Cryptography
CS 555

Topic 2: Evolution of Classical Cryptography
Lecture Outline

- Basics of probability
- Vigenere cipher.
- Attacks on Vigenere: Kasisky Test and Index of Coincidence
- Cipher machines: The Enigma machine.

- Required readings:
  - Katz & Lindell: 1.1 to 1.3
- Recommended readings
  - The Code Book by Simon Singh
Begin Math

\[
\begin{array}{c}
2 \\
+2 \\
4 \\
\end{array}
\]
Random Variable

Definition

A discrete random variable, \( X \), consists of a finite set \( \mathcal{X} \), and a probability distribution defined on \( \mathcal{X} \). The probability that the random variable \( X \) takes on the value \( x \) is denoted \( \Pr[X = x] \); sometimes, we will abbreviate this to \( \Pr[x] \) if the random variable \( X \) is fixed. It must be that

\[
0 \leq \Pr[x] \quad \text{for all} \quad x \in \mathcal{X}
\]

\[
\sum_{x \in \mathcal{X}} \Pr[x] = 1
\]
Example of Random Variables

- Let random variable $D_1$ denote the outcome of throw one dice (with numbers 0 to 5 on the 6 sides) randomly, then $\mathcal{D} = \{0, 1, 2, 3, 4, 5\}$ and $\Pr[D_1=i] = 1/6$ for $0 \leq i \leq 5$

- Let random variable $D_2$ denote the outcome of throw a second such dice randomly

- Let random variable $S_1$ denote the sum of the two dices, then $\mathcal{S} = \{0, 1, 2, \ldots, 10\}$, and
  - $\Pr[S_1=0] = \Pr[S_1=10] = 1/36$
  - $\Pr[S_1=1] = \Pr[S_1=9] = 2/36 = 1/18$
  - ...

- Let random variable $S_2$ denote the sum of the two dices modulo 6, what is the distribution of $S_2$
Definitions

Assume $X$ and $Y$ are two random variables, then we define:

- **joint probability**: $\Pr[x, y]$ is the probability that $X$ takes value $x$ and $Y$ takes value $y$.
- **conditional probability**: $\Pr[x|y]$ is the probability that $X$ takes on the value $x$ given that $Y$ takes value $y$.

\[
\Pr[x|y] = \frac{\Pr[x, y]}{\Pr[y]}
\]

- **independent random variables**: $X$ and $Y$ are said to be independent if $\Pr[x,y] = \Pr[x]\Pr[y]$, for all $x \in \mathcal{X}$ and all $y \in \mathcal{Y}$. 

Examples

- Joint probability of \( D_1 \) and \( D_2 \) for \( 0 \leq i, j \leq 5 \), \( \Pr[D_1=i, D_2=j] = ? \)
- What is the conditional probability \( \Pr[D_1=i \mid D_2=j] \) for \( 0 \leq i, j \leq 5 \)?
- Are \( D_1 \) and \( D_2 \) independent?

- Suppose \( D_1 \) is plaintext and \( D_2 \) is key, and \( S_1 \) and \( S_2 \) are ciphertexts of two different ciphers, which cipher would you use?
Examples to think after class

- What is the joint probability of $D_1$ and $S_1$?
- What is the joint probability of $D_2$ and $S_2$?

- What is the conditional probability $\Pr[S_1=s \mid D_1=i]$ for $0 \leq i \leq 5$ and $0 \leq s \leq 10$?
- What is the conditional probability $\Pr[D_1=i \mid S_2=s]$ for $0 \leq i \leq 5$ and $0 \leq s \leq 5$?

- Are $D_1$ and $S_1$ independent?
- Are $D_1$ and $S_2$ independent?
Bayes’ Theorem

If \( P[y] > 0 \) then

\[
P[x \mid y] = \frac{P[x]P[y \mid x]}{P[y]}
\]

Corollary

X and Y are independent random variables iff \( P[x\mid y] = P[x] \), for all \( x \in X \) and all \( y \in Y \).
End Math
Ways to Enhance the Substitution Cipher against Frequency Analysis

• Using nulls
  – e.g., using numbers from 1 to 99 as the ciphertext alphabet, some numbers representing nothing and are inserted randomly

• Deliberately misspell words
  – e.g., “Thys haz thi ifekkt off diztaughting thi ballans off frikwenseas”

• Homophonic substitution cipher
  – each letter is replaced by a variety of substitutes

• These make frequency analysis more difficult, but not impossible
Towards the Polyalphabetic Substitution Ciphers

- Main weaknesses of monoalphabetic substitution ciphers
  - In ciphertext, different letters have different frequency
  - Each letter in the ciphertext corresponds to only one letter in the plaintext letter

- Idea for a stronger cipher (1460’s by Alberti)
  - Use more than one cipher alphabet, and switch between them when encrypting different letters
  - As result, frequencies of letters in ciphertext are similar

- Developed into a practical cipher by Vigenère (published in 1586)
The Vigenère Cipher

Treat letters as numbers: [A=0, B=1, C=2, ..., Z=25]

Number Theory Notation: \( \mathbb{Z}_n = \{0, 1, ..., n-1\} \)

Definition:

Given \( m \), a positive integer, \( P = C = (\mathbb{Z}_{26})^n \), and \( K = (k_1, k_2, ..., k_m) \) a key, we define:

Encryption:

\[ e_k(p_1, p_2, ..., p_m) = (p_1+k_1, p_2+k_2, ..., p_m+k_m) \pmod{26} \]

Decryption:

\[ d_k(c_1, c_2, ..., c_m) = (c_1-k_1, c_2-k_2, ..., c_m-k_m) \pmod{26} \]

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
Security of Vigenere Cipher

• Vigenere masks the frequency with which a character appears in a language: one letter in the ciphertext corresponds to multiple letters in the plaintext. Makes the use of frequency analysis more difficult.

• Any message encrypted by a Vigenere cipher is a collection of as many shift ciphers as there are letters in the key.
Vigenere Cipher: Cryptanalysis

- Find the length of the key.
- Divide the message into that many simple substitution encryptions.
- Solve the resulting simple substitutions.
  - how?
How to Find the Key Length?

- For Vigenere, as the length of the keyword increases, the letter frequency shows less English-like characteristics and becomes more random.

- Two methods to find the key length:
  - Kasisky test
  - Index of coincidence (Friedman)
Kasisky Test

- Note: two identical segments of plaintext, will be encrypted to the same ciphertext, if the they occur in the text at the distance \( \Delta \), \( \Delta \equiv 0 \pmod{m} \), \( m \) is the key length.

- Algorithm:
  - Search for pairs of identical segments of length at least 3
  - Record distances between the two segments: \( \Delta_1, \Delta_2, \ldots \)
  - \( m \) divides \( \gcd(\Delta_1, \Delta_2, \ldots) \)
Example of the Kasiski Test

Key  K I N G K I N G K I N G K I N G K I N G K I N G K I N G
PT  the sun and the man in the moon
CT  D P R Y E V N T N B U K W I A O X B U K W W B T
Index of Coincidence (Friedman)

**Informally:** Measures the probability that two random elements of the n-letters string $x$ are identical.

**Definition:**
Suppose $x = x_1x_2...x_n$ is a string of $n$ alphabetic characters. Then $I_c(x)$, the index of coincidence is:

$$I_c(x) = P(x_i = x_j)$$

when $i$ and $j$ are uniformly randomly chosen from $[1..n]$
Index of Coincidence (cont.)

• Consider the plaintext $x$, and $f_0, f_1, \ldots f_{25}$ are the frequencies with which A, B, \ldots Z appear in $x$ and $p_0, p_1, \ldots p_{25}$ are the probabilities with which A, B, \ldots Z appear in $x$.
  • That is $p_i = f_i / n$ where $n$ is the length of $x$

• We want to compute $I_c(x)$.

• Given frequencies of all letters in an alphabet, index of coincidence is a feature of the frequencies
  • It does not change under substitution
Index of Coincidence (cont.)

- We can choose two elements out of the string of size \( n \) in \( \binom{n}{2} \) ways.
- For each \( i \), there are \( \binom{f_i}{2} \) ways of choosing the elements to be \( i \).

\[
I_C(x) = \frac{\sum_{i=0}^{S} \binom{f_i}{2}}{\binom{n}{2}} = \frac{\sum_{i=0}^{S} f_i (f_i - 1)}{n(n-1)} \approx \frac{\sum_{i=0}^{S} f_i^2}{n^2} = \sum_{i=0}^{S} p_i^2
\]
Index of Coincidence of English

- For English, \( S = 25 \) and \( p_i \) can be estimated

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

\[
I_c (x) = \sum_{i=0}^{i=25} p_i^2 = 0.065
\]
Finding the Key Length

\[ y = y_1 y_2 \ldots y_n, \]  
assum m is the key length,  
write \( y \) vertically in an \( m \)-row array.

\[
\begin{bmatrix}
  y_1 & y_{m+1} & \cdots & y_{n-m+1} \\
  y_2 & y_{m+2} & \cdots & y_{n-m+2} \\
  \vdots & \vdots & \ddots & \vdots \\
  y_m & y_{2m} & \cdots & y_n \\
\end{bmatrix}
\]
Finding out the Key Length

- If $m$ is the key length, then the text "looks like" English text

\[ I_c(y_i) \approx \sum_{i=0}^{i=25} p_i^2 = 0.065 \quad \forall 1 \leq i \leq m \]

- If $m$ is not the key length, the text "looks like" random text and:

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

- Basic idea: if the key in Vigenere cipher is very long, then the attacks won’t work
- Implementation idea: multiple rounds of substitutions
- A machine consists of multiple cylinders
  - Each character is encrypted by multiple cylinders
  - Each cylinder has 26 states, at each state it is a substitution cipher
  - Each cylinder rotates to change states according to different schedule
Rotor Machines

- A m-cylinder rotor machine has $26^m$ different substitution ciphers
  - $26^3 = 17576$
  - $26^4 = 456,976$
  - $25^5 = 11,881,376$
Earliest Enigma Machine

- **Use 3 scramblers (motors):** 17,576 substitutions

- **3 scramblers can be used in any order:** 6 combinations

- **Plug board:** allowed 6 pairs of letters to be swapped before the encryption process started and after it ended.
History of the Enigma Machine

- Patented by Scherius in 1918
- Widely used by the Germans from 1926 to the end of second world war
- First successfully broken by the Polish’s in the thirties by exploiting the repeating of the message key
- Then broken by the UK intelligence during the WW II
Coming Attractions …

• Information-Theoretic secrecy (Perfect secrecy), One-Time Pad

• Recommended reading for next lecture: Katz and Lindell: Chapter 2