CHAPTER 7

Historical Ciphers

Chapter Goals

- To explain a number of historical ciphers, such as the Caesar cipher and the substitution cipher.
- To show how these historical ciphers can be broken because they do not hide the underlying statistics of the plaintext.
- To introduce the concepts of substitution and permutation as basic cipher components.
- To introduce a number of attack techniques, such as chosen plaintext attacks.

7.1. Introduction

An encryption algorithm, or cipher, is a means of transforming plaintext into ciphertext under the control of a secret key. This process is called encryption or encipherment. We write

$$c = e_k(m),$$

where

- m is the plaintext,
- e is the cipher function,
- k is the secret key,
- c is the ciphertext.

The reverse process is called decryption or decipherment, and we write

$$m = d_k(c).$$

Note that the encryption and decryption algorithms e, d are public: the secrecy of m given c depends totally on the secrecy of k.

The above process requires that each party needs access to the secret key. The key needs to be known to both sides, but needs to be kept secret. Encryption algorithms which have this property are called *symmetric cryptosystems* or secret key cryptosystems. There is a form of cryptography which uses two different types of key; one is publicly available and used for encryption whilst the other is private and used for decryption. These latter types of cryptosystems are called *asymmetric cryptosystems* or *public key cryptosystems*, and we shall return to them in a later chapter.

Usually in cryptography the communicating parties are denoted by A and B. However, often one uses the more user-friendly names of Alice and Bob. But you should not assume that the parties are necessarily human; we could be describing a communication being carried out between two autonomous machines. The eavesdropper, bad girl, adversary or attacker is usually given the name Eve.

In this chapter we shall present some historical ciphers which were used in the pre-computer age to encrypt data. We shall show that these ciphers are easy to break as soon as one understands the statistics of the underlying language, in our case English. In Chapter 9 we shall study this

N.P. Smart, *Cryptography Made Simple*, Information Security and Cryptography, DOI 10.1007/978-3-319-21936-3_7

relationship between how easy the cipher is to break and the statistical distribution of the underlying plaintext in more detail.

Letter	Freq. (%)	Letter	Freq. (%)
A	8.2	N	6.7
B	1.5	O	7.5
C	2.8	P	1.9
D	4.2	Q	0.1
E	12.7	R	6.0
F	2.2	S	6.3
G	2.0	T	9.0
H	6.1	U	2.8
I	7.0	V	1.0
J	0.1	W	2.4
K	0.8	X	0.1
L	4.0	Y	2.0
M	2.4	Z	0.1

TABLE 7.1. English letter frequencies

FIGURE 7.1. English letter frequencies

The distribution of English letter frequencies is described in Table 7.1, or graphically in Figure 7.1. As one can see, the most common letters are E and T. It often helps to know second-order statistics about the underlying language, such as which are the most common sequences of two or three letters, called bigrams and trigrams. The most common bigrams in English are given by Table 7.2, with the associated approximate frequencies. The most common trigrams are, in decreasing order,

THE, ING, AND, HER, ERE, ENT, THA, NTH, WAS, ETH, FOR.

Armed with this information about English we are now able to examine and break a number of historical ciphers.

7.2. Shift Cipher

We first present one of the earliest ciphers, called the shift cipher. Encryption is performed by replacing each letter by the letter located a certain number of places further on in the alphabet. So for example if the key was three, then the plaintext A would be replaced by the ciphertext D, the letter B would be replaced by E and so on. The plaintext word HELLO would be encrypted as the ciphertext KHOOR. When this cipher is used with the key three, it is often called the Caesar cipher, although in many books the name Caesar cipher is sometimes given to the shift cipher with

Bigram	Freq. (%)	Bigram	Freq. (%)
TH	3.15	HE	2.51
AN	1.72	IN	1.69
ER	1.54	RE	1.48
ES	1.45	ON	1.45
EA	1.31	TI	1.28
AT	1.24	ST	1.21
EN	1.20	ND	1.18

TABLE 7.2. English bigram frequencies

any key. Strictly this is not correct since we only have evidence that Julius Caesar used the cipher with the key three.

There is a more mathematical explanation of the shift cipher which will be instructive for future discussions. First we need to identify each letter of the alphabet with a number. It is usual to identify the letter A with the number 0, the letter B with number 1, the letter C with the number 2 and so on until we identify the letter Z with the number 25. After we convert our plaintext message into a sequence of numbers, the ciphertext in the shift cipher is obtained by adding to each number the secret key k modulo 26, where the key is a number in the range 0 to 25. In this way we can interpret the shift cipher as a *stream cipher*, with keystream given by the repeating sequence

$$k, k, k, k, k, k, \ldots$$

This keystream is not very random, which results in it being easy to break the shift cipher. A naive way of breaking the shift cipher is to simply try each of the possible keys in turn, until the correct one is found. There are only 26 possible keys so the time for this exhaustive key search is very small, particularly if it is easy to recognize the underlying plaintext when it is decrypted.

We shall show how to break the shift cipher by using the statistics of the underlying language. Whilst this is not strictly necessary for breaking this cipher, later we shall see a cipher that is made up of a number of shift ciphers applied in turn and then the following statistical technique will be useful. Using a statistical technique on the shift cipher is also instructive as to how statistics of the underlying plaintext can arise in the resulting ciphertext. Take the following example ciphertext, which since it is public knowledge we represent in blue.

GB OR, BE ABG GB OR: GUNG VF GUR DHRFGVBA:
JURGURE 'GVF ABOYRE VA GUR ZVAQ GB FHSSRE
GUR FYVATF NAQ NEEBJF BS BHGENTRBHF SBEGHAR,
BE GB GNXR NEZF NTNVAFG N FRN BS GEBHOYRF,
NAQ OL BCCBFVAT RAQ GURZ? GB QVR: GB FYRRC;
AB ZBER; NAQ OL N FYRRC GB FNL JR RAQ
GUR URNEG-NPUR NAQ GUR GUBHFNAQ ANGHENY FUBPXF
GUNG SYRFU VF URVE GB, 'GVF N PBAFHZZNGVBA
QRIBHGYL GB OR JVFU'Q. GB QVR, GB FYRRC;
GB FYRRC: CREPUNAPR GB QERNZ: NL, GURER'F GUR EHO;
SBE VA GUNG FYRRC BS QRNGU JUNG QERNZF ZNL PBZR
JURA JR UNIR FUHSSYRQ BSS GUVF ZBEGNY PBVY,
ZHFG TVIR HF CNHFR: GURER'F GUR ERFCRPG
GUNG ZNXRF PNYNZVGL BS FB YBAT YVSR;

One technique used in breaking the previous sample ciphertext is to notice that the ciphertext still retains details about the word lengths of the underlying plaintext. For example the ciphertext

letter N appears as a single letter word. Since the only common single-letter words in English are A and I we can conclude that the key is either 13, since N is thirteen letters on from A in the alphabet, or 5, since N is five letters on from I in the alphabet. Hence, the moral here is to always remove word breaks from the underlying plaintext before encrypting using the shift cipher. But even if we ignore this information about the word break, we can still break this cipher using frequency analysis.

We compute the frequencies of the letters in the ciphertext and compare them with the frequencies obtained from English which we saw in Figure 7.1. We present the two bar graphs one above each other in Figure 7.2 so you can see that one graph looks almost like a shift of the other graph. The statistics obtained from the sample ciphertext are given in blue, whilst the statistics obtained from the underlying plaintext language are given in red. Note, we do not compute the red statistics from the actual plaintext since we do not know this yet, we only make use of the knowledge of the underlying language.

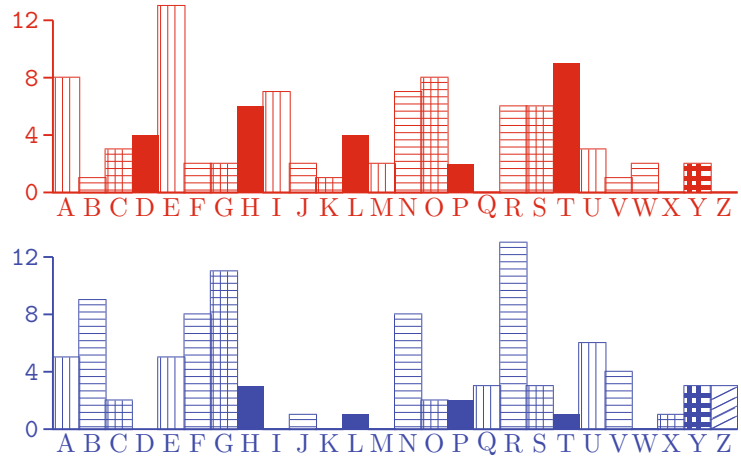

FIGURE 7.2. Comparison of plaintext and ciphertext frequencies for the shift cipher example

By comparing the two bar graphs in Figure 7.2 we can see by how much we think the blue graph has been shifted compared with the red graph. By examining where we think the plaintext letter E may have been shifted, one can hazard a guess that it is shifted by one of

$$2, 9, 13 \text{ or } 23.$$

Then by trying to deduce by how much the plaintext letter A has been shifted we can guess that it has been shifted by one of

$$1, 6, 13 \text{ or } 17.$$

The only shift value which is consistent appears to be the value 13, and we conclude that this is the most likely key value.

One may ask whether there is a more scientific way of performing the above comparison of bar graphs. Indeed there is, using something called the statistical distance. Let X and Y be random variables distributed according to distributions D_1 and D_2; we let V denote the support of X and Y (i.e. the set of values which can occur for X or Y with non-zero probability). We then define the statistical distance (actually the *total variation distance*, as there are many different statistical distances one can define) by

$$\Delta[X, Y] = \frac{1}{2} \sum_{u \in V} \left| \Pr_{X \leftarrow D_1}[X = u] - \Pr_{Y \leftarrow D_2}[Y = u] \right|.$$

To apply this to our example we let X denote the probabilities of letters occuring in English, i.e. the probabilities from Table 7.1, and we let Y_k denote the probabilities obtained from the ciphertext but shifted by the key value k. So we have twenty-six different distributions Y_k to compare to the fixed distribution X. The one which has the smallest distance is the one most likely to be the key.

Applying this method in this example we find the statistical distances given in Table 7.3. The value for the key 13 is significantly smaller than the values for the other keys; thus we can conclude (using this more scientific method) that the key is 13.

k	$\Delta(X, Y_k)$	k	$\Delta(X, Y_k)$
0	48.4	13	10.8
1	44.6	14	44.8
2	44.0	15	57.0
3	49.5	16	55.3
4	53.2	17	47.0
5	52.9	18	48.5
6	46.0	19	49.1
7	53.9	20	45.3
8	52.7	21	56.4
9	43.8	22	51.6
10	51.3	23	47.5
11	56.8	24	43.8
12	46.7	25	45.2

TABLE 7.3. Statistical distance between X and Y_k for the shift cipher example

We can now decrypt the ciphertext, using this key. This reveals that the underlying plaintext is the following text from Shakespeare's *Hamlet*:

> To be, or not to be: that is the question:
> Whether 'tis nobler in the mind to suffer
> The slings and arrows of outrageous fortune,
> Or to take arms against a sea of troubles,
> And by opposing end them? To die: to sleep;
> No more; and by a sleep to say we end
> The heart-ache and the thousand natural shocks
> That flesh is heir to, 'tis a consummation
> Devoutly to be wish'd. To die, to sleep;
> To sleep: perchance to dream: ay, there's the rub;
> For in that sleep of death what dreams may come
> When we have shuffled off this mortal coil,
> Must give us pause: there's the respect
> That makes calamity of so long life;

7.3. Substitution Cipher

The main problem with the shift cipher is that the number of keys is too small; we only have 26 possible keys. To increase the number of keys the *substitution cipher* was invented. To write down a key for the substitution cipher we first write down the alphabet, and then a permutation of the

alphabet directly below it. This mapping gives the substitution we make between the plaintext and the ciphertext

Plaintext alphabet	ABCDEFGHIJKLMNOPQRSTUVWXYZ
Ciphertext alphabet	GOYDSIPELUAVCRJWXZNHBQFTMK

Encryption involves replacing each letter in the top row by its value in the bottom row. Decryption involves first looking for the letter in the bottom row and then seeing which letter in the top row maps to it. Hence, the plaintext word HELLO would encrypt to the ciphertext ESVVJ if we used the substitution given above.

The number of possible keys is equal to the total number of permutations on 26 letters, namely the size of the group S_{26}, which is

$$26! \approx 4.03 \cdot 10^{26} \approx 2^{88}.$$

Since, as a rule of thumb, it is only feasible to run a computer on a problem which takes under 2^{80} steps we can deduce that this large key space is far too large to enable a brute force search even using a modern computer. Despite this we can break substitution ciphers using statistics of the underlying plaintext language, just as we did for the shift cipher.

Whilst the shift cipher can be considered as a stream cipher since the ciphertext is obtained from the plaintext by combining it with a keystream, the substitution cipher operates much more like a modern block cipher, with a block length of one English letter. A ciphertext block is obtained from a plaintext block by applying some (admittedly simple in this case) key-dependent algorithm.

Substitution ciphers are the ciphers commonly encountered in puzzle books; they have an interesting history and have occurred many times in literature. See for example the Sherlock Holmes story *The Adventure of the Dancing Men* by Arthur Conan Doyle; the plot of this story rests on a substitution cipher where the ciphertext characters are taken from an alphabet of "stick men" in various positions. The method of breaking the cipher as described by Holmes to Watson in this story is precisely the method we shall adopt below.

Example: We give a detailed example, which we make slightly easier by keeping in the ciphertext details about the underlying word spacing used in the plaintext. This is only for ease of exposition; the techniques we describe can still be used if we ignore these word spacings, although more care and thought is required. Consider the ciphertext

XSO MJIWXVL JODIVA STW VAO VY OZJVCO'W LTJDOWX KVAKOAXJTXIVAW VY SIDS XOKSAVLVDQ IAGZWXJQ. KVUCZXOJW, KVUUZAIKTXIVAW TAG UIKJVOLOKXJVAIKW TJO HOLL JOCJOWOAXOG, TLVADWIGO GIDIXTL UOGIT, KVUCZXOJ DTUOW TAG OLOKXJVAIK KVUUOJKO. TW HOLL TW SVWXIAD UTAQ JOWOTJKS TAG CJVGZKX GONOLVCUOAX KOAXJOW VY UTPVJ DLVMTL KVUCTAIOW, XSO JODIVA STW T JTCIGLQ DJVHIAD AZUMOJ VY IAAVNTXINO AOH KVUCTAIOW. XSO KVUCZXOJ WKIOAKO GOCTJXUOAX STW KLVWO JOLTXIVAWSICW HIXS UTAQ VY XSOWO VJDTAIWTXIVAW NIT KVLLTMVJTXINO CJVPOKXW, WXTYY WOKVAGUOAXW TAG NIWIXIAD IAGZWXJITL WXTYY. IX STW JOKOAXLQ IAXJVGZKOG WONOJTL UOKSTAIWUW YVJ GONOLVCIAD TAG WZCCVJXIAD OAXJOCJOAOZJITL WXZGOAXW TAG WXTYY, TAG TIUW XV CLTQ T WIDAIYIKTAX JVLO IA XSO GONOLVCUOAX VY SIDS-XOKSAVLVDQ IAGZWXJQ IA XSO JODIVA.

XSO GOCTJXUOAX STW T LTJDO CJVDJTUUO VY JOWOTJKS WZCCVJXOG MQ IAGZWXJQ, XSO OZJVCOTA ZAIVA, TAG ZE DVNOJAUOAX JOWOTJKS OWXTMLIWSUOAXW TAG CZMLIK KVJCVJTXIVAW. T EOQ OLOUOAX VY XSIW IW XSO WXJVAD LIAEW XSTX XSO GOCTJXUOAX STW HIXS XSO KVUCZXOJ, KVUUZAIKTXIVAW, UIKJVOLOKXJVAIKW TAG UOGIT IAGZWXJIOW IA XSO MJIWXVL JODIVA . XSO TKTGOUIK JOWOTJKS CJVDJTUUO IW VJDTAIWOG IAXV WONOA DJVZCW, LTADZTDOW

TAG TJKSIXOKXZJO, GIDIXTL UOGIT, UVMILO TAG HOTJTMLO KVUCZXIAD, UTKSIAO LOTJAIAD, RZTAXZU KVUCZXIAD, WQWXOU NOJIYIKTXIVA, TAG KJQCXVDJTCSQ TAG IAYVJUTXIVA WOKZJIXQ.

We can compute the following frequencies for single letters in the above ciphertext.

Letter	Freq. (%)	Letter	Freq. (%)	Letter	Freq. (%)
A	8.6995	B	0.0000	C	3.0493
D	3.1390	E	0.2690	F	0.0000
G	3.6771	H	0.6278	I	7.8923
J	7.0852	K	4.6636	L	3.5874
M	0.8968	N	1.0762	O	11.479
P	0.1793	Q	1.3452	R	0.0896
S	3.5874	T	8.0717	U	4.1255
V	7.2645	W	6.6367	X	8.0717
Y	1.6143	Z	2.7802		

In addition we determine that the most common bigrams in this piece of ciphertext are

TA, AX, IA, VA, WX, XS, AG, OA, JO, JV,

whilst the most common trigrams are

OAX, TAG, IVA, XSO, KVU, TXI, UOA, AXS.

Since the ciphertext letter O occurs with the greatest frequency, namely 11.479, we can guess that the ciphertext letter O corresponds to the plaintext letter E. We now look at what this means for two of the common trigrams found in the ciphertext

- The ciphertext trigram OAX corresponds to E * *.
- The ciphertext trigram XSO corresponds to * * E.

We examine similar common trigrams in English, which start or end with the letter E. We find that three common ones are given by ENT, ETH and THE. Since in the ciphertext trigrams we have one letter, X, in the first position in one and the last position in the other, we look for a similar letter in the English trigrams. We can conclude that it is highly likely that we have the correspondence

- X = T,
- S = H,
- A = N.

Even after this small piece of analysis we find that it is much easier to understand what the underlying plaintext should be. If we focus on the first two sentences of the ciphertext we are trying to break, and we change the letters for which we think we have found the correct mappings, then we obtain:

THE MJIWTVL JEDIVN HTW VNE VY EZJVCE'W LTJDEWT KVNKENTJTTIV NW VY HIDH TEKHNVLVDQ INGZWTJQ. KVUCZTEJW, KVUUZNIKTTIVNW TNG UIKJVELEKTJVNIKW TJE HELL JECJEWENTEG, TLVNDWIGE GIDITTL UEGIT, KVUCZTEJ DTUEW TNG ELEKTJVNIK KVUUEJKE.

Recall, this was after the four substitutions

O = E, X = T, S = H, A = N.

We now cheat and use the fact that we have retained the word sizes in the ciphertext. We see that since the letter T occurs as a single ciphertext letter we must have

T = I or T = A.

The ciphertext letter T occurs with a probability of 8.0717, which is the highest probability left, hence we are far more likely to have

$$\mathrm{T} = \mathrm{A}.$$

We have already considered the most popular trigram in the ciphertext so turning our attention to the next most popular trigram we see that it is equal to TAG which we suspect corresponds to the plaintext AN*. Therefore it is highly likely that G = D, since AND is a popular trigram in English. Our partially decrypted ciphertext is now equal to

THE MJIWTVL JEDIVN HAW VNE VY EZJVCE'W LAJDEWT KVNKENTJATIV NW VY HIDH TEKHNVLVDQ INDZWTJQ. KVUCZTEJW, KVUUZNIKATIVNW AND UIKJVELEKTJVNIKW AJE HELL JECJEWENTED, ALVNDWIDE DIDITAL UEDIA, KVUCZTEJ DAUEW AND ELEKTJVNIK KVUUEJKE.

This was after the six substitutions

$$\mathrm{O} = \mathrm{E},\ \mathrm{X} = \mathrm{T},\ \mathrm{S} = \mathrm{H},$$
$$\mathrm{A} = \mathrm{N},\ \mathrm{T} = \mathrm{A},\ \mathrm{G} = \mathrm{D}.$$

We now look at two-letter words which occur in the ciphertext:

- IX
  This corresponds to the plaintext *T. Therefore the ciphertext letter I must be one of the plaintext letters A or I, since the only common two-letter words in English ending in T are AT and IT. We already have worked out what the plaintext character A corresponds to, hence we must have I = I.
- XV
  This corresponds to the plaintext T*. Hence, we must have V = O.
- VY
  This corresponds to the plaintext O*. Hence, the ciphertext letter Y probably corresponds to one of F, N or R. We already know the ciphertext letter corresponding to N. In the ciphertext the probability of Y occurring is 1.6, but in English we expect F to occur with probability 2.2 and R to occur with probability 6.0. Hence, it is more likely that Y = F.
- IW
  This corresponds to the plaintext I*. Therefore, the plaintext character W must be one of F, N, S and T. We already have F, N, T, hence W = S.

All these deductions leave the partial ciphertext as

THE MJISTOL JEDION HAS ONE OF EZJOCE'S LAJDEST KONKENTJATIONS OF HIDH TEKHNOLODQ INDZSTJQ. KOUCZTEJS, KOUUZNIKATIONS AND UIKJOELEKTJONIKS AJE HELL JECJESENTED, ALONDSIDE DIDITAL UEDIA, KOUCZTEJ DAUES AND ELEKTJONIK KOUUEJKE.

This was after the ten substitutions

$$\mathrm{O} = \mathrm{E},\ \mathrm{X} = \mathrm{T},\ \mathrm{S} = \mathrm{H},\ \mathrm{A} = \mathrm{N},\ \mathrm{T} = \mathrm{A},$$
$$\mathrm{G} = \mathrm{D},\ \mathrm{I} = \mathrm{I},\ \mathrm{V} = \mathrm{O},\ \mathrm{Y} = \mathrm{F},\ \mathrm{W} = \mathrm{S}.$$

Even with half the ciphertext letters determined it is now quite easy to understand the underlying plaintext, taken from the website of the University of Bristol Computer Science Department circa 2001. We leave it to the reader to determine the final substitutions and recover the plaintext completely.

7.4. Vigenère Cipher

The problem with the shift cipher and the substitution cipher was that each plaintext letter always encrypted to the same ciphertext letter. Hence underlying statistics of the language could be used to break the cipher. For example it was easy to determine which ciphertext letter corresponded

to the plaintext letter E. From the early 1800s onwards, cipher designers tried to break this link between the plaintext and ciphertext.

The substitution cipher we used above was a mono-alphabetic substitution cipher, in that only one alphabet substitution was used to encrypt the whole alphabet. One way to solve our problem is to take a number of substitution alphabets and then encrypt each letter with a different alphabet. Such a system is called a polyalphabetic substitution cipher.

For example we could take

Plaintext alphabet	ABCDEFGHIJKLMNOPQRSTUVWXYZ
Ciphertext alphabet one	TMKGOYDSIPELUAVCRJWXZNHBQF
Ciphertext alphabet two	DCBAHGFEMLKJIZYXWVUTSRQPON

Then we encrypt the plaintext letters in odd-numbered positions encrypt using the first ciphertext alphabet, whilst we encrypt the plaintext letters in even-numbered positions using the second alphabet. For example, the plaintext word HELLO would encrypt to SHLJV, using the above two alphabets. Notice that the two occurrences of L in the plaintext encrypt to two different ciphertext characters. Thus we have made it harder to use the underlying statistics of the language. If one now does a naive frequency analysis one no longer obtains a common ciphertext letter corresponding to the plaintext letter E.

Essentially we are encrypting the message two letters at a time, hence we have a block cipher with block length two English characters. In real life one may wish to use around five rather than just two alphabets and the resulting key becomes very large indeed. With five alphabets the total key space is

$$(26!)^5 \approx 2^{441},$$

but the user only needs to remember the key which is a sequence of

$$26 \cdot 5 = 130$$

letters. However, just to make life hard for the attacker, the number of alphabets in use should also be hidden from his view and form part of the key. But for the average user in the early 1800s this was far too unwieldy a system, since the key was too hard to remember.

Despite its shortcomings the most famous cipher during the nineteenth century was based on precisely this principle. The *Vigenère cipher*, invented in 1533 by Giovan Battista Bellaso, was a variant on the above theme, but the key was easy to remember. When looked at in one way the Vigenère cipher is a polyalphabetic block cipher, but when looked at in another, it is a stream cipher; which is a natural generalization of the shift cipher.

The description of the Vigenère cipher as a block cipher takes the description of the polyalphabetic cipher above but restricts the possible ciphertext alphabets to one of the 26 possible cyclic shifts of the standard alphabet. Suppose five alphabets were used, this reduces the key space down to

$$26^5 \approx 2^{23}$$

and the size of the key to be remembered to a sequence of five numbers between 0 and 25.

However, the description of the Vigenère cipher as a stream cipher is much more natural. Just like the shift cipher, the Vigenère cipher again identifies letters with the numbers $0, \ldots, 25$. The secret key is a short sequence of letters (e.g. a word) which is repeated again and again to form a keystream. Encryption involves adding the plaintext letter to a key letter. Thus if the key is SESAME, encryption works as follows,

```
THISISATESTMESSAGE
SESAMESESAMESESAME
LLASUWSXWSFQWWKASI
```

Again we notice that A will encrypt to a different letter depending on where it appears in the message.

But the Vigenère cipher is still easy to break using the underlying statistics of English. Once we have found the length of the keyword, breaking the ciphertext is the same as breaking the shift cipher a number of times.

Example: As an example, suppose the ciphertext is given by

UTPDHUG NYH USVKCG MVCE FXL KQIB. WX RKU GI TZN, RLS BBHZLXMSNP KDKS; CEB IH HKEW IBA, YYM SBR PFR SBS, JV UPL O UVADGR HRRWXF. JV ZTVOOV YH ZCQU Y UKWGEB, PL UQFB P FOUKCG, TBF RQ VHCF R KPG, OU KFT ZCQU MAW QKKW ZGSY, FP PGM QKFTK UQFB DER EZRN, MCYE, MG UCTFSVA, WP KFT ZCQU MAW KQIJS. LCOV NTHDNV JPNUJVB IH GGV RWX ONKCGTHKFL XG VKD, ZJM VG CCI MVGD JPNUJ, RLS EWVKJT ASGUCS MVGD; DDK VG NYH PWUV CCHIIY RD DBQN RWTH PFRWBBI VTTK VCGNTGSF FL IAWU XJDUS, HFP VHCF, RR LAWEY QDFS RVMEES FZB CHH JRTT MVGZP UBZN FD ATIIYRTK WP KFT HIVJCI; TBF BLDPWPX RWTH ULAW TG VYCHX KQLJS US DCGCW OPPUPR, VG KFDNUJK GI JIKKC PL KGCJ IAOV KFTR GJFSAW KTZLZES WG RWXWT VWTL WP XPXGG, CJ FPOS VYC BTZCUW XG ZGJQ PMHTRAIBJG WMGFG. JZQ DPB JVYGM ZCLEWXR: CEB IAOV NYH JIKKC TGCWXF UHF JZK.

WX VCU LD YITKFTK WPKCGVCWIQT PWVY QEBFKKQ, QNH NZTTW IRFL IAS VFRPE ODJRXGSPTC EKWPTGEES, GMCG
TTVVPLTFFJ; YCW WV NYH TZYRWH LOKU MU AWO, KFPM VG BLTP VQN RD DSGG AWKWUKKPL KGCJ, XY OPP KPG ONZTT ICUJCHLSF KFT DBQNJTWUG. DYN MVCK ZT MFWCW HTWF FD JL, OPU YAE CH LQ! PGR UF, YH MWPP RXF CDJCGOSF, XMS UZGJQ JL, SXVPN HBG!

There is a way of finding the length of the keyword, which is repeated to form the keystream, called the *Kasiski test*. First we need to look for repeated sequences of characters. Recall that English has a large repetition of certain bigrams or trigrams and over a long enough string of text these are likely to match up to the same two or three letters in the key every so often. By examining the distance between two repeated sequences we can guess the length of the keyword. Each of these distances should be a multiple of the keyword, hence taking the greatest common divisor of all distances between the repeated sequences should give a good guess as to the keyword length.

Let us examine the above ciphertext and look for the bigram WX. The gaps between some of the occurrences of this bigram are 9, 21, 66 and 30, some of which may have occurred by chance, whilst some may reveal information about the length of the keyword. We now take the relevant greatest common divisors to find,

$$\gcd(30, 66) = 6, \text{ and } \gcd(3, 9) = \gcd(9, 66) = \gcd(9, 30) = \gcd(21, 66) = 3.$$

We are unlikely to have a keyword of length three so we conclude that the gaps of 9 and 21 occurred purely by chance. Hence, our best guess for the keyword is that it is of length six.

Now we take every sixth letter and look at the statistics just as we did for a shift cipher to deduce the first letter of the keyword. We can now see the advantage of using the histograms or statistical distance to break the shift cipher earlier. If we used the naive method and tried each of the 26 keys in turn we could still not detect which key is correct, since every sixth letter of an English sentence does not produce an English sentence. Thus we need to resort to using histograms or the statistical distance used earlier.

The relevant bar charts for every sixth letter starting with the first are given in Figure 7.3. We look for the possible locations of the three peaks corresponding to the plaintext letters A, E and T. We see that this sequence seems to be shifted by two positions in the blue graph compared with the red graph. Hence we can suspect that the first letter of the keyword is C, since C corresponds to a shift of two. Computing the statistical distance between the frequency of letters in English, and those of every sixth letter in the ciphertext shifted by a key k, produces the results in Table 7.4. Which indeed confirms our guess that the first letter of the keyword is C.

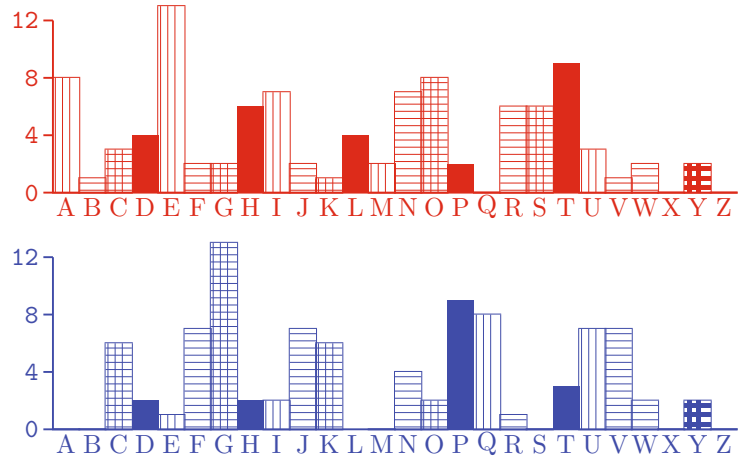

FIGURE 7.3. Comparison of plaintext and ciphertext frequencies for every sixth letter of the Vigenère example, starting with the first letter

k	$\Delta(X, Y_k)$	k	$\Delta(X, Y_k)$
0	60.7	13	40.9
1	42.8	14	47.2
2	12.4	15	50.4
3	45.0	16	46.8
4	59.5	17	41.5
5	52.6	18	45.7
6	48.0	19	55.8
7	47.8	20	54.0
8	50.5	21	46.2
9	47.1	22	47.8
10	54.7	23	48.4
11	53.5	24	43.2
12	47.8	25	53.8

TABLE 7.4. Statistical distance between X and Y_k for every sixth letter letter in the Vigenère example, starting with the first letter

We perform a similar step for every sixth letter, starting with the second one. The resulting bar graphs are given in Figure 7.4. Using the same technique we find that the blue graph appears to have been shifted along by 17 spaces, which corresponds to the second letter of the keyword being equal to R. Computing the statistical distance in Table 7.5 again confirms this guess.

Continuing in a similar way for the remaining four letters of the keyword we find the keyword is

CRYPTO.

The underlying plaintext is then found to be the following text from *A Christmas Carol* by Charles Dickens.

Scrooge was better than his word. He did it all, and infinitely more; and to Tiny Tim, who did not die, he was a second father. He became as good a friend, as good a master, and as good a man,

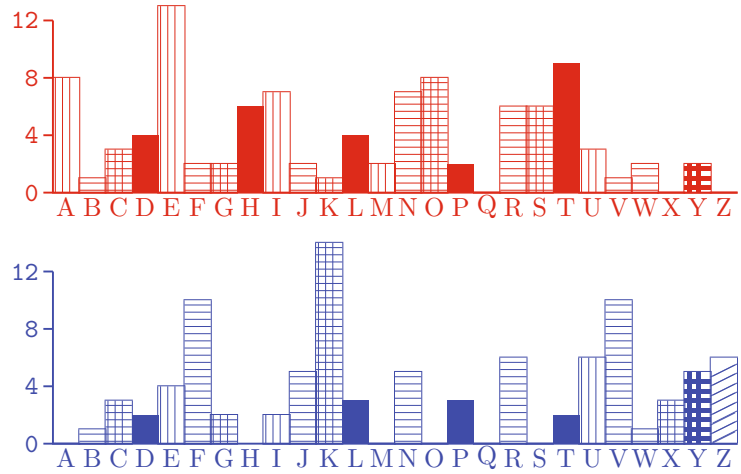

FIGURE 7.4. Comparison of plaintext and ciphertext frequencies for every sixth letter of the Vigenère example, starting with the second letter

k	$\Delta(X, Y_k)$	k	$\Delta(X, Y_k)$
0	54.6	13	39.5
1	46.2	14	55.1
2	38.3	15	59.3
3	45.1	16	53.2
4	52.6	17	17.8
5	48.3	18	47.1
6	34.6	19	53.4
7	45.3	20	53.6
8	52.4	21	44.0
9	51.3	22	49.4
10	44.6	23	48.6
11	53.2	24	48.1
12	47.1	25	52.2

TABLE 7.5. Statistical distance between X and Y_k for every sixth letter letter in the Vigenère example, starting with the second letter

as the good old city knew, or any other good old city, town, or borough, in the good old world. Some people laughed to see the alteration in him, but he let them laugh, and little heeded them; for he was wise enough to know that nothing ever happened on this globe, for good, at which some people did not have their fill of laughter in the outset; and knowing that such as these would be blind anyway, he thought it quite as well that they should wrinkle up their eyes in grins, as have the malady in less attractive forms. His own heart laughed: and that was quite enough for him.

He had no further intercourse with Spirits, but lived upon the Total Abstinence Principle, ever afterwards; and it was always said of him, that he knew how to keep Christmas well, if any man alive possessed the knowledge. May that be truly said of us, and all of us! And so, as Tiny Tim observed, God bless Us, Every One!

7.5. A Permutation Cipher

The ideas behind substitution-type ciphers forms part of the design of modern symmetric systems. For example, later we shall see that both DES and AES make use of a component called an S-Box, which is simply a substitution. The other component that is used in modern symmetric ciphers is based on permutations.

Permutation ciphers have been around for a number of centuries. Here we shall describe the simplest, which is particularly easy to break. We first fix a permutation group S_n for a small value of n, and a permutation $\sigma \in S_n$. It is the value of σ which will be the secret key. As an example suppose we take

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 1 & 3 & 5 \end{pmatrix} = (1243) \in S_5.$$

Now take some plaintext, say

Once upon a time there was a little girl called Snow White.

We break the text into chunks of five letters, and remove capitalisations,

onceu ponat imeth erewa salit tlegi rlcal ledsn owwhi te.

We first pad the message, with some random letters, so that we have a multiple of five letters in total

onceu ponat imeth erewa salit tlegi rlcal ledsn owwhi teahb.

Then we take each five-letter chunk in turn and swap the letters around according to our secret permutation σ. With our example permutation, we obtain

coenu npaot eitmh eewra lsiat etgli crall dlsdn wohwi atheb.

We then remove the spaces, so as to hide the value of n, producing the ciphertext

coenunpaoteitmheewralsiatetglicralldlsdnwohwiatheb.

However, breaking a permutation cipher is easy with a chosen plaintext attack, assuming the group of permutations used (i.e. the value of n) is reasonably small. To attack this cipher we mount a chosen plaintext attack, i.e. the attacker selects a plaintext of their choosing and asks for the encryption of it. In this specific example, they ask one of the parties to encrypt the message

abcdefghijklmnopqrstuvwxyz,

to obtain the ciphertext

cadbehfigjmknlorpsqtwuxvyz.

We can then deduce that the permutation looks something like

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & \dots \\ 2 & 4 & 1 & 3 & 5 & 7 & 9 & 6 & 8 & 10 & 12 & 14 & 11 & 13 & 15 & \dots \end{pmatrix}.$$

We see that the sequence repeats (modulo 5) after every five steps and so the value of n is probably equal to five. We can recover the key by simply taking the first five columns of the above permutation.

Chapter Summary

- Many early ciphers can be broken because they do not successfully hide the underlying statistics of the language.

- Important principles behind early ciphers are those of substitution and permutation.
- Ciphers can either work on blocks of characters via some keyed algorithm or simply consist of adding some keystream to each plaintext character.

Further Reading

The best book on the history of ciphers is that by Kahn. It is a weighty tome, so those wishing a more rapid introduction should consult the book by Singh. The book by Churchhouse also gives an overview of a number of historical ciphers.

R. Churchhouse. *Codes and Ciphers. Julius Caesar, the Enigma and the Internet.* Cambridge University Press, 2001.

D. Kahn. *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet.* Scribner, 1996.

S. Singh. *The Codebook: The Evolution of Secrecy from Mary, Queen of Scots to Quantum Cryptography.* Doubleday, 2000.