.

Certainly, $\mathbb{Z}$ is *not* a field since only 1 and -1 have inverses in $\mathbb{Z}$. Similarly, the ring $\mathbb{Z}_m$ is not generally a field, but it follows from Theorem 13.3.1 that

when p is a prime $\mathbb{Z}_p$ is a field.

Of course, the most familiar field is the field $\mathbb{R}$ of real numbers. But, as explained in Chapter 9, its definition and properties are far from being elementary. Fortunately, in Discrete Mathematics the more amenable finite fields such as $\mathbb{Z}_p$ are just as important as is the field $\mathbb{R}$ in Calculus.

In fact there are other finite fields besides the fields $\mathbb{Z}_p$ (p prime), and we shall study their construction and properties in the next chapter. The following *Example* contains a construction of one such field.

Example Let F be a field and let $S_2(F)$ denote the set of 2×2 matrices over F of the form

$$M = \begin{bmatrix} x & y \\ -y & x \end{bmatrix}, \qquad (x, y \in F).$$

Prove that

(i) $S_2(F)$ is a ring with respect to the usual matrix operations,

(ii) multiplication in $S_2(F)$ is commutative,

(iii) $S_2(F)$ is a field when $F = \mathbb{Z}_3$ but not when $F = \mathbb{Z}_5$.

Solution (i) The crucial fact is that $S_2(F)$ is closed under addition and multiplication. To prove this, we calculate that

$$\begin{bmatrix} x_1 & y_1 \\ -y_1 & x_1 \end{bmatrix} + \begin{bmatrix} x_2 & y_2 \\ -y_2 & x_2 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 & y_1 + y_2 \\ -(y_1 + y_2) & x_1 + x_2 \end{bmatrix},$$

$$\begin{bmatrix} x_1 & y_1 \\ -y_1 & x_1 \end{bmatrix} \times \begin{bmatrix} x_2 & y_2 \\ -y_2 & x_2 \end{bmatrix} = \begin{bmatrix} x_1x_2 - y_1y_2 & x_1y_2 + x_2y_1 \\ -(x_1y_2 + x_2y_1) & x_1x_2 - y_1y_2 \end{bmatrix},$$

and remark that the matrices on the right-hand side are indeed of the required form. We should remark also that the additive and multiplicative identities

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

are both in $S_2(F)$, and that if M is in $S_2(F)$ then so is $-M$. The remaining ring axioms are consequences of the standard properties of matrix algebra. For example, matrix addition and multiplication are always associative, and the distributive laws hold.

(ii) In general, matrix multiplication is not commutative. But when the matrices are in $S_2(F)$, we can verify that the commutative property does hold. We calculated the product of two such matrices in one order above; in the other order we have

$$\begin{bmatrix} x_2 & y_2 \\ -y_2 & x_2 \end{bmatrix} \times \begin{bmatrix} x_1 & y_1 \\ -y_1 & x_1 \end{bmatrix} = \begin{bmatrix} x_2x_1 - y_2y_1 & x_2y_1 + y_2x_1 \\ -(x_2y_1 + y_2x_1) & x_2x_1 - y_2y_1 \end{bmatrix},$$

which, by virtue of the field axioms for F, is the same as before.

(iii) Suppose that

$$\begin{bmatrix} x & y \\ -y & x \end{bmatrix} \quad \text{has inverse} \quad \begin{bmatrix} a & b \\ -b & a \end{bmatrix}.$$

Then

$$\begin{bmatrix} a & b \\ -b & a \end{bmatrix} \times \begin{bmatrix} x & y \\ -y & x \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

so that $ax - by = 1$, $ay + bx = 0$. Solving formally for a and b we obtain

$$a = x(x^2 + y^2)^{-1}, \qquad b = -y(x^2 + y^2)^{-1}.$$

Since F is a field, the element $x^2 + y^2$ will have a multiplicative inverse in F unless it is zero. If x and y are both zero, then we have the zero matrix, and we

do not require an inverse matrix. We have to ask as to whether it is possible that $x^2 + y^2 = 0$ in F when x and y are not both zero.

In $\mathbb{Z}_3$ we can verify explicitly that $x^2 + y^2 \neq 0$ when $(x, y) \neq (0, 0)$:

x	0	0	1	1	1	2	2	2
y	1	2	0	1	2	0	1	2
$x^2 + y^2$	1	1	1	2	2	1	2	2.

We conclude that every non-zero matrix in $S_2(\mathbb{Z}_3)$ has an inverse, and so $S_2(\mathbb{Z}_3)$ is a field.

On the other hand, in $\mathbb{Z}_5$

$$1^2 + 2^2 = 0,$$

and so the matrix with $x = 1$, $y = 2$ has no inverse in $S_2(\mathbb{Z}_5)$, and we do not have a field. □

Exercises 22.3

1 In the *Example* we showed that $S_2(\mathbb{Z}_3)$ is a field. It has nine elements.

(i) Denote the elements of $S_2(\mathbb{Z}_3)$ by O, I, A_1, $A_2, \ldots,$ A_7, where O, I denote the additive and multiplicative identities, and A_1, $A_2, \ldots,$ A_7 are the remaining elements in some order.

(ii) Write out the group table for the additive group.

(iii) Show that the additive group is not cyclic.

(iv) Write out the group table for the multiplicative group.

(v) Show that the multiplicative group is cyclic.

2 By finding a suitable generator, show that the multiplicative group of the field $\mathbb{Z}_{23}$ is cyclic.

3 Suppose x and y are elements of a field such that $xy = 0$. Prove that $x = 0$ or $y = 0$.

22.4 Polynomials

In elementary algebra the word *polynomial* is used to describe expressions such as

$$x^2 + 4x + 3, \quad 7x^4 + 2x^2 + 3x + 1.$$

We do not usually trouble ourselves about the meaning of the symbol x, since the context signifies what is intended. For instance, if we are asked to solve the *equation*

$$x^2 + 4x + 3 = 0,$$

then it is understood that x is to be replaced by a suitable number so that the statement becomes a valid equality between numbers.

When we compute with polynomials it becomes clear that the symbols x, x^2, $x^3, \ldots$ can be regarded simply as labels for the positions of the coefficients. In order to compute the sum of two polynomials, we add the corresponding coefficients of each x^i. In order to find the coefficient of x^i in the product of two polynomials, we find the product of the coefficient of x^j in the first one and the coefficient

of x^{i-j} in the second one, and add these products for $j = 0, 1, \ldots, i$. Such considerations lead to the conclusion that the important thing about a polynomial is its sequence of coefficients.

It is convenient to rely on the foregoing observations when we have to construct formal definitions about polynomials. Suppose that R is a ring with commutative multiplication. A finite sequence

$$(a_0, a_1, a_2, \ldots, a_n)$$

of elements of R is said to be a **polynomial** with the coefficients in R. Usually we represent this polynomial in traditional form, by writing it as

$$a(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

The set of all polynomials with coefficients in R is denoted by $R[x]$. The polynomials of the form (a_0) are known as **constant** polynomials, and they may be identified in the obvious way with the elements of the ring R.

Suppose we are given two polynomials

$$a(x) = a_0 + a_1 x + \cdots + a_n x^n, \qquad b(x) = b_0 + b_1 x + \cdots + b_m x^m.$$

There is no loss of generality in supposing that $n \geq m$, and if $n > m$ we shall set $b_{m+1} = b_{m+2} = \cdots = b_n = 0$. We define the **sum** $a(x) + b(x)$, and the **product** $a(x)b(x)$ of the polynomials in the following way:

$$\begin{aligned} a(x) + b(x) &= (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n, \\ a(x)b(x) &= a_0 b_0 + (a_0 b_1 + a_1 b_0)x \\ &\quad + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + \cdots + a_n b_m x^{n+m}. \end{aligned}$$

More formally, $s(x) = a(x) + b(x)$ is the polynomial $(s_0, s_1, \ldots, s_n)$ given by

$$s_i = a_i + b_i \qquad (0 \leq i \leq n),$$

and $p(x) = a(x)b(x)$ is the polynomial $(p_0, p_1, \ldots, p_{n+m})$ given by

$$p_i = a_0 b_i + a_1 b_{i-1} + \cdots + a_i b_0 \quad (0 \leq i \leq n + m).$$

It follows from the definitions that when the coefficients of $a(x)$ and $b(x)$ belong to a ring R the coefficients of their sum and product belong to the same ring. For example, if $a(x)$ and $b(x)$ are the members of $\mathbb{Z}_3[x]$ given by

$$a(x) = 1 + 2x + 2x^2, \qquad b(x) = 2 + x,$$

then

$$\begin{aligned} a(x)+b(x) &= (1+2)+(2+1)x+(2+0)x^2 \\ &= 2x^2; \\ a(x)b(x) &= (1\times 2)+(1\times 1+2\times 2)x+(2\times 2+2\times 1)x^2+(2\times 1)x^3 \\ &= 2+2x+2x^3. \end{aligned}$$

In general, $R[x]$ is closed under addition and multiplication. With a lot of tedious checking, one can prove that $R[x]$ is in fact a ring with commutative multiplication, provided that R is a ring with commutative multiplication. We shall accept this fact without explicit justification.

We shall also use the standard notational conventions for dealing with polynomials. Thus the coefficient 1 is usually suppressed: for example, we write $2 + x$ instead of $2 + 1x$ for the polynomial $b(x)$ considered above. When a coefficient is zero, we suppress both it and the corresponding power of x (as for example, with the coefficient of x^2 in $a(x)b(x)$ above). Finally, we often write a polynomial with its **leading coefficient** first, that is, in the form

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_n \neq 0$. When $a_n = 1$ we say that the polynomial is **monic**.

Exercises 22.4

1 Compute the sum and product of the following polynomials in $\mathbb{Z}_5[x]$:

(i) $x^3 + 3x^2 + 2x + 1$ and $x^2 + 4x + 2$;
(ii) $x^4 + x^2 + 2$ and $x^3 + 4x + 1$.

2 Show that the number of multiplications required to compute the product of the polynomials

$$a_n x^n + \cdots + a_0,$$

and

$$b_n x^n + \cdots + b_0,$$

by direct use of the definition, is $O(n^2)$. (Remarkably this is *not* the most efficient method.)

3 Suppose

$$a(x) = a_0 + \cdots + a_n x^n$$

and

$$b(x) = b_0 + \cdots + b_m x^m$$

are polynomials in $\mathbb{Z}[x]$. Write a program to compute the coefficients c_i $(0 \leq i \leq n + m)$ in the product $a(x)b(x)$.

4 Show that, in $\mathbb{Z}_7[x]$,

$$(x + 1)^7 = x^7 + 1.$$

For which values of m is it true that $(x + 1)^m = x^m + 1$ in $\mathbb{Z}_m[x]$?

5 Let

$$f(x) = f_0 + f_1 x + \cdots + f_k x^k$$

be an element of the ring $\mathbb{Z}_p[x]$, where p is a prime. Denote the product of n factors $f(x)$ by $f(x)^n$ and the result of replacing x by x^n in $f(x)$ by $f(x^n)$. Show that

(i) $f(x)^p = f(x^p)$,
(ii) $f(x)^{p^r} = f(x^{p^r})$ for all $r \geq 1$.

[Hint: for part (i) use the multinomial theorem and Ex. 12.3.6; for part (ii) use the principle of induction.]

22.5 The division algorithm for polynomials

In elementary algebra we are taught how to find the quotient and remainder when one polynomial is 'divided into' another. The working is usually displayed as in the following example.

$$\begin{array}{r} x^2 + x - 4 \\ x^2+3x+2 \overline{\big)\, x^4+4x^3+x^2+3x+4} \\ \underline{x^4+3x^3+2x^2} \\ -x^3-x^2+3x \\ \underline{x^3+3x^2+2x} \\ -4x^2+x+4 \\ \underline{-4x^2-12x-8} \\ 13x+12 \end{array}$$

Here $x^2 + 3x + 2$ is divided into $x^4 + 4x^3 + x^2 + 3x + 4$, and the quotient and remainder are $x^2 + x - 4$ and $13x + 12$, respectively. Explicitly we have

$$x^4 + 4x^3 + x^2 + 3x + 4 = (x^2 + 3x + 2)(x^2 + x - 4) + (13x + 12).$$

This is an equation in $\mathbb{Z}[x]$, the ring of polynomials with coefficients in $\mathbb{Z}$. It is fairly clear that essentially the same method will work more generally, and we shall now investigate the details of the general case.

The **degree** of a polynomial is the largest value of d for which the coefficient of x^d is non-zero. For example, the degree of $x^2 + 3x + 2$ is 2. We shall write deg $f(x)$ for the degree of the polynomial $f(x)$, noting that according to our definition deg 0 is not defined (where 0 denotes the zero polynomial). It is convenient for technical reasons to treat the zero polynomial as a special case, and we shall therefore be content to leave its degree undefined.

In general, the degree does not always have the properties which our familiarity with polynomials over $\mathbb{Z}$ or $\mathbb{R}$ leads us to expect. For instance, the degree of the sum of two polynomials may be strictly less than the degree of either polynomial. In $\mathbb{Z}_3[x]$ we have the example

$$a(x) = x^2 + x + 1, \qquad b(x) = 2x^2 + x + 1,$$
$$a(x) + b(x) = 2x + 2,$$

where the degree of $a(x) + b(x)$ is 1, whereas both $a(x)$ and $b(x)$ have degree 2.

It can also happen that the degree of the product of two polynomials is strictly less than the sum of their degrees. For if

$$\begin{aligned} f(x) &= f_n x^n + f_{n-1}x^{n-1} + \cdots + f_0, \\ g(x) &= g_m x^m + g_{m-1}x^{m-1} + \cdots + g_0 \end{aligned}$$

then the coefficient of x^{n+m} in the product $f(x)g(x)$ is $f_n g_m$, and this can be zero even if $f_n \neq 0$ and $g_m \neq 0$. For example, in $\mathbb{Z}_6$ we have $2 \times 3 = 0$, and so in $\mathbb{Z}_6[x]$ we have

$$(2x^2 + x + 4)(3x + 1) = 5x^2 + x + 4,$$

where the degree of the product is strictly less than the sum of the degrees of the factors. However, when the coefficients belong to a field F we have the important property that

$$uv = 0 \quad \Rightarrow \quad u = 0 \text{ or } v = 0$$

(Ex. 22.3.3). Clearly this implies that in $F[x]$

$$\deg f(x)g(x) = \deg f(x) + \deg g(x).$$

The next theorem is the analogue in $F[x]$ of the division theorem for integers. The fact that the degree in $F[x]$ satisfies the equation given above is a vital part of the proof.

Theorem 22.5 Let F be a field and suppose that $a(x)$ and $b(x)$ are polynomials in $F[x]$, with $b(x) \neq 0$. Then there are unique polynomials $q(x)$ and $r(x)$ in $F(x)$ such that

$$a(x) = b(x)q(x) + r(x),$$

where either $\deg r(x) < \deg b(x)$ or $r(x)$ is the zero polynomial.

Proof We shall suppose that $b(x)$ is given, and use induction on the degree of $a(x)$. If $\deg a(x) < \deg b(x)$ then the conditions are satisfied by taking $q(x) = 0$ and $r(x) = a(x)$, since

$$a(x) = b(x) \times 0 + a(x),$$

and by assumption $\deg a(x) < \deg b(x)$.

If $\deg a(x) \geq \deg b(x)$ we make the induction hypothesis that the result is true for all polynomials whose degree is strictly less than the degree of $a(x)$. Suppose

$$a(x) = a_{d+k}x^{d+k} + \cdots + a_0, \qquad b(x) = b_d x^d + \cdots + b_0,$$

where $a_{d+k} \neq 0$, $b_d \neq 0$, and $k \geq 0$. Let

$$\bar{a}(x) = a(x) - a_{d+k}b_d^{-1}x^k b(x).$$

The coefficient of x^{d+k} in $\bar{a}(x)$ is

$$a_{d+k} - (a_{d+k}b_d^{-1})b_d = 0,$$

and so $\deg \bar{a}(x) < \deg a(x)$. It follows from the induction hypothesis that there are polynomials $\bar{q}(x)$ and $r(x)$ such that

$$\bar{a}(x) = b(x)\bar{q}(x) + r(x),$$

where either $\deg r(x) < \deg b(x)$ or $r(x) = 0$. Thus, if we put

$$q(x) = \bar{q}(x) + a_{d+k}b_k^{-1}x^k,$$

then it follows that

$$a(x) = b(x)q(x) + r(x),$$

as required. Hence the induction step is verified and the result is true for all values of $\deg a(x)$.

In order to show that $q(x)$ and $r(x)$ are unique, suppose that

$$a(x) = b(x)q_1(x) + r_1(x) = b(x)q_2(x) + r_2(x),$$

where either $\deg r_1(x) < \deg b(x)$ or $r_1(x) = 0$, and either $\deg r_2(x) < \deg b(x)$ or $r_2(x) = 0$. Then

$$b(x)(q_1(x) - q_2(x)) = r_2(x) - r_1(x).$$

The left-hand side is zero if $q_1(x) = q_2(x)$; otherwise its degree is at least equal to that of $b(x)$. On the other hand, if the right-hand side is not zero then its degree is strictly less than that of $b(x)$. Hence both sides must be zero and $q_1(x) = q_2(x)$ and $r_1(x) = r_2(x)$. □

It is worth remarking that the construction of $\bar{a}(x)$ in the proof is precisely how we proceed in practical 'long division'. In the example at the beginning of the section, we begin with

$$a(x) = x^4 + 4x^3 + x^2 + 3x + 4, \qquad b(x) = x^2 + 3x + 2$$

and at the first step obtain

$$\bar{a}(x) = x^3 - x^2 + (3x + 4).$$

Of course when the calculation is set out in the usual style, we do not 'bring down' the terms $3x$ and 4 until they are required in the working.

Exercises 22.5

1 Find the quotient and remainder when $x^3 + x^2 + 1$ is divided by $x^2 + x + 1$ in $\mathbb{Z}_2[x]$.

2 Find the quotient and remainder when $x^2 + 2x + 3$ is divided into $x^5 + x^4 + 2x^3 + x^2 + 4x + 2$ in $\mathbb{Z}_5[x]$.

3 Repeat Ex. 2 when the polynomials are regarded as members of $\mathbb{Z}[x]$ and comment on the relationship between the two results.

4 Without carrying out any further long division, write down the quotient and remainder when $x^2 + 2x + 3$ is divided into $x^5 + x^4 + 2x^3 + x^2 + 4x + 2$ in $\mathbb{Z}_7[x]$; and in $\mathbb{Z}_{73}[x]$.

5 Let F be a field. Show that the polynomial $p(x)$ has an inverse in the ring $F[x]$ if and only if $p(x)$ is a non-zero constant polynomial.

22.6 The Euclidean algorithm for polynomials

Now that we have a division algorithm for $F[x]$, analogous to the familiar one for $\mathbb{Z}$, we can proceed with some definitions and theorems about divisibility and factorization of polynomials.

We say that $g(x)$ is a **divisor** (or **factor**) of $f(x)$ in $F[x]$ if there is a polynomial $h(x)$ in $F[x]$ such that $f(x) = g(x)h(x)$. Given any two polynomials $a(x)$ and $b(x)$ in $F[x]$, we say that $d(x)$ is a **greatest common divisor** (**gcd**) of $a(x)$ and $b(x)$ if

(i) $d(x)$ is a divisor of $a(x)$ and $b(x)$, and
(ii) any divisor of $a(x)$ and $b(x)$ is also a divisor of $d(x)$.

According to the definition there is not, in general, a unique gcd of two given polynomials. If $d_1(x)$ and $d_2(x)$ are two gcd's, then $d_1(x) = \alpha d_2(x)$ for some constant polynomial α (Ex. 22.6.4). So there will be just one gcd which is monic (that is, which has its leading coefficient equal to 1), and we can, if we wish, specify this as *the* gcd. But it is often convenient not to make this restriction.

In order to calculate a gcd of $a(x)$ and $b(x)$ in $F[x]$ we use the *Euclidean algorithm*. The method follows the familiar pattern

$$\begin{aligned} a(x) &= b(x)q_0(x) + r_0(x) \\ b(x) &= r_0(x)q_1(x) + r_1(x) \\ r_0(x) &= r_1(x)q_2(x) + r_2(x) \\ &\cdots \\ r_{n-2}(x) &= r_{n-1}(x)q_n(x) + r_n(x) \\ r_{n-1}(x) &= r_n(x)q_{n+1}(x). \end{aligned}$$

The arguments that we used in Chapter 8 can be adapted, with minor changes, to show that $r_n(x)$ is a gcd of $a(x)$ and $b(x)$.

By rearranging the equations we can express $r_n(x)$ in the form

$$\lambda(x)a(x) + \mu(x)b(x),$$

where $\lambda(x)$ and $\mu(x)$ are polynomials in $F[x]$. Furthermore, if $d(x)$ is *any* gcd of $a(x)$ and $b(x)$ then $d(x)$ is a constant multiple of $r_n(x)$ (by Ex. 22.6.4 again). So we have the following important result.

Theorem 22.6 Let F be a field and suppose $d(x)$ is a gcd of the polynomials $a(x)$ and $b(x)$ in $F[x]$. Then there are polynomials $\lambda(x)$ and $\mu(x)$ in $F[x]$ such that

$$d(x) = \lambda(x)a(x) + \mu(x)b(x). \qquad \square$$

Example Find a gcd of

$$a(x) = x^3 + 2x^2 + x + 1$$

and

$$b(x) = x^2 + 5$$

in $\mathbb{Z}_7[x]$, and express it in the form $\lambda(x)a(x) + \mu(x)b(x)$ with $\lambda(x)$ and $\mu(x)$ in $\mathbb{Z}_7[x]$.

Solution (Remember that the coefficients are in $\mathbb{Z}_7$.) The first step is to divide x^2+5 into x^3+2x^2+x+1

$$\begin{array}{r} x+2 \\ x^2+5 \overline{\big|\, x^3+2x^2+\ x+1} \\ \underline{x^3\qquad +5x} \\ 2x^2+3x+1 \\ \underline{2x^2\qquad +3} \\ 3x+5. \end{array}$$

That is

$$x^3+2x^2+x+1=(x^2+5)(x+2)+(3x+5).$$

The next step is to divide $3x+5$ into x^2+5

$$\begin{array}{r} 5x+\ 1 \\ 3x+5 \overline{\big|\, x^2\qquad +5} \\ \underline{x^2+4x} \\ 3x+5 \\ \underline{3x+5} \\ 0. \end{array}$$

Thus the remainder is zero and we have

$$x^2+5=(3x+5)(5x+1).$$

So $3x+5$ is a gcd. On rearranging the first equation we obtain

$$\begin{aligned} 3x+5 &= (x^3+2x^2+x+1)-(x+2)(x^2+5) \\ &= (x^3+2x^2+x+1)+(6x+5)(x^2+5), \end{aligned}$$

which is of the required form with $\lambda(x)=1$ and $\mu(x)=6x+5$. □

Exercises 22.6

1 Find the monic gcd of x^3+x^2+x+1 and x^2+2 in $\mathbb{Z}_3[x]$ and express the result in the form

$$\lambda(x)(x^3+x^2+x+1)+\mu(x)(x^2+2),$$

where $\lambda(x)$ and $\mu(x)$ are polynomials in $\mathbb{Z}_3[x]$.

2 Find the monic gcd of $x^4+2x^3+x^2+4x+2$ and x^2+3x+1 in $\mathbb{Z}_5[x]$.

3 In $\mathbb{Z}_2[x]$ find the monic gcd of

(i) x^4+1 and x^2+1,
(ii) x^5+1 and x^2+1,
(iii) x^9+1 and x^6+1.

Can you find a general rule for the monic gcd of x^n+1 and x^m+1 in $\mathbb{Z}_2[x]$?

4 Suppose $d_1(x)$ and $d_2(x)$ are gcd's of $a(x)$ and $b(x)$ in $F[x]$, where F is a field. Prove that

(i) $d_1(x)$ is a divisor of $d_2(x)$, and $d_2(x)$ is divisor of $d_1(x)$;
(ii) if $d_1(x)=\alpha(x)d_2(x)$ and $d_2(x)=\beta(x)d_1(x)$ then $\deg\alpha(x)=\deg\beta(x)=0$, so $\alpha(x)$ and $\beta(x)$ are constant polynomials.

22.7 Factorization of polynomials in theory

In Chapter 8 we proved that every integer $n \geq 2$ has a unique factorization into primes. In this section we investigate the analogous result for $F[x]$.

First, we note that the existence of non-zero constant polynomials allows trivial factorizations of any polynomial. This is because a non-zero constant α has an inverse β in F, which is also its inverse in $F[x]$, so that

$$f(x) = \alpha \times (\beta f(x))$$

is a factorization of $f(x)$ in $F[x]$. Clearly, we wish to exclude such trivial factorizations from the theory. For this reason, we define a polynomial $f(x)$ in $F[x]$ to be **irreducible** if it is not a constant polynomial and if, whenever

$$f(x) = g(x)h(x) \quad \text{in } F[x],$$

then either $g(x)$ or $h(x)$ is a constant polynomial. The irreducible polynomials play the same rôle in $F[x]$ as do the primes in $\mathbb{Z}$.

The proof of the following theorem follows the same lines as the proof of Theorem 1.8.2, and it will be presented as a sequence of Exercises.

Theorem 22.7 Any non-constant polynomial $f(x)$ in $F[x]$ can be expressed as a product of irreducible polynomials. If there are two such factorizations

$$f(x) = g_1(x)g_2(x)\cdots g_r(x) = h_1(x)h_2(x)\cdots h_s(x)$$

then $r = s$ and we can rearrange the order of factors so that $g_i(x)$ is a constant multiple of $h_i(x)$ $(1 \leq i \leq r)$; that is, $g_i(x) = \alpha_i h_i(x)$ for some non-zero constant polynomial α_i. □

Exercises 22.7

(All polynomials in Exercises 1–5 belong to $F[x]$, where F is a field.)

1 Show that if $r(x)$ is not a constant polynomial then

$$\deg r(x)s(x) > \deg s(x).$$

2 (Existence of factorization.) Show that, if $f(x)$ is a non-constant polynomial, then either $f(x)$ is irreducible or $f(x) = g(x)h(x)$ where $g(x)$ and $h(x)$ are non-constant polynomials whose degrees are less than that of $f(x)$. Hence prove the first assertion of Theorem 22.7 by induction on the degree of $f(x)$.

3 Show that if 1 is a gcd of $r(x)$ and $s(x)$, and $r(x)$ is a divisor of $s(x)t(x)$, then $r(x)$ is a divisor of $t(x)$.

4 Show that if $r(x)$ is irreducible and $r(x)$ is a divisor of $s_1(x)s_2(x)\ldots s_k(x)$, then $r(x)$ is a divisor of $s_i(x)$ for some i in the range $1 \leq i \leq k$.

5 (Uniqueness of factorization.) Prove the uniqueness assertion in Theorem 22.7.

6 Verify that in $\mathbb{Z}_{15}[x]$

$$(x + 1)(x + 14) = (x + 4)(x + 11).$$

What is the significance of this result in relation to the theory developed above?

22.8 Factorization of polynomials in practice

The existence of a very satisfactory theorem about factorization does not mean that it is easy to find the factors in any given case. The general problem is a difficult one, but there is a simple test for finding factors of a particular kind, which we shall now describe.

A polynomial $a_1x + a_0$ with $a_1 \neq 0$ has degree 1 and is said to be a **linear** polynomial. Multiplying by the constant a_1^{-1} transforms the polynomial into the form $x - \alpha$, where $\alpha = -a_1^{-1}a_0$, and we shall take this as the standard form because there is a useful test which tells us when $x - \alpha$ is a factor of $f(x)$ in $F[x]$.

Suppose that

$$f(x) = f_n x^n + f_{n-1}x^{n-1} + \cdots + f_0,$$

and for each α in F let

$$f(\alpha) = f_n\alpha^n + f_{n-1}\alpha^{n-1} + \cdots + f_0.$$

Since F is a field, the expression $f(\alpha)$ is an element of F, and we say that it is obtained by **evaluating** $f(x)$ at α. The rule which takes α to $f(\alpha)$ is a function from F to F; it is the *polynomial function* corresponding to the polynomial $f(x)$. Although we do not need to stress the point here, there are good reasons for distinguishing between the function and the polynomial (which is simply a sequence of coefficients). One such reason is that different polynomials may be associated with the same function (Ex. 22.9.8 and Ex. 22.9.16).

Theorem 22.8.1 Let F be a field and suppose $f(x)$ is a polynomial in $F[x]$. Then $x - \alpha$ is a divisor of $f(x)$ in $F[x]$ if and only if $f(\alpha) = 0$ in F.

Proof Suppose $x - \alpha$ is a divisor of $f(x)$, so that

$$f(x) = (x - \alpha)g(x) \quad \text{in } F[x].$$

On evaluating both sides at α we obtain

$$f(\alpha) = (\alpha - \alpha)g(\alpha) = 0g(\alpha) = 0.$$

Conversely, suppose $f(\alpha) = 0$ in F. By Theorem 22.5 there are polynomials $q(x)$ and $r(x)$ in $F[x]$ such that

$$f(x) = (x - \alpha)q(x) + r(x),$$

where either $\deg r(x) < \deg (x - \alpha)$ or $r(x) = 0$. On evaluating both sides of the equation at α, and keeping in mind that $f(\alpha) = 0$ we obtain

$$0 = f(\alpha) = (\alpha - \alpha)q(\alpha) + r(\alpha) = r(\alpha).$$

Now if $\deg r(x) < \deg (x - \alpha)$ then the degree of $r(x)$ must be zero and $r(x)$ is a non-zero constant polynomial; in this case $r(\alpha)$ cannot be zero. Hence we must have $r(x) = 0$, from which it follows that $x - \alpha$ is a divisor of $f(x)$. □

Theorem 22.8.1 is known as the **factor theorem**. Before illustrating its use in practice we shall record for future reference an important theoretical consequence. If $f(x)$ is a polynomial in $F[x]$ and α is an element of F then we say that α is a root of the equation $f(x) = 0$ whenever $f(\alpha) = 0$.

Theorem 22.8.2 If F is a field and $f(x)$ is a polynomial of degree $n \geq 1$ in $F[x]$ then the equation $f(x) = 0$ has at most n roots in F.

Proof Suppose the equation has m distinct roots $\alpha_1, \alpha_2, \ldots, \alpha_m$ in F. By the factor theorem $x - \alpha_1, x - \alpha_2, \ldots, x - \alpha_m$ are divisors of $f(x)$, and furthermore they are all irreducible. Hence the factorization of $f(x)$ in $F[x]$ takes the form

$$f(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_m)g(x)$$

for some $g(x)$ in $F[x]$. Since the coefficients belong to a field F the degree of a product is the sum of the degrees of the factors, and so it follows that the degree of $f(x)$ is at least m. Equivalently, this means that the number of roots of $f(x) = 0$ is at most n. □

Now we return to the practical problem of how to find the irreducible factors of a given polynomial. The most interesting and useful cases from our point of view occur when the polynomials belong to $\mathbb{Z}_p[x]$, where $\mathbb{Z}_p$ is the field of integers modulo a prime p. In such cases we can find the linear factors in a finite number of steps, since the factor theorem tells us that we need only evaluate the polynomial at each of the p elements of $\mathbb{Z}_p$. However, there is no reason why the irreducible factors of a given polynomial should be linear. If the polynomial has a small degree then we may be able to make some progress by simple-minded methods, and we shall now discuss such methods. There is no loss of generality in assuming that the given polynomial is monic, since we can transform a given polynomial into a monic polynomial of the same degree by multiplying by a constant.

By definition, every linear polynomial is irreducible in $\mathbb{Z}_p[x]$, and so there are p monic irreducible linear polynomials $x + a_0$. If the *quadratic* polynomial $x^2 + a_1x + a_0$ is reducible in $\mathbb{Z}_p[x]$ then it has two linear factors which can be found by the factor theorem. Since there are p possible linear factors, there are $\frac{1}{2}p(p-1)$ monic reducible quadratics of the form $(x - \alpha)(x - \beta)$ with $\alpha \neq \beta$, and p of the form $(x - \alpha)^2$. There are p^2 monic quadratics in all, so the number of irreducible ones is

$$p^2 - (\tfrac{1}{2}p(p-1) + p) = \tfrac{1}{2}p(p-1).$$

In particular, we note that there is always at least one monic irreducible quadratic polynomial in $\mathbb{Z}_p[x]$. When $p = 2$ we have

$$x^2 = (x+0)^2, \qquad x^2 + 1 = (x+1)^2, \qquad x^2 + x = (x+0)(x+1),$$

but $x^2 + x + 1$ is irreducible.

If a *cubic* polynomial $x^3 + a_2x^2 + a_1x + a_0$ is reducible then it has either three linear factors or one linear and one quadratic factor. Since there is a linear factor in either case, we can use the factor theorem here again to test for reducibility.

However, when the degree of the polynomial is four or more, other methods may be required.

Example Find the irreducible factors of $x^4 + 1$ in $\mathbb{Z}_3[x]$.

Solution First we use the factor theorem to look for linear factors. Let $f(x) = x^4 + 1$; then

$$f(0) = 0^4 + 1 = 1, \qquad f(1) = 1^4 + 1 = 2, \qquad f(2) = 2^4 + 1 = 2.$$

So there are no linear factors. The only possibility remaining is a factorization into two quadratics given by

$$x^4 + 1 = (x^2 + Ax + B)(x^2 + Cx + D),$$

with A, B, C, D in $\mathbb{Z}_3$. By equating coefficients of corresponding powers of x we obtain the equations

$$\begin{aligned} &\text{(i)} && A + C && = 0, \\ &\text{(ii)} && B + D + AC && = 0, \\ &\text{(iii)} && AD + BC && = 0, \\ &\text{(iv)} && BD && = 1. \end{aligned}$$

It follows from (i) that $A = -C$ and from (iv) that $B = D$. (Why?) Taking $B = D = 1$ we obtain $AC = 1$ in (ii), which contradicts (i). Taking $B = D = 2$ we obtain $AC = 2$; then $A = 1$, $C = 2$ is a solution. Hence the required factorization is

$$x^4 + 1 = (x^2 + x + 2)(x^2 + 2x + 2). \qquad \square$$

Of course, the method given in the example is very inefficient in most cases. A great deal of work has been done on the development of efficient algorithms for finding the irreducible factors of polynomials, but many of the best methods are beyond the scope of this book.

Exercises 22.8

1 Find the irreducible factors of

(i) $x^2 + 1$ in $\mathbb{Z}_5[x]$;
(ii) $x^3 + 5x^2 + 5$ in $\mathbb{Z}_{11}[x]$;
(iii) $x^4 + 3x^3 + x + 1$ in $\mathbb{Z}_5[x]$.

2 Find all the monic irreducible quadratics in $\mathbb{Z}_3[x]$.

3 Show that the number of monic irreducible cubics in $\mathbb{Z}_p[x]$ is $\frac{1}{3}p(p^2 - 1)$, and list them when $p = 2$.

4 Let $f(x) = f_n x^n + f_{n-1}x^{n-1} + \cdots + f_0$ be a polynomial of degree n in $F[x]$, and let $\alpha \in F$. Explain how $f(\alpha)$ can be evaluated by means of the recursion

$$f_0(\alpha) = f_n, \qquad f_i(\alpha) = \alpha f_{i-1}(\alpha) + f_{n-i} \quad (1 \le i \le n).$$

What is the number of multiplications required by this method? (It is sometimes known as *Horner's method.*)

5 Suppose that $f(x)$ and α are as in Ex. 4, and $f(\alpha)$ is evaluated by computing each term $f_i\alpha^i$ individually $(0 \le i \le n)$. Find the approximate number of multiplications required if the methods for calculating α^i given in Section 14.7 are used.

23

23
Finite fields and some applications

Contents

23.1 A field with nine elements 314
23.2 The order of a finite field 315
23.3 Construction of finite fields 317
23.4 The primitive element theorem 318
23.5 Finite fields and latin squares 322
23.6 Finite geometry and designs 324
23.7 Projective planes 327
23.8 Squares in finite fields 330
23.9 Existence of finite fields 333
23.10 Miscellaneous Exercises. 336

23.1 A field with nine elements

In Section 13.5 we showed that when p is a prime it is possible to use the arithmetical properties of $\mathbb{Z}_p$ to construct a set of $p-1$ mutually orthogonal latin squares. With hindsight it is clear that the key to the construction is the fact that $\mathbb{Z}_p$ is a field. Given any field with n elements we could use the same method to construct $n-1$ mutually orthogonal latin squares. For this reason alone it is natural to ask whether it is possible to construct finite fields other than $\mathbb{Z}_p$.

In fact we have already given one example of a field which does not have a prime number of elements: the field whose elements are the nine skew-symmetric 2×2 matrices over $\mathbb{Z}_3$, discussed in the *Example*, Section 22.3. We shall begin our discussion of the general question by giving another construction of a field of order nine, which we shall denote by F_9.

The elements of F_9 will be represented by 0 and the eight polynomials of degree 0 and 1 with coefficients in the field $\mathbb{Z}_3$, that is

$$F_9 = \{0, 1, 2, x, x+1, x+2, 2x, 2x+1, 2x+2\}.$$

This set is closed under the usual rule for adding polynomials, since, for example

$$(x+1) + (2x+1) = 2 \quad \text{in } \mathbb{Z}_3[x].$$

However, the set is not closed under ordinary multiplication of polynomials, since

$$(x+1) \times (2x+1) = 2x^2 + 1 \quad \text{in } \mathbb{Z}_3[x]$$

and $2x^2 + 1$ is not one of the designated elements of F_9. For this reason we shall define a modified rule for multiplication, whereby we first calculate the ordinary product in $\mathbb{Z}_3[x]$ and then *reduce modulo* $x^2 + 1$. For example,

$$\begin{aligned} (x+1) \times (2x+1) &= 2x^2 + 1 && (\text{in } \mathbb{Z}_3[x]) \\ &= 2 + 2(x^2+1) \\ &= 2 && (\text{in } F_9). \end{aligned}$$

Similarly,

$$\begin{aligned} (2x+1) \times (x) &= 2x^2 + x \\ &= (x+1) + 2(x^2+1) \\ &= x + 1. \end{aligned}$$

It is fairly clear that these + and × operations make F_9 into a ring. The constant polynomials 0 and 1 are indeed the 0 and 1 of the ring, and the remaining axioms

can be verified quite easily. What is rather less obvious is that F_9 is actually a field, because every element except 0 has a multiplicative inverse. The sceptical reader is invited to check the following table.

Element:	1	2	x	$x+1$	$x+2$	$2x$	$2x+1$	$2x+2$
Inverse:	1	2	$2x$	$x+2$	$x+1$	x	$2x+2$	$2x+1$

The general construction of which F_9 is a special case will be explained in Section 23.3. For the moment, we shall emphasize one further property of F_9, concerning its multiplicative group. If we compute the powers of the polynomial $2x+1$ according to the rules in F_9 we obtain the following results:

$$(2x+1)^1 = 2x+1, \quad (2x+1)^2 = x, \quad (2x+1)^3 = x+1,$$
$$(2x+1)^4 = 2, \quad (2x+1)^5 = x+2, \quad (2x+1)^6 = 2x,$$
$$(2x+1)^7 = 2x+2, \quad (2x+1)^8 = 1.$$

Thus we can express all the non-zero elements of F_9 as powers of $2x+1$. In other words, the multiplicative group of F_9 is a cyclic group C_8 generated by $2x+1$:

$$F_9 \setminus \{0\} = U(F_9) = \langle 2x+1 \rangle \approx C_8.$$

In Section 23.4 we shall prove the unexpected result that the multiplicative group of *any* finite field is cyclic.

Exercises 23.1

1 Which members of F_9 can be chosen as the generators of its multiplicative group?

2 Which members of F_9 have square roots in F_9?

3 Without using explicit calculations, prove that the product of all the non-zero elements of F_9 is 2.

4 Show that the additive group of F_9 is not cyclic.

23.2 The order of a finite field

Let F be any field, finite or infinite. Since F is closed under addition, and it contains a multiplicative identity 1, the elements

$$2 = 1+1, \quad 3 = 1+1+1, \quad 4 = 1+1+1+1,$$

and so on, all belong to F. (We do not claim that they are all distinct.) For any positive integer n the sum of n 1's is an element of F; in more formal language we may say that these are the elements of the cyclic subgroup $\langle 1 \rangle$ of the additive group of F.

When F is finite Lagrange's theorem tells us that the order of $\langle 1 \rangle$ is a divisor of $|F|$. This number is called the **characteristic** of F. For example, in the field $\mathbb{Z}_p$ the subgroup $\langle 1 \rangle$ is the whole field, and so the characteristic of $\mathbb{Z}_p$ is p. In general, the characteristic of F is the smallest positive integer m such that $m = 0$ in F, where the latter m denotes the sum of m 1's in F. (When F is infinite the characteristic may or may not be defined.)

It is easy to observe that the characteristic of any finite field must be a prime number. For a number which is not prime can be written as a product $m_1 m_2$, and if we have $m_1 m_2 = 0$ in a field then either $m_1 = 0$ or $m_2 = 0$. Thus a number which is not a prime cannot be the least number for which $m = 0$. In the next theorem we use the fact that the characteristic is prime to determine the structure of the additive group of a finite field, and to show that the order of a finite field must take a very special form.

Theorem 23.2 If F is a finite field of characteristic p, then the additive group of F is isomorphic to $(C_p)^r$, the direct product of r copies of C_p. Consequently, $|F| = p^r$ for some $r \geq 1$.

Proof Given any element $f \neq 0$ in F and any positive integer n there is an element

$$nf = (1 + 1 + \cdots + 1)f = f + f + \cdots + f$$

in F, where there are n terms in the sum. These elements form the cyclic subgroup $\langle f \rangle$ of the additive group of F. Since $p = 0$ in F the only relevant values of n are $0, 1, \ldots, p-1$, and we can regard n as an element of $\mathbb{Z}_p$.

Let us say that the subset $\{f_1, f_2, \ldots, f_k\}$ of F *spans* F if any element of F can be written in the form

$$f = n_1 f_1 + n_2 f_2 + \cdots + n_k f_k \quad (n_1, n_2, \ldots, n_k \in \mathbb{Z}_p).$$

Such sets certainly exist, since the whole of F is one of them. Suppose that $\{f_1, f_2, \ldots, f_r\}$ spans F and no proper subset of it does so. Then each one of the p^r expressions

$$n_1 f_1 + n_2 f_2 + \cdots + n_r f_r \quad (n_1, n_2, \ldots, n_r \in \mathbb{Z}_p)$$

is a member of F, and each member of F is equal to such an expression. If two different expressions represent the same member of F, say

$$n_1 f_1 + n_2 f_2 + \cdots + n_r f_r = m_1 f_1 + m_2 f_2 + \cdots + m_r f_r,$$

then we can pick the first subscript i such that $n_i \neq m_i$ and rewrite the equation as

$$(n_i - m_i) f_i = (m_{i+1} - n_{i+1}) f_{i+1} + \cdots + (m_r - n_r) f_r.$$

Since $n_i - m_i \neq 0$ it has an inverse in $\mathbb{Z}_p$. By multiplying the equation by $(n_i - m_i)^{-1}$ we obtain an expression for f_i in terms of $f_{i+1}, \ldots, f_r$. Thus f_i could be eliminated from the spanning set, contrary to the assumption that no proper subset of $\{f_1, f_2, \ldots, f_r\}$ spans F.

We conclude that there is a bijection in which the elements of F correspond to the r-tuples $(n_1, n_2, \ldots, n_r)$ of elements of $\mathbb{Z}_p$. Since addition in F corresponds to addition of r-tuples, the bijection is an isomorphism of the additive group of F with the direct product of r copies of $\mathbb{Z}_p$ (regarded as a cyclic group): that is, $(C_p)^r$. □

Of course, the most immediate result of the theorem is that if a finite field exists then its order must be a prime power. We are familiar with the fields $\mathbb{Z}_p$ (of order p) and we have come across two examples of a field of order 9 (that is, 3^2). Our task now is to show that fields of order p^r do exist for any prime p and any $r \geq 1$.

Exercises 23.2

1 Tables 23.2.1(a, b) define + and × operations on the set $\{w, y, z, t\}$, and the resulting structure is a field F_4.

Table 23.2.1(a)

+	w	y	z	t
w	w	y	z	t
y	y	w	t	z
z	z	t	w	y
t	t	z	y	w

Table 23.2.1(b)

×	w	y	z	t
w	w	w	w	w
y	w	y	z	t
z	w	z	t	y
t	w	t	y	z

(i) Identify the elements 0 and 1 of F_4.
(ii) Show that the additive and multiplicative groups of F_4 are isomorphic to $C_2 \times C_2$ and C_3, respectively.
(iii) What is the characteristic of F_4?

2 Prove that the subset of a finite field F consisting of all the elements of the form $1 + 1 + \cdots + 1$ is itself a field F_0. (F_0 is called the **prime field** of F.)

3 Let F be a field of characteristic 3. Establish the following identities in F:

(i) $x^3 + y^3 = (x + y)^3$,

(ii) $(x + y)^4 + x^4 + (x - y)^4 + y^4 = 0$.

4 Show that in any field of characteristic p

$$(x + y)^p = x^p + y^p.$$

[Hint: Ex. 11.3.5.]

23.3 Construction of finite fields

The field of order 9 constructed in Section 23.1 can be viewed in a slightly more abstract way. The relation $\sim$ on $\mathbb{Z}_3[x]$ defined by

$$a(x) \sim b(x) \quad \Leftrightarrow \quad a(x) - b(x) \text{ is divisible by } x^2 + 1$$

is an equivalence relation. In fact, the polynomials

$$0, 1, 2, x, x + 1, x + 2, 2x, 2x + 1, 2x + 2$$

form a complete set of representatives for the equivalence classes, and we can take the classes, rather than their representatives, as the elements of F_9. Denoting the equivalence class of $a(x)$ by $[a(x)]$, we can define the addition and multiplication of equivalence classes in the obvious way as

$$[a(x)] + [b(x)] = [a(x) + b(x)], \qquad [a(x)][b(x)] = [a(x)b(x)],$$

and clearly these definitions correspond exactly to the naive calculations in Section 23.1.

This more general viewpoint opens the way for a more general construction of finite fields. The critical point is that the polynomial used to define the equivalence classes must be *irreducible*.

Theorem 23.3 Let $k(x)$ be an irreducible polynomial of degree r in $\mathbb{Z}_p[x]$, and let $\sim$ be the equivalence relation on $\mathbb{Z}_p[x]$ defined by

$$a(x) \sim b(x) \quad \Leftrightarrow \quad a(x) - b(x) \text{ is divisible by } k(x).$$

Then the set of equivalence classes of $\sim$ in $\mathbb{Z}_p[x]$ is a field of order p^r.

Proof It is clear that the polynomials of degree $0, 1, \ldots, r-1$ form a complete set of representatives for the classes, so there are p^r classes in all. Furthermore, it is entirely a matter of routine checking to verify that the classes (like the polynomials themselves) form a ring, and that multiplication is commutative.

The important thing is to show that every class except $[0]$ has a multiplicative inverse, and this is where the irreducibility of $k(x)$ is vital. Given any polynomial $a(x)$ in $\mathbb{Z}_p[x]$, the fact that $k(x)$ is irreducible means that the monic gcd of $a(x)$ and $k(x)$ is 1. So by Theorem 22.6 there are polynomials $f(x)$ and $g(x)$ in $\mathbb{Z}_p[x]$ such that

$$a(x)f(x) + k(x)g(x) = 1.$$

Taking equivalence classes, and noting that $[k(x)] = [0]$, we obtain

$$[a(x)][f(x)] = [1].$$

In other words, $[a(x)]$ has the inverse $[f(x)]$. □

The theorem tells us that in order to construct a field of order p^r it is only necessary to find an irreducible polynomial of degree r in $\mathbb{Z}_p[x]$. That sounds easy. Indeed, in a very down-to-earth sense it is easy, since there are tables of irreducible polynomials covering any values of p and r which are ever likely to arise in practice. But in mathematics we wish to prove such things, and unfortunately the general proof that there is at least one irreducible polynomial for *every* value of p and r is rather difficult. For that reason we shall proceed to derive some general properties of finite fields and indicate some of their applications, deferring the proof of their existence in all cases until Section 23.9.

Exercises 23.3

1 Prove that $x^3 + x^2 + 1$ is irreducible in $\mathbb{Z}_2[x]$, and hence construct a field of order 8. What is the order of its multiplicative group? Describe the group explicitly.

2 Prove that the field of order 4 constructed by using the irreducible polynomial $x^2 + x + 1$ in $\mathbb{Z}_2[x]$ is essentially the same as the field F_4 constructed in Ex. 23.2.1.

3 For which of the following primes p can we construct a field of order p^2 by using the polynomial $x^2 + 1$?

$$p = 3, 5, 7, 11, 13, 19, 23.$$

Describe the multiplicative group for the first two cases in which the field can be constructed.

4 Show that for every prime p there are fields of order p^2 and p^3. [Hint: see Section 22.8.]

23.4 The primitive element theorem

In Section 23.2 we showed that the *additive* group of a finite field of order $q = p^r$ is isomorphic to $(C_p)^r$, the direct product of r cyclic groups C_p. The *multiplicative* group of the field is a group of order $q - 1$ (since 0 is excluded), and its structure is surprisingly straightforward.

Theorem 23.4 The multiplicative group of any finite field is cyclic.

Proof Let F be a field of order q and let F^* denote its multiplicative group $F\backslash\{0\}$. If f is any element of F^* then it follows from Theorem 20.8.3 that $f^{q-1} = 1$. In other words, the equation $x^{q-1} - 1 = 0$ has $q - 1$ roots in F.

We shall show that F^* satisfies the numerical characterization of cyclic groups obtained in Theorem 20.9. Specifically, we shall prove that for each divisor d of $q - 1$ there are d elements f of F^* for which $f^d = 1$.

Suppose $dk = q - 1$; then the following equation in $F[x]$ may be verified by elementary algebra:

$$x^{q-1} - 1 = (x^d - 1)(x^{d(k-1)} + x^{d(k-2)} + \cdots + x^d + 1).$$

Let us denote the second factor by $g(x)$. Since $g(x)$ is a polynomial of degree $d(k - 1)$ the equation $g(x) = 0$ has at most $d(k - 1)$ roots in F, by Theorem 22.8.2. Similarly, the equation $x^d - 1 = 0$ has at most d roots in F. But we have established that the equation $x^{q-1} - 1 = 0$ has exactly $q - 1$ roots in F, and since $d(k - 1) + d = q - 1$ it follows that each of the subsidiary equations must have the maximum possible number of roots. In particular, $x^d - 1 = 0$ has d roots, and so there are d elements of F^* for which $f^d = 1$. Hence the result follows from Theorem 20.9. □

Recall that a group is cyclic if and only if all its elements can be expressed as powers of a single element, which is called a generator for the group. Thus if z is a generator for F^* we have

$$F^* = \{1, z, z^2, \ldots, z^{q-2}\}, \quad \text{where } z^{q-1} = 1.$$

For example, the multiplicative group of the field $\mathbb{Z}_{23}$ is a cyclic group of order 22 with generator 5, since the powers of 5 give all the non-zero elements of the field

$$5^1 = 5, \qquad 5^2 = 2, \qquad 5^3 = 10, \qquad 5^4 = 4,$$

$$\ldots$$

$$5^{19} = 7, \qquad 5^{20} = 12, \qquad 5^{21} = 14, \qquad 5^{22} = 1.$$

In general, a generator for the multiplicative group F^* is known as a **primitive element** in the field F. Using this terminology Theorem 23.4 can be restated in the form:

every finite field has a primitive element.

Despite its elegance and simplicity, the theorem has one inevitable defect: it does not tell us how to find a primitive element in any given case. Since the number of elements of order $q - 1$ in C_{q-1} is $\phi(q - 1)$, it follows that there are $\phi(q - 1)$ primitive elements in any field of order q. If we have to find one of them 'by hand' the best method is a refined form of trial-and-error.

Example Find a primitive element in the field $\mathbb{Z}_{41}$.

Solution The smallest positive integer which could conceivably represent a primitive element of $\mathbb{Z}_{41}$ is 2, so we begin by computing the powers of 2 in $\mathbb{Z}_{41}$. If 2 is a primitive element then we shall obtain all the non-zero elements; if not, we shall have some useful information. We obtain the following table:

n:	1	2	3	4	5	6	7	8	9	10
2^n:	2	4	8	16	32	23	5	10	20	40
n:	11	12	13	14	15	16	17	18	19	20
2^n:	39	37	33	25	9	18	36	31	21	1.

Hence the order of 2 is 20, rather than 40, and we conclude that 2 is not a primitive element. We could try 3, but we note from the table that 9 (that is, 3^2) is equal to 2^{15}, so

$$3^8 = 9^4 = 2^{60} = \left(2^{20}\right)^3 = 1,$$

and the order of 3 is only 8. The elements 4 and 5 are both powers of 2, and so their orders must be divisors of 20. However, the element 6 is not a power of 2, and

$$6^2 = 36 = 2^{17}.$$

The order of 2^{17} is 20 (why?), so the order of 6 is 40 and 6 is the required primitive element. □

In practice we can resort to tables giving the least positive integer which is a primitive element of the field $\mathbb{Z}_p$. These tables are available in computer algebra systems, and they cover an extensive range of values of the prime p. There are also tables which help us when the order of the field is a prime power q, rather than a prime, and we shall now discuss some aspects of this case.

The field F of order $q = p^r$ is constructed by choosing an irreducible polynomial $k(x)$ of degree r in $\mathbb{Z}_p[x]$. If we are lucky, then it will turn out that the polynomial x itself is a primitive element of F, and when this is so we say that $k(x)$ is a **primitive** irreducible polynomial.

Example Show that $x^2 + 2x + 2$ is a primitive irreducible polynomial in $\mathbb{Z}_3[x]$.

Solution First we must show that $x^2 + 2x + 2$ is irreducible. Since it is a quadratic it can only have linear factors, and we may test for these by using the factor theorem. We find that

$$0^2 + (2 \times 0) + 2 = 2, \qquad 1^2 + (2 \times 1) + 2 = 2, \qquad 2^2 + (2 \times 2) + 2 = 1,$$

so there are no linear factors and the polynomial is irreducible. Hence the polynomials

$$0, 1, 2, x, x + 1, x + 2, 2x, 2x + 1, 2x + 2,$$

represent the elements of a field, with respect to the operations of addition and multiplication modulo $x^2 + 2x + 2$. On calculating the powers of x in this field we obtain

$$x^1 = x, \quad x^2 = x + 1, \quad x^3 = 2x + 1, \quad x^4 = 2,$$
$$x^5 = 2x, \quad x^6 = 2x + 2, \quad x^7 = x + 2, \quad x^8 = 1.$$

Thus x generates the multiplicative group of the field, and x^2+2x+2 is a primitive irreducible polynomial. □

The fact that the irreducible polynomial $x^2 + 2x + 2$ is primitive means that there are advantages in using this polynomial to construct a field of order 9. For example, we can use the table of powers of x to facilitate multiplication. Thus, to multiply $x + 1$ and $2x + 1$ we proceed as follows:

$$(x + 1) \times (2x + 1) = x^2 \times x^3 = x^5 = 2x.$$

On the other hand, the polynomial $x^2 + 1$ in $\mathbb{Z}_3[x]$, which we used to construct F_9 in Section 23.1, is not really a good choice since it is not primitive. Computing powers of x in F_9 we find that $x^4 = 1$, so the polynomial x is not a primitive element in F_9.

These remarks lead naturally to the final theoretical question about finite fields. We know that if we are given a prime power $q = p^r$ we can use any irreducible polynomial of degree r in $\mathbb{Z}_p[x]$ to construct a field of order q. What is the connection between the fields constructed by using different polynomials?

At this point the reader who has faith in the beauty of mathematics will surely guess: they are all the same. And indeed it is so. We have shown that any two fields of order q have isomorphic additive groups and isomorphic multiplicative groups; so it is hardly surprising that the fields themselves are isomorphic, in the sense that there is a bijection from one to the other which is simultaneously an isomorphism of the additive and multiplicative groups. Of course, from a constructive point of view it is more important to know that a field with q elements exists than it is to know that the field is unique. For this reason we shall be content with giving a formal proof of the existence only (in Section 23.9).

In summary, the theory of finite fields turns out to be remarkably simple.

Any finite field has prime power order $q = p^r$.
There is essentially just one field of order q.
The additive group of the field is $(C_p)^r$.
The multiplicative group of the field is C_{q-1}.

We shall use the notation $\mathbb{F}_q$ for the unique field of order q. These fields are often known as *Galois fields*, after Évariste Galois (1811–1832), and sometimes they are denoted by the symbol $GF(q)$. Of course when q is itself a prime p the field $\mathbb{F}_p$ (or $GF(p)$) is simply the familiar field $\mathbb{Z}_p$, whose elements are the integers modulo p.