Cryptography (and Information Security)
6CCS3CIS / 7CCSMCIS

Prof. Luca Viganò

Department of Informatics
King's College London, UK

First term 2020/21

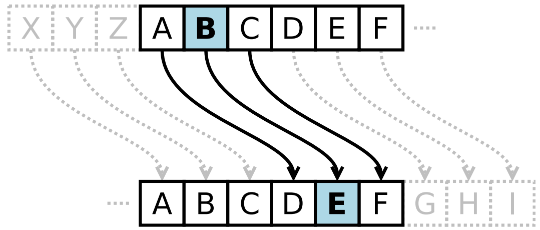
Lecture 3.1: Substitution ciphers — Caesar cipher and mono-alphabetic substitution ciphers

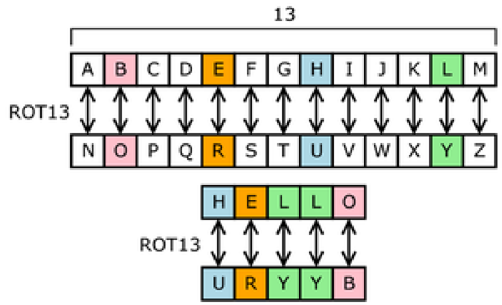
Table of contents I

Substitution ciphers

A substitution cipher is one in which the letters of plaintext are replaced by other letters or by numbers or symbols.

If the plaintext is viewed as a sequence of bits, then substitution involves replacing plaintext bit patterns with ciphertext bit patterns.





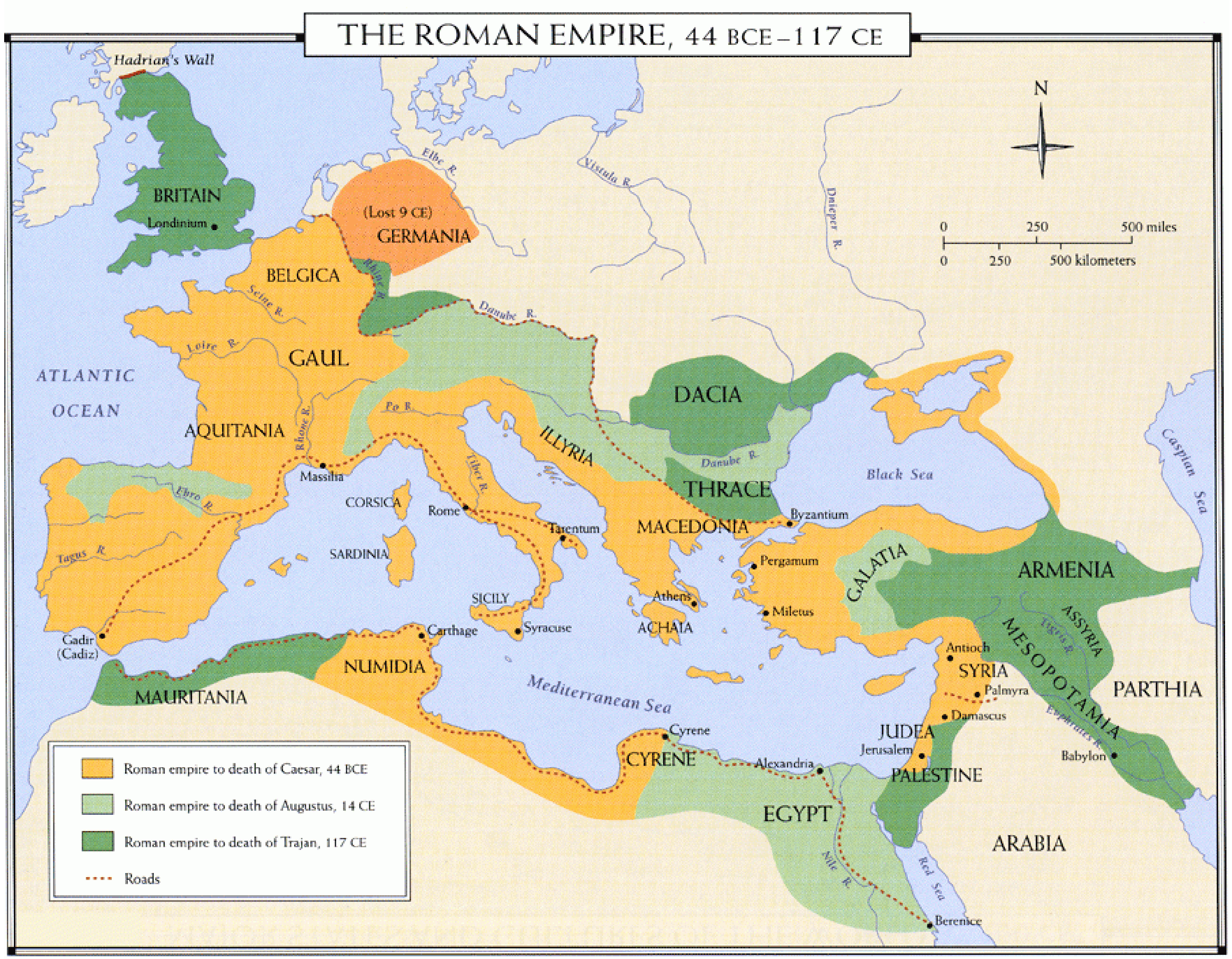
- Some simple substitution ciphers:
  - `KHOOR ZRUOG` = `HELLO WORLD`
    Caesar cipher: each plaintext character is replaced by character 3 to the right modulo 26.
  - `Jnf vg n pne be n png V fnj ?` =
    `Was it a car or a cat I saw ?`
    ROT13: shift each letter by 13 places.
    Under Unix-like systems:
    `tr a-zA-Z n-za-mN-ZA-M`
  - `1-24-4 1-24-4` = `BYE BYE`
    Alphanumeric: substitute numbers for letters.

How hard are these to cryptanalyze? Caesar? General?

Table of contents I

Caesar cipher

- “Earliest” known, simple, substitution cipher, used by Julius Caesar (see his ‘De Bello Gallico’, but also the Asterix comics!).



Caesar cipher

- “Earliest” known, simple, substitution cipher, used by Julius Caesar (see his ‘De Bello Gallico’, but also the Asterix comics!).
- Implemented by cipher disks to encrypt a letter with the third letter to the right, i.e., replace each letter of the alphabet with the letter standing 3 places further down the alphabet (wrapping around):

```
plain:   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
cipher:  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
```

Example:

```
plain:   meet me after the toga party
cipher:  PHHW PH DIWHU WKH WRJD SDUWB
```

- Mathematically, give each letter a number:

a	b	c	d	e	f	g	h	i	j	k	l	m
0	1	2	3	4	5	6	7	8	9	10	11	12

n	o	p	q	r	s	t	u	v	w	x	y	z
13	14	15	16	17	18	19	20	21	22	23	24	25

then: $C = E(3, P) = (P + 3) \mod 26$

- In general, for $K \in \{1, \ldots, 25\}$:

$$C = E(K, P) = (P + K) \mod 26$$
$$P = D(K, C) = (C - K) \mod 26$$

- We count modulo 26 because from 0 to 25 there are 26 numbers (0, 1, 2, ... 25), so we can only count up to 25 and when we reach 26, we start again from 0.
- This is exactly the same as the way our watches work: we count up to 12 and then, 13 is the same as 1, and so on

but we could also equally consider a watch that counts up to 24

Why is brute-force cryptanalysis possible?

- Three important characteristics of this problem enabled us to use a brute-force cryptanalysis:
  1. The encryption and decryption algorithms are known.
  2. There are only 25 keys to try.
  3. The language of the plaintext is known and easily recognizable.
- What generally makes brute-force cryptanalysis impractical is the use of an algorithm that employs a large number of keys, e.g.
  - triple DES algorithm makes use of a 168-bit key, giving a key space of 2^{168} ($> 3.7 \times 10^{50}$) possible keys.
- If language of plaintext is unknown, then output may not be recognizable, e.g.
  - L WRSL QRQ DYHYDQR QLSRWL = i topi non avevano nipoti (which is Italian for “the mice had no grandsons”)

  There could also be two possible sensible plaintexts in two different languages, e.g.
  - the ciphertext AMBC can be decrypted to the English plaintext CODE (with key 24) or to the Sanskrit SETU, which means “bridge” in English (with key 8)

Substitution ciphers: a little bit of history (& geography)

- **Kama Sutra cipher**:
  - Kama Sutra: a text written in the 4th century AD by the Brahmin scholar Vatsyayana, but based on manuscripts dating back to the 4th century BC.
  - The Kama-sutra recommends that women should study 64 arts, including cooking, dressing, massage and the preparation of perfumes.
  - The list also includes some less obvious arts, including conjuring, chess, bookbinding and carpentry.
  - Number 45 on the list is **mlecchita-vikalpa**, the art of secret writing, advocated in order to help women conceal the details of their liaisons.
  - One of the recommended techniques involves randomly pairing letters of the alphabet, and then substituting each letter in the original message with its partner.

Table of contents I

Mono-alphabetic substitution ciphers

Key idea

Generalise Caesar cipher by allowing an arbitrary substitution.

- **Permutation** of a finite set S of elements: an ordered sequence of all elements of S, each element appearing exactly once.

  6 permutations of $S = \{a, b, c\}$: abc, acb, bac, bca, cab, cba

  In general: $n!$ permutations of a set of n elements
  (1$^{\text{st}}$ element can be chosen in 1 of n ways, 2$^{\text{nd}}$ in $n - 1$ ways, etc.).
- If the "cipher" line of a Caesar cipher can be any permutation of the 26 alphabetic characters, then there are 26! ($> 4 \times 10^{26}$) possible keys.

Such an approach is referred to as a **mono-alphabetic substitution cipher**, because a single cipher alphabet (mapping from plain alphabet to cipher alphabet) is used per message.

Mono-alphabetic substitution ciphers

Mathematically, mono-alphabetic substitution ciphers are defined as follows:

- Let $\mathcal{K}$ be the set of all permutations on the alphabet $\mathcal{A}$.
  Define for each $e \in \mathcal{K}$ an encryption transformation E_e on strings $m = m_1 m_2 \cdots m_n \in \mathcal{M}$ as

$$E_e(m) = e(m_1)e(m_2)\cdots e(m_n) = c_1 c_2 \cdots c_n = c$$

- To decrypt c, compute the inverse permutation $d = e^{-1}$ and

$$D_d(c) = d(c_1)d(c_2)\cdots d(c_n) = m$$

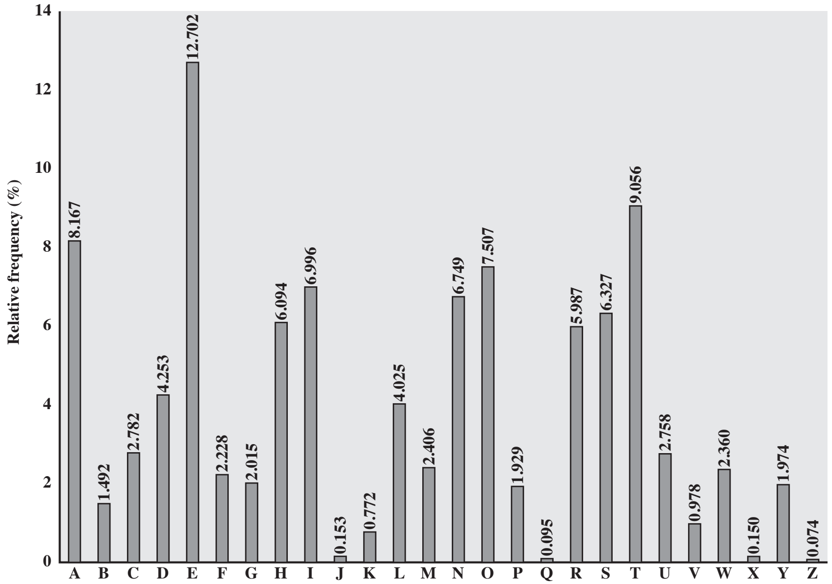
- E_e is a **mono-alphabetic substitution cipher**.

Example:

Plain:	`ABCDEFGHIJKLMNOPQRSTUVWXYZ`
Cipher:	`DKVQFIBJWPESCXHTMYAUOLRGZN`
Plaintext:	`IFWEWISHTOREPLACELETTERS`
Ciphertext:	`WIRFRWAJUHYFTSDVFSFUUFYA`

(In)security of substitution ciphers

- Key spaces are typically huge.
  26 letters $\Rightarrow 26! = 4 \times 10^{26}$ possible keys.
- This looks quite secure, doesn't it? Wrong!
- Easy to crack using frequency analysis (letters, digram, etc.).
- Frequencies for English based on data-mining books/articles:



- Easy to apply, except for short, atypical texts, e.g.,

  From Zanzibar to Zambia and Zaire, ozone zones make zebras run zany zigzags.

  $\Rightarrow$ More sophistication required to mask statistical regularities.

Example

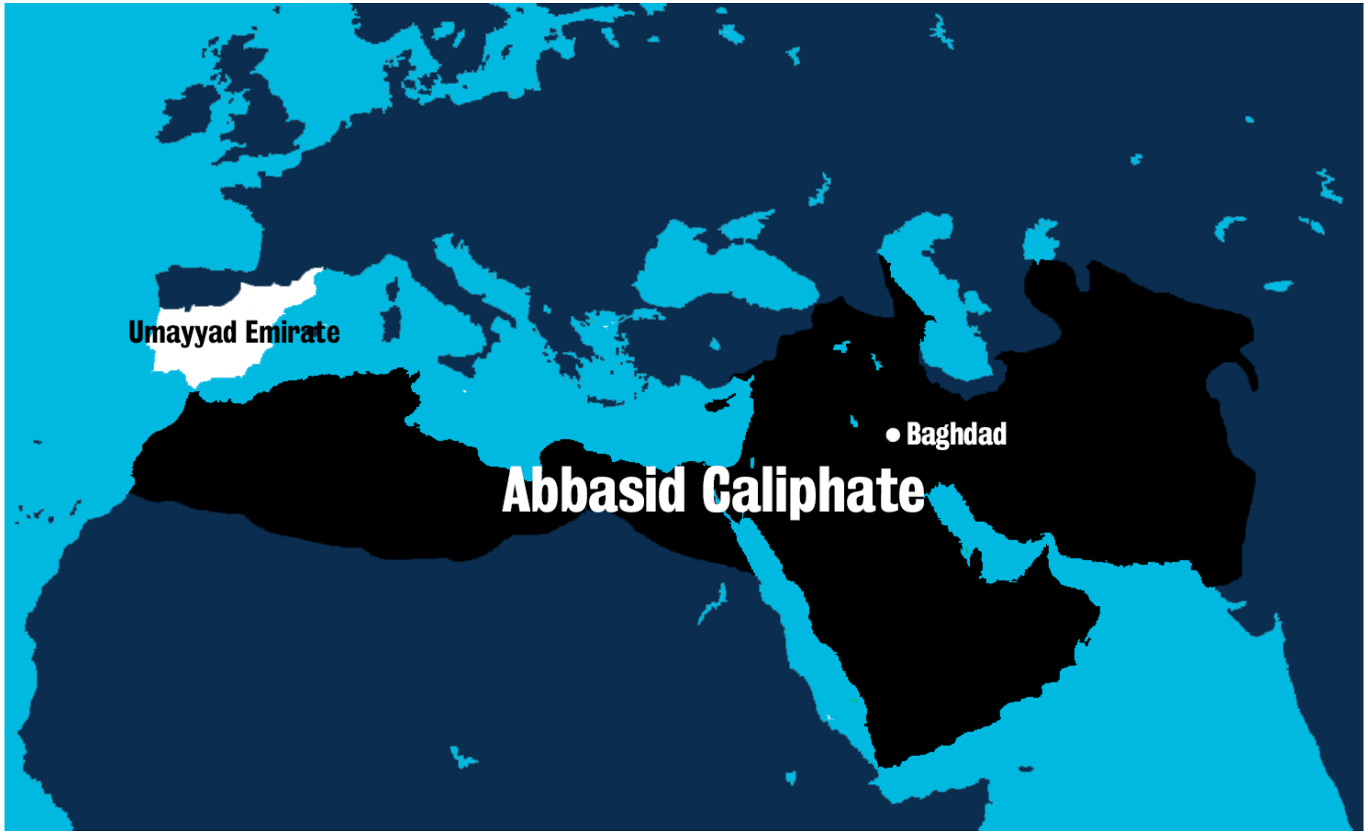
Given ciphertext:
UZQSOVUOHXMOPVGPOZPEVSGZWSZOPFPESXUDBMETSXAIZ
VUEPHZHMDZSHZOWSFPAPPDTSVPQUZWYMXUZUHSX
EPYEPOPDZSZUFPOMBZWPFUPZHMDJUDTMOHMQ

- Count relative letter frequencies.
- Since **P** and **Z** occur most frequently, guess they correspond to **E** and **T** respectively.
- Count relative **digram** (a sequence of 2 letters, a.k.a. **digraph**) frequencies.
- Since **ZW** occurs most frequently, guess it corresponds to **TH** (which is the digram occurring most frequently in English).
- Hence **ZWP** is **THE**.
- By proceeding with trial and error finally get:

IT WAS DISCLOSED YESTERDAY THAT SEVERAL INFORMAL BUT DIRECT CONTACTS HAVE BEEN MADE WITH POLITICAL REPRESENTATIVES OF THE VIET CONG IN MOSCOW

Cryptanalysis: a little bit of history (& geography)

- The Abbasid caliphate (or dynasty), started in 750 AC, heralded golden age of Islamic civilisation (arts and sciences flourished).
- A wealthy and peaceful society, which relied on an effective system of administration, and in turn the administrators relied on secure communication achieved through the use of encryption.



Cryptanalysis: a little bit of history (& geography)

- The Abbasid caliphate (or dynasty), started in 750 AC, heralded golden age of Islamic civilisation (arts and sciences flourished).
- A wealthy and peaceful society, which relied on an effective system of administration, and in turn the administrators relied on secure communication achieved through the use of encryption.
  - Many administrative manuals, such as the tenth-century Adab al-Kuttab (“The Secretaries’ Manual”), include sections devoted to cryptography — mainly monoalphabetic substitution ciphers.
- Invention of **cryptanalysis** required scholarship in many disciplines, including mathematics, statistics, linguistics and religion:
  - Theologians established the chronology of Muhammad’s revelations in the Quran by counting the **frequencies** of words contained in each revelation and considering that certain words had evolved relatively recently.
  - They also analysed individual letters, and in particular they discovered that some letters are more common than others (e.g., a and l are the most common in Arabic).