Elementary Cryptanalysis

1. Cryptanalytic Attacks

Kerckhoff's principle: The opponent knows the cryptosystem used with all its details.

Technically, this means that the opponent knows the exact specifications of all cryptographic algorithms.

Generally, we consider two types of cryptographic attacks:

- a passive attack – this is an attack in which the opponent only observes the exchanged messages; it can be assumed, as well, that the opponent has an access to a "black box" which enciphers messages (but cannot decipher them); the passive attack is the most simple model of an attack and threatens only the confidentiality of the messages.

- an active attack – this is an attack in which the opponent tries to modify the transmitted messages; such an attack threatens not just the cofidentiality, but also the integrity and the authenticity of the messages.

Attack models:

(1) ciphertext only attack – the opponent observes only the ciphertext

(2) known plaintext attack – the opponent possesses a string of plaintext and the corresponding ciphertext

(3) chosen plaintext attack – the opponent has temporary access to the encryption machinery

(4) adaptive chosen plaintext attack

(5) chosen ciphrtext attack – the opponent has temporary access to the decryption machinery

(6) adaptive chosen ciphertext attack

2. Cryptanalysis of general substitution ciphers

Consider a general substitution cipher.

In such a cipher each symbol goes to a fixed symbol, so that the statistical properties of the text are preserved. This allows the use of statistical methods.

In what follows, we assume that the plaintext is a meaningful text in English.

We consider a ciphertext only attack. The general scheme of the attack is the following:

(1) investigate the statistical characteristics of the ciphertext;

(2) compare these characteristics with the corresponding characteristics of a typical English text;

(3) these characteristics should be "close".

The resemblance should grow bigger with the increse of the available cryptotext.

The most obvious characteristic is the frequency of the letters of the alphabet.

letter	frequency	letter	frequency	letter	frequency
A	0.082	J	0.002	S	0.063
B	0.015	K	0.008	T	0.091
C	0.028	L	0.040	U	0.028
D	0.043	M	0.024	V	0.010
E	0.127	N	0.067	W	0.023
F	0.022	O	0.075	X	0.001
G	0.019	P	0.019	Y	0.020
H	0.061	Q	0.001	Z	0.001
I	0.070	R	0.060		

For the sake of convenience these are partitioned into three classes.

High frequency		Middle frequency		Low frequncy	
E	0.127	D	0.043	G	0.018
T	0.091	L	0.040	B	0.015
A	0.082	C	0.028	V	0.010
O	0.075	U	0.028	K	0.008
I	0.070	M	0.024	J	0.002
N	0.067	W	0.023	Q	0.001
S	0.063	F	0.022	X	0.001
H	0.061	Y	0.020	Z	0.001
R	0.060	P	0.020		

It is useful to know the most frequent digraphs and trigraphs (pairs and triples of letters).

Digraphs: TH HE IN ER AN RE ED ON ES ST
EN AT TO NT HA ND OU EA NG AS
OR TI IS ET IT AR TE SE HI OF

Trigraphs: THE ING AND HER ERE ENT
THA NTH WAS ETH FOR DTH

TH	2161	ED	890	OF	731	THE	1771	TER	232
HE	2053	TE	872	IT	704	AND	483	RES	219
IN	1550	TI	865	AL	681	TIO	384	ERE	212
ER	1436	OR	861	AS	648	ATI	287	CON	206
RE	1280	ST	823	HA	646	FOR	284	TED	187
ON	1232	AR	764	NG	630	THA	255	COM	185
AN	1216	ND	761	CO	606				
ET	1029	TO	756	SE	595				
AT	1019	NT	743	ME	573				
ES	917	IS	741	DE	572				

Consider the three cryptotexts below obtained from the same plaintext of length 165 obtained by

(1) general (simple) substitution;

(2) Vigenére with a key of length $m' = 3$;

(3) Vigenére with a key of length $m'' = 6$.

Plaintext:	`thepathoftherighteousmanisbesetonallsidesbytheinequitie`
Ciphertext 1:	`OINMLOIFUOINAPBIONFVHRLYPHSNHNOFYLKKHPGNHSXOINPYNTVPOPN`
Ciphertext 2:	`WVKSOZKCLWVKUWMKHKRAYPOTLGHHGKWCTDZRVWJHGHBHNHWTHEALHOH`
Ciphertext 3:	`VPTWEKJWUALVTQVOXVQCETEEKAQLWVVWCHPCUQSLWSABWLMEGYJPXZG`

Plaintext:	`softheselfishandthetyrannyofevilmanblessedishewhointhen`
Ciphertext 1:	`HFUOINHNKUPHILYGOINOXALYYXFUNCPKRLYSKNHHNGPHINJIFPYOINY`
Ciphertext 2:	`GUIHNHGKOTOVVGQRZKSZBFGQBERTKYWRPOTEZKVGKGWYKSCKCOQHNHB`
Ciphertext 3:	`ADMXYGATSJZUPPUHKJMIFVRPVNVJVXQATEEDTTZWVFQGOINJWXUXYGV`

Plaintext:	`ameofgoodwillshepherdstheweakthroughthevalleyofdarkness`
Ciphertext 1:	`LRNFUBFFGJPKKHINMINAGHOINJNLDOIAFVBIOINCLKKNXFUGLADYNHH`
Ciphertext 2:	`GPSUIUURRCLZRVVKSVKURYWVKZSGNHNUCAJVZKSBDZRHMUIRGUYTHGY`
Ciphertext 3:	`PTIFHODVHNKTAZLVRPTYHJVPTDIRMBWYSLIPIOIMCTALCFLHPYOEGAH`

The letter frequencies of these texts are presented in the following table:

	(1)	(2)	(3)		(1)	(2)	(3)
A	5	3	9	N	25	5	3
B	3	5	2	O	14	7	5
C	2	6	5	P	11	3	10
D	2	2	4	Q	0	3	6
E	0	3	8	R	3	11	3
F	11	1	4	S	3	7	4
G	6	13	6	T	1	7	12
H	15	17	7	U	6	9	5
I	17	3	7	V	3	11	15
J	3	2	7	W	0	9	9
K	9	17	4	X	4	0	6
L	10	4	8	Y	10	6	6
M	2	2	5	Z	0	9	4

Consider the first ciphertext. The high frequency letters are:

N	25	F	11
I	17	P	11
H	15	L	10
O	14	Y	10
		K	9

In addition the trigraph `OIN` appears 8 times and the digraphs `OI` and `IN` – 9 nd 10 times, respectively. Hence we should have (with a high probability)

$$\texttt{OIN} \rightarrow \texttt{the}.$$

This implies:

Ciphertext:	`OINMLOIFUOINAPBIONFVHRLYPHSNHNOFYLKKHPGNHSXOINPYNTVPOPN`
Plaintext:	`the**th**the***hte*********e*et********e***the**e***t*e`

Ciphertext:	`HFUOINHNKUPHILYGOINOXALYYXFUNCPKRLYSKNHHNGPHINJIFPYOINY`
Plaintext:	`***the*e****h***thet********e********e**e***he*h***the*`

Ciphertext:	`LRNFUBFFGJPKKHINMINAGHOINJNLDOIAFVBIOINCLKKNXFUGLADYNHH`
Plaintext:	`**e***********he*he***the*e**th****hthe****e********e**`

The most frequent bigrams are:

IN	10	FU	5	HI	3	KK	3
OI	9	LY	4	HN	3	PH	3
NH	6			IO	3		

Let us try to idetify H.

If it is a vowel, then it is one of a,o,i; if it is a consonat thenn it is one of n,s,r.

In group 28 we have NHH which is one of eaa, eoo, eii, neither of which looks very probable.

If H is a consonant then the very last word ends up with enn, err, or ess. The last possibility looks the most probable. Hence we have so far

$$\mathtt{O} \to \mathtt{t}, \mathtt{I} \to \mathtt{h}, \mathtt{N} \to \mathtt{e}, \mathtt{H} \to \mathtt{s}.$$

Ciphertext:	OINMLOIFUOINAPBIONFVHRLYPHSNHNOFYLKKHPGNHSXOINPYNTVPOPN
Plaintext:	the**th**the***hte**s****s*eset*****s**es**the**e***t*e

Ciphertext:	HFUOINHNKUPHILYGOINOXALYYXFUNCPKRLYSKNHHNGPHINJIFPYOINY
Plaintext:	s**these***sh***thet*******e********esse**she*h***the*

Ciphertext:	LRNFUBFFGJPKKHINMINAGHOINJNLDOIAFVBIOINCLKKNXFUGLADYNHH
Plaintext:	**e**********she*he**sthe*e**th****hthe****e********ess

The pair `FU` appears 5 times; `F` is of high frequency, and `U` is middlefrequency letter. If we discrad the bigrams containing letters that have been already identified we are left with:

`in`, `an`, `on`, `ar`, `no`, `ng`, `of`

At least one of `F,U` is a vowel since `OIFUOI` $\rightarrow$ `th**th`. So it looks plausible that `FU` $\rightarrow$ `of`.

Further `LY` appears 4 times, whereas `YL` appears 2 times.

It is not very probable that `L` is a vowel (`ML` in the first group. Furthermore `L` and `Y` are not both vowels (`LYY` in group 16). Hence `LY,YL` is one of

$$\{\texttt{an,na}\}, \{\texttt{ar,ra}\}, \{\texttt{in,ni}\}, \{\texttt{ir,ri}\}.$$

We shall take the first possibility. Thus we get

Ciphertext:	OINMLOIFUOINAPBIONFVHRLYPHSNHNOFYLKKHPGNHSXOINPYNTVPOPN
Plaintext:	the*athofthe***hteo*s*an*s*esetona**s**es**the*ne***t*e
Ciphertext:	HFUOINHNKUPHILYGOINOXALYYXFUNCPKRLYSKNHHNGPHINJIFPYOINY
Plaintext:	softhese***shan*thet**ann*ofe****an**esse**she*ho*nthen
Ciphertext:	LRNFUBFFGJPKKHINMINAGHOINJNLDOIAFVBIOINCLKKNXFUGLADYNHH
Plaintext:	a*eof*oo*****she*he**sthe*ea*th*o**hthe*a**e*of*a**ness

Since

$$\texttt{PYONYLRNFU} \to \texttt{*nthena*eof} \to \texttt{inthenameof}$$

we check $\texttt{P} \to \texttt{i}$, $\texttt{R} \to \texttt{m}$.

Ciphertext:	`OINMLOIFUOINAPBIONFVHRLYPHSNHNOFYLKKHPGNHSXOINPYNTVPOPN`
Plaintext:	`the*athofthe*i*hteo*s*anis*esetona**si*es**theine**it*e`

Ciphertext:	`HFUOINHNKUPHILYGOINOXALYYXFUNCPKRLYSKNHHNGPHINJIFPYOINY`
Plaintext:	`softhese**ishan*thet**ann*ofe*i*man**esse*ishe*hointhen`

Ciphertext:	`LRNFUBFFGJPKKHINMINAGHOINJNLDOIAFVBIOINCLKKNXFUGLADYNHH`
Plaintext:	`ameof*oo**i**she*he**sthe*ea*th*o**hthe*a**e*of*a**ness`

Since `A` is a high frequency consonant, it must be `r`. In addition we have

$$\texttt{OXALLYX} \rightarrow \texttt{t*rann*} \rightarrow \texttt{tyranny}$$

$$\texttt{FLYKKHPGNH} \rightarrow \texttt{ona**si*es} \rightarrow \texttt{onallsides}$$

$$\texttt{HIMMINGAH} \rightarrow \texttt{she*herds} \rightarrow \texttt{shepherds}$$

This implies

$$\texttt{X} \rightarrow \texttt{y}, \texttt{K} \rightarrow \texttt{l}, \texttt{N} \rightarrow \texttt{d}, \texttt{M} \rightarrow \texttt{p}$$

and we can easily construct the plaintext and the key.

If we new that an affine cipher was used then this would simplify immensly the cryptanalysis.

We have

$$\mathtt{T}(19) \to \mathtt{o}(14), \mathtt{H}(7) \to \mathtt{i}(8), \mathtt{E}(4) \to \mathtt{n}(13).$$

This gives the system

$$\begin{aligned} 19a + b &\equiv 14 \pmod{26} \\ 7a + b &\equiv 8 \pmod{26} \\ 4a + b &\equiv 13 \pmod{26} \end{aligned}$$

whence $a = 7\ b = 11$, and we get the same permutation on the alphabet.

3. Cryptanalysis of the Vigenére cipher

Consider a sequence of letters from which we randomly select a pair. The probability for these two letters to coincide is $\sum_\alpha \left(\frac{1}{26}\right)^2 = \frac{1}{26} \approx 0.0385$.

Now let us select two letters from a potentially infinite text in English. The probability that these two letters are the same is $\sum_\alpha p^2(\alpha) \approx 0.065$.

We can conclude that comparing two texts enciphered by the same general substutution the expected number of coincidences is 7 in 100; for text obtained by different substitutions this number is 4 in 100.

Concider a ciphertext $c_0 c_1 \ldots c_{n-1}$ of length n enciphered using Vigenère's cipher.

Denote by f_α the number of appearences of α in the ciphertext. The probability to select two identical letters is

$$I_C = \frac{\sum_\alpha f_\alpha(f_\alpha - 1)}{n(n-1)}. \tag{1}$$

The number I_C is called index of coincidences.

Assume m is the length of the used key. Write down the ciphertext in m rows as follows

$$\begin{array}{llll} c_0 & c_m & c_{2m} & \cdots \\ c_1 & c_{m+1} & c_{2m+1} & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ c_{m-1} & c_{2m-1} & c_{3m-1} & \cdots \end{array} \tag{2}$$

The symbols in the same row are enciphered by the same simple substitution (same alphabet).

Count in two ways the expected number of pairs identical symbols in the ciphertext.

On one hand, this number is

$$\frac{1}{2}\sum_{\alpha} f_{\alpha}(f_{\alpha}-1) = \frac{1}{2}n(n-1)I_C. \tag{3}$$

On the other hand let us first choose a symbol from the cryptotext (we have n choices), and then arbitrarily a second symbol.

If the second symbol is in the row containing the first one then the probability of coincidence is ≈ 0.065.

If the second symbol is in another row the probability is ≈ 0.038.

Hence we expect $\approx \frac{1}{2}n(\frac{n}{m}-1)\times 0.065$ pairs of identical symbols appearing in the same row and $\approx \frac{1}{2}n(n-\frac{n}{m})\times 0.038$ pairs of identical symbols appearing in different rows. Hence

$$\frac{1}{2}n(n-1)I_C \approx \frac{1}{2}n(\frac{n}{m}-1) \times 0.065 + \frac{1}{2}n(n-\frac{n}{m}) \times 0.038. \tag{4}$$

This implies

$$m \approx \frac{0.027n}{I_C(n-1) - 0.038n + 0.065}. \tag{5}$$

This formula is not very useful since it does not give an exact result for large values of m.

m	1	2	5	10	∞
I_C	0.065	0.052	0.043	0.041	0.038

Plaintext	`codebreakingisthemostimportantformofsecretintellig`
Ciphertext 1:	`FRGHEUHDNLQJLVWKHPRVWLPSRUWDQWIRUPRIVHFUHWLQWHOOLJ`
Ciphertext 2:	`OOBQBPQAIUNEUSRTEKASRUMNARRMNRROPYODEEADERUNRQLJUG`
Plaintext	`enceintheworldtodayitproducesmuchmoreandmuchmoretr`
Ciphertext 1:	`HQFHLQWKHZRUOGWRGDBLWSURGXFHVPXFKPRUHDQGPXFKPRUHWU`
Ciphertext 2:	`CZCCUNRTEUARJPTMPAWUTNDOBGCCEMSOHKARCMNBYUATMMDERD`
Plaintext	`ustworthyinformationthanspiesandthisintelligenceex`
Ciphertext 1:	`XVWZRUWKBLQIRUPDWLRQWKDQVSLHVDQGWKLVLQWHOOLJHQFHHA`
Ciphertext 2:	`UQFWMDTFKILROPYARUOLFHYZSNUEQMNBFHGEILFEJXIEQNAQEV`
Plaintext	`ertsgreatinfluenceuponthepoliciesofgovernmentsyeti`
Ciphertext 1:	`HUWVJUHDWLQIOXHQFHXSRQWKHSROLFLHVRIJRYHUQPHQWVBHWL`
Ciphertext 2:	`QRREGPQARUNDXUCZCCGPMZTFQPMXIAUEQAFEAVCDNKQNREYCFI`
Plaintext	`thasneverhadachronicler`
Ciphertext 1:	`WKDVQHYHUKDGDFKURQLFOHU`
Ciphertext 2:	`RTAQZETQRFMDYOHPANGOLCD`

D	C	%	D	C	%	D	C	%
1	9	4.054054	42	26	14.364461	83	5	3.571429
2	3	1.357466	43	6	3.333333	84	7	5.035971
3	14	6.363636	44	2	1.117318	85	5	3.623188
4	10	4.566210	45	7	3.932584	86	4	2.919708
5	13	5.963303	46	10	5.649718	87	9	6.617647
6	19	8.755760	47	9	5.113636	88	3	2.222222
7	5	2.314815	48	8	4.571429	89	9	6.716418
8	9	4.186047	49	7	4.022988	90	12	9.022556
9	11	5.140187	50	9	5.202312	91	14	10.606061
10	8	3.755869	51	3	1.744186	92	3	2.290076
11	9	4.245283	52	5	2.923977	93	4	3.076923
12	10	4.739336	53	4	2.352941	94	2	1.550388
13	7	3.333333	54	9	5.325444	95	7	5.468750
14	10	4.784689	55	7	4.166667	96	5	3.937008
15	13	6.250000	56	9	5.389222	97	7	5.555556
16	8	3.864734	57	16	9.638554	98	4	3.200000
17	8	3.883495	58	4	2.424242	99	4	3.225806
18	10	4.878049	59	6	3.658537	100	2	1.626016
19	4	1.960784	60	15	9.202454	101	5	4.098361
20	9	4.433498	61	7	4.320988	102	16	13.223140

21	11	5.445545	62	5	3.105590	103	4	3.333333
22	8	3.980099	63	6	3.750000	104	1	0.840336
23	4	2.000000	64	7	4.402516	105	9	7.627119
24	13	6.532663	65	7	4.430380	106	6	5.128205
25	10	5.050505	66	9	5.732484	107	7	6.034483
26	10	5.076142	67	6	3.846154	108	10	8.695652
27	13	6.632653	68	7	4.516129	109	3	2.631579
28	4	2.051282	69	4	2.597403	110	4	3.539823
29	8	4.123711	70	8	5.228758	111	12	10.714286
30	12	6.217617	71	4	2.631579	112	7	6.306306
31	12	6.250000	72	11	7.284768	113	3	2.727273
32	7	3.664921	73	7	4.666667	114	6	5.504587
33	19	10.000000	74	3	2.013423	115	7	6.481481
34	7	3.703704	75	6	4.054054	116	5	4.672897
35	8	4.255319	76	1	0.680272	117	11	10.377358
36	14	7.486631	77	6	4.109589	118	2	1.904762
37	9	4.838710	78	10	6.896552	119	6	5.769231
38	7	3.783784	79	6	4.166667	120	3	2.912621
39	8	4.347826	80	3	2.097902	121	4	3.921569
40	4	2.185792	81	13	9.154930	122	8	7.920792
41	7	3.846154	82	7	4.964539	123	8	8.000000

In the table below we present the shifts for which we have large percentage of coincidences (in this case more than 8%).

coincidences	displacement	decomposition
8.76%	6	2×3
10.00%	33	11×3
14.36%	42	$7 \times 2 \times 3$
9.64%	57	19×3
9.20%	60	$5 \times 2^2 \times 3$
9.15%	81	3^4
9.02%	90	$5 \times 2 \times 3^2$
10.60%	91	13×7
13.22%	102	$17 \times 2 \times 3$
8.70%	108	$2^2 \times 3^2$
10.71%	111	37×3
10.38%	117	13×3^2
8.00%	123	41×3

Another technique (due to Kasiski):

Find identical sequences in the ciphertext of length at least 3.

Find the distance between them.

If these identical ciphertexts result from identical plaintexts encrypted with the same part of the keysequence then the distance should be divisible by the key length.

letter sequence	distance	decomposition
PQA	150	$2 \times 5^2 \times 3$
RTE	42	$2 \times 7 \times 3$
ROPY	81	3^4
DER	57	19×3
RUN	117	13×3^2
UNR	12	$2^2 \times 3$
CZCC	114	$2 \times 19 \times 3$
MNB	42	$2 \times 7 \times 3$
ARU	42	$2 \times 7 \times 3$
UEQ	54	2×3^3

Compute the incidence of coincidences for the subsequences:

- Y_1 consisting of the symbols in positions 1,4,7,...
- Y_2 consisting of the symbols in positions 2,5,8,...
- Y_3 consisting of the symbols in positions 3,6,9,...

$$I_C(Y_1) = 0.0717117, \quad I_C(Y_2) = 0.0636801, \quad I_C(Y_3) = 0.0640504.$$