COMP3260/6360 Data Security

Lecture 4

Prof Ljiljana Brankovic

School of Electrical Engineering and Computer Science

Lecture Overview

1. Unicity Distance
2. Symmetric Cipher Model
3. Kerckhoffs' Laws
4. Codes and Ciphers
5. Transposition Ciphers
6. Breaking Transposition Ciphers
7. Substitution Ciphers
8. Breaking Substitution Ciphers

Introduction to Information Theory

- Information Theory - not in the Textbook.
  (*Based on "Cryptography and Data Security" by D. Denning [1]*)

- Chapter 2 textbook

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

Unicity Distance

Shannon measured the secrecy of a cipher in terms of the key equivocation $H_C(K)$ for a given ciphertext C:

$$H_C(K) = \sum_{C} p(C) \sum_{K} p_C(K) \log_2 \frac{1}{p_C(K)}$$

where $p_C(K)$ is the probability of K given C.

If $H_C(K)=0$ then there is no uncertainty and the cipher is theoretically breakable given enough resources.

As the length of the ciphertext increases, the equivocation usually decreases.

Unicity Distance

The **unicity distance** is the smallest N such that $H_C(K)$ is close to 0, that is, it is the amount of ciphertext needed to uniquely determine the key.

A cipher is **unconditionally secure** if $H_C(K)$ never approaches 0, regardless how large N is.

Most ciphers are too complex to determine the probabilities needed to calculate the unicity distance.

Approximating Unicity Distance

However, for ciphers which are close to a random cipher model it is possible to find a good approximation of the unicity distance.

Assume that each plaintext and ciphertext message comes from a finite alphabet of L symbols.

Then there are 2^{RN} possible messages of length N, where $R=\log_2 L$.

These messages are partitioned into two subsets:

- a set of 2^{rN} **meaningful** messages, each with probability $(1/2^{rN})=2^{-rN}$
- a set of $2^{RN}-2^{rN}$ **meaningless** messages, each with probability 0

Approximating Unicity Distance

Assume that there are $2^{H(K)}$ keys, where H(K) is the key entropy, i.e., the number of bits in the key.

The key entropy is also called the **entropy of the system**: it is a measure of the size of the key space ***K***, approximated by $H(K)=\log_2|\boldsymbol{K}|$.

For example, a cryptosystem with a 64-bit key has an entropy of 64 bits.

The probability of each key is $p(K)=1/2^{H(K)}=2^{-H(K)}$

Approximating Unicity Distance

A **random cipher** is a cipher in which for each key K and ciphertext C, the decipherment $D_K(C)$ is an independent random variable uniformly distributed over all messages (both meaningful and meaningless).

Consider the ciphertext $C=E_K(M)$ for given K and M.

A **spurious key decipherment** (false solution) arises whenever encipherment under another key K' could produce C for the same message M or for another meaningful message M'.

Approximating Unicity Distance

For every correct solution to a particular ciphertext, there are $2^{H(K)}-1$ remaining keys, each having the same probability of producing a spurious key decipherment.

Because each plaintext message is equally likely, the probability of getting a meaningful message (and so a false solution) is
$q = (2^{rN}/2^{RN}) = 2^{(r-R)N} = 2^{-DN}$

The expected number F of false solutions is
$F = (2^{H(K)}-1)q = (2^{H(K)}-1)2^{-DN} \cong 2^{H(K)-DN}$

Approximating Unicity Distance

The number of false solutions is sufficiently small to break the cipher when

$\log_2 F = H(K) - DN = 0$

and so $N = H(K)/D$ is taken as an approximation to the unicity distance, the amount of text necessary to break the cipher.

Approximating Unicity Distance

Note that in a theoretically unbreakable cipher the number of possible keys is as large as the number of messages of a given length N so that

$H(K) = \log_2(2^{RN}) = RN$ and
$H(K) - DN = (R-D)N = rN \neq 0.$

Approximating Unicity Distance

Example. Consider a simple substitution cipher with a shift of K positions, 0 ≤ K ≤ 25. What is the unicity distance?

N=4.6/3.2 = 1.5 characters.

The estimate doesn't seem plausible: no substitution cipher can be broken with just one or two characters.

Usefulness of Unicity Distance

There are two reasons why the estimate is not very good.

- The estimate D=3.2 applies only to reasonable long messages.
- The cipher is not a good approximation to the random cipher model since most ciphertexts are not produced by meaningful messages (e.g., QQQQ) an so the decipherments are not uniformly distributed over the entire message space.

Usefulness of Unicity Distance

- The random cipher model gives a lower bound of the amount of ciphertext needed to break a cipher, a particular cipher will have a unicity distance at least H(K)/D.
- Note that the unicity distance gives the number of characters required to uniquely determine the key but it does not indicate the computational difficulty of finding it.

Usefulness of Unicity Distance

- A cipher may be computationally infeasible to break even if it is theoretically possible to break it with a small amount of ciphertext (e.g., AES).

- On the other hand, many substitution ciphers which use longer keys and have much greater unicity distance than AES are relatively simple to break when enough ciphertext is intercepted.

Usefulness of Unicity Distance

The unicity distance N is inversely proportional to the redundancy D.

As D approaches 0, an otherwise trivial cipher becomes unbreakable.

Usefulness of Unicity Distance

Example: Suppose M is a 6-digit integer enciphered 351972 using a Caser-type substitution cipher with key K, $0 \le K \le 9$, and that all possible 6-digit integers are equally likely. How much ciphertext is needed to break the cipher?

Answer: Such a cipher cannot be broken because there is no redundancy - no matter how much ciphertext is available.

Obstructing Cryptanalysis

Natural languages have an inherent redundancy which can be exploited to solve many ciphers by statistical analysis: frequency distribution of letters, ciphertext repetition, probable words.

Obstructing Cryptanalysis

Protecting against statistical analysis can be provided by removing some of the redundancy of the language before encryption, using data compression.

Obstructing Cryptanalysis

Confusion involves substitutions that make the relationship between the key and the ciphertext as complex as possible.

Diffusion involves transformations that dissipate the statistical properties of the plaintext across the ciphertext.

Obstructing Cryptanalysis

Many modern ciphers provide confusion and diffusion through complex enciphering transformations over large blocks of data.

Introduction to Cryptography

- Symmetric encryption, or conventional / secret-key / single-key:
  - sender and recipient share a common key
  - all classical encryption algorithms are secret-key
  - was only type prior to invention of public-key in 1970's

- Public-key encryption:
  - sender's and recipient's keys are neither the same nor easily derived from each other
  - has advantage of not having to exchange keys

- In what follows we will refer to symmetric encryption, unless stated otherwise

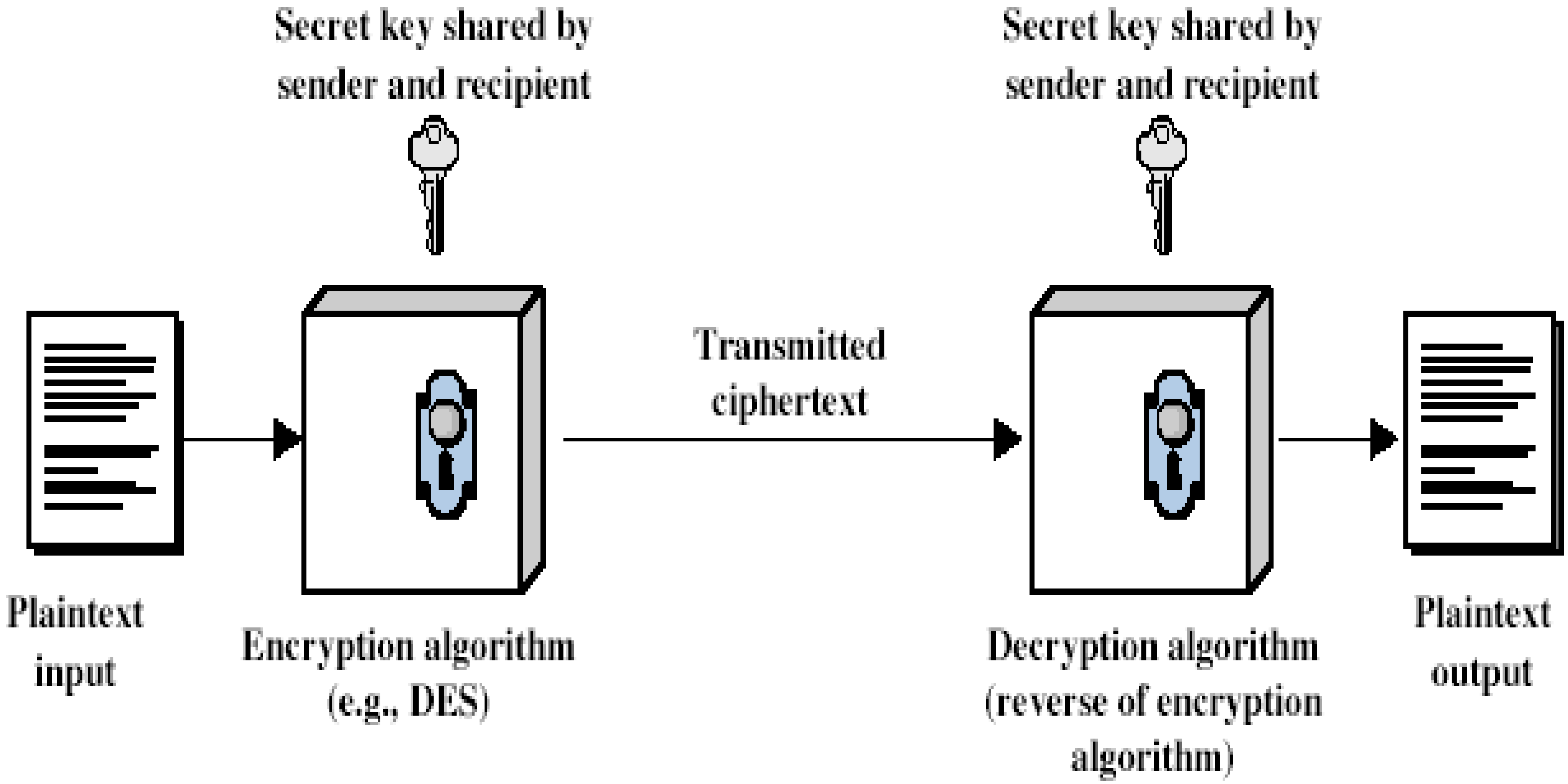
Basic Terminology

- **Plaintext** - the original message
- **Ciphertext** - the code (“encrypted”) message
- **Cipher** - algorithm for transforming plaintext to ciphertext
- **Key** - information used in cipher known only to sender/receiver

Basic Terminology

- **Enciphering (encrypting)** - converting plaintext to ciphertext
- **Deciphering (decrypting)** - recovering plaintext from ciphertext
- **Cryptography** - study of encryption principles/methods
- **Cryptanalysis (codebreaking)** - the study of principles/ methods of deciphering ciphertext *without* knowing the key
- **Cryptology** = Cryptography + Cryptanalysis

Symmetric Cipher Model


Secret key shared by sender and recipient
Secret key shared by sender and recipient
Transmitted ciphertext
Plaintext input
Encryption algorithm (e.g., DES)
Decryption algorithm (reverse of encryption algorithm)
Plaintext output

Requirements

- Two requirements for secure use of symmetric encryption:
  - a strong encryption algorithm
  - a secret key known only to sender / receiver

    $Y = E_K(X)$

    $X = D_K(Y)$

- The security of an encryption system should only depend on the secrecy of the key and not the secrecy of the encryption algorithm.

- Implies a secure channel to distribute key.

Kerckhoffs' law

(Auguste Kerckhoffs, Professor of Linguistics and cryptographer, 1835 - 1903)

A cryptosystem should be secure even if everything about the system, except the key, is public knowledge.

Kerckhoffs' laws

In 1883 Kerckhoffs published six principles of practical cipher design:

1. The system should be, if not theoretically unbreakable, unbreakable in practice.

2. Compromise of the system should not inconvenience the correspondents.

3. The key should be rememberable without notes and should be easily changeable.

4. The cryptograms should be transmittable by telegraph.

5. The apparatus or documents should be portable and operable by a single person.

6. The system should be easy, neither requiring knowledge of a long list of rules nor involving mental strain.

Kerckhoffs' law

Shannon's maxim: “The enemy knows the system.”

Bruce Schneier: “Kerckhoffs' principle applies beyond codes and ciphers to security systems in general: every secret creates a potential failure point. Secrecy, in other words, is a prime cause of brittleness—and therefore something likely to make a system prone to catastrophic collapse. Conversely, openness provides ductility.”

Security through Obscurity

Security through obscurity (security by obscurity [3]) uses secrecy of the encryption algorithm to ensure security.

Problems:

- Experience shows that secret algorithm designs are eventually disclosed either through reverse engineering or by leaked information. Thus if the system has weaknesses it cannot be subsequently used.
- The more secrets a system has, the less secure it is [3].
- If the algorithm is kept secret, the opportunities for security reviews and improvements are limited [3].

Cryptography

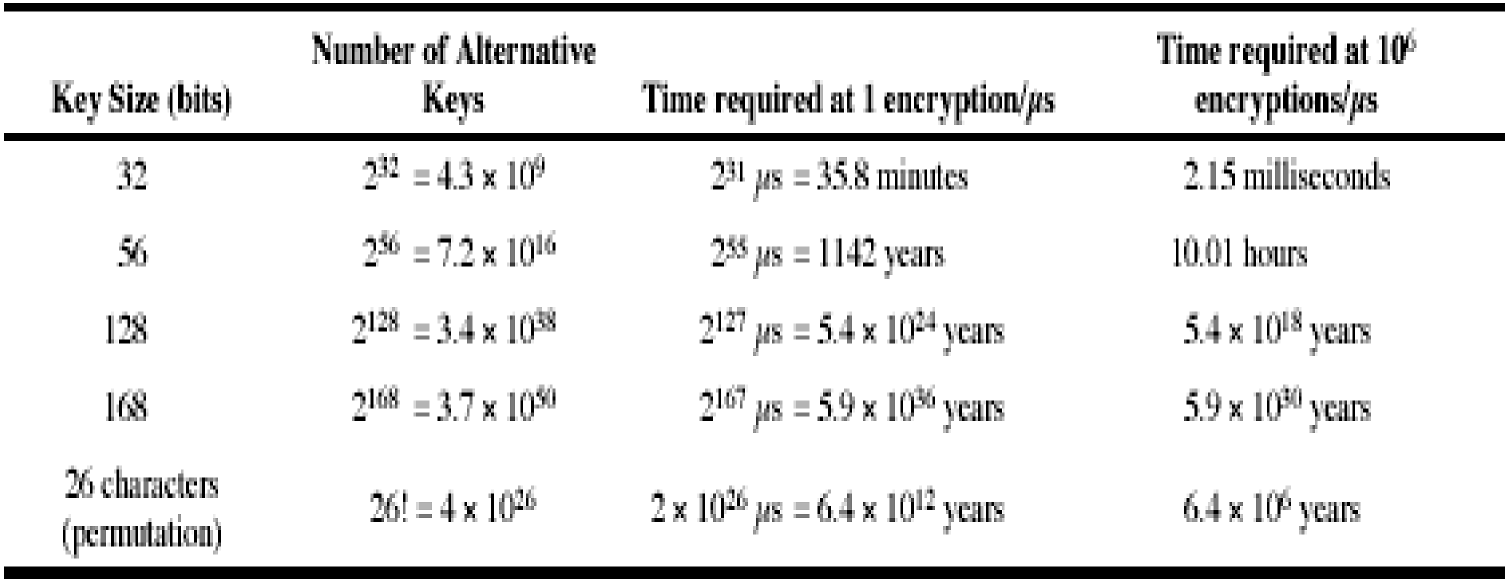
- Can characterize by:
  - type of encryption operations used
    - substitution / transposition / product
  - number of keys used
    - single-key or secret or conventional / two-key or public
  - way in which plaintext is processed
    - block / stream

Types of Cryptanalytic Attacks

- **Ciphertext only**
  - only know algorithm / ciphertext, statistical attack, can identify plaintext
- **Known plaintext**
  - know/suspect plaintext & ciphertext to attack cipher
- **Chosen plaintext**
  - select plaintext and obtain ciphertext to attack cipher
- **Chosen ciphertext**
  - select ciphertext and obtain plaintext to attack cipher
- **Chosen text**
  - select either plaintext or ciphertext to en/decrypt to attack cipher

Brute Force Search

- always possible to simply try every key
- most basic attack, proportional to key size
- assume either know / recognise plaintext



Key Size (bits)	Number of Alternative Keys	Time required at 1 encryption/μs	Time required at 10^6 encryptions/μs
32	$2^{32} = 4.3 \times 10^9$	2^{31} μs = 35.8 minutes	2.15 milliseconds
56	$2^{56} = 7.2 \times 10^{16}$	2^{55} μs = 1142 years	10.01 hours
128	$2^{128} = 3.4 \times 10^{38}$	2^{127} μs = 5.4×10^{24} years	5.4×10^{18} years
168	$2^{168} = 3.7 \times 10^{50}$	2^{167} μs = 5.9×10^{36} years	5.9×10^{30} years
26 characters (permutation)	$26! = 4 \times 10^{26}$	2×10^{26} μs = 6.4×10^{12} years	6.4×10^6 years

Cryptography

Cryptography is the art (science, study) of writing in secret letters.

Secret writing:

1. Steganography
2. Cryptography

Steganography (concealment systems) hide the real message in covering messages which themselves look real, or attempt to hide even the existence of a message (e.g., invisible ink, microdots).

Cryptography does not conceal the existence of a message, only its meaning.

Codes

Cryptographic systems:

- code systems
- cipher system

Codes are mappings which are semantic in nature and which map letters, words, and/or entire messages into encoded text by means of a predefined table.

Advantage: by correctly designing a code, it is possible to make the encoded text appear to be a message of entirely different meaning.

Disadvantage: the need for a substitution table (or code-book) entry for every possible message severely restricts the types of messages which can be encoded.

Codes

For general computer systems using coding techniques to achieve security is:

- too restrictive (usually impossible to predict types of messages)
- for general communication the code-book would have to be very large and kept in a very safe place - impractical for computer systems.

Ciphers are more flexible that codes.

Ciphers

Classical ciphers fall into one of the following categories:

- transposition ciphers, where the characters in the plaintext are simply rearranged
- substitution ciphers, where each character (or a group of characters) is substituted by another character (or a group of characters); substitution ciphers can be divided into:
  - monoalphabetic
  - homophonic
  - polyalphabetic
  - polygrams

More Definitions

- **Unconditional security**
  - no matter how much computer power is available, the cipher cannot be broken since the ciphertext provides insufficient information to uniquely determine the corresponding plaintext
- **Computational security**
  - given limited computing resources the cipher cannot be broken – eg, time needed for calculations is greater than age of universe

Transposition Ciphers

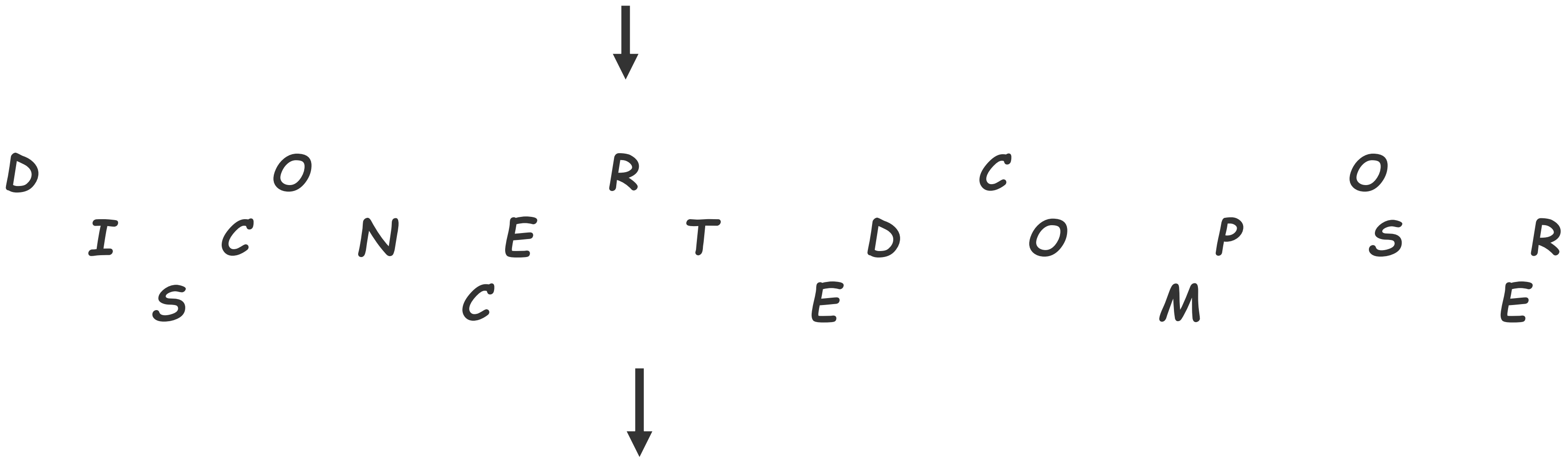
Transposition ciphers rearrange characters according to some scheme often using some geometric figure.

Recall that to encipher, we need an enciphering algorithm and an enciphering key.

The 'figure' and the 'writing-in' and 'talking-off' methods correspond to enciphering algorithm, while some parameter that determines the figure corresponds to the enciphering key.

Transposition Ciphers

Example. **DISCONCERTED COMPOSER**



DORCOICNETDOPSRSCEME

The algorithm: arrange letters of the plaintext in in rail-like way and read off by rows

The key: the 'rail' depth (in this case 3).

Columnar Transposition

Columnar transposition:

- plaintext is written into a matrix by rows
- ciphertext is obtained by taking off the columns in some order

Example: Using 6 columns, the plaintext SYDNEY OLIMPIC GAMES is written by rows as

S	Y	D	N	E	Y
O	L	Y	M	P	I
C	G	A	M	E	S

If the columns are taken off in the order 6-5-2-4-1-3 the resulting ciphertext is

YISEPEYLGNMMSOCDYA.

Periodic Transpositions

Every transposition cipher is a ***permutation*** of the plaintext with some ***period*** *d*. The period of the permutation can be as long as the message but usually it is shorter. Why?

Let Z_d be the set of integers *{1,2,…,d}* and let $f : Z_d \rightarrow Z_d$ be a permutation over Z_d.. Then the **key** is *f*. To encipher, successive blocks of *d* characters are permuted according to *f*.

A plaintext message $M = m_1m_2 \ldots m_dm_{d+1} \ldots m_{2d} \ldots$

is enciphered as $E_k(M) = m_{f(1)}m_{f(2)} \ldots m_{f(d)}m_{d+f(1)} \ldots m_{d+f(d)} \ldots$

Decipherment uses the inverse permutation.

Periodic Transpositions

Example. Suppose $d=6$ and f is the permutation

i	1	2	3	4	5	6
f(i)	6	5	4	3	2	1

Then the plaintext SYDNEY OLYMPIC GAMES is enciphered as YENDYSIPMYLOSEMAGC.

Periodic permutation ciphers can be implemented efficiently on a computer.

Breaking Transposition Ciphers

To recognise that a ciphertext was produced by a transposition cipher: Compare the relative frequencies of the letters in the ciphertext with the expected frequencies for the plaintext.

Transposition ciphers are broken by anagramming (the process of restoring a disarranged set of letters into their original positions).

Tables of frequency distributions for diagrams and trigrams are used in the anagramming process.

Frequency Distribution of Letters in English Text

Char	Percent	
A	8.0	****************
B	1.5	***
C	3	******
D	4.0	********
E	13.0	**************************
F	2.0	****
G	1.5	***
H	6.0	************
I	6.5	*************
J	0.5	*
K	0.5	*
L	3.5	*******
M	3.0	*******
N	7.0	**************
O	8.0	****************
P	2.0	****
Q	0.2	
R	6.5	*************
S	6.0	************
T	9.0	******************
U	3.0	******
V	1.0	**
W	1.5	***
X	0.5	*
Y	2.0	****
Z	0.2	

English Diagrams and Trigrams

The most frequent pairs of letters (diagrams) in English on a relative scale of 1 to 10:

Diagram	Frequency	Diagram	Frequency
TH	10.00	HE	9.05
IN	7.17	ER	6.65
RE	5.92	ON	5.70
AN	5.63	EN	4.76
AT	4.72	ES	4.24
ED	4.12	TE	4.04
TI	4.00	OR	3.98
ST	3.81	AR	3.54
ND	3.52	TO	3.50
NT	3.44	IS	3.43
OF	3.38	IT	3.26
AL	3.15	AS	3.00

English Diagrams and Trigrams

The most frequent trigrams in English:

ENT
ION
AND
ING
IVE
TIO
FOR
OUR
THI
ONE

Unicity Distance of a Permutation Cipher

How much ciphertext is needed to break a permutation cipher with period d ? Unicity distance of a permutation cipher with period d :

$$N=H(K)/D = (\log_2 (d\,!))/D$$

Sterling's approximation for large d: $d\,! \approx (d/e)^d(2\pi d)^{1/2}$.
Then $\log_2 (d\,!) \approx d \log_2 (d/e)$ and

$$N = (d \log_2(d/e)) / 3.2 = 0.3\, d \log_2(d/e)$$

Example: If the period is $d=27$, then d/e is about 10 and $\log_2(d/e)$ is about 3.2 so $N=27$.

Unicity Distance of a Permutation Cipher

The following table shows the period and the associated Unicity distance.

d	N
3	0.122804
4	0.66877
5	1.31885
6	2.05608
7	2.86579

Substitution Ciphers

Substitution ciphers can be divided into:

- monoalphabetic
- homophonic
- polyalphabetic
- polygrams

A monoalphabetic substitution cipher replaces each character of the plaintext alphabet *A* with the corresponding character of the ciphertext alphabet *C*. Usually *C* is a simple rearrangement of the lexicographic order of the characters in *A*.

Substitution Ciphers

Suppose *A* is a *n*-character alphabet

$$\{a_0\ , a_1, \ldots , a_{n-1}\}.$$

Then *C* is a *n*-character alphabet

$$\{f(a_0), f(a_1), \ldots , f(a_{n-1})\}$$

where $f : A \rightarrow C$ is a one-to-one mapping of each character of *A* to the corresponding character of *C*.

To encipher, simply rewrite the message using the corresponding characters of the ciphertext language:

$$E_k(M) = f(m_1)f(m_2) \ldots$$

Substitution Ciphers

Example.

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

Substitution Ciphers

Such a ciphertext alphabet is called a ***keyword mixed alphabet***.

In the example above the key of the cipher is SYDNEY OLYMPIC GAMES. The repeated letters in the key are dropped and after the key the remaining letters appear in alphabetic order.

The message M = DOWN ELEVATOR is encrypted as

$E_k(M)$ = NFVB EGEUSRFK

Substitution Ciphers

Ciphers based of **shifted alphabets** shift the letters of the alphabet by *k* positions to the right, modulo the size of the alphabet:

$f(x) = (x+k) \bmod n$

where *n* is the size of the alphabet *A*, *x* denotes a letter of *A* by its position, and *k* is the key.

Substitution Ciphers

More complex transformations use multiplication:

$f(x) = kx \bmod n$

where *k* and *n* are relatively prime so that the mapping is one-to-one. Here *k* is the key.

Substitution Ciphers

Example. If *k* = 9 and *A* is the English alphabet

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

Affine Transformations

Affine transformation combines addition with multiplication to get

$f(x) = (xk_1 + k_0) \bmod n$

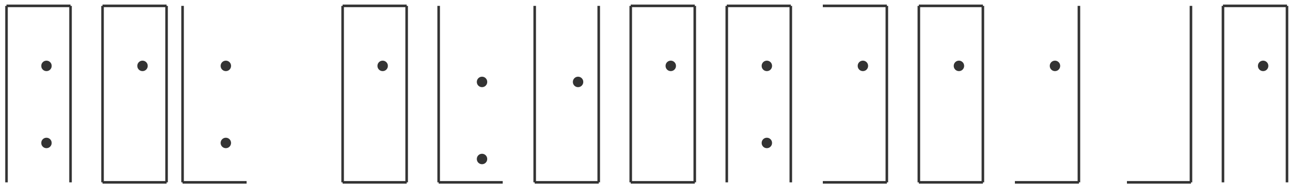
where k_1 and n are relatively prime.

In general, we can have polynomial transformations of any degree t:

$f(x) = (x^t k_t + x^{t-1}k_{t-1} + \ldots + xk_1 + k_0) \bmod n$

Note: Using nonstandard ciphertext alphabets doesn't increase the difficulty of breaking the cipher. Why?

A Churchyard cipher engraved on a tombstone in Trinity Churchyard, New York, 1794:



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		

A similar cipher was also engraved on a tombstone in St. Paul's Churchyard, New York, in 1796. The first published solution to this cipher appeared in the New York Herald in 1896 - over 100 years later.

Why did it take so long to break this cipher?

Breaking Substitution Cipher

Example. Find the number of letters needed to break general substitution alphabets of size n.

The number of possible keys is n! (that is the number of ways of arranging the n letters of the alphabet).

If all keys are equally likely then the unicity distance is

$$N = H(K) / D = (\log_2 n!) / D$$

For English, $N = (\log_2 26!) / 3.2 = 88.4 / 3.2 = 27.6$

That means that usually at least 28 letters are needed to break these ciphers. That explains the difficulty in solving the Churchyard ciphers (only about 15 characters).

Breaking Substitution Cipher

Ciphers based on polynomial transformations have smaller unicity distances.

For shifted alphabets the number of possible keys is only 26 and the unicity distance is

$$N \cong (\log_2 26\ )\ /\ 3.2 \cong 1.5$$

Breaking Substitution Cipher

Simple substitution ciphers are easy to break in a ciphertext only attack using single letter frequency analysis: comparing the letter frequencies in a given ciphertext with the expected frequencies to match the ciphertext letters with the plaintext letters.

Diagram and trigram distributions can also be used.

Ciphers based on shifted alphabets are extremely easy to break because each ciphertext letter is a constant distance from its corresponding plaintext letter.

Breaking Substitution Cipher

Ciphers based on affine transformations $f(x) = (xk_1 + k_0) \bmod n$ are more difficult to break BUT if a set of t correspondences between plaintext letters m_i and ciphertext letters c_i , $1 \leq i \leq t$, are known (or suspected) then it may be possible to find k_1 and k_0 by solving the following system of equations:

$(m_1k_1 + k_0) \bmod n = c_1$

.

.

.

. $(m_tk_1 + k_0) \bmod n = c_{t.}$

……..

Breaking Substitution Cipher

Example. Suppose we have the following possible correspondences.

Plaintext	E (4)	J (9)	N (13)
Ciphertext	K (10)	T (19)	V (21)

That gives the equations

$(4k_1 + k_0) \bmod 26 = 10$

$(9k_1 + k_0) \bmod 26 = 19$

$(13k_1 + k_0) \bmod 26 = 21$

The solutions of the first two equations is $k_1 = 7$ and $k_0 = 8$. Note that we must check that the third equation is also satisfied. What would it mean if the third equation is not satisfied?

Note that in general we may need more than 2 equations to solve for k_0 and k_1, as equations of the form $ak \bmod 26 = c$ have multiple solutions when a divides 26.

Breaking Substitution Cipher

Cryptanalysis of a general simple substitution cipher:

- Brute force attacks: try all 26! decipherments - if 1 decipherment per microsecond, it would take more that 10^3 years!

- Instead use a single letter frequency analysis - diagram and trigram distributions are also helpful.

Next Week

1. Homophonic ciphers - Beale ciphers
2. Polyalphabetic substitution ciphers
   a) Vigenere cipher
   b) Beaufort Cipher
   c) Variant Beaufort Cipher
3. Breaking periodic polyalphabetic ciphers
   a) Index of Coincidence
   b) Kasiski Method
4. Running Key ciphers, Rotor Machines and One-Time Pads
5. Polygram Substitution Ciphers - Playfair Ciphers

Chapter 3 textbook “Classical Encryption Techniques”

References

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

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

3. Bruce Schneier. “Secrecy, Security, and Obscurity”, Crypto-Gram Newsletter, May 15, 2002, http://www.schneier.com/crypto-gram-0205.html#1 last accessed on March 2014.