COMP3260
Data Security

Lecture 5

Prof Ljiljana Brankovic

Lecture Overview

1. Polyalphabetic substitution ciphers
   a) Vigenere cipher
   b) Beaufort Cipher
   c) Variant Beaufort Cipher

2. Breaking periodic polyalphabetic ciphers
   a) Index of Coincidence
   b) Kasiski Method

3. Running Key ciphers

4. Rotor Machines

5. One-Time Pads

6. Polygram Substitution Ciphers
   a) Playfair Ciphers

Classical Ciphers

- Chapter 3 textbook “Classical Encryption Techniques”
- These lecture notes (based on the text, “Cryptography and Data Security” by D. Denning [2], lecture notes by M. Miller and other sources)

  Note that in-text references and quotes are omitted for clarity of the slides. When you write as essay or a report it is very important that you use both in-text references and quotes where appropriate.

Polyalphabetic substitution ciphers

Polyalphabetic substitution ciphers conceal the single-letter frequency distribution by using multiple substitution.

The development of polyalphabetic ciphers began with Leon Battista Alberti (1404-1472), the father of Western cryptography. (He was also an artist, architect, writer, poet, priest, linguist and philosopher.)

In 1568, Alberti published a description of a 'cipher disk' that defined multiple substitutions. There were 20 letters in the outer circle, the so-called ***stabilis***, (there was no H, K, Y, J, U and W) and the numbers 1-4. In the movable inner circle, the so-called ***mobilis***, there were randomly placed letters of English alphabet plus &

(see https://en.wikipedia.org/wiki/Alberti_cipher_disk#/media/File:Alberti_cipher_disk.JPG)

Polyalphabetic substitution ciphers

Most polyalphabetic substitution ciphers are ***periodic*** substitution ciphers with period d. Given d cipher alphabets C_1, C_2, $\ldots\mathrm{Cd}$, let f_i : A Ci be a mapping from the plaintext alphabet A to the i^{th} cipher alphabet c_i, $1 \leq i \leq d$.

A plaintext message

$$M\ =\ m_1 \ldots mdm_d\,{}_{+\,1}\,{}_{\ldots}\,m_{2d} \ldots$$

is enciphered by repeating the sequence of mappings

$$f_1(m_1) \ldots f_d(md) f_1(m_{d+1}) \ldots f_d(m_{2d}) \ldots$$

In the special case when $d = 1$, the cipher is equivalent to the ***monoalphabetic*** substitution cipher.

Vigenere cipher

In ***Vigenere cipher*** the key K is a sequence of letters $K = k_1 k_2 \dots k_d$, where k_i gives the amount of shift in the i^{th} alphabet, that is,

$$f_i(x) = (x + k_i) \, mod \, n$$

***Example 1*:** Suppose the key is $K = BAND$ (that is, $K = 1\ 0\ 13\ 3$). Then the message $M = RENA\ ISSA\ NCE$ is enciphered as

$$C = E_k\,(M) = SEAD\ JSFD\ OCR$$

K	=	B	A	N	D		B	A	N	D		B	A	N
M	=	R	E	N	A		I	S	S	A		N	C	E
C	=	S	E	A	D		J	S	F	D		O	C	R

Beaufort Cipher

Beaufort cipher uses the substitution $f_i\,(x) = (k_i\ - \ x) mod\, n$

Beaufort cipher reverses the letters in the alphabet and then shifts them to the right by $k_i + 1$ positions:

$$f_i(x)\ =\ [(n-1)\ - x\ +\ (k_i\ +\ 1)]\, mod\, n$$

The same function is used for decipherment:

$$f_i^{-1}(\,c\,)\ =\ (k_i\ -\ x)\, mod\, n$$

Variant Beaufort Cipher

Variant Beaufort cipher uses the substitution $f_i(x)\ =\ (x\ -\ k_i)\, mod\, n$

Variant Beaufort cipher is the inverse of the Vigenere cipher; it is equivalent to a Vigenere cipher with key $(n\ -\ k_i)$.

Breaking periodic polyalphabetic ciphers

Recall that polyalphabetic substitution ciphers are harder to break than monoalphabetic ciphers because they conceal the single letter frequency distribution of the plaintext, while monoalphabetic ciphers preserve this distribution.

The unicity distance for periodic polyalphabetic ciphers is

$$N = \frac{H(K)}{D} = \frac{\log_2(s^d)}{D} = \frac{\log_2 s}{D} d$$

where d is the period and s is the number of possible keys for each simple substitution.

Thus, if N ciphertext characters are required to break the individual substitution ciphers, then dN characters are required to break the complete cipher.

For example, for a Vigenere cipher with period d, the number of keys for each simple substitution is $s = 26$ and

$$N = \frac{\log_2 s}{D} d \approx \frac{4.7}{3.2} d \approx 1.5d$$

Breaking periodic polyalphabetic ciphers

To break a periodic polyalphabetic cipher, a cryptanalyst must first determine the period of the cipher.

There are two helpful tools for determining the period of the cipher:

- Index of Coincidence
- Kasiski method

Index of Coincidence

The index of coincidence (IC) was introduced in the 1920s by William Friedman. It measures the variation in the frequencies of the letters in the ciphertext.

If the period of the cipher is 1 (i.e., a monoalphabetic cipher) then there will be considerable variation in letter frequencies (same as in the plaintext, that is, English text), and IC will be high. As the period increases, the variation is gradually eliminated and the IC will be low.

To derive IC, we shall first define a **measure of roughness** (MR), which gives the variation of the frequencies of individual characters relative to a uniform distribution.

$$MR = \sum_{i=0}^{n-1} (p_i - \frac{1}{n})^2$$

where p_i is the probability that an arbitrary chosen character in a random ciphertext is the i^{th} character a_i in the alphabet ($i = 0, \dots, n-1$).

Note that $\sum_{i=0}^{n-1} p_i = 1$

Index of Coincidence

For English letters we have

$$MR = \sum_{i=0}^{25} (p_i - \frac{1}{26})^2$$

$$= \sum_{i=0}^{25} {p_i}^2 - \frac{2}{26} \sum_{i=0}^{25} p_i + 26(\frac{1}{26})^2 =$$

$$= \sum_{i=0}^{25} {p_i}^2 - \frac{2}{26} + \frac{1}{26}$$

$$= \sum_{i=0}^{25} {p_i}^2 - 0.038$$

MR ranges from 0 for a flat distribution (infinite period), to 0.028 for English text and ciphers with period 1.

Note that $MR + 0.038 = \sum_{i=0}^{25} {p_i}^2$ is the probability that two arbitrarily chosen letters from the random ciphertext are the same.

Index of Coincidence

Let F_i be the frequency of the i^{th} letter of English ($i = 0, \ldots 25$); then

$$\sum_{i=0}^{25} F_i = N$$

The total number of pairs of letters in the ciphertext of length N is $\frac{N(N-1)}{2}$.

The number of pairs containing just i^{th} letter is $\frac{F_i(F_i-1)}{2}$.

The IC is defined to be the probability that two letters chosen at random from the given ciphertext are the same.

$$IC = \frac{\sum_{i=0}^{25} F_i(F_i - 1)}{N(N-1)}$$

The above is the estimate of $\sum_{i=0}^{25} {p_i}^2$ and the IC is an estimate of $MR + 0.038$.

The IC ranges from 0.038 for a flat distribution (infinite period) to 0.066 for a period of 1.

Index of Coincidence

The following table shows the expected value of *IC* for several values of period *d*.

d	1	2	3	4	5	10	large
IC	0.066	0.052	0.047	0.045	0.044	0.041	0.038

IC is a statistical measure, and it doesn't always reveal the period exactly. It rather provides a clue whether a cipher is monoalphabetic, polyalphabetic with small period or polyalphabetic with large period.

Kasiski Method

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.

Example 2. Consider the plaintext TO BE OR NOT TO BE enciphered with a Vigenere cipher with key HAM:

M	=	T	O	B	E	O	R	N	O	T	T	O	B	E
K	=	H	A	M	H	A	M	H	A	M	H	A	M	H
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, $1 \le j \le m$, the period is likely to be some number that divides most of the m intervals.

Example 3

We shall use *IC* and Kasiski method to analyse the following ciphertext.

```
ZHYME ZVELK OJUBW CEYIN CUSML RAVSR YARNH CEARI UJPGP VARDU
QZCGR NNCAW JALUH GJPJR YGEGQ FULUS QFFPV EYEDQ GOLKA LVOSJ
TFRTR YEJZS RVNCI HYJNM ZDCRO DKHCR MMLNR FFLFN QGOLK ALVOS
JWMIK QKUBP SAYOJ RRQYI NRNYC YQZSY EDNCA LEILX RCHUG IEBKO
YTHGV VCKHC JEQGO LKALV OSJED 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
```

Example 3: The frequency distribution

```
Char      Percent
A         4.0       ********
B         0.9       **
C         6.1       ************
D         2.0       ****
E         4.9       **********
F         3.5       *******
G         4.0       ********
H         3.2       ******
I         3.5       *******
J         4.6       *********
K         5.2       **********
L         5.8       ************
M         3.2       ******
N         4.6       *********
O         4.0       ********
P         2.0       ****
Q         3.8       ********
R         8.7       *****************
S         4.3       *********
T         2.0       ****
U         3.5       *******
V         4.0       ********
W         1.7       ***
X         0.6       *
Y         6.1       ************
Z         3.8       ********
```

$$IC = .04343$$

Example 3

The $IC = .04343$ indicates that this is a polyalphabetic cipher with a period of about 5.

```
ZHYME ZVELK OJUBW CEYIN CUSML RAVSR YARNH CEARI UJPGP VARDU
QZCGR NNCAW JALUH GJPJR YGEGQ FULUS QFFPV EYEDQ GOLKA LVOSJ
TFRTR YEJZS RVNCI HYJNM ZDCRO DKHCR MMLNR FFLFN QGOLK ALVOS
JWMIK QKUBP SAYOJ RRQYI NRNYC YQZSY EDNCA LEILX RCHUG IEBKO
YTHGV VCKHC JEQGO LKALV OSJED 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
```

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; the only common divisor of 51 and 72 is 3 - the period is almost certainly 3.

Running Key Ciphers

In a running key cipher, the key is as long as the plaintext. The key is typically a text in a well-known book, and is specified by the title of the book and starting position (for example, Chapter 2, Paragraph 3). The cipher is typically substitution based on shifted alphabet (e.g., a nonperiodic Vigenere cipher).

Example 4: The key is a text starting with “The second cipher…” and the plaintext starts with “The treasure is buried…”.

M:	THETREASUREISBURIED
K:	THESECONDCIPHERISAN
C:	*MOILVGOFXTMXZFLZAEQ*

Running Key Ciphers

Although a running key cipher uses a key as long as the message, it is not unbreakable. Friedman (1918) observed that a large proportion of letters in the ciphertext comes from the encipherment where both key and plaintext letters fall in the high frequency category.

Example 5: In our previous example, 12 out of 19 ciphertext pairs come from high frequency pairs:

M:	THETREASUREISBURIED
K:	THESECONDCIPHERISAN
C:	*MOILVGOFXTMXZFLZAEQ*

6 of the remaining 7 pairs have either the plaintext or the key letter belonging to the high frequency category.

To break the cipher we start with the assumption that all ciphertext letters correspond to high frequency pairs. In this way we reduce the number of initial possibilities for each pair, and then we use diagram and trigram distributions to verify the initial guesses and determine the actual pairs.

Running Key Ciphers

Example 6: We consider the first three ciphertext letters in the previous example (MOI), and we examine the possible pairs for each of the three letters. For M we get (the high frequency pairs are underlined):

Plaintext	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
Key	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
Ciphertext	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M

The high frequency pairs for all three letters are:

M: E-I, I-E, T-T

O: A-O, O-A, H-H

I: A-I, I-A, E-E, R-R

There are $3 * 3 * 4 = 36$ possible combinations of pairs. Many of them produce highly unlikely trigrams. Some of the trigrams are shown below. Trigram THE occurring in both plaintext and key is the most likely.

plaintext: EAA EAI … THE … THR

key: IOI IOA …THE … THR

ciphertext: MOI MOI …MOI … MOI

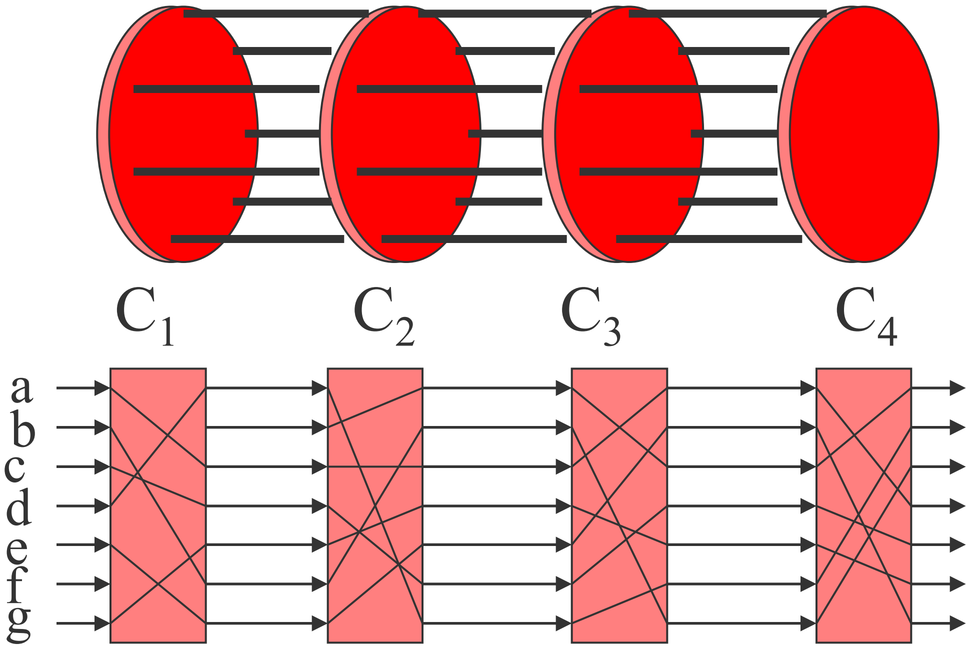
Rotor Machines

Rotor machines are used to implement polyalphabetic ciphers with a long period. A Rotor machine consists of a collection of cylinders that can rotate independently of each other.

Each cylinder has:

- 26 input pins on its front face, one for each letter in the alphabet
- 26 output pins on its rear face.

Each input pin is wired to a unique output pin. Thus each cylinder encodes a fixed permutation of the alphabet. After encoding a character in the plaintext, a cylinder is rotated; this changes the relative position of the cylinder and its neighbours.



Rotor Machines

The rotor machine encryption depends on:

- fixed permutations inside each cylinder
- initial position of each cylinder
- the rule by which the cylinders are rotated.

For a Rotor machine consists of k cylinders, the fixed permutation (mapping) inside cylinder i is defined by $f_i(a)$ and j_i denotes the position of cylinder i, then the mapping of cylinder i is defined by:

$$F_i(a) = (fi(a - ji)\; mod\; 26 + ji)\; mod\; 26$$

The mapping (encipherment) of the whole Rotor machine is:

$$F(a) = F_k(Fk_{-1(}Fk_{-2(\dots}F_2(F_1(a))\;\dots)))$$

After each of the plaintext characters is enciphered, one or more of the cylinders move to a new position, changing the encipherment of the Rotor machine. A Rotor machine with k cylinders is capable of providing 26^k different encipherments; for example, if there are 4 cylinders, there are $26^4 = 456{,}976$ different encipherments.

In practice, Rotor machines provide a period as long as the plaintext.

Rotor Machines

A Rotor machine Enigma, used by Germans in World War II, was pretty complex and included a plugboard that permuted the plaintext, and a reflecting rotor that caused each rotor to encrypt each plaintext letter twice. Enigma rotated its cylinders according to the following rule:

- -After each plaintext character is enciphered, the first cylinder advances to the next position;
- -after the first cylinder has reached a certain position, the second cylinder advances to its next position;
- - after the second cylinder has made the complete rotation, the third cylinder advances to its next position, and so on.

Enigma was broken during the World War II by Allies, first by Polish cryptographers. Germans kept modifying Enigma as the war progressed, and the British kept breaking the new versions.

A contributing factor to this successful cryptanalysis was the fact that Germans reused the code-books (keys), and had very stereotyped military messages, often starting with a same phrase.

One-Time Pads

Consider a substitution cipher whose key is a random sequence of characters, as long as the message. Such cipher is called one-time pad, and achieves perfect secrecy (recall that the perfect secrecy is achieved when the ciphertext provides no information about the plaintext - any ciphertext can be obtained from any plaintext using some key).

The computer implementation of one-time pad is based on the cryptographic device for telegraphic communications; the device was designed in 1917 by Gilbert Vernam, an employee of American Telephone and Telegraph Company (A.T. & T.).

The code used was Baudot code with 32 characters, where each character was represented as a combination of 5 marks and spaces, corresponding to bits 1 and 0.

A key was a nonrepeating random sequence of characters, also represented as marks and spaces (0's and 1's); the key was punched on a paper tape, and each key-tape was meant to be used more than once.

One-Time Pads

This cipher is known as Vernam cipher, and it generates a ciphertext bit stream

$C = E_k(M) = c_1c_2 \ldots$ where $c_i = (mi + k_i) \, mod \, 2, i = 1,2, \ldots$

The Vernam cipher is efficiently implemented on modern computers by taking exclusive-or of each plaintext/key bit pair: $c_i = mi \oplus k_i$

Deciphering is performed with the same operation:

$$mi = ci \oplus k_i$$

(To verify this, recall that $x \oplus x = 0$ and $x \oplus 0 = x$, for $x = 1 \, or \, 0$; thus $c_i \oplus k_i = mi \oplus k_i \oplus k_i = mi \oplus 0 = mi$)

Example 7: If the plaintext character A $(11000$ in Baudot) is enciphered under the key character D $(10010$ in Baudot), the resulting ciphertext character is:

```
M = 11000

K = 10010

C = 01010
```

One-Time Pads

If a key-tape is used more than once, the cipher is breakable, as it is equivalent to a running-key cipher.

To see why, suppose that two plaintext streams M and M' are enciphered with the same key stream K, giving ciphertext streams C and C'. Then

$c_i = mi \oplus k_i$ and $c_i' = mi' \oplus k_i$, for $i = 1, 2, \ldots$

Let C'' be the stream obtained by taking the exclusive-or of C and C'; then

$$c_i'' = ci \oplus ci' = mi \oplus k_i \oplus mi' \oplus k_i = mi \oplus mi'$$

Thus C'' corresponds to the encipherment of M under key M', which is equivalent to running-key cipher.

Army cryptologist Mayor Joseph Mauborgne suggested that each key-tape is used only once, and the one-time pad was born.

Polygram Substitution Ciphers

Polygram substitution ciphers encipher block of letters at the time, rather than a single letter; this makes cryptanalysis harder, as it destroys the single letter frequency distribution.

The Playfair cipher is a diagram substitution cipher invented in 1854 by Charles Wheatstone (it is named after Wheatstone's friend, English scientist Lyon Playfair).

The Playfair cipher was used by the British in World War I. The key is given by 5×5 matrix of 25 letters (J was not used). For example,

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

Playfair Cipher

A pair of plaintext letters $m_1 m_2$ is enciphered according to the following rules:

- If m_1 and m_2 are in the **same row**, then c_1 and c_2 are the two characters to the right of m_1 and m_2, respectively (the first column is considered to be to the right of the last column).
- If m_1 and m_2 are in the ***same column***, then c_1 and c_2 are the two characters below m_1 and m_2, respectively (the first row is considered to be below the last row).
- If m_1 and m_2 are in ***different rows and columns***, then c_1 and c_2 are the other two corners of the rectangle having m_1 and m_2 as corners, where c_1 is in m_1's row, and c_2 is in m_2's row.
- If $m_1 = m_2$, a null letter (for example, X) is inserted into the plaintext between m_1 and m_2 to eliminate the double.
- If the plaintext has an odd number of characters, a null letter is appended to the end of the plaintext.

Playfair Cipher

Example 8: Let the key be

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

and let the plaintext be RENAISSANCE.

Then the ciphertext is:

M =	RE	NA	IS	SA	NC	EX
C =	HG	WC	BH	HR	WF	GV

Next week:

1. Stream Ciphers
    - a) Self-Synchronising Stream Ciphers
    - b) Synchronous Stream Ciphers
    - c) Design Principles for Stream Ciphers
2. Block Ciphers
    - a) Feistel Block Cipher
3. Confusion and Diffusion
4. The Data Encryption Standard (DES)
    - a) DES Encryption/Decryption
    - b) Key Generation
    - c) Avalanche Effect
    - d) Completeness Effect

- Chapter 4. Block Ciphers and the Data Encryption Standard
- Chapter 8. Stream Ciphers
- These lecture notes (based on the text and “Cryptography and Data Security” by D. Denning [1])

References

1. W. Stallings. “Cryptography and Network Security”, Global edition, Pearson Education Australia, 2016.

2. D. Denning. “Cryptography and Data Security”, Addison Wesley, 1982.