JASS 05

Seminar: Algorithms for IT Security

Classical Cryptography

Ilya Saverchenko

June 6, 2005

Abstract

Cryptography is a study of secret writing. It allows two people, usually referred to as Alice and Bob, to communicate over an insecure channel in such a way that an opponent, Eve, cannot understand what is being said. For many centuries people have used different cryptographic algorithms. The most well known historical cryptosystems are:

- The Shift Cipher
- The Substitution Cipher
- The Affine Cipher
- The Vigenére Cipher
- The Hill Cipher
- The Permutation Cipher
- Rotor machines

All these ciphers were once considered to be secured. With the development of crypto-analysis techniques and tools it became clear that they are easily breakable. There are several kinds of attacks that one can use in order to break the ciphers:

- Ciphertext only attack
- Known plaintext attack
- Chosen plaintext attack
- Adaptive-chosen plaintext attack
- Chosen ciphertext attack
- Adaptive-chosen ciphertext attack

To break modern cryptosystems more sophisticated and complex techniques are used. Often they require specially designed algorithms and hardware. Yet there is also a universal technique, which sometimes may be the only feasible way of breaking a cipher. It is a so-called "rubber hose" cryptology. This method includes bribery, blackmail, etc.

Contents

1 Introduction

1.1 Introduction to Cryptography

Cryptography always played an important role in peoples lives. It is believed that basic cryptographic techniques were known and used by Egyptians around 4500 years ago. For many centuries cryptography was used mainly by states to achieve secrecy of communication, for example during military conflicts. A well known Caesar Cipher was used for that purpose. However there are a few examples when information hiding techniques were used in a bit different manner. Cryptography was widely used in religion not to offend dominant cultures or governments. According to [5] '666' - 'Number of the Beast' is a cryptographic way of concealing a dangerous reference such as the Roman Empire, or the Emperor Nero.

After the First World War developments in the area have received an additional boost. The military of Europe and other countries, understanding the importance of cryptography, started to contribute to most of developments and research projects. In a few dozen years great results were achieved. Rapid development of computers and communication systems in the 1960's stimulated further developments in the field. In this period, for example, the concept of public-key cryptography was introduced.

The fundamental goal of cryptography is to allow two people, Alice and Bob, to exchange messages over an insecure channel in such a way that their opponent, Eve - eavesdropper, cannot discover their content. In addition cryptography helps to ensure data integrity, non-repudiation, and authentication [2]. Cryptosystem is defined as follows:

Definition 1 *A cryptosystem is a five-tuple $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$, where the following conditions are satisfied:*

1. *$\mathcal{P}$ is a finite set of possible plaintexts*
2. *$\mathcal{C}$ is a finite set of possible ciphertexts*
3. *$\mathcal{K}$ is a finite set of possible keys, the keyspace*
4. *For each $k \in \mathcal{K}$, there is an encryption rule $e_k \in \mathcal{E}$ and a corresponding decryption rule $d_k \in \mathcal{D}$. Each $e_k : \mathcal{P} \to \mathcal{C}$ and $d_k : \mathcal{C} \to \mathcal{P}$ are functions such that $d_k(e_k(p)) = p$ for every plaintext element $p \in \mathcal{P}$.*

1.2 Mathematical Background

Before proceeding with study of famous historical cryptosystems modular arithmetics should be defined, since many algorithms are based on it.

Definition 2 *Suppose a and b are integers, and m is a positive integer. Then we write $a \equiv b (\mathrm{mod}\, m)$ if m divides $a - b$. The phrase $a \equiv b (\mathrm{mod}\, m)$ is called congruence and is read as "a is congruent to b modulo m", m is called modulus.*

For example 2 is congruent to 11 modulo 3, as $2 \bmod 3 = 11 \bmod 3 = 2$. 12 is congruent to -16 modulo 7, as $12 \bmod 7 = -16 \bmod 7 = 5$ ($-16 = -3 \times 7 + 5$). Arithmetic modulo m is defined in the following way:

Definition 3 *$\mathbb{Z}_m$ is defined to be a set $\{0, \ldots, m-1\}$, equipped with two operations, $+$ and $\times$. Addition and multiplication in $\mathbb{Z}_m$ work exactly like real addition and multiplication, except that the results are reduced modulo m.*

This definition satisfies most of the familiar arithmetic rules, e.g. addition is closed, multiplication is commutative, etc.

The notion of an equivalence class is also important. The equivalence class of an integer a is a set of all integers congruent to $a \bmod m$. E.g. if $m = 7$, then 9 and 16 are in the same equivalence class. If $a = nm + r$, where $0 \leq r < m$ and $n \geq 0$, then $a \equiv r (\bmod m)$. r is called the least residue of $a \bmod m$. It can be seen that any integer a congruent modulo m to a unique integer between 0 and $m - 1$. For example: $13 \equiv 3 (\bmod 5)$, or 3 is the least residue of $13 \bmod 5$.

2 Historical Cryptosystems

2.1 The Shift Cipher

The Shift Cipher is probably the most well know historical cipher. It is a monoalphabetic cipher, which is based on modular arithmetic. Ciphers are called monoalphabetic if, once a key is chosen, it maps each alphabetic character to a unique alphabetic character. Here is the formal definition of the Shift Cipher:

Definition 4 *Let* $\mathcal{P} = \mathcal{C} = \mathcal{K} = \mathbb{Z}_{26}$. *For* $0 \leq k \leq 25$, *define*

$$e_k(p) = (p + k) \bmod 26,$$

and

$$d_k(c) = (c - k) \bmod 26,$$

where $(p, c \in \mathbb{Z}_{26})$

As can be seen from the definition the cipher has only 26 distinct keys. The famous Caesar Cipher is a plain Shift Cipher with $k = 3$. In order to encipher a message using the Shift Cipher one has to first choose a key. Using the table below a plaintext string should be converted to a string of integers. The next step is to add value of the key to each integer reducing it modulo 26. And at last sequence of integers should be converted to a ciphertext string.

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

Decryption works in the similar way. The difference is that during description one should subtract value of the key instead of adding it.

In order to encrypt a word "julius" using the Shift Cipher with key $k = 3$ one, as described above, would convert the plaintext to a sequence of integers resulting (9 20 11 8 20 18). After 3 should be added to each of the integers reducing the result modulo 26 if needed.

$$9 + 3 = 11;\ 20 + 3 = 23;\ 11 + 3 = 14;\ 8 + 3 = 11;\ 20 + 3 = 23;\ 18 + 3 = 21$$

The resulting integer string is (12 23 14 11 23 21). Converting it to ciphertext gives "MX-OLXV".

2.2 The Substitution Cipher

The Substitution Cipher is another monoalphabetic cipher.

Definition 5 *Let $\mathcal{P} = \mathcal{C} = \mathbb{Z}_26$. $\mathcal{K}$ consists of all possible permutations of the 26 symbols 0, 1,..., 25. For each permutation $k \in \mathcal{K}$, define*

$$e_k(p) = k(p),$$

and

$$d_k(c) = k^{-1}(c),$$

where k^{-1} is the inverse permutation to k.

The cipher is one of the oldest known ciphers. Its key space consists out of 26! keys. Yet as will be demonstrated later it is fairly easy to break it using basic cryptoanalysis techniques. To encrypt a plaintext message one has to substitute all letters in the original text with the corresponding ciphertext letters, using a permutation function. In the case of English language a permutation function can be described by a mapping as given below.

a	b	c	d	e	f	g	h	i	j	k	l	m
M	I	B	A	U	P	E	G	Z	S	C	Y	W
n	o	p	q	r	s	t	u	v	w	x	y	z
Q	F	D	R	T	V	X	H	O	K	J	L	N

Using the mapping a word "secret" can be encrypted to "VUBTUX". In order to decipher a ciphertext message one has to use the inverse function.

2.3 The Vigenère Cipher

Unlike the cryptosystems described earlier, the Vigenére Cipher is a polyalphabetic cipher. That means it can map an alphabetic character to several others. The cipher is named after Blaise de Vigenére who lived in 16th century. However it was first described by Giovan Batista Belaso in 1553. The cipher is formally defined as follows:

Definition 6 *Let m be a positive integer. Define $\mathcal{P} = \mathcal{C} = \mathcal{K} = (\mathbb{Z}_{26})^m$. For a key $k = (k_1, k_2, \ldots, k_m)$, we define*

$$e_k(p_1, p_2, \ldots, p_m) = (p_1 + k_1, p_2 + p_2, \ldots, p_m + k_m),$$

and

$$d_k(c_1, c_2, \ldots, c_m) = (c_1 - k_1, c_2 - k_2, \ldots, c_m - k_m),$$

where all operations are performed in $\mathbb{Z}_{26}$.

As can be seen from the definition the number of possible keywords of length m is equal to 26^m. To encipher a message one should first convert the key and the plaintext message to a sequence of integers. For that purpose the table given below is used.

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

After the integer string corresponding to the original message must be split on n blocks of length m, where m is the length of the chosen key. The key is added modulo 26 to each block. At last the blocks are concatenated and converted to ciphertext. As with the Shift Cipher to decrypt a message one should subtract modulo 26 the key from each block. To demonstrate how the procedure works we will encrypt a string "attackatdown" using the keyword "cipher" of length 6. The numerical equivalent of k is (2 8 15 7 4 17). The plaintext is transformed to integer string (0 19 19 0 2 10 0 19 3 14 22 13). Since $m = 6$ we split the plaintext in two blocks and perform addition modulo 26.

0	19	19	0	2	10	0	19	3	14	22	13
2	8	15	7	4	17	2	8	15	7	4	17
2	1	8	7	6	1	2	1	18	21	0	4

Thus the ciphertext is "CBIHGBCBSVAE". In case if length of the plaintext is not divisible by length of the key, only part of the key can be used for encoding the last several characters of the original message.

To decrypt the ciphertext "CBIHGBCBSVAE" we follow the same sequence of steps. The numerical equvalent of k is (2 8 15 7 4 17). The ciphertext can be written using integers as (2 1 8 7 6 1 2 1 18 21 0 4). Now subtract value of the keyword modulo 26 from the ciphertext.

2	1	8	7	6	1	2	1	18	21	0	4
2	8	15	7	4	17	2	8	15	7	4	17
0	19	19	0	2	10	0	19	3	14	22	13

(0 19 19 0 2 10 0 19 3 14 22 13) corresponds to a string "attackatdown", which is indeed the same as the encrypted message.

2.4 The Hill Cipher

The Hill Cipher is another polyalphabetic cipher. It was invented by Lester S. Hill in 1929, thus it is much younger than the ciphers described up to now. The formal definition says:

Definition 7 *Let $m \geq 2$ be an integer. Let $\mathcal{P} = \mathcal{C} = (\mathbb{Z})^m$ and let*

$$\mathcal{K} = \{m \times m \text{ invertible matrices over } \mathbb{Z}_{26}\}.$$

For a key $K \in \mathcal{K}$, we define

$$e_K(p) = pK,$$

and

$$d_K(c) = cK^{-1},$$

where all operations are performed in $\mathbb{Z}_{26}$.

As can be seen from the definition this cipher applies a transformation to a plaintext, which is defined by a matrix K. To encipher a message one should first express a plaintext message as a sequence of integers.

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

The integer string should be split on blocks of length m so that each block would have a form of $P_n = (p_{n,1}, p_{n,2}, \ldots, p_{n,m})$. After multiplying each block P_n by a matrix K ciphertext is acquired. To decrypt one should find an inverse linear transformation K^{-1}. In other words find a matrix K^{-1} such that $KK^{-1} \bmod 26 = I$. To demonstrate how description works lets consider the following example. We are given ciphertext "CKHUMD" and key

$$K = \begin{pmatrix} 11 & 3 \\ 8 & 7 \end{pmatrix}.$$

To decrypt the ciphertext produced with the Hill Cipher, one should apply an inverse linear transformation K^{-1}.

$$K^{-1} = (11 \times 7 - 3 \times 8)^{-1} \begin{pmatrix} 7 & -3 \\ -8 & 11 \end{pmatrix} = \begin{pmatrix} 7 & 23 \\ 18 & 11 \end{pmatrix},$$

where $(11 \times 7 - 3 \times 8)^{-1}$ is reciprocal of the residue of $(11 \times 7 - 3 \times 8)^{-1} \bmod 26$.

2.5 The Permutation Cipher

Another old cryptosystem. It was described in a book by Giovanni Porta written in 1563. The cryptosystem is defined as follows:

Definition 8 *Let m be a positive integer. Let $\mathcal{P} = \mathcal{C} = (\mathbb{Z})^m$ and let $\mathcal{K}$ consist of all permutations of $\{1, 2, \ldots, m\}$. For a key k (e.g. a permutation), we define*

$$e_k(p_1, p_2, \ldots, p_m) = (p_{k(1)}, p_{k(2)}, \ldots, p_{k(m)}),$$

and

$$d_k(c_1, c_2, \ldots, c_m) = (c_{k^{-1}(1)}, c_{k^{-1}(2)}, \ldots, c_{k^{-1}(m)}),$$

where k^{-1} is the inverse permutation to k.

Let us consider an example. First one has to define a permutation such as the following:

p	1	2	3	4	5	6
$k(p)$	5	1	6	3	2	4

The inverse permutation is defined as follows:

c	1	2	3	4	5	6
$k^{-1}(c)$	2	5	4	6	1	3

Since length of key $m = 6$ the plaintext message has to be broken on n groups of length 6. If the last group consists of less than 6 letters a necessary number of dummy symbols should be appended. Next step is to rearrange each group according to the permutation defined previously.

To decipher a ciphertext message one should apply the inverse permutation.

2.6 Rotor Machines

In the beginning of twentieth century mechanical encryption devices started to be developed, in order to automate encryption/decryption process. They were called rotor machines. The machines were using a substitution cipher, which was rotated each cycle. The idea was not new. It was already used during the American Civil War. Probably the most well known rotor machine is Enigma.

The original Enigma was developed by Arthur Scherbius in 1919. During the Second World War Germans used a variation of the original device. It used three rotors chosen from a set of five. The three rotors were interconnected, so first rotor would turn the second each full iteration, and second would turn the third. In addition a number of extra mechanisms, a reflector for instance, were used to make the cipher more secure. Due to incorrect usage of the devices Allies eventually managed to break the code. The reading of information in the messages, Enigma did not protect anymore, is sometimes credited with ending the War at least a year earlier than it would have otherwise.

3 Cryptoanalysis

Goal of a cryptanalyst is to recover the original plaintext message without knowing the key being used or to deduce the key itself. The general assumption is that an opponent, Eve, knows the cryptosystem being used. This is referred to as Kerchkhoffs principle. The basic attack models are [3]:

- Ciphertext only attack – the cryptanalyst possesses a string of ciphertext. In other words $c = e_k(p)$ is given. The cryptanalyst should determine k or p. Any cryptosystem vulnerable to this type of attack is considered to be completely insecure.
- Known plaintext attack – the cryptanalyst possesses a plaintext message and corresponding ciphertext. Thus p and $c = e_k(p)$ are given. The cryptanalyst should determine k or find such function that will produce correct plaintext for any given ciphertext. It is a common goal of any kind of attack, thus it will be omitted in the further definitions.
- Chosen plaintext attack – the cryptanalyst can choose a message to be encrypted. So p_1, $c_1 = e_k(p_1)$, p_2, $c_2 = e_k(p_2)$, $\ldots$, p_n, $c_n = e_k(p_n)$ are given.
- Adaptive-chosen plaintext attack – the cryptanalyst can choose a message to be encrypted based on previously achieved results. This is possible if he has gained access to the encryption machine. So p_1, $c_1 = e_k(p_1)$, p_2, $c_2 = e_k(p_2)$, $\ldots$, p_n, $c_n = e_k(p_n)$ are given.
- Chosen ciphertext attack – the cryptanalyst can choose a ciphertext and obtain corresponding plaintext message. So c_1, $p_1 = d_k(c_1)$, c_2, $p_2 = e_k(c_2)$, $\ldots$, c_n, $p_n = d_k(c_n)$.
- Adaptive-chosen ciphertext attack – the cryptanalyst can select ciphertext to decrypt based on previously achieved results. This is possible if he has gained access to a decryption machine. So c_1, $p_1 = d_k(c_1)$, c_2, $p_2 = e_k(c_2)$, $\ldots$, c_n, $p_n = d_k(c_n)$.

3.1 Statistics of English Language

Often to break a cryptosystems one needs to know statistics of a language the plaintext message was written in. For our case English statistics is of a great interest. This kind of information is publicly available and can be easily found in internet, for example. English letter frequencies are given in the table below:

letter	prob.	letter	prob.	letter	prob.
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	0.020	P	0.019	Y	0.020
H	0.061	Q	0.001	Z	0.001
I	0.070	R	0.060		

The most common digrams are: 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.

The most common trigrams are: THE, ING, AND, HER, ERE, ENT, THA, NTH, WAS, ETH, FOR, DTH.

3.2 Cryptoanalysis of the Substitution Cipher

Using ciphertext-only attack we will brake the Substitution cipher. The ciphertext is given below:

```
BTLDXFETMDGLGMVMYFQEMQAPMVBZQMXZQEGZVXFTLXGUW
FVXBFWDYUXUQFQXUBGQZBMYMBBFHQXFPXGUVHISUBXZVC
MGQVXGUBFAUITUMCUTVXGZVIFFCXTMBUVBTLDXFETMDGL
PTFWZXVZQZXZMYMQAYZWZXUAHVUILXGUUELDXZMQVVFWU
PFHTXGFHVMQALUMTVMEFXFXGUXKUQXZUXGBUQXHTLKGUT
UZXDYMLUAMBTHBZMYTFYUZQXGUFHXBFWUFPIFXGKFTYAK
MTVBFWDYUXUAZQQZQUXUUQVZJXLXGTUUXGUIFFCBFOUTV
XGFVUMVDUBXVFPXGUGZVXFTLKGZBGKUTUWFVXVZEQZPZB
MQXXFXGUAUOUYFDWUQXFPXGUVHISUBX
```

The Substitution Cipher substitutes a single letter of the alphabet for another distinct letter. That means the text will have similar single letter statistic. The only difference will be that occurrences will be mixed in a random fashion. The first step would be to collect statistics of the given ciphertext. Below is the statistic for a single letter.

letter	prob.	letter	prob.	letter	prob.
A	0.023	J	0.003	S	0.005
B	0.054	K	0.015	T	0.054
C	0.010	L	0.030	U	0.120
D	0.026	M	0.061	V	0.066
E	0.018	N	0.000	W	0.023
F	0.090	O	0.005	X	0.118
G	0.066	P	0.020	Y	0.028
H	0.023	Q	0.059	Z	0.064
I	0.018	R	0.000		

The most common digrams are XG (16), GU (11), XF (8), QX (7), VX (7), BF (6), UX (6), ZQ (6). The most common trigrams are XGU (10), BFW, FPX, FXG, GZV, LDX, LXG, MQA, PXG, UBX, UQX, UXU, VXG - all appear three times in the text.

At this moment we are ready to make a few assumptions. U and X appear the most often in the ciphertext. We can assume that this letters correspond with E and T in the original

message. The most common digram in the ciphertext is XG. That means X = T and then G = H. THE is the most common trigram in English, so we can conclude that U = E. XF is a common digrams. We know that X = T. So XF can be TO or TI. O is a bit more frequent in English than I, so F = O.

At this stage the ciphertext looks as follows:

```
BTLDtoETMDhLhMVMYoQEMQAPMVBZQMtZQEhZVtoTLtheW
oVtBoWDYeteQoQteBhQZBMYMBBoHQtoPtheVHISeBtZVC
MhQVtheBoAeITeMCeTVthZVIooCtTMBeVBTLDtoETMDhL
PToWZtVZQZtZMYMQAYZWZteAHVeILtheeELDtZMQVVoWe
PoHTthoHVMQALeMTVMEotothetKeQtZethBeQtHTLKheT
eZtDYMLeAMBTHBZMYToYeZQtheoHtBoWeFPIothKoTYAK
MTVBoWDYeteAZQQZQeteeQVZJtLthTeetheIooCBFOeTV
thFVeMVDeBtVoPthehZVtoTLKhZBhKeTeWoVtVZEQZPZB
MQttotheAeOeYoDWeQtoPtheVHISeBt
```

We have discovered the following mappings: X = T, G = H, U = E, and F = O. Lets analyze QX and UQX. QX can be AT, NT, or IT. However if we consider UQX trigram, we can see that most likely Q = N. MQA is a common trigram. Taking into account that Q = N, we say that MQA = AND. Hence M = A and A = D. By now we know that Q = N, M = A, A = D, X = T, G = H, U = E, F = O. Proceeding in the same way it is not difficult to recover the complete message. The recovered message[1] with spaces added is:

```
cryptography has a long and fascinating history the most complete
nontechnical account of the subject is kahns the codebreakers this
book traces cryptography from its initial and limited use by the
egyptians some four thousand years ago to the twentieth century where
it played a crucial role in the outcome of both world wars completed
in nineteen sixty three the book covers those aspects of the history
which were most significant to the development of the subject
```

3.3 Cryptoanalysis of the Vigenére Cipher

The Vigenére Cipher is polyalphabetic, so additional cryptoanalysis techniques should be used. The given ciphertext is:

```
MRGFNIATXZQVFFNUXFFYBTCETYXIIXGZKACJLRGKQYEIX
OYYAUAPXYIJLHPRGVTSFPAYNNYURZOPHXWYXLFRNUTZBR
FKAHFWFZESYUWZMOLLBSBZBJHFPLXKHVIVMZTZHUIWAET
IUEDFGLXDIEXIYJIUXPNNEIXABVCINTVCIEZYYDAZGZIW
TYXJIKTRZLMFFKALGZNVKZXIIMXUUNAPGVXFUSMISKHVY
VOCRVXRIWTYXZOIRFNUXZNXLDUDPZGVHVOWMOYJERLAUG
LVTUXTHRBUQZTYTXORNKBASFFXGHQVDSHUYJSYHDYUWYX
YYKHVTUCDACAHXSEVGJIEFZGLXRSBXSYKOEPPNYAKTUAC
EFYILFWEAHCIAUALLZNXMVCKLRRHGFNXMOYUESKPM
```

As mentioned above, the Vigenére Cipher makes use of a keyword of length m. The first step is to determine the key. After that decryption of the message is easy. There are two techniques that can be employed. Namely the Kasiski test and the index of coincidence.

The Kasiski test was introduced in 1863 by a Prussian military officer Friedrich Kasiski. The method is based on the observation that two identical segments of plaintext will be

[1]The text was taken from Handbook of Applied Cryptography, by A. Menezes, P. van Oorschot, and S. Vanstone.

encrypted to the same ciphertext as long as they are δ positions apart ($\delta \equiv 0 (\mathrm{mod}\, m)$). Our goal is to find several identical pieces of text, each of length at least three, and record the distance between their starting position. m divides all of the distances $\delta_1, \delta_2, \ldots, \delta_n$. Hence m divides the greatest common divisor of all δ_i's.

In the ciphertext trigram TYX occurs 3 times. The starting positions are 25, 181, and 235. The distance between the first and the second is 156 symbols, between the first and the third 210. The gcd of these two numbers is 6, so we can assume that the keyword length is also 6.

Now we will use the index of coincidence to see if it gives the same result. The index of coincidence is defined as follows:

Definition 9 *Suppose $x = x_1, x_2, \ldots, x_n$ is a string of n alphabetic characters. The index of coincidence of x, denoted $I_c(x)$, is defined to be the probability that two random elements of x are identical.*

If we denote the frequencies of A, B, C, ..., Z in x by $f_1, f_2, f_3, \ldots, f_{25}$. We can choose two elements of x in $\begin{pmatrix} n \\ 2 \end{pmatrix}$ ways. There are $\begin{pmatrix} f_i \\ 2 \end{pmatrix}$ ways of choosing two same elements. Hence, we have the formula

$$I_c(x) = \frac{\sum_{i=0}^{25} \begin{pmatrix} f_i \\ 2 \end{pmatrix}}{\begin{pmatrix} n \\ 2 \end{pmatrix}} = \frac{\sum_{i=0}^{25} f_i(f_i - 1)}{n(n-1)}.$$

Index of coincidence of a string written in English is approximately equal to 0.065.

$$I_c(x) \approx \sum_{i=0}^{25} p_i^2 = 0.065$$

The same reasoning applies if x is a ciphertext string obtained using a monoalphabetic cipher. Now we rewrite the ciphertext c in the following way

$$\begin{array}{rcccccc} C_1 & = & c_1 & c_{m+1} & c_{2m+1} & \ldots \\ C_2 & = & c_2 & c_{m+2} & c_{2m+2} & \ldots \\ & & & \ldots & & \\ C_m & = & c_m & c_{2m} & c_{3m} & \ldots \end{array}$$

If $C_1, C_2, \ldots, C_m$ are constructed in such a way that m is the keyword length, then each $I_c(C_i)$ should be approximately equal to 0.065. On the other hand, if m is not the keyword length, $I_c(C_i)$ would be much more random. A completely random string would have

$$I_c \approx 26 \left(\frac{1}{26}\right)^2 = \frac{1}{26} = 0.038$$

Following table contains I_c for different values of m:

m	I_c
1	0.043
2	0.052; 0.051
3	0.05; 0.059; 0.045
4	0.049; 0.053; 0.052; 0.051
5	0.034; 0.05; 0.048; 0.038; 0.045
6	*0.063; 0.07; 0.083; 0.062; 0.071; 0.048*
7	0.033; 0.041; 0.038; 0.046; 0.041; 0.04; 0.047

This method also shows that $m = 6$.

To determine the keyword itself we use a method similar to the index of coincidence. Each substring C_i was produced using a monoalphabetic cipher. Thus defining a shift g we can use the following formula

$$M_g = \sum_{i=0}^{25} \frac{p_i f_{i+g}}{n}$$

where $f_1, f_2, \ldots, f_i$ denote the frequencies of A, B, ..., Z in the substring C_i, and n is the length of the substring. If g is the correct shift value, M_g would be roughly equal to 0.065. Now we have to find the most suitable value of M_g for each of the substrings.

i	M_g
1	*0.062*; 0.042; 0.033; 0.035; 0.041; 0.039; 0.030; 0.040; 0.036; 0.039; 0.026; 0.040; 0.043; 0.046; 0.038; 0.046; 0.032; 0.033; 0.042; 0.043; 0.037; 0.029; 0.047; 0.035; 0.032; 0.036
2	0.033; 0.037; 0.035; 0.035; 0.046; 0.042; 0.048; 0.040; 0.032; 0.028; 0.043; 0.040; 0.038; 0.046; 0.037; 0.026; 0.042; *0.065*; 0.037; 0.033; 0.041; 0.044; 0.029; 0.036; 0.038; 0.034
3	0.038; 0.030; 0.038; 0.029; 0.043; 0.041; 0.052; 0.034; 0.041; 0.041; 0.036; 0.033; 0.040; 0.040; 0.028; 0.050; 0.031; 0.025; 0.036; *0.073*; 0.039; 0.035; 0.034; 0.044; 0.033; 0.038
4	0.040; 0.043; 0.034; 0.047; 0.038; 0.031; 0.042; *0.064*; 0.037; 0.027; 0.030; 0.042; 0.036; 0.036; 0.038; 0.039; 0.043; 0.041; 0.040; 0.034; 0.044; 0.042; 0.040; 0.033; 0.027; 0.034
5	0.030; 0.037; 0.034; 0.030; 0.046; 0.047; 0.041; 0.036; 0.035; 0.043; 0.047; 0.035; 0.038; 0.035; 0.033; 0.036; 0.049; 0.034; 0.027; 0.044; *0.065*; 0.037; 0.026; 0.044; 0.045; 0.028
6	0.031; 0.039; 0.041; 0.041; 0.038; 0.044; 0.044; 0.034; 0.030; 0.037; 0.039; 0.036; 0.035; 0.039; 0.034; 0.034; 0.042; *0.059*; 0.043; 0.029; 0.036; 0.043; 0.037; 0.033; 0.039; 0.035

We have found the keyword, which is ARTHUR. The recovered plaintext message[2] with spaces added is:

```
many traces we found of him in the boggirt island where he had hid
his savage ally a huge drivingwheel and a shaft halffilled with
rubbish showed the position of an abandoned mine beside it were the
crumbling remains of the cottages of the miners driven away no doubt
by the foul reek of the surrounding swamp in one of these a staple
and chain with a quantity of gnawed bones showed where the animal had
been confined a skeleton with a tangle of brown hair adhering to it
lay among the debris
```

[2]The text was taken from Hound of the Baskervilles, by Arthur Conan Doyle.

3.4 Cryptoanalysis of the Hill Cipher

We will break the Hill Cipher using a known plaintext attack, since it can be difficult to break using only ciphertext. Suppose we possess m distinct plaintext-ciphertext pairs $P_j = (p_{1,j}, p_{2,j}, \ldots, p_{m,j})$ and $C_j = (c_{1,j}, c_{2,j}, \ldots, c_{m,j})$, where m is the key dimension. Let us define two $m \times m$ matrices $X = (c_{i,j})$ and $Y = (p_{i,j})$. Then $Y = Xk$. Now it is easy to find the key, $k = X^{-1}Y$, where $XX^{-1} \bmod 26 = I$.

Assume we have ciphertext IKNQYB and we know that the plaintext is cipher. Assume $m = 2$. $e_k(2, 8) = (8, 10)$, $e_k(15, 7) = (13, 16)$, and $e_k(4, 17) = (24, 1)$. Using the second and the third plaintext-ciphertext pairs, we come up with the following equation in the form $Y = Xk$

$$\begin{pmatrix} 13 & 16 \\ 24 & 1 \end{pmatrix} = \begin{pmatrix} 15 & 7 \\ 4 & 17 \end{pmatrix} k.$$

To find the key, we need inverse modulo of X

$$\begin{pmatrix} 15 & 7 \\ 4 & 17 \end{pmatrix}^{-1} = \begin{pmatrix} 5 & 1 \\ 8 & 9 \end{pmatrix}.$$

Now

$$k = \begin{pmatrix} 5 & 1 \\ 8 & 9 \end{pmatrix} \begin{pmatrix} 13 & 16 \\ 24 & 1 \end{pmatrix} = \begin{pmatrix} 11 & 3 \\ 8 & 7 \end{pmatrix}.$$

3.5 Other Types of Attacks

First is brute force attack. This kind of attack can be effectively used if keyspace of a cryptosystem allows it. For example it is easy to break the Caesar Cipher using the brute force attack, since there are only 26 available keys.

The other is "rubber hose" cryptography. It has nothing to do with cryptography. This kind of attack includes bribery, blackmail, and other alike methods. Yet it is one of the most effective kind of attacks on modern cryptosystems.

4 Conclusion

The cryptosystems described in this paper are not secure, thus they can not be used for their direct purpose. Yet they are not only interesting from the historical point of view. Many similar concepts and techniques are being used in modern cryptosystems. For example the DES (Data Encryption Standard) employs substitution and permutation operations. As we already know the same principals are used in many historical ciphers.

Furthermore study of historical ciphers helps in understanding of the basic techniques used for providing secrecy of information as well as type of attacks that can be used to break a cryptosystem.

References

[1] D. Stinson, Cryptography, Theory and Practice, Chapman & Hall/CRC, 2002

[2] A. Menezes, P. van Oorschot, S. Vanstone, Handbook of Applied Cryptography, CRC Press, 1996

[3] B. Schneier, Applied Cryptography, Protocols, Algorithms, and Source Code in C, John Wiley & Sons, inc., 1996

[4] D. Denning, Cryptography and Data Security, Addison-Wesley publishing company, Inc., 1982

[5] Cryptography, Wiki Books, http://en.wikibooks.org/wiki/Cryptography

[6] Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Main_Page