187 - Day 09

Monday, April 23, 2018 9:21 AM

Day 09 – Breaking Vigenère cipher (cont.)

Index of Coincidence

The following table give the relative frequency of the English alphabet letters in a 7834-letter sample of English writing.

Letter	Relative frequency	Letter	Relative frequency
A	0.08399	N	0.06778
B	0.01442	O	0.07493
C	0.02527	P	0.01991
D	0.04800	Q	0.00077
E	0.12150	R	0.06063
F	0.02132	S	0.06319
G	0.02323	T	0.08999
H	0.06025	U	0.02783
I	0.06485	V	0.00996
J	0.00102	W	0.02464
K	0.00689	X	0.00204
L	0.04008	Y	0.02157
M	0.02566	Z	0.00025

$$\frac{\text{freq. of } N}{7834} = p_N$$

prob. that a randomly chosen letter in the plaintext is "N".

The probability that two randomly selected letters in English are identical is given by

$$\sum_{\alpha=A}^{Z} p_\alpha^2 \approx 0.065$$

In a Vigenère cipher with sufficiently long keyword, the probabilities of seeing any letter in the ciphertext will converge to

$$\boxed{\frac{1}{26} = 0.0385}$$

Friedman Test

Definition 1 (Index of Coincidence)**.** *The* ***index of coincidence*** *(for a ciphertext), denoted I, is the probability that two randomly selected letters in the ciphertext are identical.*

Remark:

- If $I \approx 0.065$ then the cipher is more likely to be mono-alphabetic substitution.
- For poly-alphabetic substitution, $0.0385 \leq I \leq 0.065$

in ciphertext.

Theorem 1. *Let $n_0, n_1, n_2, \dots n_{24}, n_{25}$ be the respective counts of the letters $A, B, C, \dots, Y, Z$. Let $n = \sum n_i$ be the total number of letters in the text then*

$$I = \frac{1}{n(n-1)} \sum_{i=0}^{25} n_i(n_i - 1).$$

Now if an English plaintext is encrypted using a Vigenère cipher with keyword of length k, then

$$I \approx \frac{0.0385 \cdot n(k-1) + 0.065(n-k)}{k(n-1)}, \quad \text{or equivalently,}$$

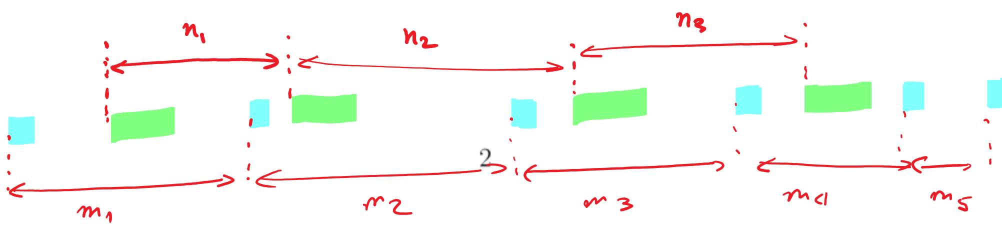
$(0.065 - 0.0385)$

different for another language.

$$\boxed{k \approx \frac{0.0265n}{(0.065 - I) + n(I - 0.0385)}.}$$

Kasiski Test

The **Kasiski Test** is another way of estimating the length of the keyword for Vigenère cipher. It obtains possible keyword lengths from the **gcd of the spacing between repeated letter groups** in the ciphertext.



$$k \approx \gcd(n_1, \dots, n_3, m_1, m_2, \dots, m_5)$$

Cryptanalysis of Vigenère cipher

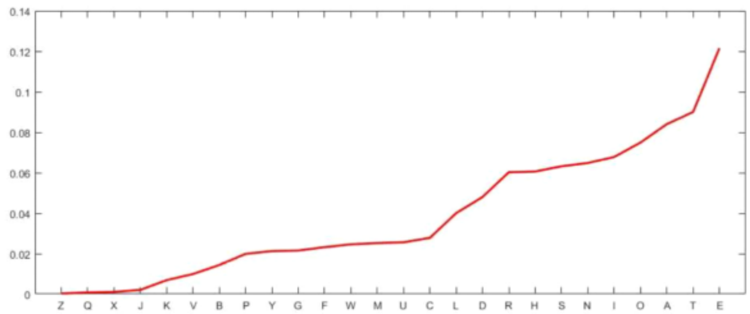
Remark: Both Friedman and Kasiski Tests only give the keyword length, but not the keyword itself. Firthermore, they are not very accurate when the ciphertext is small (usually less than 400 characters).

bad.

The **signature of English** is the graph of letter frequency distribution of English when we sort these frequencies in increasing order.

To draw the signature:

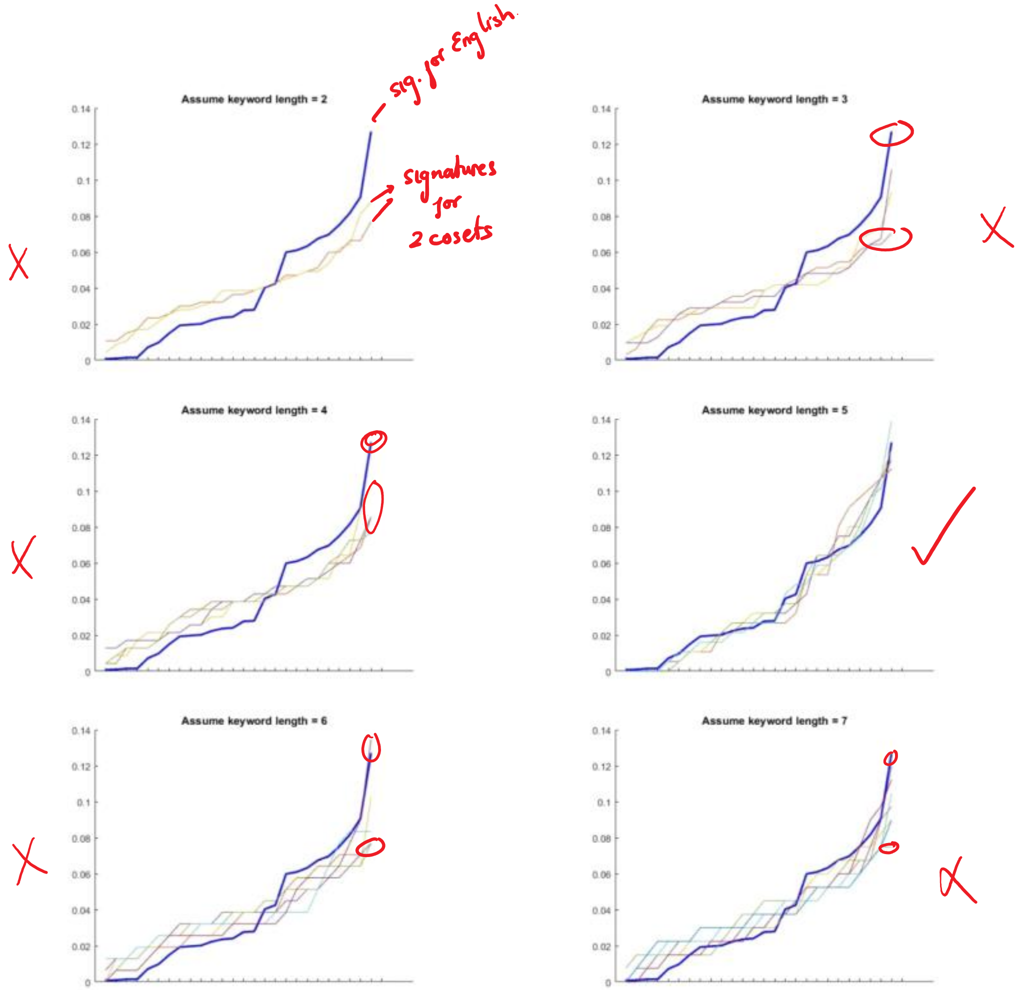
- Obtain the frequency count for each letter
- sort the list in increasing order
- plot these points



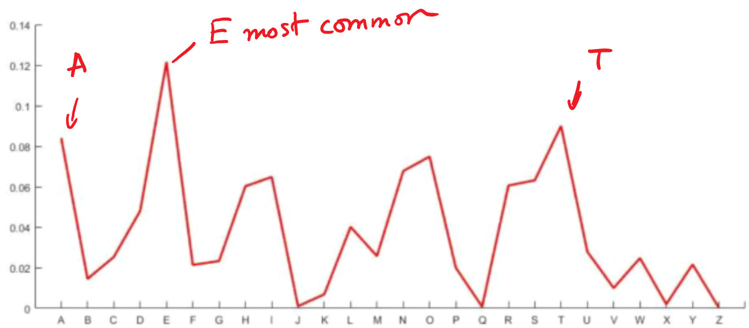
A **coset** are all the letters of the Vigenère ciphertext that are encrypted by the same letter of the keyword.

Remark: A coset of the Vigenère ciphertext has the same encryption as a Caesar shift cipher.

- pick several values for keyword length k.
- draw the signature for all cosets
- If k is the correct value then the graphs will behave similar to the signature of English.


Assume keyword length = 2
Assume keyword length = 3
Assume keyword length = 4
Assume keyword length = 5
Assume keyword length = 6
Assume keyword length = 7
sig. for English
signatures for 2 cosets

The **scrawl of English** is the graph of letter frequency distribution of English in alphabetical order.



- Scrawl of each coset must be of the same shape as above, except that the coset's scrawls may be shifted.

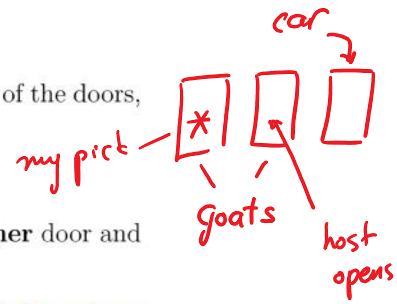
- To get the keyword: try to match the scrawl of each coset to that of English and record the shift distance

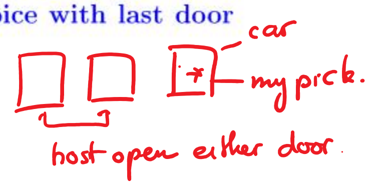
- look at the applet on Assigment page.

Monty Hall Problem

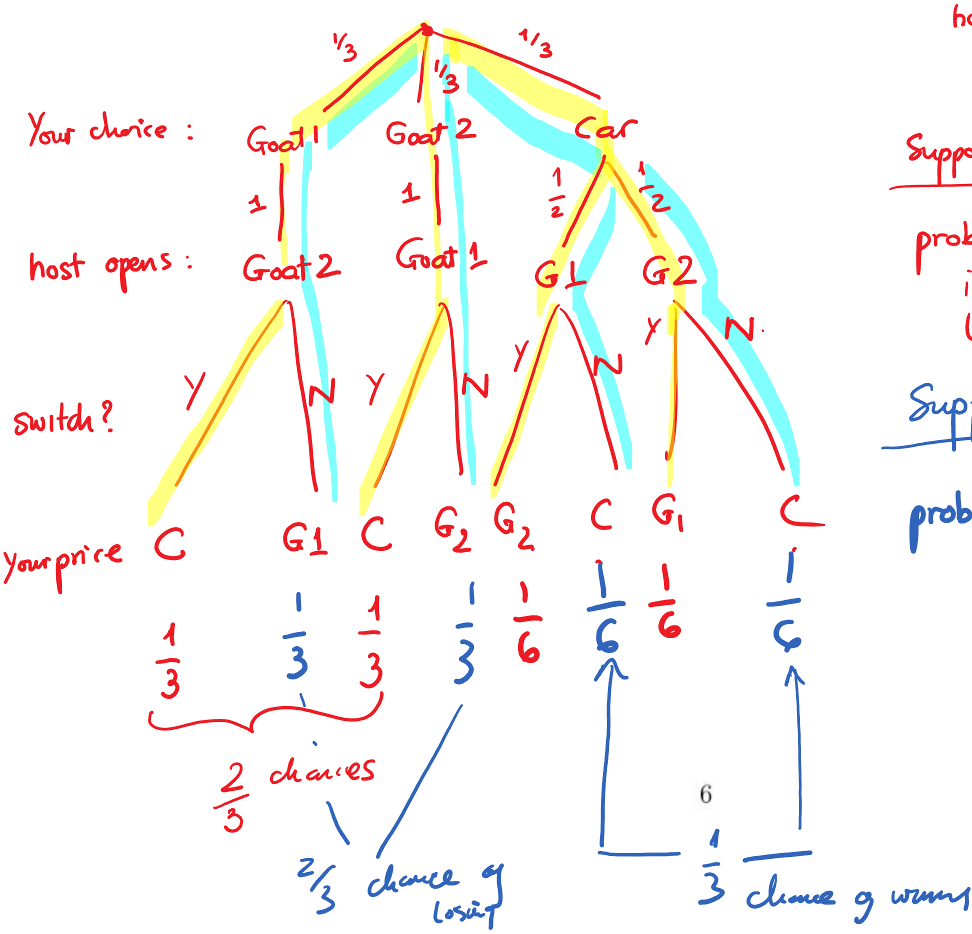
Consider the following game:

> Suppose there are three doors. A car is hidden behind one of the doors, and the other two have goats.
>
> As a player, you pick one of the doors.
>
> The host, who **knows where the car is**, will open **another** door and reveal a goat.
>
> Now you can choose whether to **swap your choice with last door** or **stay with original choice**





What is your best strategy?



Suppose you switch:

prob. that you win if you switch $= \frac{2}{3}$

Suppose you stay:

prob. that you win if you stay $= \frac{1}{3}$

Definition 2. *The **conditional probability** of an event B is the probability that this event will occur, given the knowledge that another event A has already occurred.*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \text{ and } A)}{\mathbb{P}(A)},$$

assuming that $\mathbb{P}(A) > 0$.

Refer to the table for the frequency of character pairs in English language.

Take an English text, pinpoint a letter at position λ

λ

prob. that λ is "A" ?

$\mathbb{P}(\lambda = \text{"A"}) = p_A = 0.08399$

Now if I know the letter μ to the left of λ - how will the prob. change?

μ λ

What is the prob. $\lambda = \text{"A"}$ if μ is:

- $\mu = \text{"L"}$?
- $\mu = \text{"E"}$?
- $\mu = \text{"Q"}$?

use conditional probability & today table.