Lecture 5: Classical cryptography

Thomas Johansson

Classical cryptography

Throughout history, secret writing became an established problem and the area was named **cryptology**.

Up to very recently, cryptology was primarily concerned with military and diplomatic applications and distinguished between two disciplines

cryptography which deals with development of systems for secret writing,

cryptanalysis which analyze existing systems in order to break them.

Classical cryptography

- The solutions to different cryptographic problems are refered to as *cryptographic primitives*.
- The primitive "symmetric encryption scheme"

A Model of a Cryptosystem

Let $\mathcal{P}$ be a finite set which is called the *alphabet*.

We often use the english letters as our alphabet, we number them as $\mathsf{a} = 0$, $\mathsf{b} = 1$, $\ldots$, $\mathsf{z} = 26$. This gives $\mathcal{P} = \mathbb{Z}_{26}$, and $|\mathcal{P}| = 26$.

Cryptographic convention: use the names Alice, Bob, Caesar, and Eve (Eve is considered to be the enemy).

A Model of a Cryptosystem

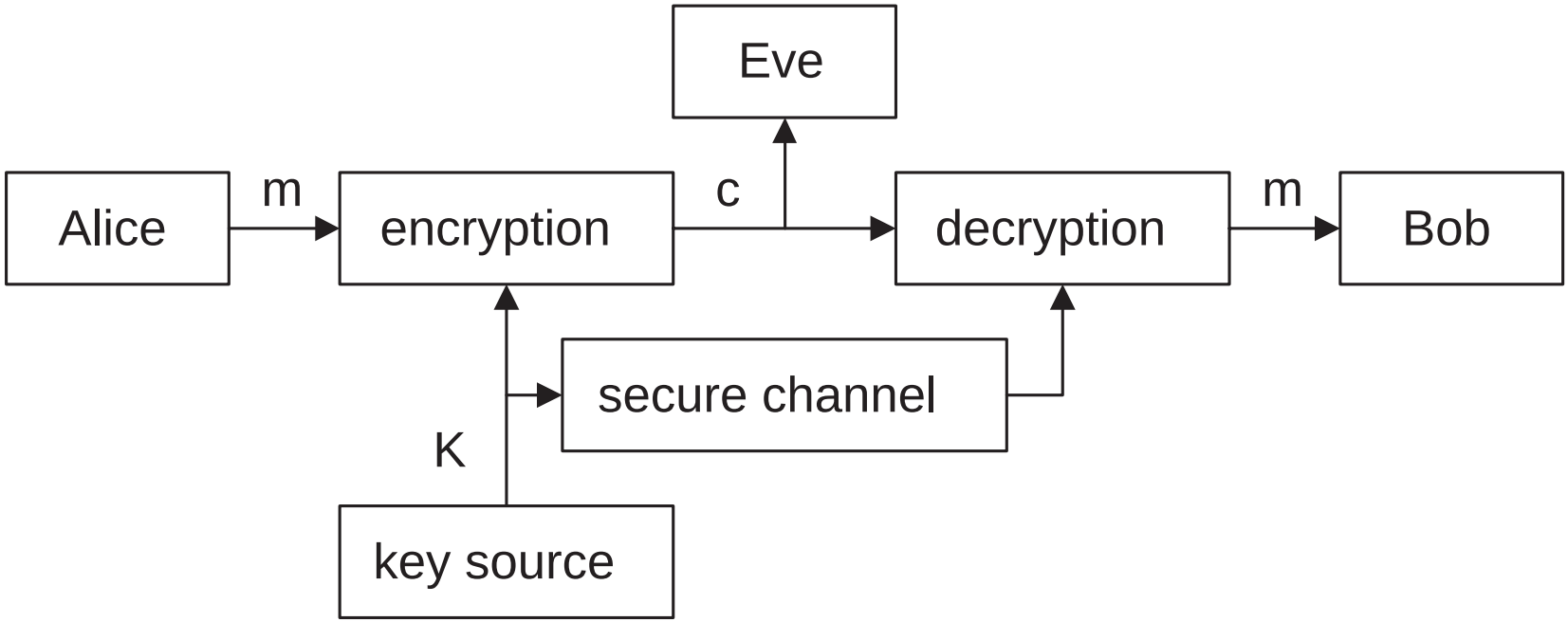

Figure: The Shannon model for symmetric encryption

A Model of a Cryptosystem

- The *plaintext* $\mathbf{m}$ consists of the produced symbols in vector form, i.e.,

$$\mathbf{m} = (m_1, m_2, \ldots, m_n),$$

- the *encryption function* $E_k()$ calculating $\mathbf{c} = E_k(\mathbf{m})$
- The key is taken from a set $\mathcal{K}$ of possible keys. Each key describes a certain function $E_k : \mathcal{P}^n \rightarrow \mathcal{C}^{n'}$
- each function $E_k()$ must be **invertible** (*injective*)

A Model of a Cryptosystem

- a *decryption function* $D_k()$ that decrypts encrypted messages back to its original form.
- $$D_k(E_k(\mathbf{m})) = \mathbf{m}, \quad \forall \mathbf{m} \in \mathcal{P}^n. \tag{1}$$
- the set of transformations $\{E_k()|k \in \mathcal{K}\}$ is known to the enemy, as well as the probability distribution for the selected key and the plaintext. *Kerckhoff's principle*.

A Model of a Cryptosystem

Different kind of attacks that can be applied.

- *Ciphertext only attacks*. A ciphertext only attack is an attack in which the enemy has access to a ciphertext $\mathbf{c}$ and tries to recover either the plaintext $\mathbf{m}$ or the secret key k.
- *Known plaintext, ...*

The Caesar cipher

- shifted each letter 3 steps. The letter a became the letter d, b became e, et.
- A generalization is to shift not three but k positions, where $k \in \{0, 1, \ldots, 25\}$ is the secret key. This cryptosystem is usually called the *Caesar cipher*.
- $$E_k(m) = m + k \pmod{26}.$$
- The decryption function is given by $$D_k(m) = m - k.$$

Cryptanalysis of the Caesar cipher

- *Exhaustive key search*
- ciphertext `wklvlvdphvvdjhwrbrx`.

k	plaintext
0	wklvlvdphvvdjhwrbrx
1	vjkukucoguucigvqaqw
2	uijtjtbnfttbhfupzpv
3	thisisamessagetoyou
4	sghrhrzldrrzfdsnxnt
5	rfgqgqykcqqyecrmwms
6	qefpfpxjbppxdbqlvlr
7	pdeoeowiaoowcapkukq
8	ocdndnvhznnvbzojtjp
9	nbcmcmugymmuaynisio
10	mablbltfxlltzxmhrhn
11	lzakaksewkksywlgqgm
12	kyzjzjrdvjjrxvkfpfl
13	jxyiyiqcuiiqwujeoek
14	iwxhxhpbthhpvtidndj
15	hvwgwgoasggoushcmci

The simple substitution cipher

- $E_k()$ operates on individual characters, but now we consider an arbitrary permutation of the alphabet as the key.
- the set of all permutations on $\mathbb{Z}_{26}$, written as

$$\{\pi_1, \pi_2, \ldots, \pi_{|\mathcal{K}|}\}.$$

- $$E_K(m) = \pi_k(m). \tag{2}$$

- $D_k(c) = \pi_k^{-1}(c)$.
(examples)

Cryptanalysis of the simple substitution cipher

- the number of different permutations on $\mathbb{Z}_{26}$ is $26! > 4 \cdot 10^{26}$. Too large!
- Instead, expoit the statistical nature of the plaintext source!

Cryptanalysis of the simple substitution cipher

- *r-grams*, $(m_i, m_{i+1}, \ldots, m_{i+r})$
- Single letters:

a	0.0804	j	0.0016	s	0.0654
b	0.0154	k	0.0067	t	0.0925
c	0.0306	l	0.0414	u	0.0271
d	0.0399	m	0.0253	v	0.0099
e	0.1251	n	0.0709	w	0.0192
f	0.0230	o	0.0760	x	0.0019
g	0.0196	p	0.0200	y	0.0173
h	0.0549	q	0.0011	z	0.0009
i	0.0726	r	0.0612		

- First approach: count the number of occurrences of the different letters in the ciphertext. Clearly, the most common letter in the ciphertext $\mathbf{c}$ is likely to correspond to the letter `e` in the plaintext.
- Better: instead of 1-grams make use of 2-grams or 3-grams.
- Our source is not *memoryless*. Large dependence between consecutive letters!
- Example: If the source has generated the two letters `wa`, the next letter can be for example `f`, `g`, `i`, etc., but the otherwise so common letter `e` can not occur.

The most common 2-grams and 3-grams and their corresponding probabilities.

th	0.0270	on	0.0154	ed	0.0111
he	0.0257	an	0.0152	te	0.0109
in	0.0194	en	0.0129	ti	0.0108
er	0.0180	at	0.0127	or	0.0108
re	0.0160	es	0.0115	st	0.0103
the	0.0215	for	0.0036	ere	0.0027
and	0.0060	tha	0.0032	con	0.0026
tio	0.0048	ter	0.0029	ted	0.0023
ati	0.0036	res	0.0027		

Example

Assume that we have received the following ciphertext:

```
xsftbiwmqooxwdxssfmtxmaibcsfiisctfzsfmibcixiiczc
awbxosbxamqiwqsaxjfqscixssaxnxiibxijcznlqlcmixnq
emtfmiczmxiufinbqvwfixsoxwlxrfmtxmflvqzixmiafwuq
jcznibclnwiczfqewvxifcmifmzqqlioqqmcibzccbxadfmx
ssnxoxrcmcawbclqxmcawqdisnuqesafigcibxiwbcoxwibc
wfwiczqdibcgqnfmrxmwxwobqsqjcaibctfzsofibibcixii
czcawbxosobqoxwibcaxetbiczqdibclxfaobqbxacwuxvca
          dzqlibcvfzxicwibcfmiczmdzqomca
```

Different letters, the most frequent ones are `i` 46 times, `c` 40 times, `x` 37 times, etc.

Frequency count on 3-grams, which is much more powerful: The most frequent 3-grams : `ibc` occuring 11 times and `icz` occuring 7 times

- This gives us $\pi(\mathtt{t}) = \mathtt{i}$, $\pi(\mathtt{h}) = \mathtt{b}$, and $\pi(\mathtt{e}) = \mathtt{c}$.
- `ibc` corresponds to a plaintext 3-gram of the form `te*`
- Conclusion: $\pi(\mathtt{r}) = \mathtt{z}$.
- start decrypting the ciphertext

  `a***ht*****a**a*****a**the**tt*e**r***thetattere**ha**ha***`

The unknown last letter in the sequence `*thetattere*` is a `d` (tattered). This gives $\pi(\mathtt{d}) = \mathtt{a}$.

This gives us some additional information and if we continue to do a "partial" decryption we find later in the plaintext the sequence `**a*theda**hter**the*...` .

We guess that the word `daughter` is present, giving $\pi(\mathtt{o}) = \mathtt{q}$, $\pi(\mathtt{f}) = \mathtt{d}$.

Polyalphabetic ciphers

Polyalphabetic ciphers operate differently on various portions of the plaintext.

Ssimplest case: a number of mono-alphabetic ciphers with different keys are used sequentially and then cyclically repeated.

The Vigenére cipher

Cyclically uses t Caesar ciphers, where t is called the *period* of the cipher.

The encryption function maps the plaintext $\mathbf{m} = m_1, m_2, \ldots$ to the ciphertext $\mathbf{c} = c_1, c_2, \ldots$, through

$$\mathbf{c} = E_{\mathbf{k}}(m_1, m_2, \ldots, m_t), E_{\mathbf{k}}(m_{t+1}, m_{t+2}, \ldots, m_{t+t}), \ldots,$$

where

$$E_{\mathbf{k}}(m_1, m_2, \ldots, m_t) = (m_1 + k_1, m_2 + k_2, \ldots, m_t + k_t), \tag{3}$$

and the key $\mathbf{k}$ consists of t characters $\mathbf{k} = (k_1, k_2, \ldots, k_t)$.

The key $\mathbf{k}$ is often chosen as a word (*keyword*).

Example

The plaintext youmustvisitmetonight is to be encrypted using a Vigenére cipher with period 4, where $\mathbf{k} =$ lucy.

y	o	u	m	u	s	t	v	i	s	i	t	m	e	t	o	...
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
l	u	c	y	l	u	c	y	l	u	c	y	l	u	c	y	...
j	i	w	k	f	m	v	t	t	m	k	r	x	y	v	m	...

The resulting ciphertext is jiwkfmvttmkrxyvm...

Cryptanalysis of the Vigenére cipher

Cryptanalysis of a polyalphabetic substitution cipher is split into two different problems,

1. determine the period of the cipher
2. reconstruct the t different substitution alphabets that have been used.

Kasiski's method

- observation: repeated portions of plaintext encrypted with the same portion of the keyword results in identical ciphertext segments.
- expects the number of characters between the beginning of repeated ciphertext segments to be a multiple of the keyword length.
- Compute the greatest common factor of all such distances between identified repeated segments.

Example

Assume that we have the following ciphertext

vyckbygecgxukgftkmuzlvtjgcyibngcwhagtkntkaqughbfuvajkwdlrqrm
uzlvtjcebdfcgprtqlqirwxhuvsshjcebdfcgprtqlqwntkaqujtgyvyegxs
vvpiaspkftfwujyvxshrggequvajkwupqixedvadfwurtpbdcsjtmzgkfgxw
nflvkwrvyixvuvojxfevqxgljzqekgdccbtjgkwebuccmumzgncpdfgjqdyl
jvndeqccnwttgkgrthrimpvzyzrwtkorjadwanmgwp

Distance of muzlvtj is 42. The other two distances are 96 and 24 respectively.

The gcd is 6.

Second part: Find the different Caesar ciphers

- After t has been determined, we know that $m_i, m_{i+t}, m_{i+2t}, \ldots$ all have been encrypted with the same substitution cipher.
- Split the ciphertext characters into t different multisets, each containing the letters encrypted by a mono-alphabetic substitution cipher.
- The key of each mono-alphabetic substitution cipher can be determined through the statistics of 1-grams together with some trial and error.

Example cont'

$$S_1 = \{v, g, k, u, g, g, t, \ldots\},$$
$$S_2 = \{y, e, g, z, c, c, k, \ldots\},$$
$$\vdots$$
$$S_6 = \{y, u, m, j, n, g, a, \ldots\},$$

Counting the different letters in S_1 we get

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
1	0	5	3	1	0	10	1	0	2	1	0	0	1	1	1	4	2	0	1

Example cont'

By cyclically shifting this frequency count and looking for the best match compared to the known distribution we find it at $k_1 = 2$ (or $k_1 = \mathsf{c}$).

Doing the same procedure for all six multisets we finally reach the key $\mathbf{k} = \texttt{crypts}$ and the ciphertext can be decrypted as

`the vigenere cipher using a relatively short period is insecure but by using ....`

Index of coincidence

- Let p_i be the unknown probability of the character i in the ciphertext, $i = \mathtt{a}, \mathtt{b}, \ldots, \mathtt{z}$.
- The *measure of roughness*, MR, measures the deviation of the distribution of ciphertext characters from a flat frequency distribution as follows:

$$MR = \sum_{i=\mathtt{a}}^{\mathtt{z}} (p_i - \frac{1}{26})^2 = \sum_{i=\mathtt{a}}^{\mathtt{z}} p_i^2 - \frac{1}{26}. \quad (4)$$

Index of coincidence

- The minimum value of MR is $MR_{\min} = 0$, corresponding to a flat distribution.
- The maximum value will occur when $\sum_{i=\mathtt{a}}^{\mathtt{z}} p_i^2$ is maximized, which in our case corresponds to a mono-alphabetic cipher.

$$MR_{\max} = 0.0667 - 0.0385 = 0.0282, \tag{5}$$

- MR will vary with the period t, reaching $MR_{\max}$ for $t = 1$ and MR_{min} for $t = \infty$.

Index of coincidence

- The MR cannot be computed directly, because the distribution (and t) is unknown
- but it can be estimated

IC

Let f_i denote the number of appearances of letter i, $i = \mathtt{a}, \mathtt{b}, \ldots, \mathtt{z}$, in a ciphertext of length n.

The IC is defined as the probability that two arbitrary chosen character from the *given* ciphertext are the same, i.e.,

$$IC = \frac{\sum_{i=\mathtt{a}}^{\mathtt{z}} f_i(f_i - 1)}{n(n-1)}. \tag{6}$$

IC is an *estimate* of $\sum_{i=\mathtt{a}}^{\mathtt{z}} p_i^2$, and it will also provide an estimate of $MR + 1/26$. We can express this as $E(IC) = MR + 1/26$, where $E(IC)$ is the expectation of (6).

IC

t	1	2	3	4	5	7	∞
E(IC)	0.066	0.052	0.047	0.044	0.042	0.041	0.038

Example: Calculating IC for the ciphertext in the previous example gives $IC = 0.042$, indicating a period of 5. As we know that the period is 6, ...

Ciphertext autocorrelation - a better way to find the period

With a given ciphertext $\mathbf{c} = c_1, c_2, \ldots, c_n$, we count the number of occurrences $c_i = c_{i+t^*}$ in the interval $i = 1, \ldots, n - t^*$, for different values of t^*. The lowest value of t^* with number of occurrences around $0.066n$ is with high probability the period t.

Improved if we divide the letters in t^* multisets by selecting every $t^{*\text{th}}$ letter and consider all possible pairs of letters within each multiset.

Other important classical ciphers

The Vernam cipher: (or the one-time-pad)

a ciphertext of fixed length n encrypted by a Vigenére cipher, where the period of the key is chosen to be $t = n$.

we prove that the Vernam cipher is totally secure (i.e. unbreakable) if the key is chosen uniformly at random among all $\mathbb{Z}_{26}^n$ possible values.

Other important classical ciphers

The simple transposition cipher: (or permutation cipher)

Transposition cipher with a period t involves grouping the plaintext into blocks of t consecutive characters. The key is a permutation π on the positions within the block,($t!$ possible keys).

The encryption function maps the plaintext $\mathbf{m} = m_1, m_2, \ldots$ to the ciphertext $\mathbf{c} = c_1, c_2, \ldots$, through

$$c = E_k(m_1, m_2, \ldots, m_t), E_k(m_{t+1}, m_{t+2}, \ldots, m_{t+t}), \ldots,$$

where

$$E_k(m_1, m_2, \ldots, m_t) = m_{\pi(1)}, m_{\pi(2)}, \ldots, m_{\pi(t)}.$$

Other important classical ciphers

Decryption is done through the inverse permutation,

$$D_k(c_1, c_2, \ldots, c_t) = c_{\pi^{-1}(1)}, c_{\pi^{-1}(2)}, \ldots, c_{\pi^{-1}(t)}.$$

Let the key be $\pi = (25134)$, meaning that $\pi(1) = 2, \pi(2) = 5, \ldots$. Then $E_k(m_1, m_2, \ldots, m_5) = (m_2, m_5, m_1, m_3, m_4)$.

Assume that $\mathbf{m} = \texttt{findingthetreasurecanonlybedoneby} \ \ldots$. Then $E_k(\texttt{findi}) = \texttt{iifnd}$, $E_k(\texttt{ngthe}) = \texttt{genth}$, etc., and the ciphertext is

$$\mathbf{c} = \texttt{iifndgenthrstearauecoynnl} \ldots$$

Decryption is done by the inverse permutation $\pi^{-1} = (31452)$, and $D_k(c_1, c_2, \ldots, c_5) = (c_3, c_1, c_4, c_5, c_2)$. Decrypting gives $D_k(\texttt{iifnd}) = \texttt{findi}$, etc.

Other important classical ciphers - The Hill cipher

The Hill cipher acts on t-grams from $\mathbb{Z}_{26}$ through a key which is an invertible $t \times t$ matrix $A = [a_{ij}]$.

The encryption function maps the plaintext $\mathbf{m} = m_1, m_2, \ldots$ to the ciphertext $\mathbf{c} = c_1, c_2, \ldots$, through
$c = E_k(m_1, m_2, \ldots, m_t), E_k(m_{t+1}, m_{t+2}, \ldots, m_{t+t}), \ldots$, where

$$E_k(m_1, m_2, \ldots, m_t) = (m_1, m_2, \ldots, m_t)A.$$

Decryption involves using A^{-1}.

Other important classical ciphers - Rotor-based machines

Rotor-based machines: Polyalphabetic substitution ciphers implemented by a class of rotor-based machines were the dominant cryptographic tool in World War II.

The most well known rotor-based machine is the *Enigma* (used by the Germans in World War II).

Boris Hagelin also made several designs of rotor-based machines. One of the, called M-209, was extensively used by the US army during the 1940s.