Basic methods of cryptography

J.C.A. van der Lubbe

Basic Methods of Cryptography

Basic Methods of Cryptography

Jan C.A. VAN DER LUBBE
Associate Professor
Information Theory Group
Department of Electrical Engineering
Delft University of Technology

Translated by Steve Gee

VSSD

Published by:
VSSD
Leeghwaterstraat 42, 2628 CA Delft, The Netherlands
tel. +31 15 27 82124, telefax +31 15 27 87585, e-mail: hlf@vssd.nl
internet: http://www.vssd.nl/hlf

ISBN Ebook 978-90-6562-240-2
NUR 983

Keywords: cryptography.

Contents

Preface

As a result of current technological developments, the computer can now be found in all layers of our society and the possibilities for communication have grown immensely. At present, information is being communicated and processed automatically on a large scale. There are numerous examples: medical or fiscal computer files, automatic banking, video-phone, pay-tv, facsimiles, tele-shopping, global computer networks, etc. All these examples increasingly require measures for secure storage and transportation of the information. There are many reasons for this growing need. Protection of the information may be necessary to guard economic interests, to prevent fraud, to guarantee the privacy of the citizen, etc.

Cryptology is the science which is concerned with methods of providing secure storage and transportation of information in its widest sense.

In this book we will cover the fundamentals of secure storage and transportation of information, as they are currently being developed and used. The objective of this book is to allow the reader to become acquainted which the various possibilities of cryptology, and also with the impossibilities and necessary conditions involved in the use of cryptology.

This book is written for anyone who is in some way or other involved in protecting information processing and communication: engineers, system designers, application programmers, information analysts, security officers, EDP-auditors, etc.

This book has resulted from lectures given by the author to students of the Faculties of Electrical Engineering, Technical Mathematics and Informatics, Systems Engineering and Policy Analysis and Applied Physics of the Delft University of Technology and from the course in cryptology provided by TopTech Studies, which is responsible for the post-doctoral courses of the Delft University of Technology, and of which the author is the director.

The author wishes to thank dr.ir. J.H. Weber for his assistance during the lectures in cryptology at the Delft University of Technology and also all his TopTech cryptology course colleagues (in particular ir. R.E. Goudriaan of

the International Nederlanden Bank), as they have taught the author a great deal about the practical aspects of the use of cryptology.

Delft J.C.A. van der Lubbe
May 1997

Abstract

Chapter 1 focuses mainly on the role of cryptology within the total field of security. We will examine the various objectives of security and an initial summary of the available cryptographic methods is provided.

In Chapter 2 we will deal with the more classical forms of cipher systems, such as the transposition and the substitution ciphers. In addition, we will also take a look at the methods employed by cryptanalysts ('hackers') for cracking existing security measures.

In many cases, the strength of a cryptographic algorithm depends almost entirely on the obtainable level of security. However, since the term 'security' is itself far from clear, in Chapter 3 we will first deal with the concept of security, using terms from the field of information theory, and we will also pay attention to how security can be achieved.

One of the currently most popular cryptographic algorithms, which is based on enciphering with secret keys, is the DES algorithm. The principles of this algorithm are explained in Chapter 4.

Chapter 5 focuses on the use of shift registers for providing pseudorandom sequences, which can be used for generating keys as well as enciphering bit streams. In this chapter we will also study the term 'randomness'.

Chapter 6 is concerned with so-called public key systems; cryptographic algorithms with a secret and a public key. The RSA algorithm is an important example of such a system.

Chapter 7 deals with other types of cryptographic protection concerned with authentication and integrity. These items involve techniques which enable us to determine whether a transmitted message is intact and whether a message purported to be transmitted by some entity was really transmitted by that entity. Amongst other things we will examine digital signatures and zero knowledge techniques for identification.

In general ,we can say that no matter how good our cryptographic algorithms may be, the overall security always relies on the extent to which the secret keys remain secret. Chapter 8 therefore looks at the problem of key management, which is concerned with securely generating, distributing, etc., keys, as well as the specific aspects of the security of networks.

Finally, there are two appendices. Appendix A explains Shannon's measure of information and is meant for those who are not yet acquainted with the fundamentals of information theory. Appendix B covers several specific techniques for encrypting imagery.

Notation

A_d	Number of residues with d elements
C	Ciphertext
$C(\tau)$	Autocorrelation
CI	Coincidence index
CI′	Pure estimator of CI
d	Part of the secret key of RSA; number of elements of the residue
δ	Hamming distance
$DK(.)$	Decipherment with a symmetric algorithm using a key K
$dS_X(.)$	Decipherment with an asymmetric algorithm using a secret key S_X
D_L	Redundancy in a text of length L
e	Part of the public key of RSA
E	Expansion operation of DES
$E(.)$	Expectation
$EK(.)$	Encipherment with a symmetric algorithm using a key K
$eP_X(.)$	Encipherment with an asymmetric algorithm using a public key P_X
ε	Number of elements of an alphabet
ϕ	Euler totient function
$f(.)$	Characteristic polynomial
$f(.,.,.)$	Feedback function of a shift register
$F(.,.)$	Cipher function DES
$G(.)$	Generating function
χ	Chi-test
$h(.)$	Hash-code
H	Hypothesis
$H(.)$	Marginal information measure
$H(K/C)$	Key equivocation
$H(M/C)$	Message equivocation
$H(K/M,C)$	Key appearance equivocation
I	Identification sequence
IP	Initial permutation

IV	Initial vector
J	Jacobi symbol
K	Secret key of a symmetric algorithm
kDES	First k bits of the result of an encipherment using DES
K_i	Subkey of DES
K_{ij}	Key for Diffie–Hellmann protocol
KS	Session key
l	Length of a run
L	Length of a message
LC	Linear complexity profile
m	Number of sections of a shift register
$mgK(.)$	Result of a MAC using key K
M	Plaintext, original message
MK	Master key
mod	Modulo addition
N	Length of (pseudo)random sequence
$oK(.)$	Result of a one-way function using key K
p	Period of a shift register sequence; prime number
P_X	Pubic key of X for an asymmetric algorithm
Pe	Error probability
PeD	Error probability distance
q	Prime number
r	Total number of runs in a binary sequence
R	Random sequence
S	Knapsack sum; security event
S_X	Secret key of X for an asymmetric algorithm
s_i	Element of a shift register sequence; secret number for zero-knowledge techniques
T	Period
T_K	Encipherment transformation
TK	Terminal key
UD	Unicity distance
var	Variance

1

Introduction to cryptology

1.1 Cryptography and cryptanalysis

The title of this book contains the word *cryptography*. Cryptography is an area within the field of *cryptology*. The name cryptology is a combination of the Greek *cruptos* (= hidden) and *logos* (= study, science). Therefore, the word cryptology literally implies the science of concealing. It comprises the development of methods for *encrypting* messages and signals, as well as methods for *decrypting* messages and signals. Thus, cryptology can be divided into two areas: *cryptography* and *cryptanalysis*.

Cryptography can be defined more specifically as the area within cryptology which is concerned with techniques based on a secret key for concealing or enciphering data. Only someone who has access to the key is capable of deciphering the encrypted information. In principle this is impossible for anyone else to do.

Cryptanalysis is the area within cryptology which is concerned with techniques for deciphering encrypted data without prior knowledge of which key has been used. This is more commonly known as 'hacking'.

It is evident that cryptography and cryptanalysis are very closely related. One is only able to design good (sturdy) cryptographic algorithms when sufficient knowledge of the methods and tools of the cryptanalysists is available. The person responsible for the implementation of this type of security measure must therefore obtain this knowledge and be aware of the methods of a potential intruder. Obviously, successful cryptanalysis requires at least a fundamental insight into cryptographic algorithms and methods.

This book will focus mainly on cryptography.

A first impression of what a cryptographic algorithm does is given by considering the following situation, which also offers the opportunity of

introducing some notation. Suppose *A* (the transmitter) wishes to send an enciphered message, i.e. secret code, to *B* (the receiver). Often, the original text or *plaintext* is simply denoted by the letter *M* of message and the encrypted message, referred to as the *ciphertext*, by the letter *C*. A possible method is for *A* to use a secret key *K* for *encrypting* the message *M* to the ciphertext *C*, which can then be transmitted and decrypted by *B*, assuming that B also possesses the secret key *K*. This is illustrated in Figure 1.1. *EK* represents the encryption of the message with the aid of *K*; the *decryption* of the message is represented by *DK*. Hereafter we will use the following notation:

$C = EK(M)$
(i.e. original text *M* is encrypted to ciphertext *C* with the secret key *K*);

$M = DK(C)$
(i.e. ciphertext *C* is decrypted to the original text *M* with the secret key *K*).

An example of what occurs at the transmitter and receiver is given by Figure 1.2. It is up to the reader to find the correct key. This should prove not too great a problem for those who can use a word-processor .

Figure 1.1. Cipher system.

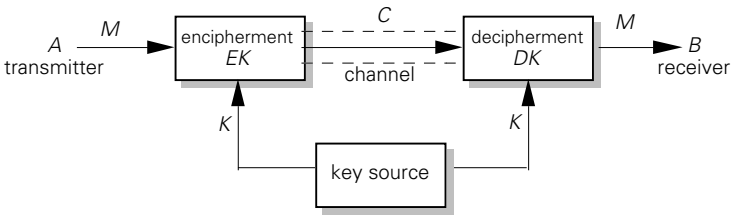
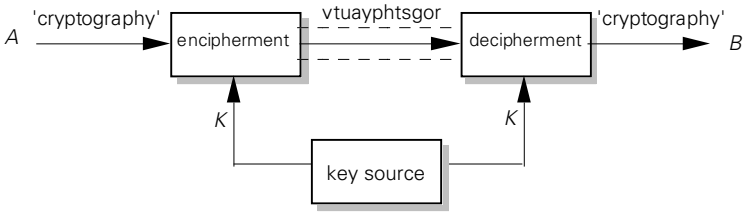


Figure 1.2. Example of encipherment and decipherment.



1.2 Aspects of security

Before proceeding with a description of the methods used in cryptography, we will first pay attention to the position and use of cryptography within the total concept of security. Here, three aspects play an important role, as illustrated in Figure 1.3.

The first question to be considered in practice is what purpose the security measures must serve. This unavoidably leads to some means of adequate threat analysis, which should provide a clear picture of what must be protected against whom or what.

Subsequently, the available means of security must be considered. This involves answering questions such as: How? With which security measures?

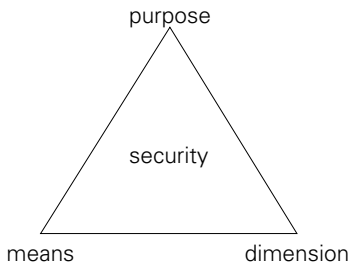
The third important aspect shown in Figure 1.3 has been labelled dimension. By this, we mean whether the security measures are designed for the prevention of or the correction of the damage caused by a security breach. We will return to this later.

The division of Figure 1.3 into three aspects can be extended to lower levels. If, for instance, we consider the purpose of the security measures, we can draw up a list of numerous possibilities against which security measures must be taken. Several examples are given below:

(*a*) reading or tapping data;
(*b*) manipulating and modifying data;
(*c*) illegal use of (computer) networks;
(*d*) corrosion of data files;
(*e*) distortion of data transmission;
(*f*) disturbance of the operation of equipment or systems.

The main issue of item (*a*) is *secrecy* and *confidentiality*. Confidentiality has always played an important role in diplomatic and military matters. Often information must be stored or transferred from one place to another,

Figure 1.3. Aspects of security.



without being exposed to an opponent or enemy. Another example of how encipherment can be used to guarantee secrecy is in the communication between police patrols and the control room. Conversations are scrambled so that it is extremely difficult for outsiders to extract any relevant information from the transmitted messages. It is even conceivable that the simple fact of whether or not a message has been transmitted must also remain confidential. In this case, dummy messages can be transmitted in order to camouflage the real message.

So-called *key management* is also closely related to confidentiality. This area deals with generating, distributing and storing keys. It is obvious that no matter how strong a cryptographic algorithm may be, the final effectiveness of the algorithm depends largely on the obtainable secrecy level of the key. Once an intruder has managed to acquire a copy of the key, in principle, he is capable of decrypting the enciphered message. Therefore, key management must be considered as an essential element of the entire security plan.

Items (*b*)–(*d*) are primarily concerned with *reliability*. Take electronic banking, for instance. A bank requires some means of guaranteeing the authenticity of a financial transaction to prevent the wrongful withdrawl of large sums of money. Often the expression *integrity* is used as a measure of the genuineness of the data. Also, computer networks must be protected against intruders and unauthorised users. When one receives a fax message from a person *A*, one likes to be sure the fax was indeed written by *A* and that *A* is truly *A* and not an impostor. This is, in fact, an example of *authentication*, i.e. giving legal validity to the identity of the transmitter and determining the origin of the data. The above examples cover all aspects of security which are concerned primarily with reliability.

In items (*e*) and (*f*), a different aspect of the security of the information, its coninuity, is considered. Here, the data must be protected against deliberate disruption during its transmission and storage.

We can therefore distinguish between three different aims of security (see Figure 1.4). This also applies to the type of security used for a specific purpose (see Figure 1.5).

The phrases introduced here are self-evident. We can speak of *physical security* when a system is protected against the physical entry of an intruder, for instance by using metal containers, certain plastics or temperature or vibration sensors. However, in this book we will deal with *hardware* and *software based security*; i.e. cryptographic algorithms and methods.

No matter how high the standard of physical and hardware and software security of a system, the safety of the information cannot be guaranteed

without sufficient organisational measures. So-called *organisational security* ensures that conditions are created which allow the physical and hardware and software security measures to be fully effective. If, in practice, certain security measures prove complicated or confusing to the user, naturally this will introduce the risks of neglect and carelessness. Therefore, one must always bear in mind that human beings will always be present somewhere in the chain, regardless of the level of automation of a security system.

Figure 1.4. Purposes of security.

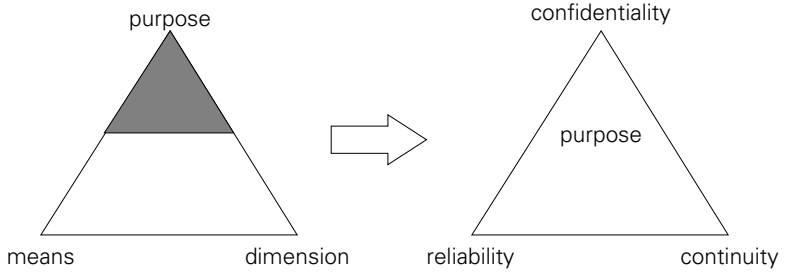


Figure 1.5. Types of security.

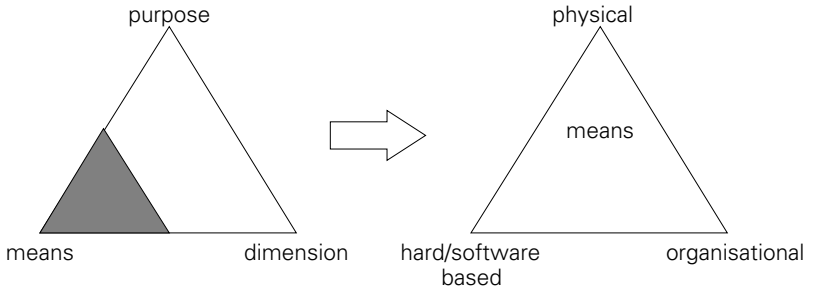
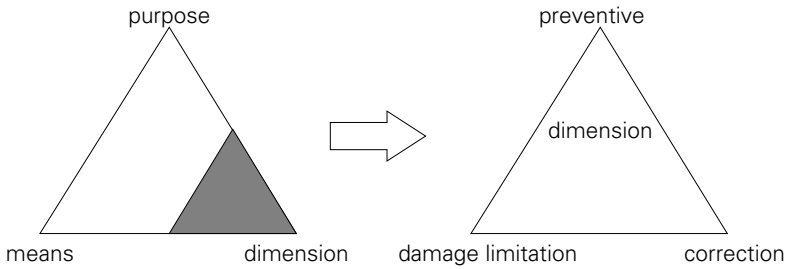


Figure 1.6. Dimensions of security.



The last aspect we will consider here is the dimension of the security. Again, we can divide this into three (see Figure 1.6). In the first instance, cryptography is applied as a *preventive* measure. We are attempting to minimise the chance of anything happening to the transmitted information. This can be accomplished by using strong cryptographic algorithms and protocols and installing adequate physical and organisational measures. However, by definition, *absolute security measures* do not exist. In practice, the chance of a mishap occurring can be minimised, but can never be reduced to zero. Therefore, another aspect of the dimension of security is *damage limitation*. If the chance of a mishap occurring is not zero, we can at least ensure that the resulting damage remains as limited as possible. For example, if someone manages to break into a computer file, it is possible to ensure that he gains access to only a small part of the entire file. Or if a key falls into the wrong hands, it must never be possible to decrypt all messages with that key, but only a fraction of them, etc. The final aspect of the dimension of security is *correction*. If something happens to the encrypted information, it must be possible to correct this quickly. For instance, there must always be some means available of rendering a key useless, just in case the key falls into the hands of an unauthorised person. Also, when vital information is damaged, it must be possible to reconstruct this information easily.

Obviously, in any practical situation a trade-off between the listed aspects of security must be made.

In addition, the economic facet of the security measures has to be taken into consideration. This is the relation between the desired level of security, the value of that which is being secured, or of that against which is being secured, and the investments necessary to obtain the desired level of security or to gain access to the information.

In the preceding text several applications of cryptography were mentioned. These can be divided into two groups, i.e. applications related to the storage of information and those related to the transportation of information.

Nowadays information is mostly stored in computer systems, on either disc or magnetic tape. The method of storage is often public knowledge and only the key is kept secret. As this type of data is usually stored for a considerable length of time, a cryptanalytic attack is attractive; the cryptanalyst can take his time finding the key. Consequently, this situation requires a relatively high level of security.

On the other hand, when the data are transmitted (e.g. TV, satellite), they are available to the cryptanalyst for only a short period of time and, in addition, the key can easily be changed regularly. Obviously, the

cryptanalyst can record a transmitted message, but this does not necessarily help him to decrypt other transmitted messages if the key is frequently altered.

Moreover, communicated messages are often meaningful for only a short period of time, as the information ages or becomes obsolete (e.g. news, weather information, etc.). For this reason, such data communication usually requires a lower level of security and therefore also lower investment.

The costs of the security measures must also be viewed in a different light. Consider, for instance, cable television. The cable company will obviously profit from good security measures which prevent as much illegal viewing as possible. However, the price of decrypting the TV signal must still remain reasonable, for both the cable company and the consumer. Clearly, most consumers, who can be regarded as honest subscribers, will only pay a limited contribution towards security measures they did not ask for themselves. Furthermore, the level of security is sufficient if a potential viewer must make a larger investment than the ordinary subscription to be able to view the programmes illegally.

1.3 Cryptanalytic attacks

Let us consider a cryptographic algorithm which requires the use of a secret key. With regard to Figure 1.1, we can speak of a cryptanalytic attack when an intruder tries to discover the contents of a message or the secret key by other means than straightforward random attempts. Clearly, the intruder will find it more interesting if he can discover the key itself, rather than occasionally disclosing the plaintext, as then, hopefully, other ciphertexts can be decrypted as well. One method of finding the correct key is by simply trying all possibilities, until the correct key is found. This is called an *exhaustive key search*. However, this is not really a cryptanalytic attack in the true sense, as generally we expect a cryptanalyst to behave more 'intelligently'.

As far as real cryptanalytic attacks are concerned, we can distinguish between three types, depending on the level of information available to the cryptanalyst, see Figure 1.7. These three types are:

(*a*) an attack based solely on the ciphertext: (*ciphertext-only-attack*);
(*b*) an attack based on a given plaintext and the corresponding ciphertext: (*known-plaintext-attack*);
(*c*) an attack based on a chosen plaintext and corresponding ciphertext: (*chosen-plaintext-attack*).

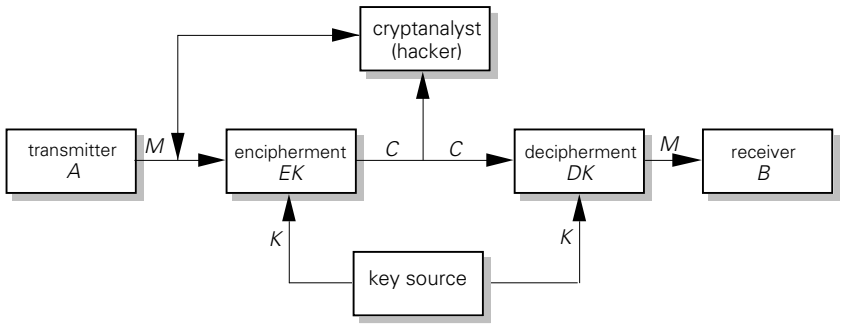
When the attack is based on the ciphertext alone, the cryptanalyst only has access to the encrypted signal. With the use of the necessary statistics and by analysing apparent patterns in the signal the cryptanalyst must attempt to decipher the hidden message (plaintext) and more importantly, the key. This is often the case when analysing enciphered speech, tapping car telephones, etc. It is clear that this situation is the least favourable for a cryptanalyst.

A far more favourable situation can be found when the cryptanalyst can obtain information on the corresponding plaintext, in addtion to what he already knows about the ciphertext. If a relation can be found between a certain part of the ciphertext and the plaintext, then this knowledge may be used to decrypt other sections of the ciphertext, or even to find the key. In order to obtain information on both the ciphertext and the plaintext the cryptanalyst must gain access to (a part of) the cipher system or its users. Every financial transaction, for instance, contains information on the payer and the payee. If a cryptanalyst has inside information on how the information on the parties involved is enciphered in the message, he can attempt to decipher the remaining part of the ciphertext.

The most favourable situation is one in which the cryptanalyst can select a certain plaintext and generate the corresponding ciphertext. By choosing the correct plaintext and corresponding ciphertext, he can decipher parts of the text which are still encrypted, or even find the key. For example, a word processor which stores files in an encrypted form is an easy target for a chosen-plaintext-attack.

Ideally, a cryptographic system must be able to withstand all three types of attack, although in practice this is often difficult to realise. A system which appears capable of resisting ciphertext-only-attacks, may prove sensitive to know chosen-plaintext-attacks. However, a system which can withstand a chosen-plaintext-attack is usually regarded as of a higher standard than a system which can only stand an attack based on the ciphertext alone.

Figure 1.7. Cryptanalytic attack.



The above text focuses mainly on cryptanalytic attacks which form a breach of confidentiality. Attacks which affect the reliability (integrity and authenticity) will be described in Chapter 7.

2

Classical cipher systems

2.1 Introduction

In this chapter we will examine several classical cipher systems, which are often based on methods of encipherment with a very long history. Nowadays, though, most of these methods have become less popular; they were frequently used during the second world war, but since computers have become available to cryptanalysts, their applicability has diminished. However, this does not imply that a description of classical cipher systems is merely of historical interest. On the contrary, although the classical cipher systems are rarely used on their own, they are often incorporated in more modern crypto-systems, either in a cascaded form, or in combination with other methods.

We can distinguish two types of classical cipher system:
– transposition systems;
– substitution systems.

A transposition cipher is based on changing the sequence of the characters of the plaintext; the characters themselves remain unchanged. A substitution cipher does not alter the order of the characters of the plaintext, but replaces the original characters with others.

We will examine these two cipher systems in more detail in the following sections.

2.2 Transposition ciphers

In the previous section we mentioned that a transposition cipher only alters the order of the characters of the plaintext. This is performed in blocks of characters and is demonstrated by the following example.

plaintext: THE MEETING HAS BEEN POSTPONED UNTIL NEXT MONTH
divided into
blocks: THEME ETING HASBE ENPOS TPONE DUNTI LNEXT MONTH
ciphertext: MEETH NGIET BESHA OSPEN NEOTP TINDU XTELN THNMO

The plaintext of this example is divided into blocks of five letters. We can say that here, the period is equal to 5. Within each block the order of the letters is changed according to the key 4 5 3 1 2, so, with regard to the original block, the 4th and 5th characters have been exchanged with the 1st and 2nd and the 3rd remains in the same position.

Often a so-called key-word is used so that the key is easily remembered. Here, a suitable key-word is, for instance, 'stock'. The key is given by the alphabetical order of the letters of the key-word.

The encryption of a message with a transposition cipher can, in fact, be regarded as the transposition of the columns of a matrix, as is demonstrated by the following example. The blocks of letters are now placed below each other, rather than next to each other.

plaintext: THE MEETING HAS BEEN POSTPONED UNTIL NEXT MONTH
key-word: STOCK, KEY 45312

THEME	MEETH
ETING	NGIET
HASBE	BESHA
ENPOS	OSPEN
TPONE	NEOTP
DUNTI	TINDU
LNEXT	XTELN
MONTH	THNMO

ciphertext: MEETH NGIET BESHA OSPEN NEOTP TINDU XTELN THNMO

The ciphertext is obtained by changing the columns according to the alphabetical order of the letters of the key-word.

Consider a message M with a total length of $L = nT$ letters, in which T is the period and n a positive integer. If the message is divided into blocks of length T, it can be represented by:

$$M^{nT} = [a_1,a_2,\ldots,a_T][a_{T+1},\ldots,a_{2T}]\ldots[a_{(n-1)T+1},\ldots,a_{nT}],$$

or, by introducing matrix notation:

$$M^{nT} = \begin{bmatrix} a_1 & a_2 & \ldots & a_T \\ a_{T+1} & a_{T+2} & \ldots & a_{2T} \\ \vdots & & & \vdots \\ a_{(n-1)T+1} & \ldots & \ldots & a_{nT} \end{bmatrix}$$

or even as:

$$\underline{M}^T = [\underline{a}_1,\underline{a}_2,\ldots,\underline{a}_T],$$

in which for every $i = 1,\ldots,T$ the column $\underline{a}_i$ is given by:

$$\underline{a}_i = (a_i,a_{T+i},\ldots,a_{(n-1)T+i})^{\mathrm{t}}.$$

The corresponding ciphertext C can be found by merely interchanging the columns of the matrix, or, in formal notation:

$$\underline{C}^T = [\underline{a}_{k(1)}\underline{a}_{k(2)},\ldots\underline{a}_{k(T)}]$$

in which k is a permutation of $(1,\ldots,T)$.

We can now calculate the total number of possible keys, which is equal to $T!$ or, more precisely, $T! - 1$, because one key will always produce a ciphertext identical to the plaintext.

Obviously, the period T must be large. Our example has a period of 5, which means that there are $5! - 1 = 119$ possibilities for the key. This value is rather small and a cryptanalyst who knows the period T will find no difficulty in quickly decrypting the ciphertext.

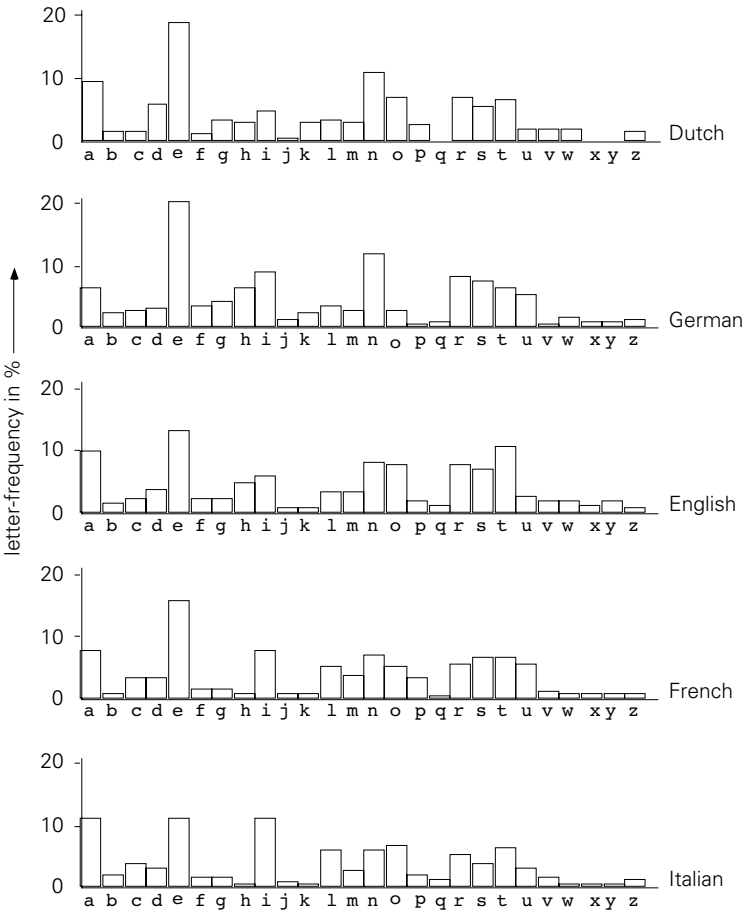
On the other hand, if a large period is used to increase the number of different keys and make it more difficult for a cryptanalyst, the rightful receiver is forced to remember a very long key, which introduces new risks.

Cryptanalysts are faced with two problems. First, they must find the period T, which involves trying all possible combinations of n and T, which satisfy $L = nT$, with L equal to the length of the message. If dummy letters have been added to the original text, then more combinations of n and T must be considered. The second problem is to find the key in a structured manner, preferably without having to try all possible permutations first.

To solve these two problems, the cryptanalyst can benefit from the linguistic characteristics of the language in which the plaintext is written. Certain letters are used more frequently than others, as can be seen in Figure

2.1, in which the relative letter-frequencies of several languages are plotted. This kind of plot can also be made for the frequency of combinations of two letters. Regarding a text in this manner, we find that vowels tend to be surrounded by consonants and vice versa. This means that vowels are generally distributed evenly throughout a text. Therefore, when the ciphertext is put into a matrix form, a period T should be chosen which produces the most even distribution of the vowels of the text across the columns of the matrix. The text can now be deciphered more easily if the columns are exchanged in such a way that the most frequently occurring letter combinations appear first.

Figure 2.1. Relative frequency of occurrence of letters for several languages.



2.3 Substitution ciphers

A substitution cipher is based on replacing the characters of the plaintext with other characters. Assuming the plaintext is based on an alphabet of 26 letters, a substitution cipher can be described by the following:

alphabet of the plaintext: $A = [a_1,\ldots,a_{26}]$
alphabet of the ciphertext: $B = [b_1,\ldots,b_{26}]$
plaintext: $a_3,\ a_{23},\ a_9,\ a_{17},\ a_4$
ciphertext: $b_3,\ b_{23},\ b_9,\ b_{17},\ b_4$

The most straightforward substitution cipher is the Caesar substitution, named after the Roman Emperor Julius Caesar (100–44 BC). The substitution alphabet is obtained by simply shifting the original alphabet a given number of characters, with respect to the original alphabet. In the following example the alphabet is shifted three places.

original alphabet *A*:

```
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
```

substitution alphabet *B*:

```
D E F G H I J K L M N O P Q R S T U V W X Y Z A B C
```

plaintext: `PLEASE CONFIRM RECEIPT`
ciphertext: `SOHDVE FRQILUP UHFHLSW`

If the characters of the plaintext alphabet and ciphertext alphabet are numbered and denoted by i and j respectively, then in the above example, for all $i = 1,\ldots,26$: $j = i + 3 \pmod{26}$. Mod 26 implies that the left part and right part of the equation may only differ by a multiple of 26. In a more general form, $j = i + t \pmod{26}$, in which t represents the number of characters the two alphabets are shifted.

An important characteristic of the Caesar substitution is the fact that the order of the characters of the substitution alphabet remains unchanged. The total number of keys is no more than 26, so this cipher can very easily be cracked; once a single letter of the ciphertext can be related to a letter of the plaintext, the system breaks down. If the message is sufficiently large, it is all the more straightforward to find such a relation; simply note the most frequently occurring letter and the chances are that this is equal to the letter e of the original plaintext, assuming that this was written in English.

If the letters of the cipher alphabet are placed in random order, instead of simply being shifted with respect to the original alphabet, the number of keys increases to 26!. This makes decipherment considerably more difficult than in the case of Caesar substitution. For example:

original alphabet *A*:

```
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
```

substitution alphabet *B*:

```
E S T V F U Z G Y X B H K W C I R J A L M P D Q O N
```

plaintext: `PLEASE CONFIRM RECEIPT`
ciphertext: `IHFEAF TCWUYJK JFTFYIL`

It is even possible for the substitution alphabet to consist of entirely different symbols from the original alphabet, as for instance in Figure 2.2.

Clearly, the key of this example is far too complicated to be remembered easily and therefore key-sentences are often used. Each time a new letter appears in the key-sentence, it is added to the substitution alphabet . When all the different letters of the sentence have been recorded, the remaining letters of the alphabet are placed behind this list in their usual order. Consider the following example:

key-sentence: `THE MESSAGE WAS TRANSMITTED AN HOUR AGO`

original alphabet *A*:

```
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
```

substitution alphabet *B*:

```
T H E M S A G W R N I D O U B C F J K L P Q V X Y Z
```

plaintext: `PLEASE CONFIRM RECEIPT`
ciphertext: `CDSTKS EBUARJO JSESRCL`

Despite the 26! possibilities, finding the correct key will still be relatively easy, since languages generally contain a high level of redundancy. Also, the commonest letters, such as e, t, n, r, o, a, etc., can always be found with comparatively little effort by considering the letter-frequency distribution.

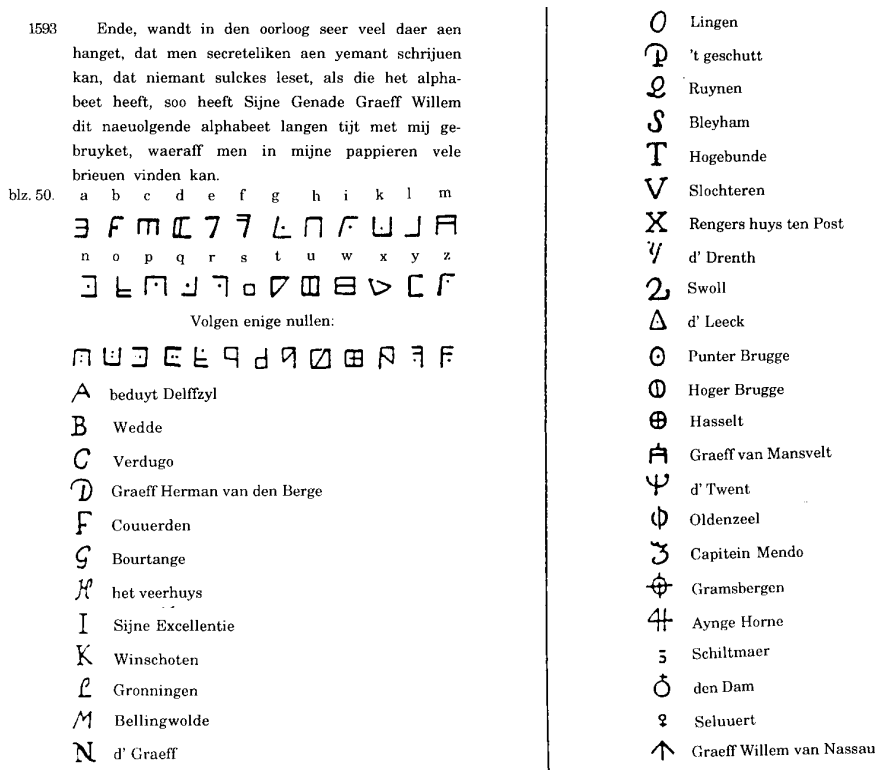
We can conclude that, in general, substitution methods as described above are not very resistant to attacks, since the characteristics of the language can still be extracted from the ciphertext. This can be avoided by applying more

than one substitution cipher. This procedure is referred to as a *polyalphabetical substitution*, as opposed to the *monoalphabetical substitution* of the examples above. A well-known example of a polyalphabetical substitution is the Vigenère system, which was devised in France in 1568 by Blaise de Vigenère. This system uses a different Caesar substitution for each letter. For example, the first letter is shifted by 10 positions, the second by 17, etc.

Encryption based on the Vigenère system is often performed with the aid of a so-called *Vigenère table* (see Table 2.1) and a key-word. The top row of the Vigenère table consists of the letters of the plaintext alphabet and the first column contains the letters of the key-word. A text can be enciphered using the following procedure.

The key-word is repeated below the plaintext as in the example below. A letter of the ciphertext is equal to the letter located at the intersection of the column designated by the letter of the plaintext and the row designated by the letter of the key.

Figure 2.2. Excerpt of a cipher as used for the communication between the Dutch Viceroy Willem Lodewijk and his commander Fredrich von Vernou.



plaintext:	PLEASE CONFIRM RECEIPT
key:	CRYPTO CRYPTOC RYPTOCR
ciphertext:	RCCPLS EFLUBFO HCRXWRK

Clearly, a given letter of the plaintext is represented by different letters in the ciphertext, depending on the letters of the key-word, thus concealing linguistic characteristics more effectively than any of the previous methods.

The number of monoalphabetical substitutions on which the Vigenère system is based is equal to the length of the key word. Here, the number of monoalphabetical substitutions is five and consequently, five rows of the table have been used. Obviously, if a cryptanalyst can discover the length of the key-word, this knowledge will be of great help in finding a solution to the cryptogram.

Someone using this system will generally attempt to employ as many different rows of the table as possible. One way of ensuring this is to use the plaintext itself, in addition to the key-word. This is demonstrated by the following example. The key is constructed by placing the letters of the plaintext itself after the key-word CRYPTO. The major problem of transposition and substitution ciphers is successfully to hide the statistical

Table 2.1. The Vigenère table.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

parameters of the text. One solution is to ensure that the characters of the ciphertext have a uniform distribution. The characters of the plaintext can be coded in a certain manner, e.g. according to the Huffmann coding, to obtain a uniform distribution. This distribution is preserved when a transposition or substitution cipher is applied to the plaintext.

2.4 The Hagelin machine

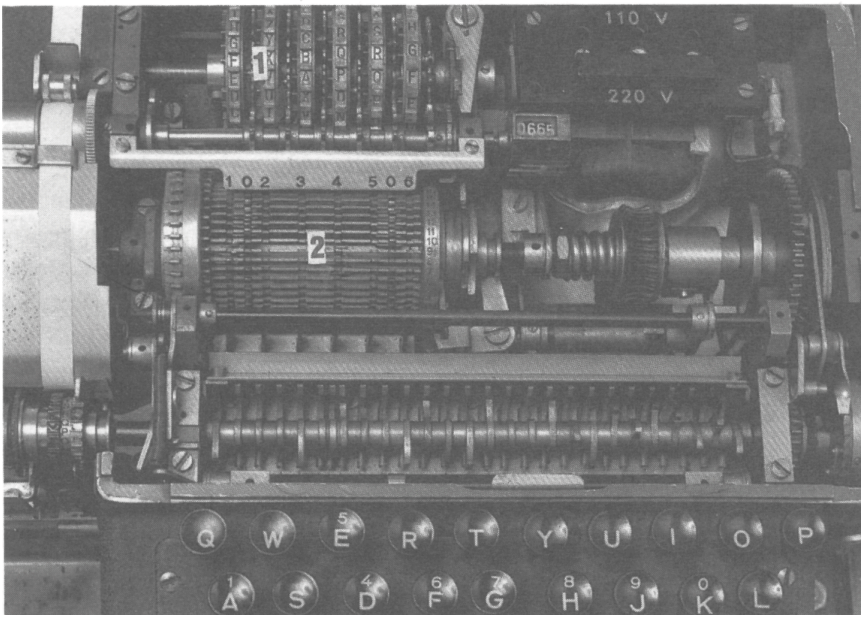
In the 1930s, the Swede Boris Hagelin invented a machine which is capable of generating enciphered text based on polyalphabetical substitutions (see Figure 2.3). This machine is called the Hagelin cryptograph (or, M-209 machine) and was used by the American army until approximately 1950.

The cryptograph employs a polyalphabetical substitution method which relies on the so-called *square of Beaufort* (see Table 2.2), which is comparable to the Vigenère table. The substitution is given by:

$$j = t + 1 - i \pmod{26},$$

in which i is the alphabetical position of the plaintext character, t the row of the Beaufort table and j the alphabetical position of the encrypted symbol.

Figure 2.3. The Hagelin machine. The lower photograph clearly shows the coding wheels (1) and drum (2) (photographs by Facilitair Bedrijf TU Delft, Photographic Service).



Example
Assume the plaintext is equal to the sequence of letters 'SECRET' and that the following values of t are used: 2, 15, 8, 7, 3 and 1. We then find:

$$\begin{aligned}
&\text{S:} \quad i = 19, \quad t = 2 \quad \Rightarrow j = 10 \Rightarrow \text{J} \\
&\text{E:} \quad i = 5, \quad t = 15 \quad \Rightarrow j = 11 \Rightarrow \text{K} \\
&\text{C:} \quad i = 3, \quad t = 8 \quad \Rightarrow j = 6 \Rightarrow \text{F} \\
&\text{R:} \quad i = 18, \quad t = 7 \quad \Rightarrow j = 16 \Rightarrow \text{P} \\
&\text{E:} \quad i = 5, \quad t = 3 \quad \Rightarrow j = 25 \Rightarrow \text{Y} \\
&\text{T:} \quad i = 20, \quad t = 1 \quad \Rightarrow j = 8 \Rightarrow \text{H}
\end{aligned}$$

The ciphertext is 'JKFPYH', which corresponds to the text obtained with the aid of Table 2.2. △

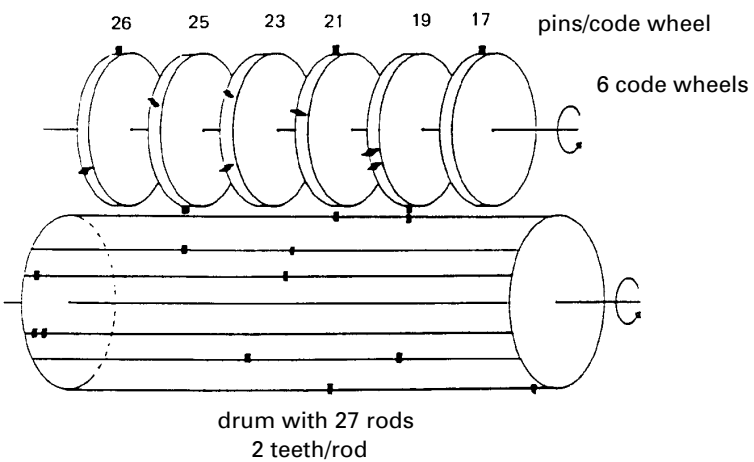
An ingenious mechanism in the Hagelin machine determines the value of t and thus selects an alphabet of the Beaufort square for the corresponding encryption. The Hagelin machine contains a drum which is constructed from 27 rods. Two movable teeth are mounted on each rod. The teeth can occupy eight possible positions, two of which inactivate the teeth. The remaining six positions are located opposite the six code wheels. The code wheels are equipped with respectively 26, 25, 23, 21, 19 and 17 pins which are set in either an active or a passive position.

The encryption of a letter is performed during a single revolution of the drum. As the teeth of the drum pass the active pins of the code wheels, the number of passing teeth is registered at the contact points. The resulting value is, in fact, always equal to t and determines which row of the Beaufort square will be used. Before the following letter is encrypted, the code wheels are rotated over a single position, thus moving the active pins, so subsequently different teeth of the drum are counted. The following example demonstrates this process.

Table 2.2. Square of Beaufort.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
0,26	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A
1,27	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B
2	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C
3	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D
4	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E
5	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F
6	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G
7	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H
8	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I
9	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J
10	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K
11	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L
12	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M
13	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O	N
14	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P	O
15	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q	P
16	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R	Q
17	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S	R
18	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T	S
19	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U	T
20	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V	U
21	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W	V
22	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X	W
23	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y	X
24	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z	Y
25	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	Z

Figure 2.4. Code wheels and drum of the Hagelin machine.



Example

Table 2.3(a) shows the position of the teeth in relation to the six code wheels for each of the 27 rods: a 1 indicates the presence of a tooth at that location, 0 means no tooth. The positions of the active (1) and passive (0) pins on each of the code wheels are given in Table 2.3(b).

Assuming the contact points are set to position 1 on the code wheels, we can see from Table 2.3(b) that only code wheels 3 and 5 have active pins at this position. As the drum and rods are rotated, the number of teeth passing code wheels 3 and 5 is counted. In this example this is equal to 1 + 9 = 10 (consider the number of ones in the third and fifth rows of Table 2.3(a)). This number determines which row of the Beaufort table ($t = 10$) is used for encryption. For the next letter, the code wheels are rotated one position to position 2, resulting in active pins at positions 1, 4 and 5, as can be seen in Table 2.3(b). Returning to Table 2.3(a) we now find $t = 10 + 3 + 9 = 22$, and so forth. In this manner, all the values of t can be found. However, there is one restriction: two teeth on the same rod will be registered as one single tooth, to ensure that t cannot assume a value larger than 26.

The word CRYPTO is encrypted as follows:

Table 2.3. Hagelin-machine: (a) position of teeth on drum; (b) active pins on code wheels.

(a)

code wheel no. \ rod no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0
6	0	0	0	0	1	0	0	0	0	1	1	1	1	1	0	1	0	0	0	0	1	1	1	1	1	1	1

(b)

code wheel no. \ positions on code wheel	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	0	0	1	1	1	1	0	1	0	1	1	0	0	0	1	1	0	0	1	0	1	0	0	1	1	1]
2	0	1	1	0	1	0	0	0	1	0	0	1	1	0	0	1	0	1	1	0	1	0	0	0	1]	
3	1	0	0	1	1	1	1	1	1	0	1	0	0	0	1	0	1	0	1	1	0	1	0]			
4	0	1	1	0	1	1	0	0	0	1	0	0	1	0	1	1	0	1	0	1	1]					
5	1	1	1	0	0	1	0	1	0	1	0	0	1	0	1	0	0	0	1]							
6	0	0	1	0	0	1	1	1	0	1	0	1	1	0	1	0	0]									

C: $i = 3$, $t = 10$ $\Rightarrow j = 8$ $\Rightarrow$ H
R: $i = 18$, $t = 22$ $\Rightarrow j = 5$ $\Rightarrow$ E
Y: $i = 25$, $t = 26$ $\Rightarrow j = 2$ $\Rightarrow$ B
P: $i = 16$, $t = 5$ $\Rightarrow j = 16$ $\Rightarrow$ P
T: $i = 20$, $t = 15$ $\Rightarrow j = 22$ $\Rightarrow$ V
O: $i = 15$, $t = 22$ $\Rightarrow j = 8$ $\Rightarrow$ H. △

The number of pins on the code wheels has been chosen so that they have no common factors. Therefore it will take no less than $26 \times 25 \times 23 \times 21 \times 19 \times 17 = 101.405.850$ revolutions before all the code wheels have returned to their initial positions. Thus, 101 405 850 letters can be encrypted before the machine repeats the same encryption pattern. This is a maximum value since certain configurations of the active and inactive pins can result in an earlier repetition of the encryption pattern.

The key to the encryption is, in fact, given by the positions of the teeth on the rods and the positions of the pins on the code wheels, so with this knowledge, we can calculate the total number of possibilities for the keys. In all, there are $26 + 25 + 23 + 21 + 19 + 17 = 131$ pins, which can assume one of two settings (active and inactive). Hence, there are 2^{131} possible configurations for the code wheels.

Let us now consider a rod. Each rod has two teeth and zero, one or two teeth are located opposite the six code wheels. There is only one way of locating zero teeth opposite the code wheels and that is by setting both teeth to inactive. There are six ways of placing one tooth opposite the code wheels (either at code wheel 1 or code wheel 2, etc.) and finally, two teeth can be positioned in $\binom{6}{2} = 15$ ways. Therefore, there are $1 + 6 + 15 = 22$ possible ways of fixing two teeth to a rod. Thus, the computation of the total number of configurations of the drum with its 27 rods is equivalent to calculating in how many ways 27 objects can be selected from a set of 22. Obviously, after each selection, the object is returned to the set, as some of the 27 rods will be identical. The solution to this combinatorial problem can be found by drawing the same number of columns as objects and placing a cross × in the corresponding column. For instance, if we wish to select five objects from a set of 4, this may look something like this:

Object 1	Object 2	Object 3	Object 4
× /	/	× × /	× ×

where the delimiter / is used to separate the columns. A shorter notation is ×//××/××. Two slashes directly after one another is interpreted as an object

not being selected. Each possibility is entirely defined by a given arrangement of the five crosses and three slashes. The five crosses can be placed in 8 (= 5 + 3) positions in $\binom{8}{5}$ ways. Or, if n objects are selected from a set of m objects, this can be done in $\binom{m+n-1}{n}$ ways. Thus, we find for the total number of possible configurations of the drum:

$$\binom{48}{27} = \frac{48!}{21!\,27!} = 2.23 \times 10^{13}.$$

The total number of keys is therefore:

$$2^{131} \times 2.23 \times 10^{13} = 6.07 \times 10^{52} \text{ keys.}$$

No matter how large this figure may be, the Hagelin machine will eventually fail to withstand cryptanalytic attacks. The six code wheels will assume a certain position, in which the number of active pins is counted. The configuration of the drum, however, will remain unchanged, so that in fact, the letter of the ciphertext is determined by the position of the code wheels. There are $2^6 = 64$ possible positions of the code wheels with respect to the drum. Since 64 cannot be divided by 26, the values for t will exhibit a non-uniform distribution, which is a potential weakness of this method.

A ciphertext-only-attack can provide information on the positions of the active and passive pins of the code wheels. Considering code wheel 6, which has 17 pins, we can write the ciphertext as a matrix, with the first 17 letters in the first row, the second 17 letters in the second, etc. Since code wheel 6 returns to the same position after 17 letters, all the letters in column i will be enciphered by the same active or passive pin. Assuming column i was encrypted by an active pin, then a second column j will show the same distribution if this was also encrypted by an active pin. The same holds for the case in which columns i and j were both encrypted by passive pins. Here we have assumed that the influence of the other code wheels on columns i and j is entirely random. Therefore, by looking at the distributions in the columns, a reasonable idea of which pins are active or passive can be obtained. By arranging the ciphertext in a 19-column matrix, we can estimate which columns are related to the active and passive pins of code wheel 5, and so on for the remaining code wheels.

In practice, 1000–2000 letters of the ciphertext prove sufficient to be able to find the relative positions of the pins of the code wheels. For a known-plaintext-attack, only approximately 50–100 letters are needed.

2.5 Statistics and cryptanalysis

Cryptanalysts has several statistical tools at their disposal for finding the key or plaintext from a ciphertext, or, for instance, for determining whether two columns have the same frequency distribution, as in the case of the Hagelin machine. In this section we will examine several statistical tests which play an important role in cryptanalysis and demonstrate the process of decipherment with a simple example.

Coincidence index (CI)

If we consider a totally random text, constructed from an alphabet of 26 letters, then each letter will have the same probability of occurrence, equal to 1/26. Suppose we have a second random text, which we place beneath the first. We may then wonder how great the chance is of finding two identical letters one above the other. Since each letter exhibits a random character, the probability of finding for example two a's together is equal to $(1/26)^2$. Obviously, this also applies to two b's etc., which results in the total probability of finding two of the same letters together of:

$$(1/26)^2 + (1/26)^2 \ldots + (1/26)^2 = 26 \times (1/26)^2 = 1/26 = 0.0385.$$

However, for an English text, as opposed to a random text, we find that the probabilities of occurrence of the letters are not the same. In English, approximately: $p(a) = 0.082$, $p(b) = 0.015$ … etc. Now, the calculation of the probability of finding two identical letters together yields:

$$(0.082)^2 + (0.015)^2 + \ldots = 0.0661.$$

This value is larger than in the case of a random text. This is referred to as the coincidence index and is generally defined according to the following expression:

$$\mathrm{CI} = \sum_{i=1}^{n} p_i^2, \tag{2.1}$$

in which n is the size of the alphabet and p_i the probability of occurrence of the ith symbol of the alphabet.

In the preceding section we found that for a random and an English text the CI was equal to 0.0385 and 0.0661, respectively. Every language is characterised by a specific value of CI, as is shown in Table 2.4.

In cryptanalysis, the calculation of CI can prove worthwhile in several ways. In the case of a monoalphabetical substitution the letters of the

plaintext are replaced by other letters. This does not influence the statistical parameters of the text and therefore neither the value of CI of the plaintext nor the ciphertext. We can use this information to test whether a text was enciphered by means of a monoalphabetical or a polyalphabetical substitution. If the CI value of a text corresponds roughly to that of the language in which the text is written, so that:

$$\mathrm{CI}(\text{plaintext}) = \mathrm{CI}(\text{ciphertext}),$$

then it is likely that a monoalphabetical substitution has been used. However, if the value of CI of the ciphertext turns out to be considerably lower, a polyalphabetical substitution could have been used. Since a polyalphabetical substitution will tend to conceal the statistical parameters of a text the value of CI will approach that of a random text.

In addition, a calculation of the CI can also provide insight into the probability of two different ciphertexts, say C_1 and C_2, being encrypted by the same method. If this is the case, then $\mathrm{CI}(C_1) \approx \mathrm{CI}(C_2)$. At the end of this section, we will present an example which relies on this principle.

Since in practice ciphertexts have only a finite length, the value we find for the CI from the ciphertext will always differ from the theoretical value. For this reason, we often use an estimation, CI′, instead of CI. A suitable choice of CI′ is given by:

$$\mathrm{CI}' = \sum_{i=1}^{n} x_i(x_i - 1)/L(L - 1) \tag{2.2}$$

in which L represents the length of the ciphertext and x_i the number of occurrences of symbol i in the ciphertext. The probability that in a ciphertext of length L, symbol i will occur x_i times can be described by a binomial distribution:

Table 2.4. Values of the CI for various languages.

English	0.0661
French	0.0778
German	0.0762
Italian	0.0738
Japanese	0.0819
Russian	0.0529
random text	0.0385

$$p(x_i) = \binom{L}{x_i} p_i^{x_i}(1-p_i)^{L-x_i}. \tag{2.3}$$

Assuming L is sufficiently large, we can derive the following expressions for the expectation and the variance of x_i:

$$E(x_i) = Lp_i, \tag{2.4}$$

$$\text{var}(x_i) = Lp_i(1-p_i). \tag{2.5}$$

These two expressions can be used to prove that CI′ is a pure estimator of CI.

Theorem 2.1
Let the CI be defined as:

$$\text{CI} = \sum_{i=1}^{n} p_i^2,$$

and let

$$\text{CI}' = \sum_{i=1}^{n} x_i(x_i-1)/L(L-1),$$

then CI′ is a pure estimator of CI. This implies that

$$E(\text{CI}') = \text{CI}. \tag{2.6}$$

Proof
We can write:

$$\begin{aligned} E(\text{CI}') &= E\left[\sum_i x_i(x_i-1)/L(L-1)\right] \\ &= \sum_i E[x_i(x_i-1)/L(L-1)]. \end{aligned}$$

Since $\text{var}(x_i) = E(x_i^2) - [E(x_i)]^2$ we find for $E(\text{CI}')$:

$$\begin{aligned} E(\text{CI}') &= \sum_i E\left[(x_i^2 - x_i)/L(L-1)\right] \\ &= \sum_i \left[E(x_i^2) - E(x_i)\right]/L(L-1) \\ &= \sum_i \left\{\text{var}(x_i) + [E(x_i)]^2 - E(x_i)\right\}/L(L-1). \end{aligned}$$

Substitution of the previously found formulae for $E(x_i)$ and $\mathrm{var}(x_i)$, eqs. (2.4) and (2.5), respectively, in this expression, leads to the following result:

$$\begin{aligned}\mathrm{E(CI')} &= \sum_i [Lp_i(1-p_i) + L^2p_i^2 - Lp_i]/L(L-1)\\ &= \sum_i [Lp_i - Lp_i^2 + L^2p_i^2 - Lp_i]/L(L-1)\\ &= \sum_i p_i^2 = \mathrm{CI},\end{aligned}$$

which is exactly the result we were looking for. □

Kasiski test

It may happen that at different places in the ciphertext identical sequences of letters appear. These repeated patterns are interesting as they can provide information on periodicity within the text. Consider the following example, in which the plaintext is transformed to a ciphertext with a given key:

plaintext:	`REQUESTS ADDITIONAL TEST …`
key:	`TELEXTEL EXTELEXTEL EXTE …`
ciphertext:	`CAVKTBLT EUQWSWJGEA LTBL …`

The plaintext contains the letter sequence EST twice. Since for both cases, the same section of the key is used for encryption, the resulting letter sequence TBL in the ciphertext is also the same for both cases. This is caused by the fact that the sequences EST are positioned exactly a multiple number of the key length, or period, apart. Clearly, the distance between identical letter sequences can tell us something about the period of an encrypted text. We can find this value by determining the most frequently occurring common factor.

Example

A given ciphertext contains letter sequences repeated at the following distances:

	distance	
`PQA`	150	$= 2 \times 5^2 \times 3$
`RET`	42	$= 2 \times 7 \times 3$
`FRT`	10	$= 2 \times 5$
`ROPY`	81	$= 3^4$

DER $57 = 19 \times 3$
RUN $117 = 13 \times 3^2$

Since a factor 3 appears as the most common factor, we can state that the period of the ciphertext is most probably 3. △

Chi test

The chi-test offers a straightforward means of comparing two frequency distributions. The following sum is calculated, in which p_i represents the uncertainty of the occurrence of symbol i with the first distribution and q_i the uncertainty for the second distribution.

$$\chi = \sum_{i=1}^{n} p_i q_i. \tag{2.7}$$

It is evident that when the two frequency distributions are similar, the value of χ will be higher than when the two distributions are dissimilar.

Assume we have two ciphertexts C_1 and C_2, which are both the result of a Caesar substitution. The alphabet of the first is shifted by t_1 letters; that of the second by t_2. If $t_1 = t_2$, i.e. C_1 and C_2 are encrypted with the same Caesar substitution, then χ will be large, since the statistics of C_1 will exhibit large similarities with those of C_2. Correspondingly, when $t_1 \neq t_2$, χ will be small.

Besides being used to determine whether the same or different substitutions have been employed, χ can also be used to reduce a poly-alphabetical substitution to a monoalphabetical substitution. This is illustrated by the following example.

Example

A given plaintext is transposed to a ciphertext with the Vigenère table and the key-word RADIO:

plaintext: `EXECUTE THESE COMMANDS`
key: `RADIORA DIORA DIORADIO`
ciphertext: `VXHKIKE WPSJE FWADAQLG`

In order to convert the ciphertext back to the plaintext, it is written in a matrix whose number of columns corresponds to the length of the key word:

R	A	D	I	O
V	X	H	K	I
K	E	W	P	S
J	E	F	W	A
D	A	Q	L	G

The original plaintext can be retrieved by deciphering the first column with row *R* of the Vigenère table, the second column with row *A*, the third with row *D*, etc.

The decipherment can also be performed in a different way. Consider the letters of the key-word and their relative distances to the first letter, in this case the letter R. This yields the following values:

R	A	D	I	O
0	9	12	17	23

By replacing the letters of column 2 of the ciphertext with the letters which are located 9 positions earlier in the alphabet, the letters of column 3 with letters which are located 12 places earlier, etc., we obtain the following table:

V	O	V	T	L
K	V	K	Y	V
J	V	T	F	D
D	R	E	U	J

Now only row *R* of the Vigenère table is used for deciphering the entire ciphertext, instead of five different rows of the table.

We have now, in fact, reduced the problem of decrypting a ciphertext based on a polyalphabetical substitution to deciphering a text based on a monoalphabetical substitution. △

This procedure for converting a polyalphabetical substitution to a monoalphabetical substitution is only effective when the distances between the letters of the key word are known. Usually, though, a cryptanalyst does not have this information. The chi test, however, offers a means of finding some indication of these distances. In the above example we saw that the letters of the columns of the ciphertext were shifted such that they could all be deciphered with one and the same substitution cipher. Since the columns are encrypted with the same monoalphabetical substitution, the value of χ calculated for two columns will be high. Now all that remains for the cryptanalyst is to shift the letters of a column repeatedly, until the highest value for χ calculated for a given column and the first one is found. It is then

most likely that the columns can be deciphered with one and the same key row.

This section has given several examples of the statistical tools which are available to the cryptanalyst. Finally we will now give an example of a practical analysis of a ciphertext.

An example of cryptanalysis

Two ciphertexts are intercepted, shortly after one another:

Cipher text 1:

```
k o o m m a c o m o q e g l x x m q c c k u e y f c u r
y l y l i g z s x c z v b c k m y o p n p o g d g i a z
t x d d i a k n v o m x h i e m r d e z v x b m z r n l
z a y q i q x g k k k p n e v h o v v b k k t c s s e p
k g d h x y v j m r d k b c j u e f m a k n t d r x b i
e m r d p r r j b x f q n e m x d r l b c j h p z t v v
i x y e t n i i a w d r g n o m r z r r e i k i o x r u
s x c r e t v
```

Cipher text 2:

```
z a o z y g y u k n d w p i o u o r i y r h h b z x r c
e a y v x u v r x k c m a x s t x s e p b r x c s l r u
k v b x t g z u g g d w h x m x c s x b i k t n s l r j
z h b x m s p u n g z r g k u d x n a u f c m r z x j r
y w y m i
```

As these two ciphertexts were received within a short time of each other, it is very possible that they were both enciphered according to the same method. This hypothesis can be verified by calculating CI′ for both texts:

$$\mathrm{CI}'(C_1) = 0.0421,$$

$$\mathrm{CI}'(C_2) = 0.0445.$$

These values correspond well enough to assume safely that both texts were indeed enciphered according to the same method.

The values of CI′ lie somewhere between that of a random text and, for instance, an English text. This would indicate a polyalphabetical substitution. In order to determine whether this is actually the case, we can use the Kasiski test, for which we must find repeated letter sequences in the ciphertext and their distances. The results are given below:

Kasiski test ciphertext 1

ak	70
bc	35, 84
cj	35
ck	22
dr	21, 22
em	13, 70
et	29
gd	63
ia	7, 115
ie	70
kk	11, 12
kn	70
ma	126
mr	21, 42, 49
mx	88
ne	56
om	5, 58, 117
pn	49
rd	21, 49
re	12
rr	41
sx	161
tv	36
ue	106
vb	63
vv	65
xb	60
xc	161
xd	98
xy	53
yl	2
zr	105
zt	109
zv	37
akn	70
bcj	35
emr	70
iem	70
mrd	21, 49
sxc	161
emrd	70
iemr	70
iemrd	70

Kasiski test ciphertext 2

bx	28
cm	67
cs	21
dw	56
gz	32
hb	63
lr	28
rx	14
sl	28
uk	48
xc	21
xm	18
xs	3
zx	84
slr	28
xcs	21

When these distances are resolved as products of prime numbers, the number 7 appears relatively frequently. It is therefore probable that the period is equal to 7 and that 7 alphabets were used for encryption. We can now write the ciphertexts in matrix form.

```
k o o m m a c
o m o q e g l
x x m q c c k
u e y f c u r
y l y l i g z
s x c z v b c
k m y o p n p
o g d g i a z
t x d d i a k
n v o m x h i
e m r d e z v
x b m z r n l
z a y q i q x
g k k k p n e
v h o v v b k
k t c s s e p
k g d h x y v
j m r d k b c
j u e f m a k
n t d r x b i
e m r d p r r
j b x f q n e
m x d r l b c
j h p z t v v
i x y e t n i
i a w d r g n
o m r z r r e
i k i o x r u
s x c r e t v
z a o z y g y
u k n d w p i
o u o r i y r
h h b z x r c
e a y v x u v
r x k c m a x
s t x s e p b
r x c s l r u
k v b x t g z
u g g d w h x
m x c s x b i
k t n s l r j
z h b x m s p
u n g z r g k
u d x n a u f
c m r z x j r
y w y m i
```

Each column should correspond to a certain monoalphabetical substitution. Calculation of the values CI′ for the various columns results in:

$$
\begin{aligned}
\mathrm{CI}'(\text{column 1}) &= 0.0522,\\
\mathrm{CI}'(\text{column 2}) &= 0.0801,\\
\mathrm{CI}'(\text{column 3}) &= 0.0734,\\
\mathrm{CI}'(\text{column 4}) &= 0.0744,\\
\mathrm{CI}'(\text{column 5}) &= 0.0705,\\
\mathrm{CI}'(\text{column 6}) &= 0.0717,\\
\mathrm{CI}'(\text{column 7}) &= 0.0606.
\end{aligned}
$$

As far as the values of CI′ are concerned, it seems likey that each column has been enciphered with a monoalphabetical substitution. We can now attempt to convert the ciphertext based on a polyalphabetical substitution to a different ciphertext which can be deciphered with a single monoalphabetical substitution. First, we repeatedly shift the letters of each column, starting at position 1 and ending with position 25, and each time we calculate the corresponding value of χ, with respect to another column. At a certain point, one of these values of χ is found to be considerably higher than the others. These have been underlined below:

columns 1 and 2: 0.0388 0.0487 0.0317 0.0326 0.0274 0.0340 0.0421 0.0402 0.0321 0.0350 0.0425 0.0411 0.0662 0.0350 0.0317 0.0359 0.0491 0.0331 0.0236 0.0378 0.0345 0.0288 0.0567 0.0525 0.0302

columns 1 and 3: 0.0378 0.0274 0.0331 0.0331 0.0250 0.0581 0.0491 0.0458 0.0284 0.0383 0.0529 0.0491 0.0307 0.0250 0.0312 0.0444 0.0392 0.0359 0.0392 0.0354 0.0421 0.0657 0.0416 0.0269 0.0232

columns 1 and 4: 0.0317 0.0369 0.0364 0.0312 0.0454 0.0383 0.0558 0.0302 0.0388 0.0345 0.0520 0.0250 0.0359 0.0336 0.0477 0.0260 0.0435 0.0520 0.0406 0.0369 0.0468 0.0468 0.0326 0.0307 0.0326

columns 1 and 5: 0.0430 0.0586 0.0279 0.0274 0.0331 0.0473 0.0298 0.0359 0.0336 0.0354 0.0302 0.0491 0.0548 0.0265 0.0359 0.0406 0.0506 0.0312 0.0345 0.0336 0.0440 0.0354 0.0506 0.0411 0.0321

columns 1 and 6: 0.0382 0.0271 0.0502 0.0449 0.0295 0.0319 0.0440 0.0522 0.0372 0.0372 0.0343 0.0396 0.0391 0.0391 0.0280 0.0324 0.0454 0.0430 0.0614 0.0362 0.0324 0.0343 0.0498 0.0338 0.0275

columns 1 and 7: 0.0285 0.0444 0.0362 0.0382 0.0357 0.0353 0.0343 0.0415 0.0377 0.0483 0.0333 0.0396 0.0425 0.0300 0.0565 0.0348 0.0329 0.0348 0.0454 0.0304 0.0377 0.0324 0.0449 0.0295 0.0444

columns 2 and 3: 0.0156 0.0369 0.0288 0.0350 0.0340 0.0506 0.0222 0.0350 0.0827 0.0440 0.0307 0.0302 0.0340 0.0345 0.0340 0.0364 0.0307 0.0293 0.0373 0.0402 0.0600 0.0416 0.0378 0.0477 0.0558

columns 2 and 4: 0.0411 0.0411 0.0321 0.0317 0.0473 0.0317 0.0548 0.0421 0.0562 0.0279 0.0373 0.0227 0.0435 0.0274 0.0402 0.0397 0.0312 0.0336 0.0255 0.0775 0.0425 0.0369 0.0284 0.0629 0.0142

columns 2 and 5: 0.0284 0.0435 0.0369 0.0586 0.0227 0.0364 0.0274 0.0463 0.0340 0.0406 0.0473 0.0336 0.0288 0.0298 0.0822 0.0293 0.0321 0.0288 0.0454 0.0161 0.0458 0.0411 0.0345 0.0336 0.0345

columns 2 and 6: 0.0290 0.0295 0.0377 0.0300 0.0343 0.0874 0.0304 0.0295 0.0304 0.0473 0.0329 0.0401 0.0440 0.0314 0.0232 0.0338 0.0444 0.0304 0.0464 0.0464 0.0478 0.0420 0.0478 0.0179 0.0493

columns 2 and 7: 0.0251 0.0700 0.0348 0.0411 0.0319 0.0420 0.0150 0.0454 0.0353 0.0425 0.0396 0.0401 0.0502 0.0174 0.0589 0.0309 0.0440 0.0391 0.0377 0.0169 0.0527 0.0377 0.0329 0.0473 0.0406

columns 3 and 4: 0.0312 0.0269 0.0444 0.0321 0.0321 0.0369 0.0463 0.0236 0.0336 0.0411 0.0676 0.0440 0.0298 0.0397 0.0444 0.0246 0.0269 0.0440 0.0279 0.0326 0.0383 0.0421 0.0331 0.0468 0.0671

columns 3 and 5: 0.0454 0.0444 0.0232 0.0326 0.0529 0.0662 0.0345 0.0246 0.0449 0.0336 0.0312 0.0430 0.0454 0.0203 0.0364 0.0473 0.0548 0.0260 0.0350 0.0572 0.0406 0.0265 0.0241 0.0378 0.0317

columns 3 and 6: 0.0435 0.0425 0.0314 0.0420 0.0237 0.0295 0.0449 0.0444 0.0309 0.0367 0.0594 0.0386 0.0382 0.0367 0.0343 0.0430 0.0377 0.0300 0.0271 0.0213 0.0338 0.0464 0.0787 0.0348 0.0261

columns 3 and 7: 0.0338 0.0396 0.0411 0.0507 0.0251 0.0459 0.0502 0.0333 0.0411 0.0227 0.0304 0.0401 0.0444 0.0348 0.0478 0.0338 0.0319 0.0367 0.0556 0.0449 0.0304 0.0430 0.0338 0.0266 0.0401

columns 4 and 5: 0.0402 0.0558 0.0298 0.0312 0.0260 0.0704 0.0288 0.0548 0.0312 0.0534 0.0147 0.0317 0.0435 0.0487 0.0321 0.0232 0.0577 0.0222 0.0378 0.0350 0.0789 0.0189 0.0359 0.0236 0.0435

columns 4 and 6: 0.0459 0.0324 0.0406 0.0406 0.0502 0.0295 0.0179 0.0459 0.0251 0.0396 0.0401 0.0734 0.0232 0.0319 0.0271 0.0541 0.0396 0.0386 0.0415 0.0246 0.0242 0.0304 0.0536 0.0459 0.0565

columns 4 and 7: 0.0488 0.0411 0.0295 0.0522 0.0261 0.0401 0.0256 0.0720 0.0285 0.0469 0.0188 0.0372 0.0295 0.0435 0.0493 0.0343 0.0430 0.0256 0.0478 0.0285 0.0638 0.0271 0.0556 0.0242 0.0251

columns 5 and 6: 0.0237 0.0546 0.0343 0.0449 0.0285 0.0556 0.0295 0.0382 0.0290 0.0536 0.0444 0.0222 0.0367 0.0300 0.0338 0.0353 0.0768 0.0285 0.0304 0.0251 0.0546 0.0454 0.0435 0.0425 0.0348

columns 5 and 7: 0.0256 0.0580 0.0304 0.0333 0.0242 0.0478 0.0430 0.0324 0.0478 0.0290 0.0367 0.0217 0. 0739 0.0300 0.0502 0.0227 0.0430 0.0300 0.0372 0.0454 0.0430 0.0396 0.0237 0.0507 0.0213

columns 6 and 7: 0.0212 0.0420 0.0405 0.0400 0.0489 0.0326 0.0479 0.0212 0.0533 0.0311 0.0479 0.0365 0.0267 0.0242 0.0440 0.0484 0.0370 0.0553 0.0351 0.0272 0.0212 0.0652 0.0341 0.0474 0.0385

Based on these results, the most probable distances across which the columns have been shifted with respect to column 1 are:

column 2: 13
column 3: 22
column 4: 7
column 5: 2
column 6: 19
column 7: 15

These values are then used to convert ciphertexts C_1 and C_2 to the following text:

```
k b k t o t r o z k x g z a x k i x
e v z u r u m e n g y y u s k z o s
k y g x u r k z u v r g e o t z n k
t o t k z k k t z n i k t z a x e z
n k g s k x o i g t g a z n u x k j
m g x g r r g t v u k c x u z k g y
z u x e k t z o z r k j z n k m u r
j h a m o t z n g z y z u x e z n k
r k g j o t m s g t m k z y n u r j
u l g v o k i k u l v g x i n s k t
z c o z n g t k t i x e v z k j s k
y y g m k z n k g a z n u x j k y i
x o h k y k r g h u x g z k r e n u
c z n k r k g j o t m s g t z g i q
r k y z n k j k i x e v z o u t c k
y a m m k y z z u x k g j z n k y z
u x e o l e u a c g t z z u q t u c
n u c z n g z c g y j u t k
```

It should now be possible to decipher this text with a single monoalphabetical substitution. It is left to the reader to investigate this and to find the plaintext! △