Cryptography
CS 555

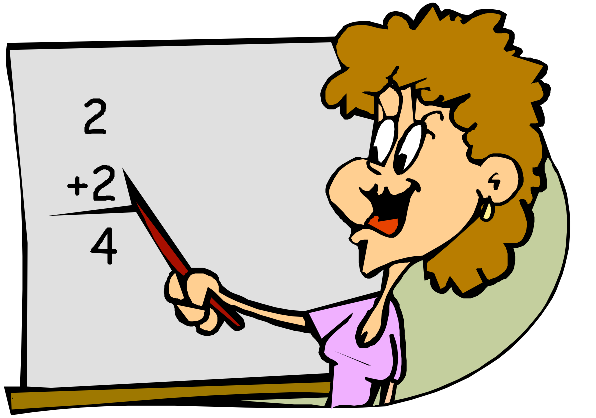
Topic 2: Evolution of Classical Cryptography

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



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

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

$$\sum_{x \in \mathcal{X}} \Pr[x] = 1$$

Example of Random Variables

- Let random variable $\mathbf{D_1}$ 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 $\mathbf{Pr}[\mathbf{D_1}=i] = 1/6$ for $0 \leq i \leq 5$
- Let random variable $\mathbf{D_2}$ denote the outcome of throw a second such dice randomly
- Let random variable $\mathbf{S_1}$ denote the sum of the two dices, then $\mathcal{S}$ ={0,1,2,…,10}, and

  $\mathbf{Pr}[\mathbf{S_1}=0] = \mathbf{Pr}[\mathbf{S_1}=10] = 1/36$

  $\mathbf{Pr}[\mathbf{S_1}=1] = \mathbf{Pr}[\mathbf{S_1}=9] = 2/36 = 1/18$

  …
- Let random variable $\mathbf{S_2}$ denote the sum of the two dices modulo 6, what is the distribution of $\mathbf{S_2}$

Relationships between Two Random Variables

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.

  $$\mathbf{Pr}[x|y] = \mathbf{Pr}[x, y] \ / \ \mathbf{Pr}[y]$$

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

Examples

- Joint probability of $\mathbf{D_1}$ and $\mathbf{D_2}$
  for $0 \le i, j \le 5$, $\mathbf{Pr}[\mathbf{D_1}=i, \mathbf{D_2}=j] = ?$
- What is the conditional probability $\mathbf{Pr}[\mathbf{D_1}=i \mid \mathbf{D_2}=j]$ for $0 \le i, j \le 5$?
- Are $\mathbf{D_1}$ and $\mathbf{D_2}$ independent?

- Suppose $\mathbf{D_1}$ is plaintext and $\mathbf{D_2}$ is key, and $\mathbf{S_1}$ and $\mathbf{S_2}$ are ciphertexts of two different ciphers, which cipher would you use?

Examples to think after class

- What is the joint probability of $\mathbf{D_1}$ and $\mathbf{S_1}$?
- What is the joint probability of $\mathbf{D_2}$ and $\mathbf{S_2}$?

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

- Are $\mathbf{D_1}$ and $\mathbf{S_1}$ independent?
- Are $\mathbf{D_1}$ and $\mathbf{S_2}$ independent?

Bayes' Theorem

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|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: $Z_n = \{0, 1, \ldots, n\text{-}1\}$

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:

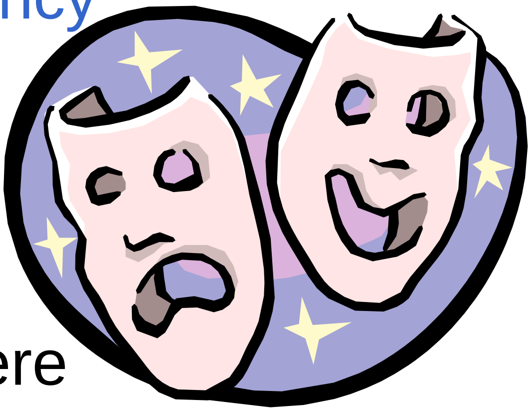
Plaintext: C R Y P T O G R A P H Y

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

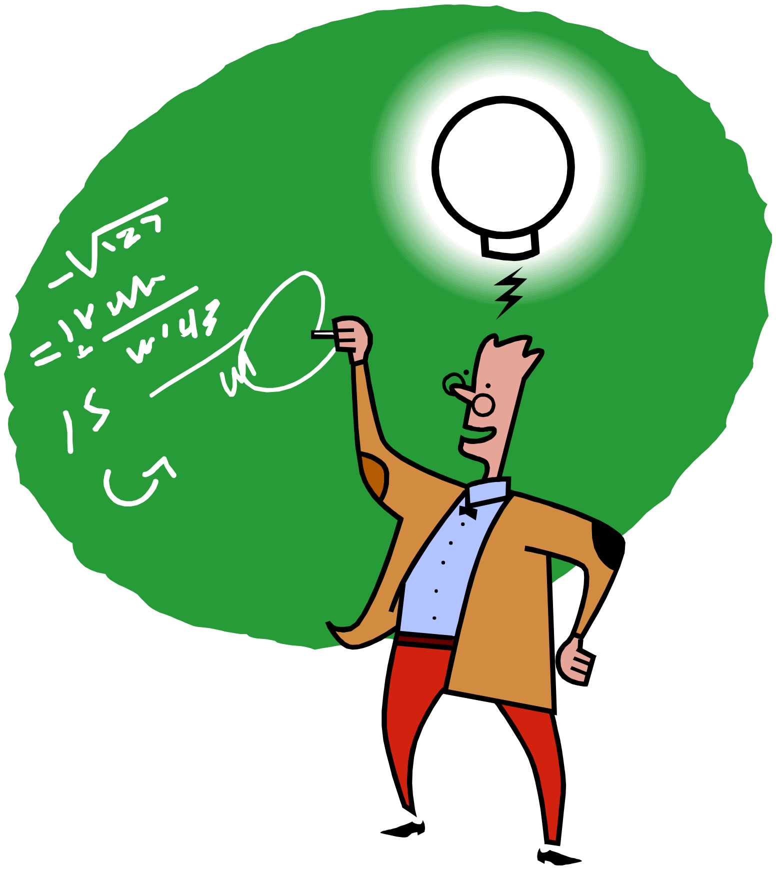
Security of Vigenere Cipher

- 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.



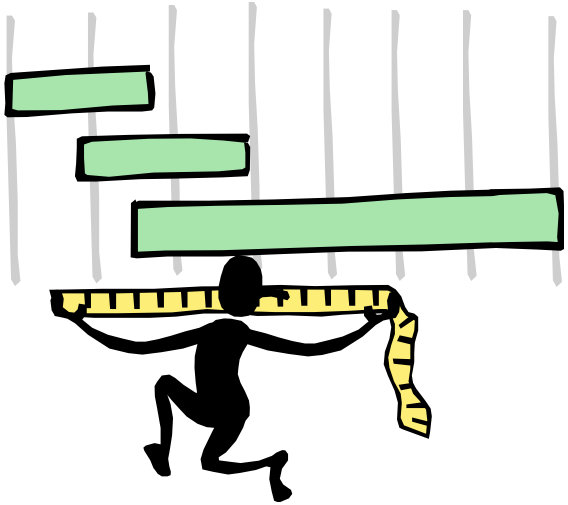
Vigenere Cipher: Cryptanalysis

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



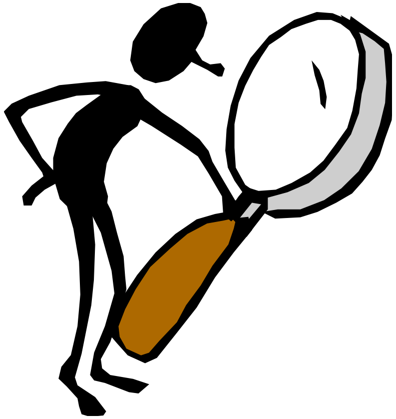
How to Find the Key Length?

- 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 Δ, ($\Delta \equiv 0 \pmod{m}$), m is the key length).
- Algorithm:
  - Search for pairs of identical segments of length at least 3
  - Record distances between the two segments: $\Delta 1$, $\Delta 2$, …
  - m divides gcd($\Delta 1$, $\Delta 2$, …)



Example of the Kasisky 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
PT	t	h	e	s	u	n	a	n	d	t	h	e	m	a	n	i	n	t	h	e	m	o	o	n
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 = $x_1x_2\ldots x_n$ 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, … f_{25} are the frequencies with which A, B, … Z appear in x and p_0, p_1, … p_{25} are the probabilities with which A, B, … 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

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		

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

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

$$\begin{array}{cc} \begin{bmatrix} y_1 & y_{m+1} & \cdots & y_{n-m+1} \\ y_2 & y_{m+2} & \cdots & y_{n-m+2} \\ \cdots & \cdots & \cdots & \cdots \\ y_m & y_{2m} & \cdots & y_n \end{bmatrix} & \begin{array}{c} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \ldots \\ \mathrm{y}_\mathrm{m} \end{array} \end{array}$$

Finding out the Key Length

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

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

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

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

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): **17576 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