(1.8) Cryptanalysis:

The science of deducing the plaintext from a ciphertext, without knowledge of the key.

(1.8.1) Attacks on encrypted messages

The objective of the following attacks is to systematically recover plaintext from ciphertext, or even more drastically, to deduce the decryption key.

1. ***A* ciphertext-only attack**: is one where the adversary (or cryptanalyst) tries to deduce the decryption key or plaintext by only observing ciphertext. Any encryption scheme vulnerable to this type of attack is considered to be completely insecure.
2. ***A* known-plaintext attack**: is one where the adversary has a quantity of plaintext and corresponding ciphertext. This type of attack is typically only marginally more difficult to mount.
3. ***A* chosen-plaintext attack**: is one where the adversary chooses plaintext and is then given corresponding ciphertext. Subsequently, the adversary uses any information deduced in order to recover plaintext corresponding to previously unseen ciphertext.
4. **An adaptive chosen-plaintext attack**: is a chosen-plaintext attack wherein the choice of plaintext may depend on the ciphertext received from previous requests.
5. ***A* chosen-ciphertext attack**: is one where the adversary selects the ciphertext and is then given the corresponding plaintext. One way to mount such an attack is for the adversary to gain access to the equipment used for decryption (but not the decryption key, which may be securely embedded in the equipment). The objective is then to be able, without access to such equipment, to deduce the plaintext from (different) ciphertext.
6. **An adaptive chosen-ciphertext attack**: is a chosen-ciphertext attack where the choice of ciphertext may depend on the plaintext received from previous requests.

(1.8.2) Some concepts on cryptanalysis:

- **Frequency**: number of appearance of the letter in the ciphertext, where the frequencies of the ciphertext letters are compared with the frequencies in Table 1 or Figure 5.

- **Repetition**: is the similar parts in the ciphertext that have length not less than three. This helps us to find the length of the key (the number of alphabets that used to enciphering in the polyalphabetic systems).
  Take the Highest Common Factor HCF between the reputations, which represent the length of the key, this method, is called the Kasiski method.

Letter	%	Letter	%
a	8.167	**n**	6.749
b	1.492	**o**	7.507
c	2.782	**p**	1.929
d	4.253	**q**	0.095
e	12.702	**r**	5.987
f	2.228	**s**	6.327
g	2.015	**t**	9.056
h	6.094	**u**	2.758
i	6.966	**v**	0.978
j	0.153	**w**	2.360
k	0.772	**x**	0.150
l	4.025	**y**	1.974
m	2.406	**z**	0.074

Table 1: English letters frequencies

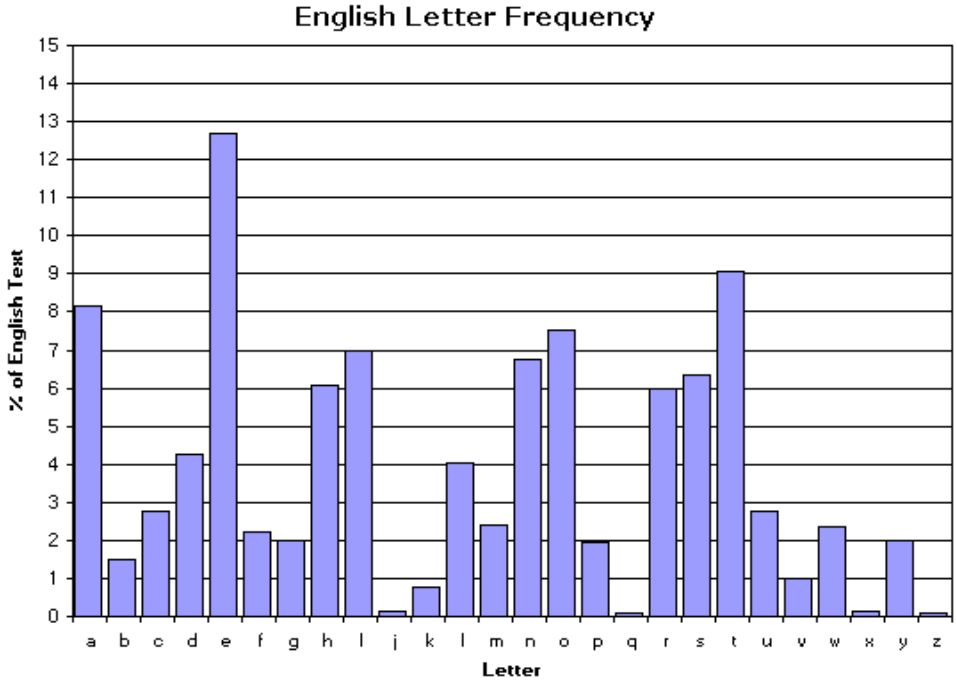


Figure 5: Histogram of English letters frequencies

- **Index of Coincidence (IC)**: is the probability that two letters selected from the text are identical, we can compute the IC from the following equation:

$$IC = \frac{\sum_{A}^{Z} f_\lambda (f_\lambda - 1)}{n(n-1)},$$

where f_λ is the frequency of the letter λ in the ciphertext and n is the length of the letter. The IC value differs from language to another. We can use the IC to discover if the message were enciphered using Monoalphabetic system or polyalphabetic system.

- **Coincidence**: is the computing of the coincidence of the ciphertexts, where two messages put one over the other, and the purpose is to discover if the two messages were enciphered using the same key. If there is 7 coincidence letters between 100 letters in the two messages then the two messages were enciphered using the same key, while if there is 4 letters coincidence between every 100 letters then they enciphered with different keys.

(1.9) Cryptanalysis examples:

First of all we must specified the type of the cipher system that was used. If the frequencies of the ciphertext are the same as the frequencies of the language then, a transposition cipher system was used; otherwise a substitution cipher system was used.

(1.9.1) Cryptanalysis of transposition cipher systems:

When we decide that a transposition cipher system were used, we put the cipher text in mXn matrix, m and n depends on the length of the received ciphertext, for example if the length is 500 then one of the possible sizes is 20X25. Then we rearrange the columns to get some known patterns such as (and, the, ion, that,…) in addition to some expected word in the message.

As we know there are two types of transposition cipher system: simple and double transposition, the cryptanalysis of the last one is more complicated because we lose the ability to find the known patterns.

(1.9.2) Cryptanalysis of substitution cipher systems:

If we know that a substitution cipher system was used, the next step is to determine whether a monoalphabetic system or polyalphabetic system was used, by using the IC of the language.

Example: A sample of ordinary English contains the following distribution of letters

Letter	Count	Letter	count
A	141	N	119
B	36	O	132
C	36	P	28
D	103	Q	1
E	188	R	95
F	37	S	64
G	34	T	182
H	102	U	59
I	123	V	13
J	4	W	55
K	18	X	3
L	56	Y	23
M	27	Z	0

What is the probability of selecting an identical pair of letters from this collection? in other word compute the IC.

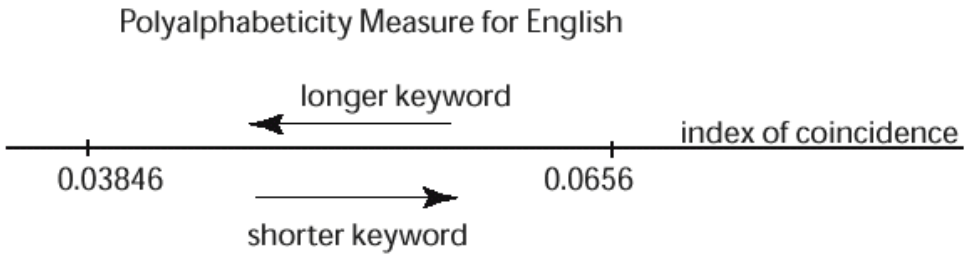
$$IC = \frac{\sum_{A}^{Z} f_\lambda (f_\lambda - 1)}{n(n-1)}$$

$$\text{IC}= \frac{141(141-1)+36(36-1)+\ldots+23(23-1)+0(0-1)}{1679(1679-1)} = \frac{184838}{2817362} \approx 0.0656 .$$

Example: What is the index of coincidence for a collection of 2600 letters consisting of 100 A 's,100 B 's,100 C 's,...,100 Z 's?

$$\text{IC}= \frac{100\cdot 99+100\cdot 99+\ldots+100\cdot 99+100\cdot 99}{2600\cdot 2599} \approx 0.0384615 .$$

As we see from the two examples above the index of coincidence of totally random (uniformly distributed) collection of letters is about 0.0385. Vigenere ciphertexts from longer keywords have a more uniform distribution of letters. For keyword length closer to 1, the index of coincidence will be larger, closer to 0.0656.



If the length of the text is n, we can quantify the connection between index of coincidence and keyword length k, (number of alphabets), where:

$$k \approx \frac{0.0265\cdot n}{(0.065 - IC) + n(IC - 0.0385)},$$

Example: A polyalphabetic ciphertext has the following letter counts.

Letter	Count	Letter	count
A	60	N	28
B	50	O	83
C	42	P	44
D	64	Q	69
E	51	R	13
F	63	S	29
G	19	T	66
H	48	U	87
I	56	V	63
J	67	W	19
K	23	X	43
L	45	Y	39
M	44	Z	67

Estimate the keyword length.

Solution: There are n=1282 letters.

$$IC=\frac{60\cdot 59+50\cdot 49+\ldots+67\cdot 66}{1282\cdot 1281}=\frac{35761}{821121}\approx 0.04355.$$

$$K=\frac{0.0265\cdot 1282}{(0.065-0.04355)+1282(0.04355-0.03846)}=5.1892.$$

Based only on this evidence, a reasonably likely keyword length is 5.

- Now, after the above tests if we conclude that a monoalphabetic cipher system was used, then:

❶ If a direct standard or reversed system were used, we compare the frequencies of the ciphertext with the frequencies of the English language, start by putting E against the letter with the higher frequency in the ciphertext, then we put the other letters sequentially.

❷ If a mixed cipher system was used (Random) then we compare the frequencies of the ciphertext with that in Table 1 and Figure 5.

For advanced analysis we can use in addition to Table 1, a table of double letter frequencies TH, HE, IN, ER, RE, ON, AN, EN,…, and triple letter frequencies THE, AND, TIO, ATI, FOR, THA, TER, RES,… and so on.

- If a polyalphabetic cipher system was used then we will use the Kasiski method to find the length of the key k (number of alphabets). Then we divide the ciphertext into k parts, each part will analyze as in ❷ above.

The Kasiski method was introduced in 1863 by the Prussian military officer Friedrich W. Kasiski. The method analysis repetitions in the ciphertext to determine the period.
For example, consider the plaintext TO BE OR NOT TO BE enciphered with a Vigenere cipher with key HAM:

K=	H	A	M	H	A	M	H	A	M	H	A	M	H
M=	T	O	B	E	O	R	N	O	T	T	O	B	E
C=	A	O	N	L	O	D	U	O	F	A	O	N	L

The ciphertext contains two occurrences of the sequence AONL 9 characters apart, and the period could be 1,3 or 9 (we know it's 3).
Repetitions in the ciphertext more than two characters long are unlikely to occur by chance. They occur when the plaintext pattern repeats at a distance equal to a multiple of the period.
If there are m ciphertext repetitions that occur at intervals I j ($1\leq j\leq m$) the period is likely to be some number that divides most of the m intervals.

Example: We shall use IC and Kasiski method to analyze the following ciphertext.

ZHYME	ZVELK	OJUBW	CEYIN	CUSML	RAVSR	YARNH	CEARI	UJPGP	VARDU
QZCGR	NNCAW	JALUH	GJPJR	YGEGQ	FULUS	QFFPV	EYED**Q**	***GOLKA***	***LVOSJ***
TFRTR	YEJZS	RVNCI	HYJNM	ZDCRO	DKHCR	MMLNR	FFLFN	***QGOLK***	***ALVOS***
JWMIK	QKUBP	SAYOJ	RRQYI	NRNYC	YQZSY	EDNCA	LEILX	RCHUG	IEBKO
YTHGV	VCKHC	JE***QGO***	***LKALV***	***OSJ***ED	WEAKS	GJHYC	LLFTY	IGSVT	FVPMZ
NRZOL	CYUZS	FKOQR	YRTAR	ZFGKI	QKRSV	IRCEY	USKVT	MKHCR	MYQIL
XRCRL	GQARZ	OLKHY	KSNFN	RRNCZ	TWUOC	JNMKC	MDEZP	IRJEJ	W

When we calculate the frequency distribution, we will find that the IC=0.04343, n=346,

$$k=\frac{0.0265\cdot 346}{(0.065-0.04343)+346(0.04343-0.03846)}=5.2659$$

The IC indicates that this is a polyalphabetic cipher with a period of about 5.
We observe that there are 3 occurrences of the sequence QGOLKALVOSJ, the first two occurrences are separated by 51 and the last two by 72 characters (start to start); the only common divisor of 51 and 72 is 3 - the period is almost certainly 3.

Example: When we calculate the IC of some ciphertext, we find that k=9.34. Also we observe that there is NYX appearance many times in the ciphertext and the distance between them are 30, 50, 90, 110, and 33.
Since these can each be factored as

30=2X3X5
50=2X5X5
90=2X3X3X5
110=2X5X11
33=3X11

there are a number of candidates for key length. 2 and 5 are popular factors among these distance followed by 3 and 11. Note that all but 33 have 2X5=10 as a factor. The cryptanalyst might then disregard 33 as a pure coincidence, and discard that data in favor of conjecture that the key length is a multiple of 2 and/or 5. Combining this with data from the Friedman test that the key approximately 9 letters long, the cryptanalyst guesses that the key is 10 letters long, and not 2 or 5 letters long.