Cryptography I

General concepts and some classical ciphers

- Basic concepts
- Attack models
- Classic ciphers: mono-alphabetic
- Vigenere cipher
- One-time-pad cipher
- Perfect secrecy

Security Goals

- Confidentiality (secrecy, privacy)
  - Assure that data is accessible to only one who are authorized to know
- Integrity
  - Assure that data is only modified by authorized parties and in authorized ways
- Availability
  - Assure that resource is available for authorized users

What is Crypto?

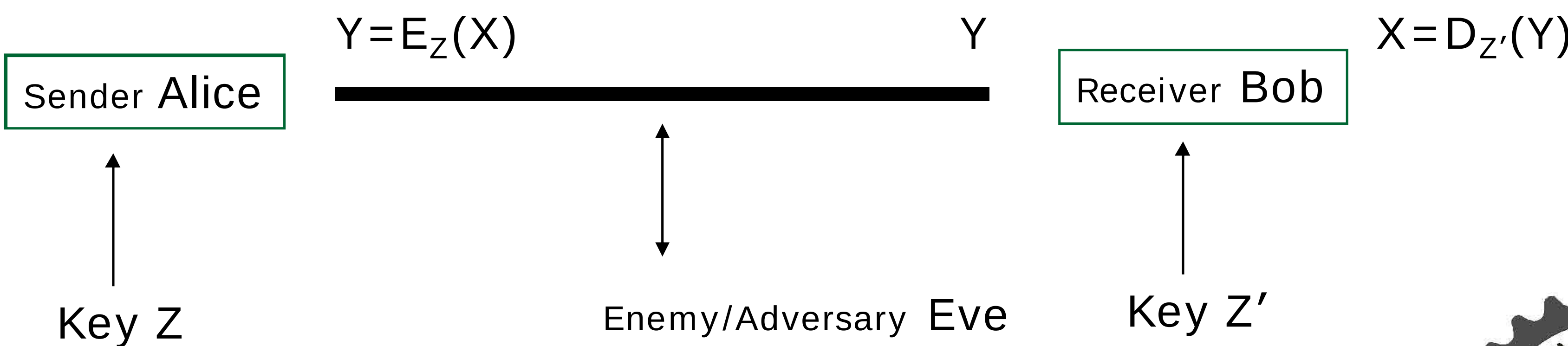
- Constructing and analyzing cryptographic protocols which enable parties to achieve security objectives
  - Under the present of adversaries.
- A protocol (or a scheme) is a suite of procedures that tell each party what to do
  - usually, computer algorithms
- Cryptographers devise and analyze protocols under Attack model
  - assumptions about the resources and actions available to the adversary
    - So, you need to think as an adversary

Terms

- **Cryptography:** the study of mathematical techniques for providing information security services.
- **Cryptanalysis:** the study of mathematical techniques for attempting to get security services breakdown.
- **Cryptology:** the study of cryptography and cryptanalysis.

Terms …

- plaintexts
- ciphertexts
- keys
- encryption
- decryption



Secret-key cryptography

- Also called: symmetric cryptography
- Use the same key for both encryption & decryption (Z=Z')
- Key must be kept secret
- Key distribution – how to share a secret between A and B very difficult

Public-key cryptography

- Also called: asymmetric cryptography
- Encryption key different from decryption key and
  - It is not possible to derive decryption key from encryption key
- Higher cost than symmetric cryptography

Is it a secure cipher system?

- **Why insecure**
  - **just break it under a certain reasonable attack model (show failures to assure security goals)**
- Why secure:
  - Evaluate/prove that under the considered attack model, security goals are assured
  - Provable security: Formally show that (with mathematical techniques) the system is as secure as a well-known secure one (usually simpler).

Breaking ciphers …

- There are different methods of breaking a cipher, depending on:
  - the type of information available to the attacker
  - the interaction with the cipher machine
  - the computational power available to the attacker

Breaking ciphers …

- **Ciphertext-only attack**:
  - The cryptanalyst knows **only the ciphertext**.
  - Goal: to find the plaintext and the key.
  - NOTE: such vulnerable is seen completely insecure
- **Known-plaintext attack**:
  - The cryptanalyst knows **one or several pairs of ciphertext and the corresponding plaintext.**
  - Goal: to find the key used to encrypt these messages
    - or a way to decrypt any new messages that use the same key (although may not know the key).

Breaking ciphers …

- **Chosen-plaintext attack**
  - The cryptanalyst **can choose a number of messages and obtain the ciphertexts for them**
  - Goal: deduce the key used in the other encrypted messages or decrypt any new messages (using that key).
- **Chosen-ciphertext attack**
  - Similar to above, but the cryptanalyst **can choose a number of ciphertexts and obtain the plaintexts.**
- Both can be **adaptive**
  - The choice of ciphertext may depend on the plaintext received from previous requests.

Models for Evaluating Security

- **Unconditional (information-theoretic) security**
  - **Assumes that the adversary has unlimited computational resources**.
  - Plaintext and ciphertext modeled by their distribution
  - Analysis is made by using probability theory.
  - For encryption systems: **perfect secrecy**, observation of the ciphertext provides no information to an adversary.

Models for Evaluating Security

- **Provable security:**
  - Prove security properties based on assumptions that it is difficult to solve a well-known and supposedly difficult problem (NP-hard …)
    - E.g.: computation of discrete logarithms, factoring
- **Computational security (practical security)**
  - Measures the amount of computational effort required to defeat a system using the best-known attacks.
  - Sometimes related to the hard problems, but no proof of equivalence is known.

Models for Evaluating Security

- **Ad hoc security (heuristic security):**
  - Variety of convincing arguments that every successful attack requires more resources than the ones available to an attacker.
  - Unforeseen attacks remain a threat.
  - **THIS IS NOT A PROOF**

Classic ciphers

Shift cipher (additive cipher)

- Key Space: [1 .. 25]
- Encryption given a key K:
  - each letter in the plaintext P is replaced with the K'th letter following corresponding number (shift right):
  - Another way: Y=X ⊕ K ➔ additive cipher
- Decryption given K:
  - shift left

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

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

P = CRYPTOGRAPHYISFUN

K = 11

C = NCJAVZRCLASJTDQFY

Shift Cipher: Cryptanalysis

- Easy, just do exhaustive search
  - key space is small (<= 26 possible keys).
  - once K is found, very easy to decrypt

General Mono-alphabetical Substitution Cipher

- The key space: all permutations of $\Sigma = \{A, B, C, \ldots, Z\}$
- Encryption given a key π:
  - each letter X in the plaintext P is replaced with $\pi(X)$
- Decryption given a key π :
  - each letter Y in the cipherext P is replaced with $\pi^{-1}(Y)$

- **Example:**

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

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

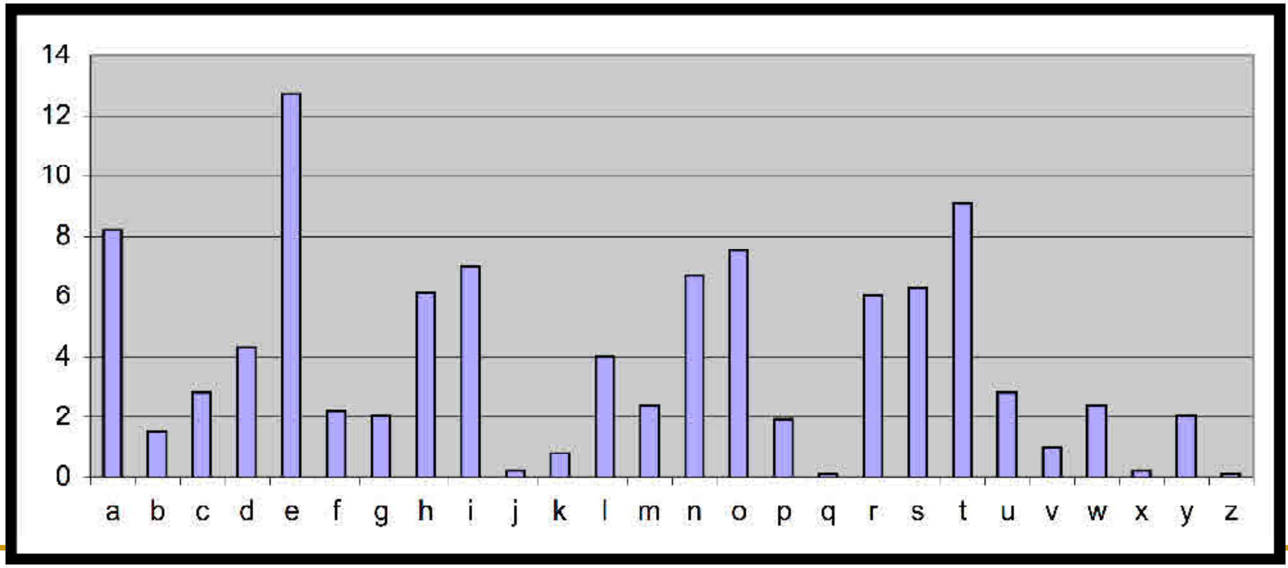
BECAUSE $\rightarrow$ AZDBJSZ

Looks secure, early days

- Exhaustive search is infeasible
  - key space size is $26! \approx 4*10^{26}$
- Dominates the art of secret writing throughout the first millennium A.D.
- Thought to be unbreakable by many back then

Cryptanalysis of Substitution Ciphers: Frequency Analysis

- Each language has certain features:
  - frequency of letters, or of groups of two or more letters.
- Substitution ciphers preserve the mentioned language features ➔ vulnerable to frequency analysis attacks



Substitution Ciphers: Cryptanalysis

- The number of different ciphertext characters or combinations are counted to determine the frequency of usage.
- The cipher text is examined for patterns, repeated series, and common combinations.
- Replace ciphertext characters with possible plaintext equivalents using known language characteristics.
- Example:

  THIS IS A PROPER SAMPLE FOR ENGLISH TEXT. THE FREQUENCIES OF LETTERS IN THIS SAMPLE IS NOT UNIFORM AND VARY FOR DIFFERENT CHARACTERS. IN GENERAL THE MOST FREQUENT LETTER IS FOLLOWED BY A SECOND GROUP. IF WE TAKE A CLOSER LOOK WE WILL NOTICE THAT FOR BIGRAMS AND TRIGRAMS THE NONUNIFORM IS EVEN MORE.

  - Observations: f_x=1 và f_A=15.

Substitution Ciphers: Cryptanalysis

- The letters in the English alphabet can be divided into 5 groups of similar frequencies

  I e

  II t,a,o,i,n,s,h,r

  III d,l

  VI c,u,m,w,f,g,y,p,b

  V v,k,j,x,q,z

- Some frequently appearing bigrams or trigrams

  Th, he, in, an, re, ed, on, es, st, en at, to

  The, ing, and, hex, ent, tha, nth, was eth, for, dth.

Example

Letter:	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Frequency:	5	24	19	23	12	7	0
Letter:	*H*	*I*	*J*	*K*	*L*	*M*	*N*
Frequency:	24	21	29	6	21	1	3
Letter:	*O*	*P*	*Q*	*R*	*S*	*T*	*U*
Frequyency:	0	3	1	11	14	8	0
Letter:	*V*	*W*	*X*	*Y*	*Z*		
Frequency:	27	5	17	12	45		

- $e \Rightarrow Z$

$f_j = 29, f_v = 27$

$f_{jcz} = 8 \rightarrow t \Rightarrow J$

$h \Rightarrow C$

- $a \Rightarrow V$

  (article a)

$J,V,B,H,D,I,L,C\ \{t,a,o,i,n,s,h,r\}$

$t,a \qquad\qquad h$

$JZB = te\ ?\ \ \{\ teo, tei, ten, ter, tes\ \}$ ➔ $n \Rightarrow B$

Substitution Ciphers: Cryptanalysis

- Observations:
  - A cipher system should not allow statistical properties of plaintext to pass to the ciphertext.
  - The ciphertext ginerated by a "good" cipher systim should be satistically indistinguishable form random text.
- Idea for a stronger cipher (1460’s by Alberti)
  - use more than one cipher alphabet, and switch between them when encrypting different letters ➔ Polyalphabetic Substitution Ciphers
  - Developed into a practical cipher by Vigenère (published in 1586)

Vigenère cipher

- **Definition**:
  - Given m, a positive integer, $P = C = (Z_{26})^n$, and $K = (k_1, k_2, \ldots, k_m)$ a key, we define:
- **Encryption**:

  $e_k (p_1, p_2 \ldots p_m) = (p_1+k_1, p_2+k_2 \ldots p_m+k_m) \pmod{26}$
- **Decryption**:

  $d_k (c_1, c_2 \ldots c_m) = (c_1-k_1, c_2-k_2 \ldots c_m- k_m) \pmod{26}$
- **Example:**

  Plaintext: C R Y P T O G R A P H Y

  Key: L U C K L U C K L U C K

  Ciphertext: N L A Z E I I B L J J I

Vigenère Cipher: Cryptanalysis

- Find the length *p* of the key: the crucial problem
- If p is known, divide the message into *p* groups of letters
  - For each fixed i=0,p-1, group # i consists of letters at positions kp + i (k=1,2,3 ...)
  - Obviously, each such group is a shift cipher encryption.
- Use frequency analysis to solve the resulting shift ciphers.

Use Index of Coincidence to find the key length

- Index of Coincidence (IC) – by Friedman,1922
- Informally: Measures the probability that two random elements of the n-letters string x are identical.
- **Definition:** Suppose $x = x_1x_2\ldots x_n$ is a string of n alphabetic characters. Then $I_c(x)$, the index of coincidence is:

$\boldsymbol{I_c(x) = \Pr\{x_i = x_j |}$ for any two x_i, x_j randomly selected from x}

- *Now to find the key length we find the p which make the average IC of each mentioned letter group become highest*

Use Index of Coincidence to find the key length

- IC can be determined by this

$$IC(x) = \frac{\sum_{i=0}^{25} f_i (f_i - 1)}{n(n-1)}$$

  - Where f_i is the appearance frequency of the alphabet's ith letter in the message *x*.

Use Index of Coincidence to find the key length

- For natural language such as English, IC is higher

Letter	p_i	Letter	p_i	Letter	p_i	Letter	p_i
A	.082	H	.061	O	.075	V	.010
B	.015	I	.070	P	.019	W	.023
C	.028	J	.002	Q	.001	X	.001
D	.043	K	.008	R	.060	Y	.020
E	.127	L	.040	S	.063	Z	.001
F	.022	M	.024	T	.091		
G	.020	N	.067	U	.028		

$$I_c(x) = \sum_{i=0}^{i=25} p_i^{\,2} = 0.065$$

Use Index of Coincidence to find the key length

Key length (p)	1	2	3	4	5	...	10
IC	0.068	0.052	0.047	0.044	0.043	...	0.041

- For a shift cipher, IC is just the same as of English plaintext
- For Vigenere cipher if we gradually increase the key length p the IC will decrease gradually (p=1 ➔ shift cipher)
- Remark: the high fluctuation of letter frequencies in natural languages (e.g. English) cause the IC become higher
  - For Vigenere, the letter frequencies becomes more equally fc
    higher key length ➔ IC become lower

Use Index of Coincidence to find the key length

- Practical approach in finding p of a given Vigenere cipher

  1. Set k=1

  2. Check if p equals k

  2.a. Devide the cipher into k letter groups as before and compute the IC of each.

  2.b. If they all are quite the same and approximately equals to 0.068 then p=k

  If they are quite different to each other and quite smaller than 0.068 then p>k

  4. Increase k by 1 and go back to step 2

$q = q_1 q_2 \ldots q_n$, m is the key length

$$
\begin{array}{cc}
\begin{bmatrix}
q_1 & q_{m+1} & \cdots & q_{n-m+1} \\
q_2 & q_{m+2} & \cdots & q_{n-m+2} \\
\cdots & \cdots & \cdots & \cdots \\
q_m & q_{2m} & \cdots & q_n
\end{bmatrix}
&
\begin{array}{c}
y_1 \\ y_2 \\ \cdots \\ y_m
\end{array}
\end{array}
$$

- If m is the key length, then the text ``looks like'' English text

$$I_c(y_i) \approx \sum_{i=0}^{i=25} p_i^{2} = 0.065 \quad \forall 1 \le \mathrm{i} \le \mathrm{m}$$

- If m is not the key length, the text ``looks like'' random text and:

$$I_c \approx \sum_{i=0}^{i=25} \left(\frac{1}{26}\right)^2 = 26 \times \frac{1}{26^2} = \frac{1}{26} = 0.038$$

One-Time Pad

Key is chosen randomly

Plaintext $X = (x_1\ x_2\ \ldots\ x_n)$

Key $K = (k_1\ k_2\ \ldots\ k_n)$

Ciphertext $Y = (y_1\ y_2\ \ldots\ y_n)$

$$e_k(X) = (x_1+k_1\ \ x_2+k_2\ \ldots\ x_n+k_n) \bmod m$$

$$d_k(Y) = (x_1-k_1\ \ x_2-k_2\ \ldots\ x_n-k_n) \bmod m$$

Example

Plaintext space = Ciphtertext space =

Keyspace = $\{0,1\}^n$

Key is chosen randomly

For example:

Plaintext is	10001011
Key is	00111001
Then ciphertext is	10110010

Main points in One-Time Pad

- The key is never to be reused
  - Thrown away after first and only use
  - If reused ➔ insecure!
- One-Time Pad uses a very long key, exactly the same length as of the plaintext
  - In old days, some suggest choose the key as texts from, e.g., a book ➔ i.e. not **randomly chosen**
    - Not One-Time Pad anymore ➔ this does not have perfect secrecy as in true One-Time-Pad and can be broken
  - Perfect secrecy means key length be at least message length
    - **Difficult in practice**!

Further remarks

- Shift ciphers are easy to break using brute force attacks (eshautive key search)
- Substitution ciphers preserve language features (in N-gram frequency) and are vulnerable to frequency analysis attacks.
- Vigenère cipher are also vulnerable to frequency analysis once the key length is found.
  - In general poly-alphabetical substitution ciphers are not that secure
- OTP has perfect secrecy if the key is chosen randomly in the message length and is used only once.

Perfect Secrecy

- Consider this ciphertext-only attack model
  - Eve can eavesdrop all the ciphertext
  - Eve has unconditional power: unlimited computation resource
- We now consider if such an enemy with computation power as of God's can always find the plaintext (or key)?
  - If yes, there is never Perfect Secrecy
  - Otherwise, then see how we can define it

Exhaustive searching

- Given a cryptogram Y created by a substitution cipher, to find the corresponding plaintext X, Eve can try all the possible key (i.e. substitution)
- However for short Y we can find multiple X which are meaningful English ➔ can't find exactly the origin X.

E.g. given Y= AZNPTFZHLKZ

- We can find at least 2 possible plaintexts
  - **Subs 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
K	B	C	D	T	E	G	I	J	M	O	L	A	Q	R	H	S	F	N	P	U	V	W	X	Z	Y

  - **Subs 2**

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
L	P	H	N	Z						K	T						A			F					E

  - **Possible plaintext**

MÃ:	A	Z	N	P	T	F	Z	H	L	K	Z
TIN 1:	m	y	s	t	e	r	y	p	l	a	y
TIN 2:	r	e	d	b	l	u	e	c	a	k	e

Remarks

- On mono-alphabetic cipher
  - For given short ciphertext, there can be multiple possible plaintext corresponding to it
  - However if the length of the ciphertext is at least 50 then there will be always only one true plaintext.
  - Thus, for sufficient long ciphertext powerful Eve will always success
- Thus, Eve can not always be successful, but when obtained enough ciphertext.
  - Chance of success is higher when more ciphertext is intercepted
- However, Perfect Secrecy can be obtained:
  - One-time pad: Eve can guess nothing, no matter how long ciphertext she can intercept

Shannon’s (Information-Theoretic) Perfect Secrecy

- Basic Idea: Ciphertext should provide no “information” about Plaintext –

  $$Pr(X) = Pr(X/Y) \;\; \forall \text{ TIN X VÀ MÃ Y}$$

  Probabilistic distribution of plaintext X is still the same after Eve has the knowledge of the corresponding ciphertext Y

- One-time pad has perfect secrecy
  - E.g., suppose that the ciphertext is “Hello”, can we say any plaintext is more likely than another plaintext?
- Theory due to Shannon, 1949.

  C. E. Shannon, “Communication Theory of Secrecy Systems”, Bell System Technical Journal, vol.28-4, pp 656--715, 1949.

On “examining” the security of a cipher

- Shannon: the concept of “unicity distance” to “measure” the security of a cipher system:
  - Unicity distance, denoted by N_0, is the minimum length of ciphertext that the powerful Eve have to obtain in order to figure out an unique plaintext appropriate for it.
  - This can be computed as $$N_0 = \frac{\log_2 E}{d}$$
    - d: *the redundancy rate of the plaintext language*.
- Example on redundancy
  - The following sentence has been shortened but can be figured out uniquely!

  Mst ids cn b xprsd n fwr ltrs, bt th xprsn s mst nplsnt

On “examining” the security of a cipher

- Shannon: the concept of “unicity distance” to “measure” the security of a cipher system:
  - Unicity distance, denoted by N_0, is the minimum length of ciphertext that the powerful Eve have to obtain in order to figure out an unique plaintext appropriate for it.
  - This can be computed as

$$N_0 = \frac{\log_2 E}{d}$$

    - d: *the redundancy rate of the plaintext language*.
- Example on redundancy

  Mst ids cn b xprsd n fwr ltrs, bt th xprsn s mst nplsnt

➔

Most ideas can be expressed in fewer letters, but the expression is most unplesant

- This proves that natural languages have redundancy

Redundancy

- **Redundancy** in information theory is the number of bits used to transmit a message minus the number of bits of actual information in the message.
  - Informally, it is the amount of wasted "space" used to transmit certain data.
  - Data compression is a way to reduce or eliminate unwanted redundancy, while checksums are a way of adding desired redundancy for purposes of error detection when communicating over a noisy channel of limited capacity.
- Redundancy can be defined as

$$\mathbf{d = R - r \text{ bits}}$$

  - where R is the *absolute rate* and r is the *true rate* of a language.
  - The **absolute rate** of a language or source is simply

$$R = \log_2 M \text{bits}$$

    where M is the size of the alphabet.
    - For English, $R = \log_2 26 \approx 4.7$ bits.
  - True rate r is the rate after the text is compressed
  - *For English*, r is 1 - 1,5 bit

On "examining" the security of a cipher

- Redundancy is reflecting, measuring the structure or the predictability of a language.
  - For English, redundancy is between 3.2 and 3.7 bits (caused by the high difference of frequencies of letters as well as bigrams trigram)
- Using unicity distance we can have a "feeling" about the security of different ciphers
  - For mono-alphabetic ciphers, we observe

$$E= |Z| = 26!$$

$$Pr(Z) =1/26!$$

$$\log_2 E = \log_2(26!) \approx 88.4 \text{ bits}$$

$$N_0 \approx 88.4 / 3.7 \approx 23.9 \text{ ký tự}$$

  - So ciphers of length at least 24 would be solved uniquely!

On "examining" the security of a cipher

- For one-time-pad:
  - X, the plaintext space = {set of English text of length k}
  - Z, the key space = {set of k-length sequence on the English alphabet}
  - If keys are selected equally randomly

$$N_0 = \log_2 E/d$$

$$E = 26^k \Rightarrow \log_2(26^k) = k*\log_2 26 \approx 4.7k$$

$$N_0 = (4.7k)/3.7 = 1.37k$$

  - Thus, Eve can intercept to all the ciphertext she want but can never find the true plaintext.

Some quiz

- Why IC decreases when the number of substitution alphabets increases?
- Can you guess of any connection between IC and redundancy