Fundamentals of Cryptography

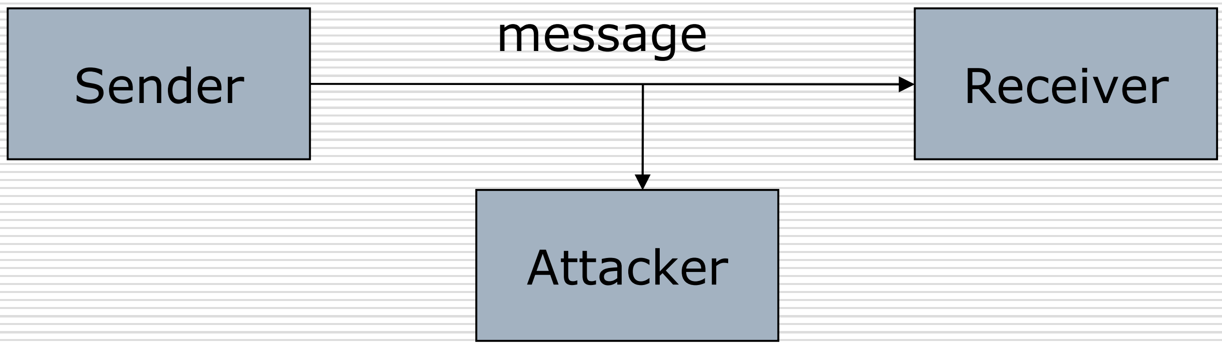
Dr. Fauzan Mirza
fauzan.mirza@seecs.edu.pk

BESE, MCS

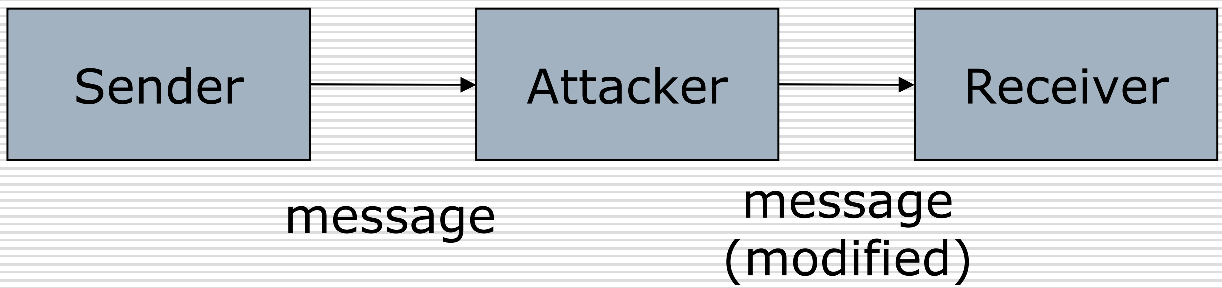
Classical Cryptology

Types of threats

- **Passive attack** (confidentiality):



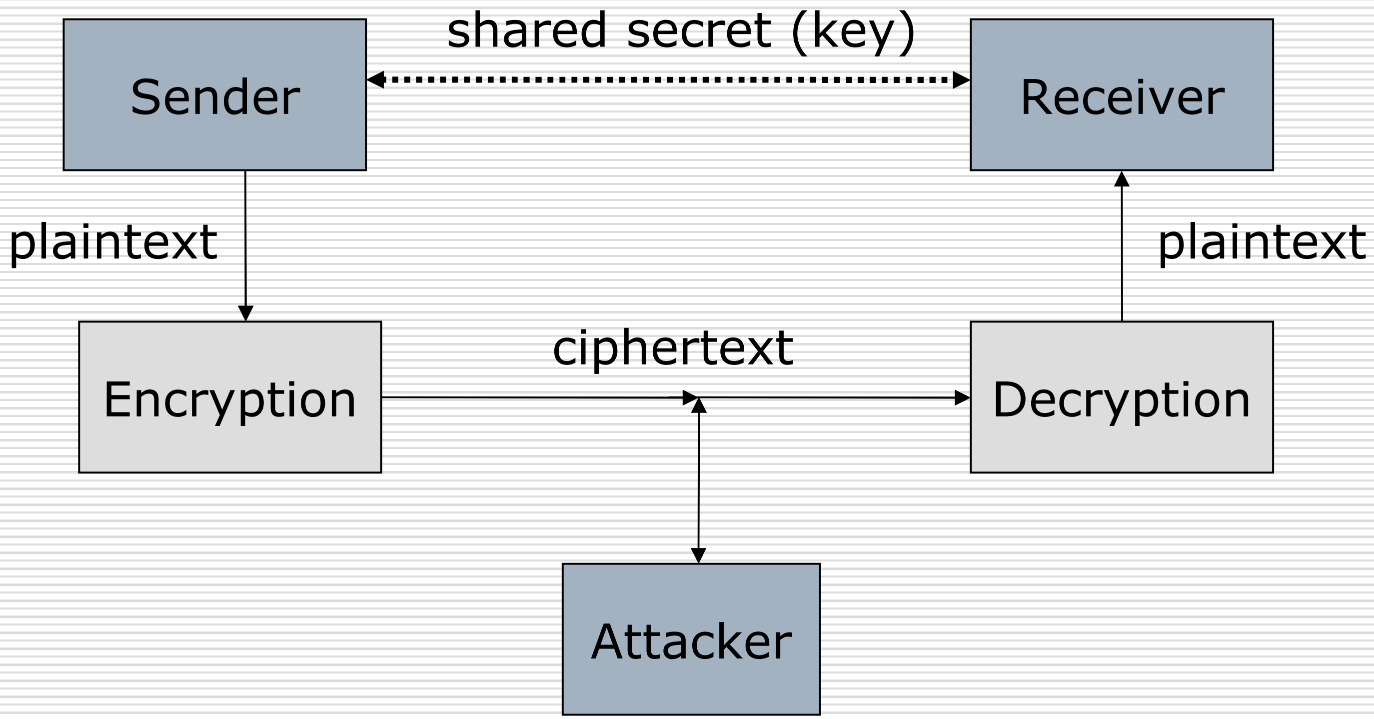
- **Active attack** (confidentiality & integrity):



Cryptography

- Cryptography means “hidden writing” (in Greek).
- It is the study of encoding meaningful information (also called the **message** or the **plaintext**) using a secret transformation function (called the **cipher**) so that nobody will understand the encoded message (called the **ciphertext**) unless they have knowledge of the cipher.
- The process of encoding plaintext to ciphertext is called **encryption**.
- The process of decoding ciphertext back to the original message (plaintext) is called **decryption**.

Conventional cryptography



Mathematical representation

- In mathematical terms, encryption and decryption can be described as follows:
  - Encryption: $c = F(p)$
  - Decryption: $p = F^{-1}(c)$
- Where the plaintext is represented by p, the ciphertext is c and the cipher is the bijective function F.
- Basically, this means that all possible ciphertexts have unique corresponding plaintexts (mathematical conditions of a bijective function are that it is *one-to-one* and *onto*)
- Both p and c are members of the set of all possible messages M and so F is a mapping $F : M \rightarrow M$

ROT-13 cipher

- The ROT-13 cipher was commonly used to hide the meaning of messages on the Internet (particularly on Usenet and E-mail).

- Each letter in the plaintext is substituted with the ciphertext letter according to the following mapping:

p	:	ABCDEFGHIJKLMNOPQRSTUVWXYZ
$F(p)$	:	NOPQRSTUVWXYZABCDEFGHIJKLM

- Example:
  - Plaintext: `THIS IS A SECRET`
  - Ciphertext: `GUVF VF N FRPERG`

- The ROT-13 cipher is an *involution* (i.e., self-inverse) so that encoding twice will result in the original message.

- This means that a separate decoding function F^{-1} is not needed.

Caesar cipher

- The Roman emperor Julius Caesar used to substitute each letter in his diplomatic communications with the letter that was three letters further along in the alphabet.

- p : `ABCDEFGHIJKLMNOPQRSTUVWXYZ`
- $F(p)$: `DEFGHIJKLMNOPQRSTUVWXYZABC`

- Plaintext: `ET TU BRUTUS`
- Ciphertext: `HW WX EUXWXV`

Caesar cipher

- The cipher used by Julius Caesar can be generalized to a function defined by a parameter k representing the number of letters that we “shift” each plaintext letter:
  $$c = F_k(p) = p + k \pmod{26}$$
- Where Julius Caesar used k=3, and ROT-13 uses k=13.
- This cipher is called the **Caesar cipher**.
- The parameter k is called the **key**.
- Example: `YVCCF Z YFGV PFL WZEU KYZJ ZEKVIVJKZEX`

Simple substitution cipher

- A generalization of the Caesar cipher, called a **simple substitution cipher** or **monoalphabetic substitution cipher**, maps plaintext letters to ciphertext letters according to a fixed mapping (the key).

- Example:
  - p : ABCDEFGHIJKLMNOPQRSTUVWXYZ
  - $F_k(p)$: QWERTYUIOPASDFGHJKLZXCVBNM

- Both the sender and receiver secretly share the key, representing the plaintext-ciphertext letter mapping, which is also called the **substitution alphabet**.

Railfence cipher

- A **transposition cipher** rearranges the plaintext letters according to a secret transformation defined by the key.
- The simplest example of this is the **railfence cipher**, in which the plaintext is written in rows of n-letter blocks (the number of columns n is the key) and then the ciphertext is read in columns
- Example:
  - Plaintext: `TRANSPOSITIONCIPHERX`
  - In this example, the key is: 5
  - Re-write as rows of 5-letter blocks:

```
TRANS
POSIT
IONCI
PHERX
```

  - Ciphertext: `TPIPROOHASNENICRSTIX`

Transposition cipher

- Problems with the railfence cipher:
  - The first and last letters of the plaintext do not move
  - The key is a number that divides the total message length
- In a **single columnar transposition** cipher, the key is a word or phrase whose letters, in alphabetic order, indicate the order of the columns as they are read
- Example:
  - Plaintext: `TRANSPOSITIONCIPHER`
  - Key is "SECRET", so re-write as rows of 6-letter blocks:

```
SECRET        CEERST
521436        123456
TRANSP        ARSNTP
OSITIO        ISITOO
NCIPHE        ICHPNE
R                R
```

  - Ciphertext: `AIIRSCSIHNTPTONRPOE`

Cryptanalysis

- **Cryptanalysis** is the study of breaking ciphers (also called **codebreaking** or **cracking**) or reading encrypted messages without knowledge of the key

- Goals:
  - Decrypt a message
  - Recover the key

- Types of attacks depend on:
  - The type of information available
  - Interaction with the cipher

Types of cryptanalytic attacks

- A **ciphertext only** attack is when an attacker has a quantity of ciphertext
  - Goal is to recover the plaintext or the key
- A **known plaintext** attack is when an attacker has a quantity of ciphertext and its corresponding plaintext
  - Goal is to recover the key
- A **chosen plaintext** attack is when an attacker can obtain a quantity of ciphertext corresponding to plaintext supplied by the attacker
  - Goal is to recover the key

Cracking the Caesar cipher

- Caesar cipher is defined by:

$$c = F_k(p) = p + k \pmod{26}$$

- There are only 26 possible values of k (the key)
- Out of these 26, only 25 values of k are valid keys (since k=0 has no effect on the plaintext)
- We can break a Caesar cipher by trying all 25 possible valid keys
- This is called an **exhaustive key search**.

Example: Exhaustive key search

- Suppose we have the ciphertext:

  TYQZCXLETZYDPNFCTEJ

- We decrypt the ciphertext by trying all 25 possible valid keys:

1. UZRADYMFUAZEQOGDUFK
2. VASBEZNGVBAFRPHEVGL
3. WBTCFAOHWCBGSQIFWHM
4. XCUDGBPIXDCHTRJGXIN
5. YDVEHCQJYEDIUSKHYJO
6. ZEWFIDRKZFEJVTLIZKP
7. AFXGJESLAGFKWUMJALQ
8. BGYHKFTMBHGLXVNKBMR
9. CHZILGUNCIHMYWOLCNS
10. DIAJMHVODJINZXPMDOT
11. EJBKNIWPEKJOAYQNEPU
12. FKCLOJXQFLKPBZROFQV

- GLDMPKYRGMLQCASPGRW

1. HMENQLZSHNMRDBTQHSX
2. INFORMATIONSECURITY
3. JOGPSNBUJPOTFDVSJUZ
4. KPHQTOCVKQPUGEWTKVA
5. LQIRUPDWLRQVHFXULWB
6. MRJSVQEXMSRWIGYVMXC
7. NSKTWRFYNTSXJHZWNYD
8. OTLUXSGZOUTYKIAXOZE
9. PUMVYTHAPVUZLJBYPAF
10. QVNWZUIBQWVAMKCZQBG
11. RWOXAVJCRXWBNLDARCH
12. SXPYBWKDSYXCOMEBSDI

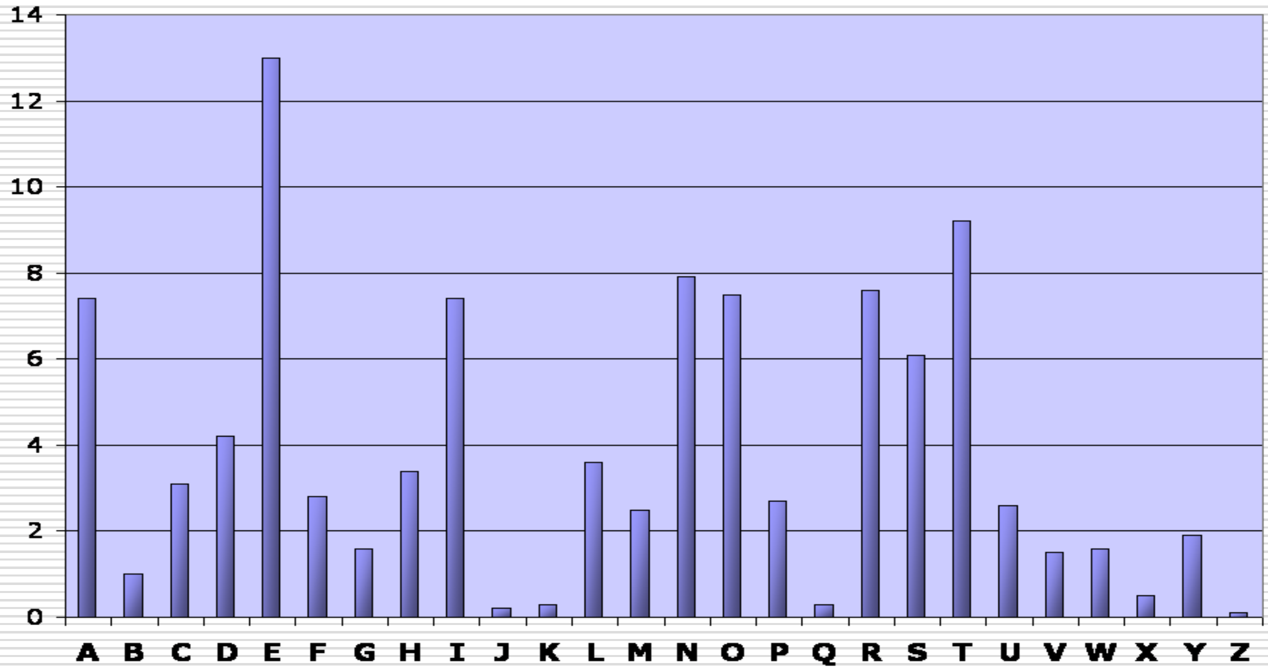
Cracking the simple substitution cipher

- There are 26! = ~4 x 10^{26} possible keys
- Exhaustive key search is not practical
- Simple substitution ciphers were considered strong for many centuries
- In 850 CE, Arab/Iraqi scientist Abu Yusuf Yaqub ibn Ishaq al-Kindi published his book “Risalah fi Istikhraj al-Mu’amma” (*A Manuscript on Deciphering Cryptographic Messages*), which contains the first ever published description of how to crack simple substitution ciphers
- The method he described is now known as **frequency analysis**

Frequency analysis

- The statistical distribution of letter frequencies of a message (text) written in any language tend towards a known letter frequency distribution profile of the language
- This is particularly true for long messages (i.e., the longer the text, the closer the letter frequency distributions match the language's letter frequency distributions)
- The simple substitution cipher preserves the letter frequency distributions of the plaintext in the ciphertext (i.e., information about the plaintext is leaked in the ciphertext)
- The attacker takes a frequency count of the ciphertext letters and tries to match them to the letter frequency distribution profile of the plaintext language

English language: Relative letter frequencies



Letter	A	B	C	D	E	F	G	H	I	J	K	L	M
Frequency	7.4	1.0	3.1	4.2	13.0	2.8	1.6	3.4	7.4	0.2	0.3	3.6	2.5
Letter	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Frequency	7.9	8.0	2.7	0.3	8.0	6.1	9.2	2.6	1.5	1.6	0.5	1.9	0.1

Example: Frequency analysis

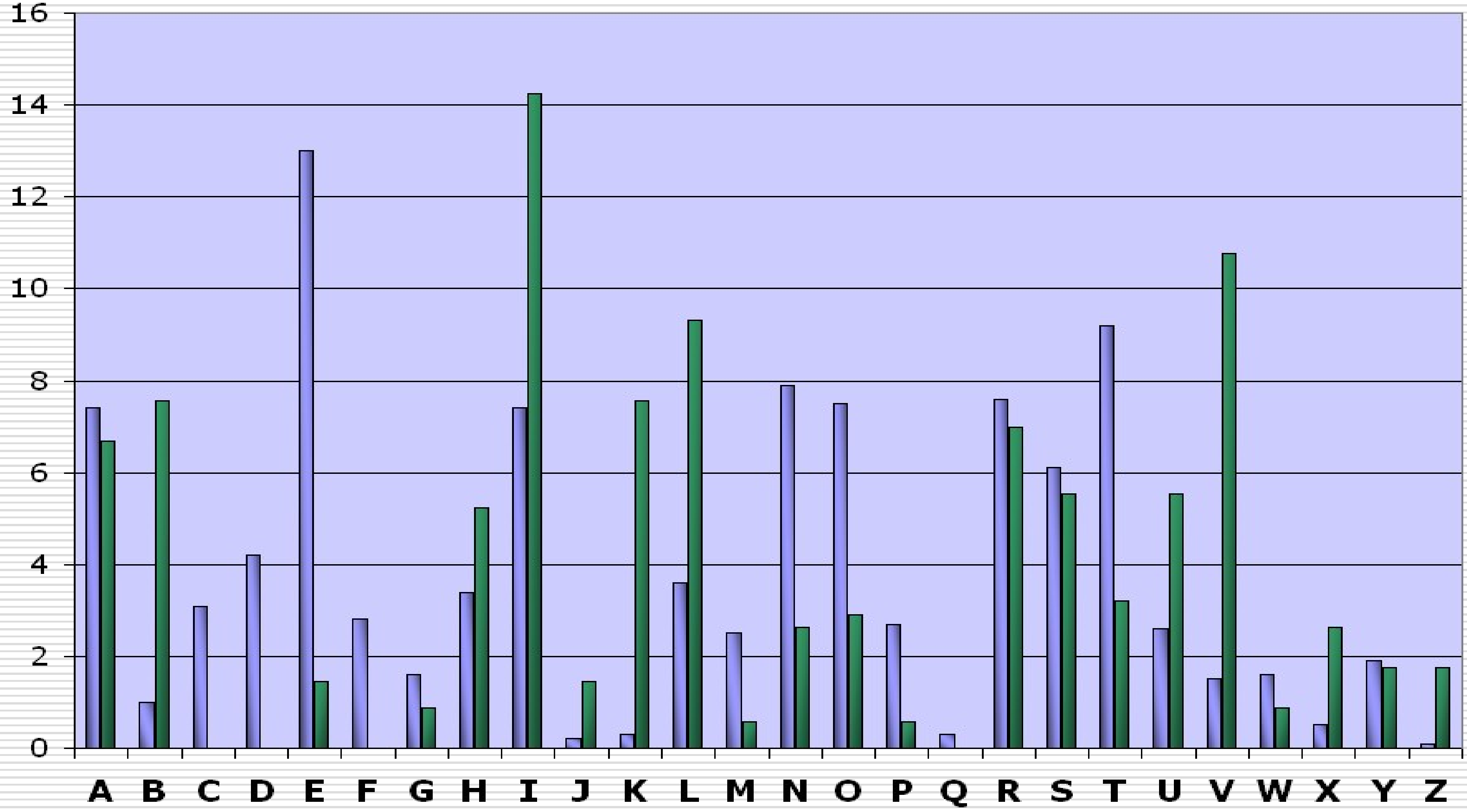
- Ciphertext:

  R jrk hbxiu lk vai vzihova ohlls lo rk rmrsvjikv ywbhtbkn. Ixise jlskbkn ai vrgiu vai ihixrvls tlzk vl vai hlyye rkt hirxiu vai ywbhtbkn. Bk vai ixikbkn, ai nivu bkvl vai ihixrvls, rkt, bo vaisi bu uljilki ihui bk vai ihixrvls -- ls bo bv zru srbkbkn varv tre -- ai nliu yrpg vl abu ohlls tbsipvhe. Alzixis, bo vaisi bu klylte ihui bk vai ihixrvls rkt bv aruk'v srbkit, ai nliu vl vai vikva ohlls rkt zrhgu wm vzl ohbnavu lo uvrbsu vl abu sllj.

- Letter frequency count (total = 344 letters):

Letter	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
Frequency	23	26	0	0	5	0	3	18	49	5	26	32	2
Letter	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
Frequency	9	10	2	0	24	19	11	19	37	3	9	6	6

Relative frequency distributions (English & ciphertext)



Example: Frequency analysis

- From the frequency distributions, we assume that:
  - The ciphertext letter **I** corresponds to the plaintext letter **E** (the most frequent letter in the English language)
  - The ciphertext letter **V** corresponds to the plaintext letter **T** (the second most frequent letter in the English language)
- Partially decrypted ciphertext (green = plaintext):

```
R jrk hbxeu lk vae vzehova ohlls lo rk
rmrsvjekv ywbhtbkn. Exese jlskbkn ae vrgeu vae
ehexrvls tlzk vl vae hlyye rkt herxeu vae
ywbhtbkn. Bk vae exekbkn, ae nevu bkvl vae
ehexrvls, rkt, bo vaese bu uljelke ehue bk vae
ehexrvls -- ls bo bv zru srbkbkn varv tre -- ae
nleu yrpg vl abu ohlls tbsepvhe. Alzexes, bo
vaese bu klylte ehue bk vae ehexrvls rkt bv
aruk'v srbket, ae nleu vl vae vekva ohlls rkt
zrhgu wm vzl ohbnavu lo uvrbsu vl abu sllj.
```

Example: Frequency analysis

- From the frequency distributions, we assume that:
  - The ciphertext letter **I** corresponds to the plaintext letter **E** (the most frequent letter in the English language)
  - The ciphertext letter **V** corresponds to the plaintext letter **T** (the second most frequent letter in the English language)
- Partially decrypted ciphertext (green = plaintext):

R jrk hbxeu lk tae tzehota ohlls lo rk rmrsvjekt ywbhtbkn. Exese jlskbkn ae trgeu tae ehexrtls tlzk tl tae hlyye rkt herxeu tae ywbhtbkn. Bk tae exekbkn, ae nevu bktl tae ehexrtls, rkt, bo taese bu uljelke ehue bk tae ehexrtls -- ls bo bt zru srbkbkn tart tre -- ae nleu yrpg tl abu ohlls tbsepthe. Alzexes, bo taese bu klylte ehue bk tae ehexrvls rkt bt aruk't srbket, ae nleu tl tae tekta ohlls rkt zrhgu wm tzl ohbnatu lo utrbsu tl abu sllj.

Example: Frequency analysis

- We can assume that the ciphertext letter **A** corresponds to the plaintext letter **H** because:
  - The digram 'TH' is the most common in the English language
  - The word "THE" is the only frequently used 3-letter English word starting with T and ending with E
- Partially decrypted ciphertext (green = plaintext):

R jrk hbxeu lk the tzehoth ohlls lo rk rmrsvjekt ywbhtbkn. Exese jlskbkn he trgeu the ehexrtls tlzk tl the hlyye rkt herxeu the ywbhtbkn. Bk the exekbkn, he nevu bktl the ehexrtls, rkt, bo these bu uljelke ehue bk the ehexrtls -- ls bo bt zru srbkbkn thrt tre -- he nleu yrpg tl hbu ohlls tbsepthe. Hlzexes, bo taese bu klylte ehue bk the ehexrtls rkt bt hruk't srbket, he nleu tl the tekth ohlls rkt zrhgu wm tzl ohbnatu lo utrbsu tl hbu sllj.

Example: Frequency analysis

- We can assume that the ciphertext letter **R** corresponds to the plaintext letter **A** because:
  - The word "THAT" is the only frequently used 4-letter English word starting with 'TH' and ending with T
  - The relative frequency of R in the ciphertext closely approximates the relative frequency of A in English
- Partially decrypted ciphertext (green = plaintext):

```
A jak hbxeu lk the tzehoth ohlls lo ak
amasvjekt ywbhtbkn. Exese jlskbkn he tageu the
ehexatls tlzk tl the hlyye akt heaxeu the
ywbhtbkn. Bk the exekbkn, he nevu bktl the
ehexatls, akt, bo these bu uljelke ehue bk the
ehexatls -- ls bo bt zau sabkbkn that tae -- he
nleu yapg tl hbu ohlls tbsepthe. Hlzexes, bo
taese bu klylte ehue bk the ehexatls akt bt
hauk't sabket, he nleu tl the tekth ohlls akt
zahgu wm tzl ohbnatu lo utabsu tl hbu sllj.
```

Example: Frequency analysis

- We can assume that the ciphertext letter **K** corresponds to the plaintext letter **N** because:
  - The words “AN” and “AT” are the only frequently used 2-letter English words starting with A
  - The relative frequency of K in the ciphertext closely approximates the relative frequency of N in English
- Partially decrypted ciphertext (green = plaintext):

A jan hbxeu ln the tzehoth ohlls lo an amasvjent ywbhtbnn. Exese jlsnbnn he tageu the ehexatls tlzn tl the hlyye ant heaxeu the ywbhtbnn. Bn the exenbnn, he nevu bntl the ehexatls, ant, bo these bu uljelne ehue bn the ehexatls -- ls bo bt zau sabnbnn that tae -- he nleu yapg tl hbu ohlls tbsepthe. Hlzexes, bo taese bu nlylte ehue bn the ehexatls ant bt haun't sabnet, he nleu tl the tekth ohlls ant zahgu wm tzl ohbnatu lo utabsu tl hbu sllj.

Example: Frequency analysis

- We assume that:
  - The ciphertext letter **T** corresponds to the plaintext letter **D** (from the word '`ant`')
  - The ciphertext letter **B** corresponds to the plaintext letter **I** (from the words '`bt`' and '`bn`')

- Partially decrypted ciphertext (green = plaintext):

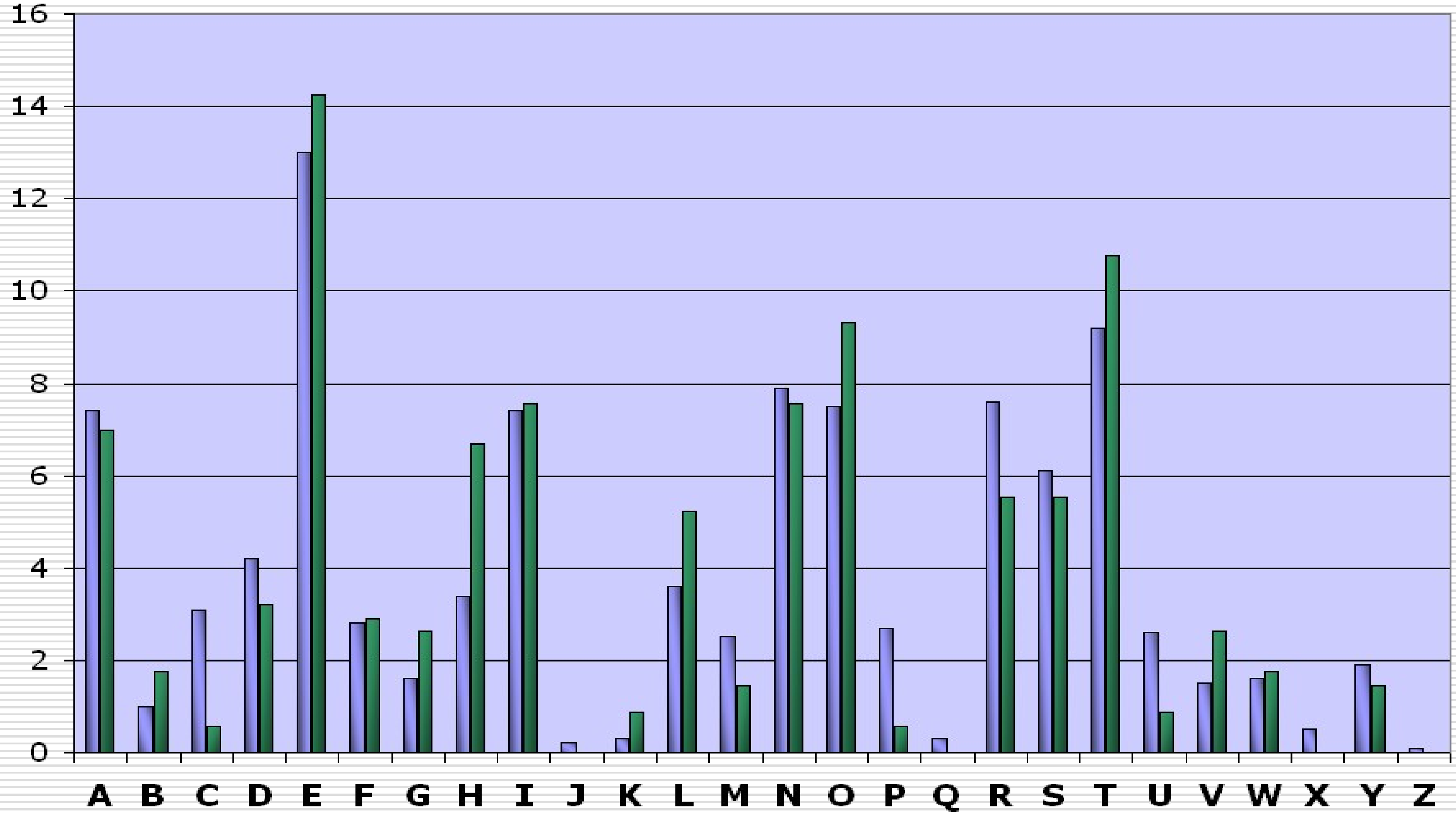
```
A jan hixeu ln the tzehoth ohlls lo an
amasijent ywihtinn. Exese jlsninn he tageu the
ehexatls dlzn tl the hlyye and heaxeu the
ywihdinn. In the exeninn, he nevu intl the
ehexatls, and, io these iu uljelne ehue in the
ehexatls -- ls io it zau saininn that dae -- he
nleu yapg tl hbu ohlls tisepthe. Hlzexes, io
taese iu nlylte ehue in the ehexatls ant it
haun't sainet, he nleu tl the tekth ohlls and
zahgu wm tzl ohinatu lo utaisu tl hiu sllj.
```

Example: Frequency analysis

- If you continue like this, completing words (using your knowledge of the English language) and matching ciphertext letters with probable plaintext letters (using the relative frequencies), you will eventually obtain a complete decryption of the message and will also have recovered the key (the substitution alphabet)

- The substitution alphabet for this example is:
  - p : ABCDEFGHIJKLMNOPQRSTUVWXYZ
  - $F_k(p)$: RYPTIONABFGHJKLMQSUVWXZDEC

Relative frequency distributions (English & plaintext)



Other English language features

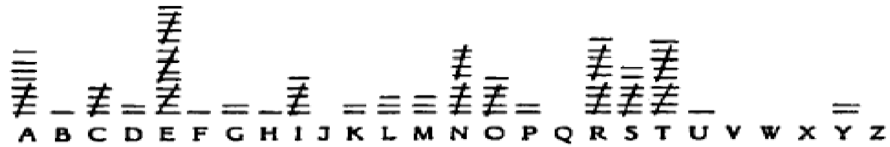
- Digram frequencies
  - Common digraphs: EN, RE, ER, NT, TH
- Trigram frequencies
  - Common trigrams: THE, ING, THA, ENT
- Vowels other than E are rarely followed by another vowel
- The letter Q is followed only by U

Polyalphabetic ciphers

- The simple substitution cipher is weak because the attacker can exploit the fact that:
  - The letter frequency distribution of the ciphertext will match the letter frequency distribution of the plaintext
  - These will generally follow the letter frequency distribution of the plaintext language
- A simple way to defeat frequency analysis is to encipher each plaintext letter with a different substitution alphabet
- The use of multiple substitution alphabets will mean that a plaintext letter can encrypt to different ciphertext letters, thus causing the letter frequency distribution to appear “flatter” (the individual letter frequencies are averaged out)
- A cipher that uses multiple substitution alphabets is called a **polyalphabetic substitution cipher**

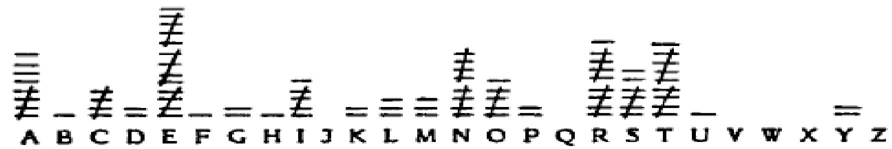
Plaintext:

AERIAL RECONNAISSANCE REPORTS ENEMY REINFORCEMENTS ESTIMATED
AT BATTALION STRENGTH ENTERING YOUR SECTOR PD CLARKE



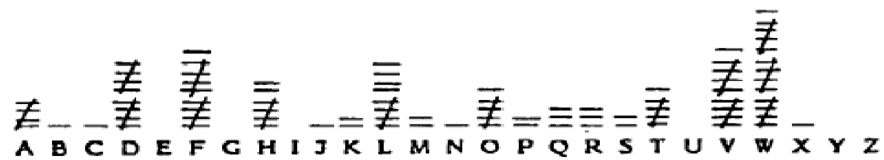
Transposition:

ANRME MTNNO ENEYM AAGGR RAPRE TLTYP IIOEN EIHOD ASRIT DOEUC
LSTNS ANNRL RASFE TSTSA ENEOS BTEER CCNRT ARRCK OEECI TEITE



Simple substitution:

LWVOL QVWAT DDLOH HLDAW VWPTV FHWDW RSVWO DNTVA WRWDF HWHFO
RLFWK LFJLF FLQOT DHFVW DMFBW DFWVO DMSTX VHWAF TVPKA QLVCW



Polyalphabetic substitution:

TARAB CZPNW TNNLL ZEFNM KLNHF OWWQM PEPVM NKRXK QNPRB FXZXE
MBXEO LFJML RWPZS GZXSS EUZYS IXWRV QZFSG FEITT HYHRW EGIKF

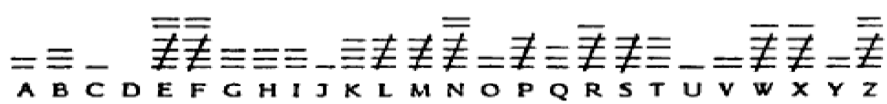


Figure 2-1. Frequency count comparison.

Vigènere cipher

- The **Vigènere cipher** is a polyalphabetic substitution cipher
- A secret word or phrase, representing the key, is agreed by the sender and receiver
- Each letter of the key is used to encrypt a plaintext letter using the Caesar cipher; each key letter represents the "shift" amount (i.e., A=0, B=1, …, Z=25).
- After the final key letter is used to encrypt a plaintext letter, the first key letter is used (again) to encrypt the next plaintext letter, and the cipher continues like this; this type of cipher is called **repeated key**.

Vigènere tableau

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

If n is the length of the key, then:

$$C_i = P_i + K_{(i + n) \bmod n}$$

Example:

Plaintext letter: T

Key letter: N

T = 19, N = 13

C = 19 + 13
= 32
= 6 (mod 26)

Ciphertext letter: G

Example: Vigènere cipher

- Plaintext: THEBOYHASTHETHEORYHE…
- Key: NUSTNUSTNUSTNUSTNUST…
- Ciphertext: GBWUBSZTFNZXGBWHESZX…

The period of a cipher

- Assuming the key is n letters long:
  - The 1st, $(n+1)$-th, $(2n+1)$-th, etc., plaintext letters are encrypted using the same substitution alphabet
  - The 2nd, $(n+2)$-th, $(2n+2)$-th, etc., plaintext letters are encrypted using the same substitution alphabet
  - In general, the i-th, $(n+i)$-th, $(2n+i)$-th, etc., plaintext letters are encrypted using the same substitution alphabet
- We say that the **period** of the cipher is n

Cracking the Vigènere cipher

- Find the key length (period)
- Write the ciphertext in n columns (where n is the period)
  - Each column represents plaintext letters that were encrypted using the same substitution alphabet
- Apply frequency analysis on each column of ciphertext letters to determine the substitution alphabet applied to the plaintext letters in that column
  - Each substitution alphabet ("shifted" alphabet) represents a key letter, so we eventually recover the entire key word/phrase

Kasiski analysis

- **Kasiski analysis** is a method to find the period of a repeated-key cipher:
  - Look for all repetitions in the ciphertext (especially long repeated sequences); note the distance between pairs of repetitions
  - The period of the cipher is probably a factor of the most frequently-noted distances between repetitions
- Example:
  - Ciphertext: `GBWUBSZTFNZXGBWHESZX`
  - Repetitions: **GBW**UBSZTFNZX**GBW**HESZX 13–1=12

    GBWUB**SZ**TFNZXGBWHE**SZ**X 18–6=12

    GBWUBSZTFN**ZX**GBWHES**ZX** 19–11=8
  - Period is very likely to be one of: 2, 4

Letter coincidence

- **Coincidence**: Picking two letters at random from a message that are identical

- Probability of picking two A's
  - Let there be n letters in the ciphertext.
  - Let there be n_a A's in the ciphertext.
  - The probability of selecting two A's at random is $\frac{n_a}{n} \times \frac{n_a - 1}{n - 1}$

Index of Coincidence

- Probability of choosing two identical letters is

$$\kappa = \frac{n_a}{n} \times \frac{n_a - 1}{n - 1} + \frac{n_b}{n} \times \frac{n_b - 1}{n - 1} + \ldots + \frac{n_z}{n} \times \frac{n_z - 1}{n - 1}$$

- Coincidence probabilities for two letters:
  - English plaintext: ~0.0667
  - Random English letters: 1/26 = ~0.0385

Index of Coincidence

- The **Index of Coincidence (IC)** is a calculation that measures the roughness of the letter distribution probabilities in a text

$$\mathrm{IC} = \frac{\sum_{i=\mathrm{A}}^{\mathrm{Z}} n_i(n_i - 1)}{N(N-1)}$$

• n_i is the number of times the letter i (A-Z) appears in the text
• N is the total number of letters in the text

- It has a number of properties that are useful to cryptanalysis:
  - It can help identify the language of plaintext
  - It can help decide whether two ciphertexts were encrypted using the same key
  - It can help determine the keylength of a repeated-key cipher

Example: Index of Coincidence

- Ciphertext (simple substitution):

  R jrk hbxiu lk vai vzihova ohlls lo rk rmrsvjikv ywbhtbkn. Ixise jlskbkn ai vrgiu vai ihixrvls tlzk vl vai hlyye rkt hirxiu vai ywbhtbkn. Bk vai ixikbkn, ai nivu bkvl vai ihixrvls, rkt, bo vaisi bu uljilki ihui bk vai ihixrvls -- ls bo bv zru srbkbkn varv tre -- ai nliu yrpg vl abu ohlls tbsipvhe. Alzixis, bo vaisi bu klylte ihui bk vai ihixrvls rkt bv aruk'v srbkit, ai nliu vl vai vikva ohlls rkt zrhgu wm vzl ohbnavu lo uvrbsu vl abu sllj.

- Index of Coincidence = 0.071903

Example: Index of Coincidence

- Ciphertext (Vigènere, polyalphabetic substitution):

  N gsg ycnxf if muy lprfxmu zdhbl gy nh sinllfrhl uhcdwvhy. Xiyjr zijgvhy ar nsdrm lar ydxiulhe xgpa ng muy dhovq tax dxnpwl gbw uhcdwvhy. Ba nzx rpwgvhy, ar awmf cfmb nzx rfwonngk, nhv, bs nzxey al fiexbhw xymw ba nzx rfwonngk -- bl ay vn otf lsbacfz gbsm quq -- ar agxf vsvx ng avm xebij wvlwvgfq. Abqworl, ay gbwkr ck gbvgwl ydlr cf muy werpsmbl sgq cl anmf'm euagrx, zx tiwl gi lar nwggb xebij tax otyek nc noh sfazunk hs mltvlk mb bal eigf

- Index of Coincidence = 0.043562

Example: Index of Coincidence

- Table of key lengths (number of alphabets) and corresponding average ICs:

Key Length	Average of ICs for each alphabet
1	0.043562
2	0.051680
3	0.044336
4	0.070315
5	0.045298
6	0.051872
7	0.042896
8	0.070460
9	0.047614
10	0.055396
11	0.042681
12	0.069011

- IC is maximum at key lengths of 4, 8, 12
- We assume that the key length is 4

Countering frequency analysis

- If Vigènere key is very long, frequency analysis will not work

- Problem: Long keys are hard to remember

- Solution: Use multiple encryption
  - Encrypting with a key m and key n is same as encryption by key whose length is the least common multiple of m and n
  - If m and n are relatively prime (i.e., their greatest common divisor is 1) then the least common multiple is mn

Advanced classical ciphers

Advanced classical ciphers:

- Operate on multiple characters of plaintext to produce multiple characters of ciphertext (i.e., operate on blocks); these are called **polygraphic ciphers**
- May apply a combination of substitution and transposition (one operation followed by another); these are called **product ciphers**

- Examples:
  - *Playfair* (invented by Charles Wheatstone, used by the British army during the second Boer war and in the first World War)
  - *ADFGVX* (invented by Colonel Fritz Nebel, used by the German army during the first World War)

- These ciphers are not secure because they can still be broken, although their cryptanalysis is considerably more complicated compared to simple substitution and Vigènere

Lessons from cryptanalysis

- The security of the cipher should depend entirely on the key (i.e., we should assume that the attacker knows how the cipher works); this is known as **Kerckhoffs' principle**
- The a large key space only makes exhaustive key search impractical; it does not mean that the cipher cannot be broken by some other method