Cristina Nita-Rotaru

CS355: Cryptography

Lecture 3: Vigenere cipher.

Towards polyalphabetic substitution ciphers

- Main weaknesses of monoalphabetic substitution ciphers
  - 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
- Giovani Battista Bellaso published it in 1553
- Developed into a practical cipher by Blaise de Vigenère and 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:

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

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

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

Cristina Nita-Rotaru

Security of Vigenere cipher

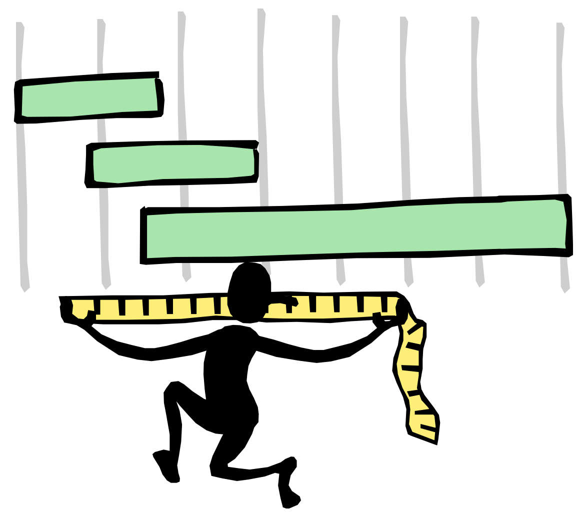
- Vigenere masks the frequency with which a character appears in a language: one letter in the ciphertext corresponds to multiple letters in the plaintext. Makes the use of frequency analysis more difficult
- Any message encrypted by a Vigenere cipher is a collection of as many shift ciphers as there are letters in the key

Vigenere cipher cryptanalysis

- Find the length of the key
- Divide the message into that many shift cipher encryptions
- Use frequency analysis to solve the resulting shift ciphers
  - 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)



History of breaking Vigenere

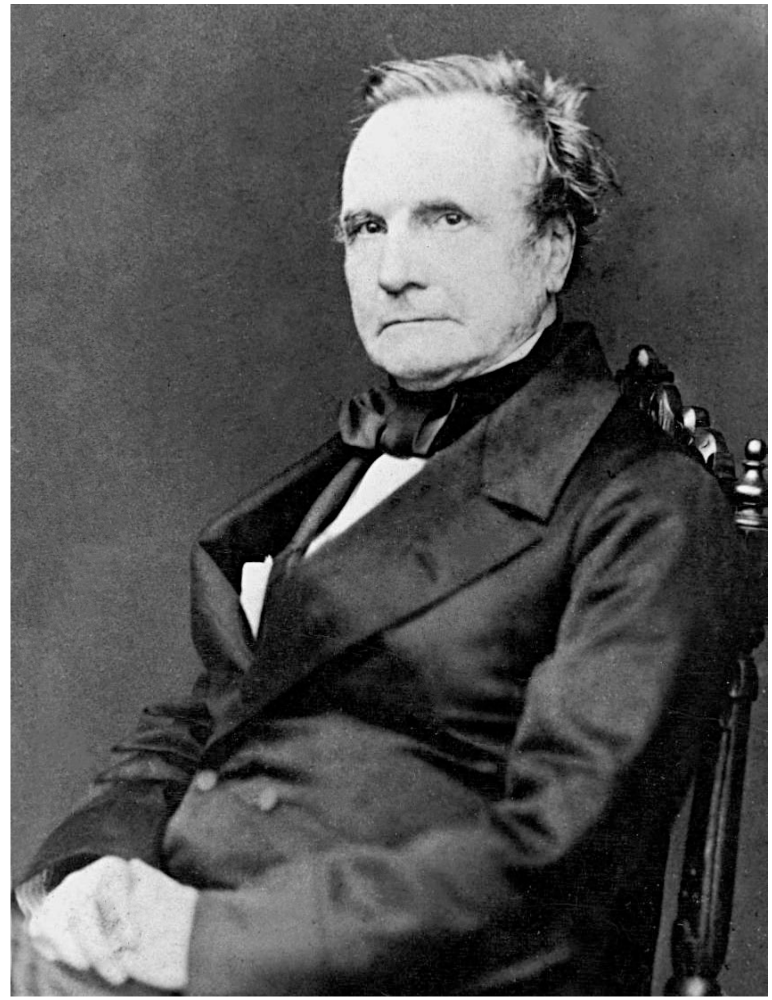
- 1596 - Cipher was published by Vigenere
- 1854 - It is believed the Charles Babbage knew how to break it in 1854, but he did not published the results
- 1863 - Kasiski showed the Kasiski examination that showed how to break Vigenere
- 1920 - Friedman published ``The index of coincidence and its applications to cryptography"

Friedrich Wilhelm Kasiski (1805 – 1881)

- German infantry officer, cryptographer and archeologist.

Charles Babbage (1791 – 1871)

- English mathematician, philosopher, inventor and mechanical engineer who originated the concept of a programmable computer.
- Considered a "father of the computer", he invented the first mechanical computer that eventually led to more complex designs.



William Frederick Friedman (1891 – 1969)

- US Army cryptographer who ran the research division of the Army's Signals Intelligence Service (SIS) in the 1930s, and parts of its follow-on services into the 1950s.
- In 1940, people from his group, led by Frank Rowlett broke Japan's PURPLE cipher machine

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: $\Delta 1, \Delta 2, \ldots$
  - m divides gcd($\Delta 1, \Delta 2, \ldots$)

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:

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

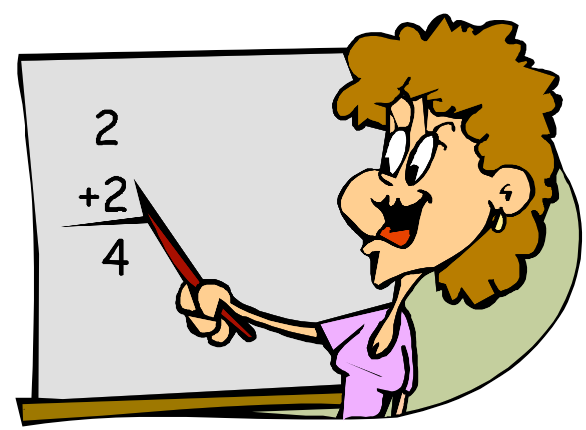
Index of coincidence (cont.)

- Reminder: binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

- Consider the plaintext x, and c_0, c_1, … c_{25} are the number of occurrences 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.
- We want to compute

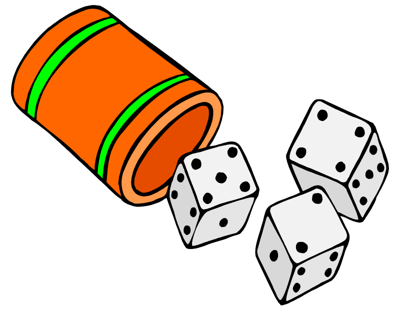
$$I_c(x) = P(x_i = x_j)$$

Begin math



Elements of probability theory

A random experiment has
an unpredictable outcome.



Definition

The sample space (S) of a random phenomenon is the set of all outcomes for a given experiment.

Definition

The event (E) is a subset of a sample space, an event is any collection of outcomes.

Cristina Nita-Rotaru

Basic axioms of probability

If E is an event, *Pr(E)* is the probability that event E occurs then

(a) $0 \leq Pr(A) \leq 1$ for any set ***A in S***
(b) $Pr(S) = 1$, where S is the sample space.
(c) If $E_1, E_2, \ldots E_n$ is a sequence of mutually exclusive events, that is $Ei \cap Ej = 0$, for all $i \neq j$ then:

$$\Pr(E_1 \cup E_2 \cup \ldots \cup E_n) = \sum_{i=1}^{n} \Pr(E_i)$$

More probabilities

If E is an event and *Pr(E)* is the probability that the event E occurs then

- *Pr(Ê)* = 1 - *Pr(E)* where Ê is the complimentary event of E
- If outcomes in S are equally like, then

  Pr(E) = |E| / |S| (where | | denotes the cardinality of the set)

Example

Random throw of a pair of dice.
What is the probability that the sum is 3?

Solution: Each dice can take six different values {1,2,3,4,5,6}. The number of possible events (value of the pair of dice) is 36, therefore each event occurs with probability 1/36.

Examine the sum: 3 = 1+2 = 2+1
The probability that the sum is 3 is 2/36.

What is the probability that the sum is 11? How about 5?

End math

Index of coincidence (cont.)

- Reminder: binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

- Consider the plaintext x, and c_0, c_1, … c_{25} are the number of occurrences 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.
- We want to compute .

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

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 $\binom{c_i}{2}$ ways of choosing the elements to be i

$$I_C(x) = \frac{\sum_{i=0}^{S}\binom{c_i}{2}}{\binom{n}{2}} = \frac{\sum_{i=0}^{S} c_i(c_i - 1)}{n(n-1)} \approx \frac{\sum_{i=0}^{S} c_i^{\,2}}{n^2} = \sum_{i-0}^{S} p_i^{\,2}$$

THIS IS AN APPROXIMATION IF N is VERY BIG

Example: IC of a string

- Consider the text **THE INDEX OF COINCIDENCE**

$$I_C(x) = \frac{\sum_{i=0}^{S} c_i(c_i - 1)}{n(n-1)}$$

- There are 21 characters, so N = 21, S = 25

I_c = (3*2+ 2*1+ 4*3+ 1*0+ 1*0+ 3*2+ 3*2+ 2*1+ 1*0+ 1*0) / 21*20 = **34/420 = 0.0809**

Example: IC of a language

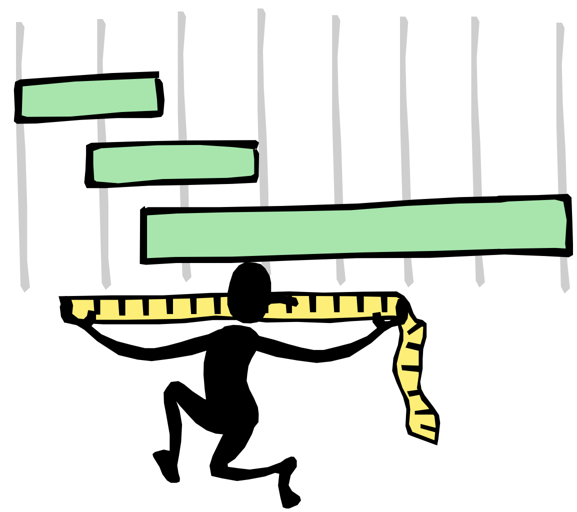
- For English, $S = 25$ and p_i can be estimated

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		

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

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)



Finding the key length

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

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

Guessing the key length

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

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

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

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

Finding the key, once key length known

- Consider vectors y_i, and look for the most frequent letter
- Check if mapping that letter to e will not result in unlikely mapping for other letters
- If that's not the case, look at the shift of the mapping, that represents the letter of the key
- Repeat for each vector

Kasisky example

- Suppose that a Kasiski analysis of the ciphertext from a Vigenere cipher identifies these seven pairs of repeated sequences of ciphertext letters.

First occurence	**8**	**20**	**38**	**48**	**59**	**72**
Second occurence	32	64	110	104	163	132

- What can you say about the length of the key used to encrypt the message?

Index of coincidence example

- A ciphertext of 100 letters was intercepted. The frequency distribution of letters of the alphabet in this ciphertext is as follows:

A 2;	B 10;	C 2;	D 5;	E 3;	F 8;	G 1;	H 2;	I 2;
J 5;	K 1;	L 1;	M 3;	N 2;	O 10;	P 1;	Q 8;	R 1;
S 8;	T 5;	U 2;	V 1;	W 3;	X 5;	Y 1;	Z 8.	

- What is the index of coincidence of this ciphertext?

Index of coincidence example

- Suppose there is a language that has only three letters: a,b,c.
  - frequency of letter a is 0.5
  - frequency of letter b is 0.3
  - frequency of letter c is 0.2

- What is the index of coincidence of the language?

Vigenere example

- EBBBLCKSYMMKTHPDLSPWLCKVJCKDSYMVLCK
- 5 21 33
- Key length dives gcd(21-5, 33-21) = gcd(16, 12) = 4
- EBBB
- LCKS
- YMMK
- THPD
- LSPW
- LCKV
- JCKD
- SYMV
- LCK

Vigenere challenge

- KVAESZXYFQZGGEVPQVRSOHVZTXYXOCG
XQVDRKALKIEKUUCAKXKOGKSSSMMKXUU
GMTEIJSSPWBGVFBZREWZWVVSTVSQZML
NGSURSFYCIINIRGGNUVTLGUWLEEEAYXQC
GLFJGBXLTRTXDXLQVPQKUEXASJFIWXDLF
TFKHJRLVVULVYHUELFJGBXLXYFVVPLFYFJ
EOGRPRXPRCRLQVPQIEZKPRFUQEMGIXB

- 10 bonus points, submit all your code you used to solve it
- Deadline is Sept. 2

Take home lessons

- Vigenère cipher is vulnerable: once the key length is found, a cryptanalyst can apply frequency analysis.