

Classic Cryptography Tutorial

Ciphertext-Only Attack on Substitution Cipher

Monoalphabetic substitution ciphers cannot protect against known plaintext and chosen plaintext attacks. Therefore, we restrict our discussion to ciphertext-only attacks, and try to prove that even under a ciphertext-only scenario they are insecure.

We will see that there are algorithmic shortcuts that help the attacker using additional information.

Monoalphabetic substitution ciphers are vulnerable to ciphertext-only attacks if the ciphertext and the distribution of the plaintext letters (i.e., in an English text) are known to the attacker.

The main observation is that the distribution of the letters is invariant to the permutation, and that each letter is permuted to another letter, whose frequency (in the ciphertext) is the same as the frequency of the letter in the original text.

Ciphertext-Only Attack on Substitution Cipher (cont.)

For example, the most frequent letter in an English text is e:

Letter	Frequency	Letter	Frequency	Letter	Frequency
e	12.31%	l	4.03%	b	1.62%
t	9.59%	d	3.65%	g	1.61%
a	8.05%	c	3.20%	v	0.93%
o	7.94%	u	3.10%	k	0.52%
n	7.19%	p	2.29%	q	0.20%
i	7.18%	f	2.28%	x	0.20%
s	6.59%	m	2.25%	j	0.10%
r	6.03%	w	2.03%	z	0.09%
h	5.14%	y	1.88%		

Ciphertext-Only Attack on Substitution Cipher (cont.)

The most frequent English word is the:

Word	Frequency	Word	Frequency	Word	Frequency
the	6.421%	a	2.092%	i	0.945%
of	4.028%	in	1.778%	it	0.930%
and	3.150%	that	1.244%	for	0.770%
to	2.367%	is	1.034%	as	0.764%

Breaking Monoalphabetic Substitutions

Exercise: Solve

```
UCZCS NYEST MVKBO RTOVK
VRVKC ZOSJM UCJMO MBRJM
VESZB SMOSJ DBKYE MJTRV
VEMPY JMOMJ AMVEM HKOVJ
KTRVK CZCQV ENMMV VMJOS
ZHVER OVEMP BSZTM MSOKN
PTJCI MZ
```

The frequency of the letters in this ciphertext:

Letter	A	B	C	D	E	F	G	H	I	J	K	L	M
Occurs	1	5	7	0	8	0	0	2	1	10	8	0	19
Letter	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Occurs	3	11	3	1	6	9	6	2	15	0	0	3	7

Vignere Cipher

Uses Caesar's cipher with various different shifts, in order to hide the distribution of the letters. The key defines the shift used in each letter in the text.

A key word is repeated as many times as required to become the same length as the plaintext. The result is added to the plaintext as follows:

```
Plaintext: vignerescipher
Key:       keykeykeykeykey
Ciphertext: FMEORCBIQMMNRIP
```

($a=0, b=1, \dots, z=25, \text{mod } 26$).

Vignere Cipher (cont.)

```
Plaintext: vignerescipher
Key:       keykeykeykeykey
Ciphertext: FMEORCBIQMMNRIP
```

Vignere is easy to break (Kasiski, 1863):

Assume we know the length of the key. We can organize the ciphertext in rows with the same length of the key. Then, every column can be seen as encrypted using Caesar's cipher, and we can find the corresponding letter by observing the distribution.

The length of the key can be found using several methods:

1. If short, try 1, 2, 3, ...
2. Find repeated strings in the ciphertext. Their distance is expected to be a multiple of the length. Compute the gcd of (most) distances.
3. Use the index of coincidence.

Vignere Cipher (cont.)

Exercise: Solve

```
KBPYU AXTGV HANWV IQNTT GZRUC ZFCXM
GGPOB LVPXQ GMHLG MAGNT LUJMG DCBAC
TEHJH EHIGC IDTWF FKAWY AAKCU BLATI
MAJMW FKULC NNRYV VXBCV WQRZX YFWNR
JQLNC ELMPM TDVMK RJMHR JMKRQ OXRTQ
WMHBA CUTXC RGYCG TBLIQ GRJMB PVQIQ
```

Distances between the triplets: 4, 8, 32, 36.

Vigenere Cipher (cont.)

We now repartition the text into quartets. This way, in each quartet, the first letter is encrypted by the same key, the second letter is encrypted by the same (but not necessarily the same as the first) key, etc.

KBPY
UAXT
GVHA
NWVI
QNTT
GZRU
CZFC
X...

Vigenere Cipher (cont.)

The frequency of the letters in this ciphertext in the first column:

Letter	A	C	F	G	H	I	J	K	N	P	Q	R	T	U	V	X	Y
Occurs	1	4	1	6	2	1	6	3	3	2	2	2	4	3	3	1	1

The frequency of the letters in this ciphertext in the second column:

Letter	A	B	C	E	G	I	M	N	O	P	Q	T	U	V	W	Z
Occurs	3	4	1	3	1	1	10	2	2	1	4	3	1	3	3	3

We can guess that in the second column $e \rightarrow M$ (i.e., the second letter of the Vigenere key is I).

In the first column, we might suspect that $e \rightarrow G/J$ (i.e., the first letter of the Vigenere key is either C or F).

Vigenere Cipher (cont.)

Problem:

Given two finite distribution vectors V and U of n elements how can we determine the shift of U with respect to V ? Given the distribution vector of the plaintext (V), and the distribution vector of the letter in some column (U), how we can determine the shift of U with respect to V (i.e., the key)?

Solution:

We score each shift of the vectors. The *distance* between V and U can be measured by $d(V, U) = \sum_{i=0}^{n-1} (V_i - U_i)^2 = \sum V_i^2 - 2 \sum V_i U_i + \sum U_i^2$, where V_i is the probability of the element i in the probability vector V . We would like a smaller distance to have a larger score, so we take the minus of the distance. Moreover, as we are interested in the relative scores of the shifts, we can eliminate the constants factors.

We define the *mutual index of coincidence* of V and U as

$$MI_c(V, U) = \sum_{i=0}^{n-1} V_i \cdot U_i$$

Vigenere Cipher (cont.)

We compute the MI_c between V and various shifts of U , and take into consideration only those with high enough MI_c value.

The difference between guessing which encrypted letter is e and checking MI_c , is that guessing e is equivalent to matching the peaks of the probability vectors, while using MI_c use more information (other entries) to find the true value.

Vigenere Cipher (cont.)

Back to our example, we define V^j to be the probability vector of the first column shifted to the left by j places, and compute the various $MI_c(V_{plaintext}, V^j)$. Where $V_{plaintext}$ is the expected probability vector related to the plaintext.

Shift	0	1	2	3	4	5	6
$MI_c(V_{plaintext}, V^j)$	0.032	0.041	0.066	0.035	0.030	0.043	0.049
Shift	7	8	9	10	11	12	13
$MI_c(V_{plaintext}, V^j)$	0.031	0.031	0.045	0.027	0.028	0.031	0.047
Shift	14	15	16	17	18	19	20
$MI_c(V_{plaintext}, V^j)$	0.036	0.048	0.041	0.046	0.037	0.037	0.035
shift	21	22	23	24	25		
$MI_c(V_{plaintext}, V^j)$	0.041	0.035	0.028	0.044	0.034		

Thus, we conclude that the first letter of the key is c .

Vigenere Cipher (cont.)

The described approach does not take into consideration biased texts, e.g., texts with "too many" z 's or only a few e 's.

Another possible approach is to compute the MI_c between pairs of columns (various positions in the key), and get the difference of their shifts. Then we can normalize the text by adding the differences of the shifts, and get a ciphertext whose all columns are shifted by the same number of letters, i.e., encrypted by a simple Caesar cipher. Finally, we solve the Caesar cipher (either by exhaustive search or using MI_c) to get the plaintext.

Finding the Key Length using the Index of Coincidence

Write the ciphertext, and below it write the ciphertext shifted by j locations; count the number of characters that are identical in the same location in both lines. The peaks are expected to be when j is a multiplication of the key length.

KBPYUAXTGVHANWVIQNTTGZRUZFCXMGGOBLVPXQGMHLGMAGNT...
 ... KBPYUAXTGVHANWVIQNTTGZRUZFCXMGGOBLVPXQGMHLGMAGNT...
 ... 00000000000000000000100000000000000000110000...

For example in the above text:

Shift	0	1	2	3	4	5	6
Index of Coincidence	180	6	4	7	18	6	5
Shift	7	8	9	10	11	12	13
Index of Coincidence	8	12	4	13	7	11	6
Shift	14	15	16	17	18	19	20
Index of Coincidence	9	2	10	7	5	4	7