1

Introduction

Some aspects of secure communication

For at least two thousand years there have been people who wanted to send messages which could only be read by the people for whom they were intended. When a message is sent by hand, carried from the sender to the recipient, whether by a slave, as in ancient Greece or Rome, or by the Post Office today, there is a risk of it going astray. The slave might be captured or the postman might deliver to the wrong address. If the message is written *in clear*, that is, in a natural language without any attempt at concealment, anyone getting hold of it will be able to read it and, if they know the language, understand it.

In more recent times messages might be sent by telegraph, radio, telephone, fax or e-mail but the possibility of them being intercepted is still present and, indeed, has increased enormously since, for example, a radio transmission can be heard by anyone who is within range and tuned to the right frequency whilst an e-mail message might go to a host of unintended recipients if a wrong key on a computer keyboard is pressed or if a 'virus' is lurking in the computer.

It may seem unduly pessimistic but a good rule is to assume that any message which is intended to be confidential *will* fall into the hands of someone who is not supposed to see it and therefore it is prudent to take steps to ensure that they will, at least, have great difficulty in reading it and, preferably, will not be able to read it at all. The extent of the damage caused by unintentional disclosure may depend very much on the time that has elapsed between interception and reading of the message. There are occasions when a delay of a day or even a few hours in reading a message nullifies the damage; for example, a decision by a shareholder to

buy or sell a large number of shares at once or, in war, an order by an army commander to attack in a certain direction at dawn next day. On other occasions the information may have long term value and must be kept secret for as long as possible, such as a message which relates to the planning of a large scale military operation.

The effort required by a rival, opponent or enemy to read the message is therefore relevant. If, using the best known techniques and the fastest computers available, the message can't be read by an unauthorised recipient in less time than that for which secrecy or confidentiality is essential then the sender can be reasonably happy. He cannot ever be *entirely* happy since success in reading some earlier messages may enable the opponent to speed up the process of solution of subsequent messages. It is also possible that a technique has been discovered of which he is unaware and consequently his opponent is able to read the message in a much shorter time than he believed possible. Such was the case with the German Enigma machine in the 1939–45 war, as we shall see in Chapter 9.

Julius Caesar's cipher

The problem of ensuring the security of messages was considered by the ancient Greeks and by Julius Caesar among others. The Greeks thought of a bizarre solution: they took a slave and shaved his head and scratched the message on it. When his hair had grown they sent him off to deliver the message. The recipient shaved the slave's head and read the message. This is clearly both a very insecure and an inefficient method. Anyone knowing of this practice who intercepted the slave could also shave his head and read the message. Furthermore it would take weeks to send a message and get a reply by this means.

Julius Caesar had a better idea. He wrote down the message and moved every letter three places forward in the alphabet, so that, in the English alphabet, A would be replaced by D, B by E and so on up to W which would be replaced by Z and then X by A, Y by B and finally Z by C. If he had done this with his famous message

VENI. VIDI. VICI.
(I came. I saw. I conquered.)

and used the 26-letter alphabet used in English-speaking countries (which, of course, he would not) it would have been sent as

YHQL. YLGL. YLFL.

Not a very sophisticated method, particularly since it reveals that the message consists of three words each of four letters, with several letters repeated. It is difficult to overcome such weaknesses in a naïve system like this although extending the alphabet from 26 letters to 29 or more in order to accommodate punctuation symbols and spaces would make the word lengths *slightly* less obvious. Caesar nevertheless earned a place in the history of *cryptography*, for the 'Julius Caesar' cipher, as it is still called, is an early example of an *encryption system* and is a special case of a *simple substitution cipher* as we shall see in Chapter 2.

Some basic definitions

Since we shall be repeatedly using words such as *digraph, cryptography* and *encryption* we define them now.

A *monograph* is a single letter of whatever alphabet we are using. A *digraph* is any pair of *adjacent* letters, thus `AT` is a digraph. A *trigraph* consists of three adjacent letters, so `THE` is a trigraph, and so on. A *polygraph* consists of an unspecified number of adjacent letters. A polygraph need not be recognisable as a word in a language but if we are attempting to decipher a message which is expected to be in English and we find the heptagraph `MEETING` it is much more promising than if we find a heptagraph such as `DKRPIGX`.

A *symbol* is any character, including letters, digits, and punctuation, whilst a *string* is any adjacent collection of symbols. The *length* of the string is the number of characters that it contains. Thus `A3£%$` is a string of length 5.

A *cipher system,* or *cryptographic system*, is any system which can be used to change the text of a message with the aim of making it unintelligible to anyone other than intended recipients.

The process of applying a cipher system to a message is called *encipherment* or *encryption.*

The original text of a message, before it has been enciphered, is referred to as *the plaintext*; after it has been enciphered it is referred to as *the cipher text*.

The reverse process to *encipherment*, recovering the original text of a message from its enciphered version, is called *decipherment* or *decryption.* These two words are not, perhaps, entirely synonymous. The intended recipient of a message would think of himself as *deciphering* it whereas an unintended recipient who is trying to make sense of it would think of himself as *decrypting* it.

Cryptography is the study of the design and use of *cipher systems* including their strengths, weaknesses and vulnerability to various methods of attack. A *cryptographer* is anyone who is involved in *cryptography.*

Cryptanalysis is the study of methods of solving *cipher systems*. A *cryptanalyst* (often popularly referred to as a *codebreaker*) is anyone who is involved in *cryptanalysis.*

Cryptographers and cryptanalysts are adversaries; each tries to outwit the other. Each will try to imagine himself in the other's position and ask himself questions such as 'If I were him what would I do to defeat me?' The two sides, who will probably never meet, are engaged in a fascinating intellectual battle and the stakes may be very high indeed.

Three stages to decryption: identification, breaking and setting

When a cryptanalyst first sees a cipher message his first problem is to discover what type of cipher system has been used. It may have been one that is already known, or it may be new. In either case he has the problem of ***identification***. To do this he would first take into account any available collateral information such as the type of system the sender, if known, has previously used or any new systems which have recently appeared anywhere. Then he would examine the preamble to the message. The preamble may contain information to help the intended recipient, but it may also help the cryptanalyst. Finally he would analyse the message itself. If it is too short it may be impossible to make further progress and he must wait for more messages. If the message is long enough, or if he has already gathered several sufficiently long messages, he would apply a variety of mathematical tests which should certainly tell him whether a code book, or a relatively simple cipher system or something more sophisticated is being used.

Having identified the system the cryptanalyst may be able to estimate how much material (e.g. how many cipher letters) he will need if he is to have a reasonable chance of ***breaking*** it, that is, knowing exactly how messages are enciphered by the system. If the system is a simple one where there are no major changes from one message to the next, such as a codebook, simple substitution or transposition (see Chapters 2 to 6) he may then be able to decrypt the message(s) without too much difficulty. If, as is much more likely, there are parts of the system that are changed from message to message he will first need to determine the parts that don't

change. As an example, anticipating Chapter 9, the Enigma machine contained several wheels; inside these wheels were wires; the wirings inside the wheels didn't change but the order in which the wheels were placed in the machine changed daily. Thus, the wirings were the fixed part but their order was variable. The breaking problem is the most difficult part; it could take weeks or months and involve the use of mathematical techniques, exploitation of operator errors or even information provided by spies.

When the fixed parts have all been determined it would be necessary to work out the variable parts, such as starting positions of the Enigma wheels, which changed with each message. This is the ***setting*** problem. When it is solved the messages can be decrypted.

So ***breaking*** refers to the encipherment system in general whilst ***setting*** refers to the decryption of individual messages.

Codes and ciphers

Although the words are often used loosely we shall distinguish between *codes* and *ciphers.* In a *code* common phrases, which may consist of one or more letters, numbers, or words, are replaced by, typically, four or five letters or numbers, called *code groups,* taken from a *code-book*. For particularly common phrases or letters there may be more than one *code group* provided with the intention that the user will vary his choice, to make identification of the common phrases more difficult. For example, in a four-figure code the word 'Monday' might be given three alternative *code groups* such as 1538 or 2951 or 7392. We shall deal with *codes* in Chapter 6.

Codes are a particular type of *cipher system* but not all *cipher systems* are *codes* so we shall use the word *cipher* to refer to methods of *encipherment* which do not use *code-books* but produce the enciphered message from the original plaintext according to some rule (the word *algorithm* is nowadays preferred to 'rule', particularly when computer programs are involved). The distinction between *codes* and *ciphers* can sometimes become a little blurred, particularly for simple systems. The Julius Caesar cipher could be regarded as using a one-page code-book where opposite each letter of the alphabet is printed the letter three positions further on in the alphabet. However, for most of the systems we shall be dealing with the distinction will be clear enough. In particular the Enigma, which is often erroneously referred to as 'the Enigma code', is quite definitely a *cipher machine* and not a *code* at all.

Historically, two basic ideas dominated cryptography until relatively recent times and many cipher systems, including nearly all those considered in the first 11 chapters of this book were based upon one or both of them. The first idea is to shuffle the letters of the alphabet, just as one would shuffle a pack of cards, the aim being to produce what might be regarded as a random ordering, permutation, or anagram of the letters. The second idea is to convert the letters of the message into numbers, taking A = 0, B = 1, ..., Z = 25, and then add some other numbers, which may themselves be letters converted into numbers, known as 'the *key*', to them letter by letter; if the addition produces a number greater than 25 we subtract 26 from it (this is known as (mod 26) *arithmetic*). The resulting numbers are then converted back into letters. If the numbers which have been added are produced by a sufficiently unpredictable process the resultant cipher message may be very difficult, or even impossible, to decrypt unless we are given the key.

Interestingly, the Julius Caesar cipher, humble though it is, can be thought of as being an example of either type. In the first case our 'shuffle' is equivalent to simply moving the last three cards to the front of the pack so that all letters move 'down' three places and X, Y and Z come to the front. In the second case the key is simply the number 3 repeated indefinitely – as 'weak' a key as could be imagined.

Translating a message into another language might be regarded as a form of encryption using a code-book (i.e. dictionary), but that would seem to be stretching the use of the word *code* too far. Translating into another language by looking up each word in a code-book acting as a dictionary is definitely not to be recommended, as anyone who has tried to learn another language knows.* On the other hand use of a little-known language to pass on messages of short term importance might sometimes be reasonable. It is said, for example, that in the Second World War Navajo Indian soldiers were sometimes used by the American Forces in the Pacific to pass on messages by telephone in their own language, on the reasonable assumption that even if the enemy intercepted the telephone calls they would be unlikely to have anyone available who could understand what was being said.

* I recall a boy at school who wrote a French essay about a traveller in the Middle Ages arriving at an inn at night, knocking on the door and being greeted with the response 'What Ho! Without.' This he translated as 'Que Ho! Sans.' The French Master, after a moment of speechlessness, remarked that 'You have obviously looked up the words in the sort of French dictionary they give away with bags of sugar.'

Another form of encryption is the use of some personal shorthand. Such a method has been employed since at least the Middle Ages by people, such as Samuel Pepys, who keep diaries. Given enough entries such codes are not usually difficult to solve. Regular occurrences of symbols, such as those representing the names of the days of the week, will provide good clues to certain polygraphs. A much more profound example is provided by Ventris's decipherment of the ancient Mycenaen script known as Linear B, based upon symbols representing Greek syllables [1.4].

The availability of computers and the practicability of building complex electronic circuits on a silicon chip have transformed both cryptography and cryptanalysis. In consequence, some of the more recent cipher systems are based upon rather advanced mathematical ideas which require substantial computational or electronic facilities and so were impracticable in the pre-computer age. Some of these are described in Chapters 12 and 13.

Assessing the strength of a cipher system

When a new cipher system is proposed it is essential to assess its strength against all known attacks and on the assumption that the cryptanalyst knows what type of cipher system, but not all the details, is being used. The strength can be assessed for three different situations:

(1) that the cryptanalyst has only cipher texts available;
(2) that he has both cipher texts and their original plaintexts;
(3) that he has both cipher and plain for texts *which he himself has chosen*.

The first situation is the 'normal' one; a cipher system that can be solved in a reasonable time in this case should not be used. The second situation can arise, for example, if identical messages are sent both using the new cipher and using an 'old' cipher which the cryptanalyst can read. Such situations, which constitute a serious breach of security, not infrequently occur. The third situation mainly arises when the cryptographer, wishing to assess the strength of his proposed system, challenges colleagues, acting as the enemy, to solve his cipher and allows them to dictate what texts he should encipher. This is a standard procedure in testing new systems. A very interesting problem for the cryptanalyst is how to construct texts which when enciphered will provide him with the maximum information on the details of the system. The format of these

messages will depend on how the encipherment is carried out. The second and third situations can also arise if the cryptanalyst has access to a spy in the cryptographer's organisation; this was the case in the 1930s when the Polish cryptanalysts received plaintext and cipher versions of German Enigma messages. A cipher system that cannot be solved even in this third situation is a strong cipher indeed; it is what the cryptographers want and the cryptanalysts fear.

Error detecting and correcting codes

A different class of codes are those which are intended to ensure the *accuracy* of the information which is being transmitted and not to hide its *content*. Such codes are known as *error detecting and correcting codes* and they have been the subject of a great deal of mathematical research. They have been used from the earliest days of computers to protect against errors in the memory or in data stored on magnetic tape. The earliest versions, such as Hamming codes, can detect and correct a *single* error in a 6-bit character. A more recent example is the code which was used for sending data from Mars by the Mariner spacecraft which could correct up to 7 errors in each 32-bit 'word', so allowing for a considerable amount of corruption of the signal on its long journey back to Earth. On a different level, a simple example of an error *detecting,* but not error *correcting*, code is the ISBN (International Standard Book Number). This is composed of either 10 digits, or 9 digits followed by the letter X (which is interpreted as the number 10), and provides a check that the ISBN does not contain an error. The check is carried out as follows: form the sum

> 1 times (the first digit) +2 times (the second digit) +3 times (the third digit) . . . and so on to +10 times (the tenth digit).

The digits are usually printed in four groups separated by hyphens or spaces for convenience. The first group indicates the language area, the second identifies the publisher, the third is the publisher's serial number and the last group is the single digit *check digit*.

The sum (known as the *check sum*) should produce a multiple of 11; if it doesn't there is an error in the ISBN. For example:

1-234-56789-X produces a check sum of

$$1(1)+2(2)+3(3)+4(4)+5(5)+6(6)+7(7)+8(8)+9(9)+10(10)$$

which is

$$1+4+9+16+25+36+49+64+81+100=385=35\times 11$$

and so is valid. On the other hand

0-987-65432-1 produces a check sum of

$$0+18+24+28+30+30+28+24+18+10=210=19\times 11+1$$

and so must contain at least one error.

The ISBN code can *detect* a *single* error but it cannot *correct* it and if there are two or more errors it may indicate that the ISBN is correct, when it isn't.

The subject of error correcting and detecting codes requires some advanced mathematics and will not be considered further in this book. Interested readers should consult books such as [1.1], [1.2], [1.3].

Other methods of concealing messages

There are other methods for concealing the meaning or contents of a message that do not rely on codes or ciphers. The first two are not relevant here but they deserve to be mentioned. Such methods are

(1) the use of secret or 'invisible' ink,
(2) the use of *microdots*, tiny photographs of the message on microfilm, stuck onto the message in a non-obvious place,
(3) 'embedding' the message inside an otherwise innocuous message, the *words* or *letters* of the secret message being scattered, according to some rule, throughout the non-secret message.

The first two of these have been used by spies; the outstandingly successful 'double agent' Juan Pujol, known as GARBO, used both methods from 1942 to 1945 [1.5]. The third method has also been used by spies but may well also have been used by prisoners of war in letters home to pass on information as to where they were or about conditions in the camp; censors would be on the look-out for such attempts. The third method is discussed in Chapter 7.

The examples throughout this book are almost entirely based upon English texts using either the 26-letter alphabet or an extended version of it to allow inclusion of punctuation symbols such as space, full stop and comma. Modification of the examples to include more symbols or numbers or to languages with different alphabets presents no difficulties *in theory*. If, however, the cipher system is being implemented on a physical device it may be impossible to change the alphabet size without redesigning it; this is true of the Enigma and Hagelin machines, as we shall see later. Non-alphabetic languages, such as Japanese, would need to be 'alphabetised' or, perhaps, treated as non-textual material as are photographs, maps, diagrams etc. which can be enciphered by using specially

designed systems of the type used in enciphering satellite television programmes or data from space vehicles.

Modular arithmetic

In cryptography and cryptanalysis it is frequently necessary to add two streams of numbers together or to subtract one stream from the other but the form of addition or subtraction used is usually not that of ordinary arithmetic but of what is known as *modular arithmetic*. In modular arithmetic all additions and subtractions (and multiplications too, which we shall require in Chapters 12 and 13) are carried out with respect to a fixed number, known as *the modulus*. Typical values of the modulus in cryptography are 2, 10 and 26. Whichever modulus is being used all the numbers which occur are replaced by their remainders when they are divided by the modulus. If the remainder is negative the modulus is added so that the remainder becomes non-negative. If, for example, the modulus is 26 the only numbers that can occur are 0 to 25. If then we add 17 to 19 the result is 10 since $17 + 19 = 36$ and 36 leaves remainder 10 when divided by 26. To denote that modulus 26 is being used we would write

$$17 + 19 \equiv 10 \pmod{26}.$$

If we subtract 19 from 17 the result (-2) is negative so we add 26, giving 24 as the result.

The symbol $\equiv$ is read as 'is congruent to' and so we would say

'36 is congruent to 10 (mod 26)' and '-2 is congruent to 24 (mod 26)'.

When two streams of numbers (mod 26) are added the rules apply to each pair of numbers separately, with no 'carry' to the next pair. Likewise when we subtract one stream from another (mod 26) the rules apply to each pair of digits separately with no 'borrowing' from the next pair.

Example 1.1

Add the stream 15 11 23 06 11 to the stream 17 04 14 19 23 (mod 26).

Solution

	15	11	23	06	11
	17	04	14	19	23
	32	15	37	25	34
(mod 26)	06	15	11	25	08

and so the result is 06 15 11 25 08.