Cryptanalysis of Classical Ciphers

Debdeep Mukhopadhyay

Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
INDIA -721302

Objectives

- **Models for Cryptanalysis**
- **Cryptanalysis of Monoalphabetic Ciphers**
- **Cryptanalysis of Polyalphabetic Ciphers**
- **Cryptanalysis of Hill Cipher**

Cryptanalysis

- **Kerckhoff's Principle:**
  - **The cryptosystem is known to the adversary.**
  - **But the key is not known to the attacker.**
  - **The secrecy of the cryptosystem lies in the key.**
- **Cryptanalysis is the art of obtaining the key.**

Models for Cryptanalysis

- **Cipher-text only: opponent possesses a string of ciphertext**
- **Known plaintext: opponent possesses a plaintext, x and the corresponding ciphertext, y.**
- **Chosen plaintext: Attacker can choose plaintext, and obtain the corresponding ciphertexts**

Models for Cryptanalysis

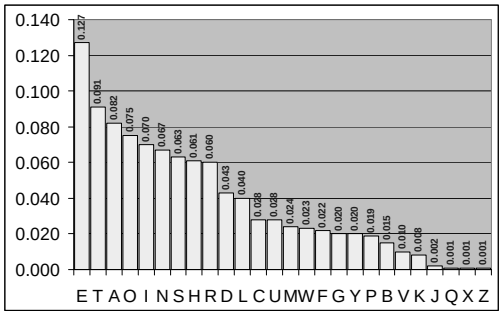
- **Chosen Ciphertext:**
  - **The opponent has temporary access to the decryption function.**
  - **He can choose ciphertexts and decrypt to obtain the corresponding plaintexts.**
- **In each case, the objective is to obtain the key.**
- **Increasing order of strength:**
  - **Ciphertext only < Known plaintext < Chosen Plaintext < Chosen Ciphertext**

Statistical analysis

- **Probabilities of occurrences of 26 letters**
  - **E, having probability about 0.120 (12%)**
  - **T,A,O,I,N,S,H,R, each between 0.06 and 0.09**
  - **D,L, each around 0.04**
  - **C,U,M,W,F,G,Y,P,B, each between 0.015 and 0.028**
  - **V,K,J,X,Q,Z, each less than 0.01**
- **30 common digrams (in decreasing order):**
  - **TH, HE, IN, ER, AN, RE,…**
- **12 common trigrams (in decreasing order):**
  - **THE, ING,AND,HER,ERE,…**

Cryptanalysis of a Monoalphabetic Cipher

- **Ciphertext-only attack**
  - **using letter frequencies in the English language (plaintext character sets)**



Cryptanalysis of Affine Cipher

- Suppose a attacker got the following cipher from an affine cipher:
  - **FMXVEDKAPHFERBNDKRXRSREFNORUDSDKDVSHVUFE DKAPRKDLYEVLRHHRH**
- Cryptanalysis steps:
  - **Compute the frequency of occurrences of letters**
    - **R: 8, D:7, E,H,K:5, F,S,V: 4**
    - **Guess the letters, solve the equations, decrypt the cipher, judge correct or not.**
    - First guess: **R←→e, D←→t,**
      - **Thus, $e_K(4)=17$, $e_K(19)=3$**
      - **Thus, 4a+b=17**
        19a+b=3

      This gives, a=6, b=19, since gcd (6,26)=2, so incorrect.

Cryptanalysis of Affine Cipher

- Next guess: **R←→e, E←→t, the result will be a=13, not correct.**
- Guess again: **R←→e, H←→t, the result will be a=8, not correct again.**
- Guess again: **R←→e, K←→t, the result will be a=3, b=5.**
  - ***K*=(3,5), $e_K(x)=3x+5 \bmod 26$, and $d_K(y)=9y-19 \bmod 26$.**
  - **Decrypt the cipher: *algorithmsarequitegeneraldefinitionsofarithmeticprocesses***
- **If the decrypted text is not meaningful, try another guess.**

- Need programming: compute frequency and solve equations
- Since Affine cipher has 12*26=312 keys, we can write a program to try all keys.

Cryptanalysis of Vigenere cipher

- **In some sense, the cryptanalysis of Vigenere cipher is a *systematic method* and can be *totally programmed*.**
- **Step 1: determine the length *m* of the keyword**
  - ***Kasis*ki test and *index of coincidence***
- **Step 2: determine $K=(k_1,k_2,\ldots,k_m)$**
  - **Determine each k_i separately.**

Kasiski test—determine keyword length *m*

- **Observation: two identical plaintext segments will be encrypted to the same ciphertext whenever they appear δ positions apart in plaintext, where $\delta \equiv 0$ mod *m.* Vice Versa.**
- **So search ciphertext for pairs of identical segments, record the distance between their starting positions, such as δ_1, δ_2,…, then *m* should divide all of δ_i's. i.e., *m* divides gcd of all δ_i's.**

Index of coincidence

- **Can be used to determine *m* as well as to confirm *m,* determined by *Kasiski* test**
- **Definition: suppose $x=x_1x_2,\ldots,x_n$ is a string of length *n*.**
- **The *index of coincidence* of *x*, denoted by $I_c(x)$, is defined to be the probability that two random elements of *x* are identical.**
  - **Denoted the frequencies of A,B,…,Z in *x* by $f_0,f_1,\ldots,f_{25}$**

$$I_c(x) = \frac{\sum_{i=0}^{25}\binom{f_i}{2}}{\binom{n}{2}} = \frac{\sum_{i=0}^{25} f_i(f_i-1)}{n(n-1)}$$

Index of coincidence (cont.)

An Important Property:

Suppose x is a string of English text, denote the expected probability of occurrences of A,B,…,Z by $p_0,p_1,\ldots,p_{25}$ with values from the frequency graph, then:

- probability that two random elements both are A is p_0^2, both are B is p_1^2,…
- then $I_c(x) \approx \sum p_i^2 = 0.082^2+0.015^2+\ldots+0.001^2=0.065$

Question: *if y is a ciphertext obtained by **shift cipher**, what is the $I_c(y)$?*

Answer: should be 0.065, because the individual probabilities will be permuted, but the $\sum p_i^2$ will be unchanged. So, this is an Invariant.

This Property is used to determine the key.

Index of coincidence (contd.)

Therefore, suppose $\boldsymbol{y}=y_1y_2\ldots y_n$ is the ciphertext from Vigenere cipher.

For any given m, divide $\boldsymbol{y}$ into m substrings:

$\mathbf{y_1}=y_1y_{m+1}y_{2m+1}\ldots$

$\mathbf{y_2}=y_2y_{m+2}y_{2m+2}\ldots$

…

$\mathbf{y_m}=y_my_{2m}y_{3m}\ldots$

if m is indeed the keyword length, then each $\mathbf{y_i}$ is a shift cipher, $I_c(\mathbf{y_i})$ is about 0.065.
otherwise, $I_c(\mathbf{y_i}) \approx 26(1/26)^2 = 0.038$.

Index of coincidence (cont.)

For purpose of verify keyword length *m*, divide the ciphertext into *m* substrings, compute the index of coincidence by for each substring. If all IC values of the substrings are around 0.065, then *m* is the correct keyword length. Otherwise *m* is not the correct keyword length.

If want to use I_c to determine correct keyword length *m*, what to do?

Beginning from *m*=2,3, … until an *m*, for which all substrings have IC value around 0.065.

Now, how to determine keyword $K=(k_1,k_2,\ldots,k_m)$? Assume *m* is given.

Determine keyword $K=(k_1,k_2,\ldots,k_m)$

- **Suppose $x=x_1,x_2,\ldots,x_n$ and $y=y_1,y_2,\ldots,y_n$ are strings of n and n' alphabetic characters respectively.**
- **The mutual index of coincidence of x and y, denoted by $MI_c(x,y)$, is the probability that a random element of x is equal to that of y.**
- **Let, the probabilities of A, B, … be $f_0,f_1,\ldots,f_{25}$ and $f_0',f_1',\ldots,f_{25}'$ respectively in *x* and *y*.**

$$MI_c(x,y)=\frac{\sum_{i=0}^{26} f_i f_i'}{nn'}$$

contd.

A	B	…	Z
p_0	p_1		p_{25}

If a k_i is used as a key:

$A+k_i$	$B+k_i$	…	$Z+k_i$
p_0	p_1		p_{25}

What is the probability that in the cryptogram a character is A?

It is the probability corresponding to $j+k_i=0 \Rightarrow j=-k_i \pmod{26}$, that is p_{-ki}

Computing $MI_c(x,y)$

- **The probability that both characters in x and y are A is thus $p_{-ki}p_{-kj}$**
- **The probability that both characters in x and y are B is thus $p_{1-ki}p_{1-kj}$**

$$MI_C(y_i, y_j) = \sum_{h=0}^{25} p_{h-k_i} p_{h-k_j} = \sum_{h=0}^{25} p_h p_{h+k_i-k_j}$$

- **This value of estimate thus depends on the difference $k_i-k_j \pmod{26}$**
- **A relative shift of l yields the same estimate as 26-l**

Mutual Index of Coincidence

- **From the table we can see that is easy to see when k_i-k_j=0**
- **So, we can always fix a y_i and modify y_j (subtracting) from 1 to 25**
- **The value to which we get a MI_c close to 0.065 will indicate the correct k_i-k_j**

k_i-k_j	MI_c
0	0.065
1	0.039
2	0.032
3	0.034
4	0.044
5	0.033
6	0.036
7	0.039
8	0.034
9	0.034
10	0.038
11	0.045
12	0.039
13	0.043

Computing the shift between two keys

Under the key k_i:

A	B	i	Z
f_0	f_1	f_i	f_{25}

Under the key k_j:

A	B	i	Z
f'_0	f'_1	f'_i	f'_{25}

if MI between the two series is 0.065 or close to it => k_i-k_j=0

If not then what?

- **Let us make $k_j=k_j+g$**

A+g	B+g	i+g	Z+g
f'_0	f'_1	f'_i	f'_{25}

So, the frequency of a character being i is f'_{i-g}

Thus, we compute the $MI_c(x,y^g)=(\Sigma f_i f'_{i-g})/nn'$

If, now we have 0.065 or close to it, $k_i=k_j+g$ or, $k_i-k_j=g$

Example (Vigenere Cipher)

- **CHREEVOAHMAERATBIAXXWTNXBEEOP HBSBQMQEQERBWRVXUOAKXAOSXXW EAHBWGJMMQMNKGRFVGXWTRZXWIAK LXFPSKAUTEMNDCMGTSXMXBTUIADNG MGPSRELXNJELXVRVPRTULHDNQWTW DTYGBPHXTFALJHASVBFXNGLLCHRZB WELEKMSJIKNBHWRJGNMGJSGLXFEYP HAGNRBIEQJTAMRVLCRREMNDGLXRRI MGNSNRWCHRQHAEYEVTAQEBBIPEEW EVKAKOEWADREMXMTBHHCHRTKDNVR ZCHRCLQOHPWQAIIWXNRMGWOIIFKEE**

Example

- CHR**EEVOAHMAERATBIAXXWTNXBEEOPHB SBQMQEQERBWRVXUOAKXAOSXXWEAHB WGJMMQMNKGRFVGXWTRZXWIAKLXFPSK AUTEMNDCMGTSXMXBTUIADNGMGPSRELX NJELXVRVPRTULHDNQWTWDTYGBPHXTFA LJHASVBFXNGLL**CHR**ZBWELEKMSJIKNBHW RJGNMGJSGLXFEYPHAGNRBIEQJTAMRVLC RREMNDGLXRRIMGNSNRW**CHR**QHAEYEVTA QEBBIPEEWEVKAKOEWADREMXMTBHH**CH R**TKDNVRZ**CHR**CLQOHPWQAIIWXNRMGWOII FKEE**

Computation of m

- **The text CHR, starts at 1, 166, 236 and 286.**
- **The distance between the first occurrence and successive ones are 165, 235 and 285.**
- **Thus m=gcd(165,235,285)=5.**
- **We verify m, by computing the IC by trying m=1, 2, 3, 4, 5**

Verifying m by Kasiski Test

- CHR**EEVOAHMAERATBIAXXWTNXBEEOP HBSBQMQEQERBWRVXUOAKXAOSXXW EAHBWGJMMQMNKGRFVGXWTRZXWIAK LXFPSKAUTEMNDCMGTSXMXBTUIADNG MGPSRELXNJELXVRVPRTULHDNQWTW DTYGBPHXTFALJHASVBFXNGLL**CHR**ZB WELEKMSJIKNBHWRJGNMGJSGLXFEYP HAGNRBIEQJTAMRVLCRREMNDGLXRRI MGNSNRW**CHR**QHAEYEVTAQEBBIPEEW EVKAKOEWADREMXMTBHH**CHR**TKDNVR Z**CHR**CLQOHPWQAIIWXNRMGWOIIFKEE

Verifying m by Kasiski Test

- CHR**EEVOAHMAERATBIAXXWTNXB EEOPHBSBQMQEQERBWRVXUOAK XAOSXXWEAHBWG**

Kasiski Test

- **A:7 M:2 U:1**
- **B:6 N:1 V:2**
- **C:1 O:4 W:4**
- **E:8 P:1 X:7**
- **G:1 Q:3**
- **H:4 R:4**
- **I:1 S:2**
- **K:1 T:2**

$I_C(x)=0.065$

This will be for all the other four rows.

If the m is anything other than 5, the $I_c(x)$ (index of co-incidence) is around 0.04

This confirms the value of m.

Now what is the key?

- **There are 313 characters in the text.**
- **It is divided into 5 rows, each having 62 characters, the last row having the remaining.**
- **Each row of the table has been shifted by the same key.**
- **So, its Index of Coincidence was 0.06**
- **Now we need to compute the shifts by the MI test.**

The decrypted Text

- **The almond tree was in tentative blossom. The days were longer often ending with magnificent evenings of corrugated pink skies. The hunting session was over with hounds and guns put away for six months. The vineyards were busy again as the well organized farmers treated their vines and more lackadaisical neighbors hurried to do the pruning they should have done in November.**

Another Example

LIOMWGFEGGDVWGHHCQUCRHRW
AGWIOWQLKGZETKKMEVLWPCZV
GTHVTSGXQOVGCSVETQLTJSUMV
WVEUVLXEWSLGFZMVVWLGYHCU
SWXQHKVGSHEEVFLCFDGVSUMPH
KIRZDMPHHBVWVWJWIXGFWLTSH
GJOUEEHHVUCFVGOWICQLTJSUX
GLW

Kasiski Test

String	First Index	Second Index	Difference
QLT	65	165	100
LTJ	66	166	100
TJS	67	167	100
JSU	68	168	100
SUM	69	117	48
VWV	72	132	60

Kasiski Test thus predicts key size is the gcd, which is 4.

Confirmation of Kasiski Test

1st string :
LWGWCRAOKTEPGTQCTJVUEGVGUQGECVPRPVJGTJEUGCJG
IC = 0.067677

2nd string :
IGGGQHGWGKVCTSOSQSWVWFVYSHSVFSHZHWWFSOHCOQSL
IC = 0.074747

3rd string:
OFDHURWQZKLZHGVVLUVLSZWHWKHFDUKDHVIWHUHFWLUW
IC = 0.070707

4th string:
MEVHCWILEMWVVXGETMEXLMLCXVELGMIMBWXLGEVVITX
IC = 0.076768

Computing the shift of each row

- **Then we perform the Mutual Index of Coincidence to obtain the actual key value.**
- **Running the test, we obtain that the key value is CODE, and the corresponding plaintext is:**

**JULIUSCAESARUSEDACRYPTOSYSTEMINHISW
ARWHICHISNOWREFERREDTOASCAESARCIPH
ERITISASHIFTCIPHERWITHTHEKEYSETTOTHRE
EEACHCHARACTERINTHEPLAINTEXTISSHIFTER
THREECHARACTERSOCREATEACIPHERTEXT**

Cryptanalysis of Hill Cipher

- **Cipher-text only attack is difficult**
  - **Large key space**
  - **Hill ciphers do not preserve the statistics of the plaintext.**
  - **Frequency analysis does not work.**
  - **For a key matrix of size m x m, a frequency analysis of size m may work, but it is very rare for the plaintext to have strings of same characters of size m.**

Known-plaintext attack

- **However known-plaintext attack is possible.**
- **Eve can create two m x m matrices, P (plaintexts) and C (ciphertext).**
- **If the key matrix is K, we have:**

  $C = P K$,

  Here every row of C and P are corresponding ciphertext/plaintext pairs.

Thus, $K = P^{-1} C$ (if P is invertible)

Example

- **Assume m=3.**
- **Some known plaintext/ciphertext pairs:**

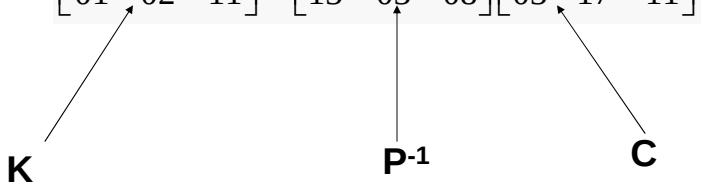
  [05 07 10] → [03 06 00]

  [13 17 07] → [14 16 09]

  [00 05 04] → [03 17 11]

Recovering the Key

$$\begin{bmatrix} 02 & 03 & 07 \\ 05 & 07 & 09 \\ 01 & 02 & 11 \end{bmatrix} = \begin{bmatrix} 21 & 14 & 01 \\ 00 & 08 & 25 \\ 13 & 03 & 08 \end{bmatrix} \begin{bmatrix} 03 & 06 & 00 \\ 14 & 16 & 09 \\ 03 & 17 & 11 \end{bmatrix}$$



Points to Ponder

- **Why does a Hill cipher disturb the frequency of the plaintext?**
- **Write a C Program to automate the Cryptanalysis of Polyalphabetic Ciphers.**

References

- **B. A. Forouzan and D. Mukhopadhyay, *"Cryptography and Network Security"*, TMH, 2nd Edition.**