Analysis of Four Historical Ciphers Against Known Plaintext Frequency Statistical Attack

International Journal of Integrated Engineering:
Special Issue 2018: Data Information Engineering, Vol. 10 No. 6 (2018) p. 183-192.
 DOI: https://doi.org/10.30880/ijie.2018.10.06.026

Analysis of Four Historical Ciphers Against Known Plaintext Frequency Statistical Attack

Chuah Chai Wen[1*], Vivegan A/L Samylingam[2], Irfan Darmawan[3], P.Siva Shamala A/P Palaniappan[4], Cik Feresa Mohd. Foozy[5], Sofia Najwa Ramli[6], Janaka Alawatugoda[7]

[1,2,4,5,6]Information Security Interest Group (ISIG), Faculty Computer Science and Information Technology University Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
E-mail: cwchuah@uthm.edu.my, svivegan92@gmail.com, {shamala, feresa, sofianajwa}@uthm.edu.my
[3]School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia
[7]Department of Computer Engineering, University of Peradeniya, Sri Lanka
E-mail: alawatugoda@eng.pdn.ac.lk

Received 28 June 2018; accepted 5August 2018, available online 24 August 2018

Abstract: The need of keeping information securely began thousands of years. The practice to keep the information securely is by scrambling the message into unreadable form namely ciphertext. This process is called encryption. Decryption is the reverse process of encryption. For the past, historical ciphers are used to perform encryption and decryption process. For example, the common historical ciphers are Hill cipher, Playfair cipher, Random Substitution cipher and Vigenère cipher. This research is carried out to examine and to analyse the security level of these four historical ciphers by using known plaintext frequency statistical attack. The result had shown that Playfair cipher and Hill cipher have better security compare with Vigenère cipher and Random Substitution cipher.

Keywords: Historical Ciphers, Hill Cipher, Playfair Cipher, Random Substitution Cipher, Vigenère Cipher, Known

1. Introduction

Cryptography is the science of writing in secret code. Cryptography was first found being used by Egyptian on circa 1900 B.C. The Egyptian practices their cryptography in their daily life. For example, a non-standard hieroglyph is used by the Egyptian to scribe in an inscription.

Later, the non-standard cryptography is standardized by using some secret writing methods. The common secret writing techniques are substitution and transposition the letters in a paragraph into unreadable text. The unreadable text of secret writing, namely ciphertext. The ciphers that used these two methods are commonly known as historical ciphers. The examples of historical ciphers are Playfair cipher, Hill cipher, Vigenèrecipher and Random Subsitution cipher.

Historical ciphers hiding a secret message which caught the eyes of adversary who try to read the secret message. Based on Kerckhoff's principle [1, 2] the cryptography algorithm is publicly known, and the security relies on the properties of the cryptography keys that it used. This means, the adversary knows the historical ciphers algorithm and the keys (encryption key and decryption key) of the historical ciphers are secret to them.

Known plaintext frequency statistical attack can be used to attack these history ciphers. Meaning that, the adversary is given the frequency statistical of plaintext such as entropy, frequency of bigram, trigram and autocorrelation. At the same time, the adversary may produce the similar frequency statistical for the corresponding ciphertext. These two set of information are used to find the decryption key for the historical ciphers.

This research aims to cryptanalyse Playfair cipher, Random substitution cipher, Hill cipher and Vigenère cipher by using known plaintext frequency statistical attack. The security level of these four historical ciphers areanalysed and are investigated. Hence, one may know which historical cipher more reliable security if compare with each other. Then, the results may be referred to protect the confidentiality of the systems which may not require intensive security such that [3 – 5]. This may help in minimizing cost of implementation and maintenance.

2. Literature Review

*Correspondingauthor: cwchuah@uthm.edu.my

Cryptography is a method of storing and transmitting data. Cryptography may provide integrity and for confidentiality to the authorized person to read and process. Cryptography has encryption and decryption process. Encryption process is define as $C= E(K_E,P)$, where C is the ciphertext, P is the plaintext, K_E is encryption key and E is the encryption process. Decryption process is define as $P = D(K_D,C)$, where K_D is the decryption key and D is the decryption process [1, 2].

Historical cipher is widely used during the middle age where the wars happen. This cipher is used to send secret messages. There are four type of ciphers are reviewed. Those ciphers are Hill Cipher [6, 7], Random Substitution Cipher [1, 2, 8], VigenèreCipher [1, 2, 9, 10] and Playfair Cipher [11, 12].

2.1 Hill Cipher

Hill cipher is invented by Lester S.Hill in 1929. This cipher used the polygraphic substitution cipher. Hill cipher also known as matrix cipher due to, using matrix to encrypt and decrypt this cipher [6, 7]. Hill cipher is applied based on linear algebra.

Encryption process of Hill Cipher has seven steps. First step, all the space and punctuation are removed. Next, the letters are capitalized. Third step, the letters are converted into number based on their position in the English letters. Fourth step is the string is divided into block of size *n*. Fifth step, the string is converted into matrix. Next, this matrix is to be multiply with the encryption key based on the formula $C= KP$. Lastly all the number in the matrix is converted into alphabets.

For example: "how are you?" is encrypted by using 2x2 matrix. The key used for encryption is $\begin{pmatrix}16 & 03\\25 & 21\end{pmatrix}$.

Step 1: The plaintext to encrypted is "how are you".

Step 2: the plaintext is capitalized "HOW ARE YOU".

Step 3: The plaintext is converted into numbers "7 14 22 0
17 4 24 14 20"

Step 4: The string is divided in to block size *n* "7 14 | 22 0
| 17 4 | 24 14 | 20 *x*|", the unknown *x* is padded. Where *x* can be any number from 0-26.

Step 5:The first four numbers are changed to the matrix $\begin{pmatrix}7 & 20\\14 & 0\end{pmatrix}$

Step 6: The matrix is multiply with the encryption key, $K=\begin{pmatrix}16 & 03\\25 & 21\end{pmatrix}$, $C=\begin{pmatrix}16 & 03\\25 & 21\end{pmatrix}\begin{pmatrix}7 & 20\\14 & 0\end{pmatrix}$

Step 7:The number in the matrix is converted into alphabet such that $C =\begin{pmatrix}24 & 14\\1 & 4\end{pmatrix}=\begin{pmatrix}Y & O\\B & E\end{pmatrix}$.

This matrix to be written in a text and it will be written as YBOE. This matrix is just for first four alphabets. The entire ciphertext is YBOEYPKKIG.

Decryption ciphertext of Hill Cipher has four steps. First step the decryption key is identified. The formula for decryption key is formula $K=K^{-1}$ *(mod 26)*. Secondly the ciphertext is converted into matrix and the letter is written in alphabet based on its position in english alphabet. Thirdly calculate $K^{-1}C=P$. Lastly, the matrix is converted into plaintext [6].

The ciphertext to decrypt is "YBOE"

Step 1: the decryption key is identified.

$D = K^{-1}$ *(mod 26)*

$D = \begin{pmatrix}16 & 03\\25 & 21\end{pmatrix}^{-1}$ *(mod 26)*

$D =$ *[(16 x 21) – (3 x 25)]*$^{-1}$ *x* $\begin{pmatrix}21 & -03\\-25 & 16\end{pmatrix}$ *(mod 26)*

$D = \begin{pmatrix}21 & 23\\1 & 16\end{pmatrix}$

Step 2: The ciphertext is converted into matrix and written in number.

$C= \begin{pmatrix}24 & 14\\1 & 4\end{pmatrix}$

Step 3: Calculate $P= DC$

$P= \begin{pmatrix}21 & 23\\1 & 16\end{pmatrix}\begin{pmatrix}24 & 14\\1 & 4\end{pmatrix}$

$P = \begin{pmatrix}7 & 20\\14 & 0\end{pmatrix}$

Step 4: the matrix is converted into plaintext

The process is proceeding for entire ciphertext until recover the plaintext which is "How are you"

2.2 Random Substitution Cipher

Random Substitution Cipher is monoalphabetic cipher. This random substitution cipher is introduced by substituting an alphabet with another alphabet randomly. The number of possible keys for this cipher is 26!. Which is almost equals to 4.03x1026. The key is too large to be brute force by modern computers. However, one may cryptanalyze the Random Substitution Cipher by analyzing its frequency of English alphabet [1, 2].

Encryption of Random Substitution Cipher has three steps. First, all the punctuation at the text are removed. Next, a key for encryption is identified. Lastly, the letters are substituted based on the key. For example: Encrypt "HOW ARE YOU?"

Step 1: The plaintext to encrypted is "HOWAREYOU"

Step 2: The key for encryption of this cryptogram which is hill cipher is identified, such that "ZACBDFEHGJILKMNOPRQTSVUWYX". This keys means "A" is substituted with "Z" and "B" is substituted with "A" and respectively.

Step 3:The letters are substituted based on the key. "HOWAREYOU" will become "HNUZRDYNS".

Decryption process for Random Substitution Cipher has three steps. First, the key of decryption is identified. Next, the ciphertext is decrypted by using the key which is invert from the encryption key. Lastly, insert the punctuation to the plaintext. For example: "HNUZRDYNS" is the ciphertext.

Step 1: Identify the keys:

K_E = "ZACBDFEHGJILKMNOPRQTSVUWYX". The alphabet is arranged and is replaced according to the placement of the alphabets.

K_D = "BDCEGFIHKJMLNOPQSRUTWVXZYA".

Step 2: The decryption key is used to find out the plaintext. Plaintext: "HOWAREYOU".

Step 3: Insert the punctuation and the spacing. The end plaintext is "HOW ARE YOU?".

2.3 Playfair Cipher

Playfair cipher is a polygraphic substitution cipher. Playfair Cipher is invented by Charlses Wheatstone on 1854. This cipher encrypts a pair of letters (diagraphs) instead of single letter. The possible key for this cipher are (25!)/(52). There are 3.8778025x1026 possible keys. These keys are lower than the Random Substitution cipher.

A key is formed for both encryption and decryption. The formation of keys requires five steps. First, draws a 5x5 matrix. This is because only 25 characters is used as a key in Playfair Cipher. Second, fill up the chosen keyword in the 5x5 matrix. Third, deducts the repeated letter in the keyword. Fourth, fill up the rest of the remaining letter according to alphabetic order which not in the keyword. Lastly, place "I" and "J" in the same box [11, 12].

For example, "JOHN" is chosen as the keyword. Then the key for this Playfair Cipher will be as shown below.

I/J	O	H	N	A
B	C	D	E	F
G	K	L	M	P
Q	R	S	T	U
V	W	X	Y	Z

To encrypt the plaintext using Playfair cipher. Firstly, the spaces of plaintext are removed. Then the lowercase letters are changed to uppercase and the letters are paired up. The letter "J" will be replaced with letter "I". Then the double letter is separated and the letter "X" to be add in first letter of the paired double letter. Whereas the second letter will take the next letter as its pair. If there is a letter that not paired up a random letter is chosen to padded with it [11, 12].

Encryption of Playfair cipher have three simple rules. The first rule is, if the both letter in digraph in same row should be replaced with the letter in the right. The second rule is, if the both letter in the digraph is in same column then it is replaced by the letter below in the same column. Lastly, if both digraph is not in same row nor column, then the substitution is based upon the intersection of the letter [11, 12].

For example, the sentence is encrypted "How Are You Uncle?". The spaces of the plaintext are removed and we got "HOWAREYOUUNCLE". Then it is to be split to digraph. "HO WA RE YO UU NC LE". Then there is a digraph "UU" which have same characters then it is split to "UX UN". The new plaintext will be "HO WA RE YO UX UN CL E". Since E is single then T is to be add randomly to make it to digraph. So the plaintext will be "HO WA RE YO UX UN CL ET".The detail encryption process is as below.

Step 1: The letter in same row to be substituted based on the key JOHN.The letter "HO" will be "NH". This is because "H" will change to "N" which is the letter at the left hand side of "H". While letter "O" will change to "H" based on the same rule.

I/J	O	H	N	A
B	C	D	E	F
G	K	L	M	P
Q	R	S	T	U
V	W	X	Y	Z

Step 2: The letter in same column is replaced by the letter below in the same column.The letter "ET" will be "MY". This is because "E" will change to the letter below it which is "M". While the letter "T" will change to "Y" due to same rule.

Step 3: The letters is not in the same row nor same column, the substitution is basedupon their intersection"WA RE YO UX UN CL" will be changed to "ZO TC WN SZ TA DK". The ciphertext is "NH ZO TC WN SZ TA DK MY".

The decryption process for there are three steps to be followed but in reverse mode. First step is both letter in same row replaced with the left letter. Both letter in the same column replaced by the letter above in the same column. At last if the letters not in same row nor same column, the substitution is based upon their intersection. After all the decryption process is done then, if there are extra characters that don't make sense in the plaintext then locate any missing J's was replaced. For example, decrypt the "NH ZO TC WN SZ TA DK MY".

Step 1: Both letter in same row replace the left letter. The pair of letter "NH" is replaced to "HO". This is because the letter "N" is moved to left and changes to be "H". The letter "H" will be "O" based on the same rule.

I/J	O	H	N	A
B	C	D	E	F
G	K	L	M	P
Q	R	S	T	U
V	W	X	Y	Z

Step 2: Both letter in same column replace the letter above in the same column.The pair of letter "MY" will replace to "ET". This is because the letter "M" will move a letter above and replaced to "E". While letter "Y" change to "T" based on same rule.

Step 3: Not in same row nor column, the substitution is based upon their intersection. "ZO TC WN SZ TA DK" will changed to "WA RE YO UX UN CL".

Lastly, clear all X and extra letters at the end. "HO WA RE YO UX UN CL ET". Then "HOWAREYOUUNCLE" will be after X and T is deducted.

2.4 Vigenère Cipher

Vigenère Cipher was first introduced by a French diplomat, Blaise de Vigenère in 1523 -1596. They use multiple number of Caesar Cipher. The number of Caesar Cipher is based on the key length. This Caesar Cipher is written in tabula recta to encrypt and decrypt a message. The number of possible key is 26*L*, where *L* is the key length [1, 2].

A keyword is selected to encrypt a plaintext using Vigenère Cipher. This keyword is repeated until it fulfill the number of words in a sentence of the plaintext. Then the letter which is encrypted are found at the column's letter which intercept with the row's letter based on tabula recta. The column's letter represents the keyword whereas the row represents the plaintext. The letters inside the tabula recta is the ciphertext. After being encrypted the ciphertext to be written. For example: encrypt the plaintext is "HOW ARE YOU"

Step 1: Let "BOY" as a keyword to encrypt the plaintext.

Step 2: The keyword of "BOY" is repeated three times to fulfil the nine letter in "HOWAREYOU".

Step 3: Tabula recta is used to find out the ciphertext where the ciphertext is "ICU BFC ZCS".

The decryption process of Vigenère Cipher is followed as below. First, uses the same key which used in encryption process. The key is repeated until it fulfils the ciphertext size. Then, finds the letter of the ciphertext inside the tabula recta and looks at the column that the ciphertext intercepted with the key. The sentence we write will be the plaintext. For Example: "ICUBFCZCS" is the ciphertext and the key used is "BOY".

Step 1: "BOY" is the decryption key.

Step 2: Repeats the key "BOY" to fulfil the ciphertext.

Step 3: Find the intercepted word using tabula recta. The plaintext is "HOW ARE YOU".

2.5 Known Plaintext Frequency Statistical Attack

Known plaintext frequency statistical attack, adversary is given the frequency statistical of plaintext such as entropy, frequency of N-gram (bigram and trigram) and autocorrelation, together withthe corresponding ciphertext. With this the adversaryneeds to deduce the key that is used to encrypt and decrypt the message. Adversary guesses the plaintext by using the information and perform the attack to the historical cipher to find the decryption key.

2.6 Index of Coincidence

Index of coincidence (I_c) is a measurement of the variation of character frequencies in text from a uniform distribution of the probability that two randomly chosen characters are equal [13]. I_c is based on the ratio of the times the character appears to the total number of characters. The number of characters appear in a text is taken and multiply with the number of characters minus one. Then divided by the ciphertext length multiply with the ciphertext length minus one. Next, all the value is sum up to get probability of two characters in the ciphertext. The equation if I_c as below. Noted that 26 is the total number of English character. The formula is, $I_c = \lim_{n\to\infty} \sum \frac{(n(n-1))}{26n(26n-1)}$. Due to the summation of both part of characters is same. Therefore : $I_c = \lim_{n\to\infty} \frac{26n(n-1)}{26n(26n-1)} = \lim_{n\to\infty} \frac{(n-1)}{(26n-1)}$ [13]. Therefore, I_c is approximately equals to 1/26 when the n is increased to infinity. It giving out information on how often the characters could theoretically appear next to each other. It explains how evenly distributed the character's frequencies are within the frequency distribution table.

2.7 Kasiski Method

Kasiski method is used to allow cryptanalyst to deduce the length of the keyword. To perform kasiski method the identical trigrams and their distance is tabulated. If the distance of these trigram match the same plaintext trigram, the distance will be the multiple of the key length. The great common divisor of the separation distance will be counted for the trigram which occur more than once. If the factor jumps out as occurring in great common divisor, then it will be the key length [1, 2].

2.8 Autocorrelation

Autocorrelation refers to the correlation of a time series with its own past and future values. The autocorrelation function is used to detect the non-randomness in data. Positive autocorrelation meant a tendency for a system remain a same state. The autocorrelation also used to identify an appropriate time series model if the data are not random. The equation of the autocorrelation as per below [14].

$$Rk = \frac{\sum_{n=1}^{N-k}(Y_n - \bar{Y})(Y_{n+k} - \bar{Y})}{\sum_{n=1}^{N}(Y_n - \bar{Y})^2} \quad (1)$$

2.9 Entropy

Entropy is computed as a function of a probability distribution. Entropy is used to find out the information on stumbling block of bits of information technology, randomness and obfuscation. Based on Shannon, Entropy is a demonstrated mathematical methods of treating communication channels, bandwidth and the effects of random noise [15]. Given the entropy as

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \quad (2)$$

p is the probability of a given message and n is the number of possible messages.

3. Methodology

Four type of historical ciphers are studied. There are Playfair cipher, Hill cipher, Vigenère cipher and Random substitution cipher. The algorithms and security properties for these four historical ciphers are examined. The cryptool is used to aid in cryptanalysing these historical ciphers.

The aim of this research is to cryptanalyze the historical ciphers by finding the decryption key. There are five major steps to complete this research: encrypt the plaintexts, analysis plaintext frequency, analysis the ciphertexts, cryptanalyse the ciphertext and result analysis as shown in Fig. 1. Firstly, two set of plaintext are encrypted using four historical ciphers with some secret random key. The plaintext and key are kept secret. At the same time, plaintext is analysed. Output of the encryption is the ciphertext. The ciphertext is analysed and cryptanalysed. Four analysis methods are used to

classified the ciphertext are entropy, frequency of histogram, frequency of N-Gram and autocorrelation.

Autocorrelation is chosen because it may use to identify theVigenèrecipher. The autocorrelation for the Vigenèrecipher have the fixed time peak. The differ between the two peaks is the key length for the Vigenèrecipher.

Entropy is chosen because it may use to identify the Hill cipher. If the entropy for the ciphertext is high and the autocorrelation of the ciphertext do not have fixed key length, then the ciphertext is encrypted using Hill Cipher. This is because this cipher is used by matrix and the information hidden is higher.

Histogram is used to identify Playfair Cipher. If the histogram has 25 alphabets and there are no Trigram, then by no choice it would be the Playfair cipher. This is because the key size is just 25 characters and there is no chance for the ciphertext to be in 26 characters.

To identified the Random Substitution cipher, one may compare the frequency diagram of the plaintext and the frequency diagram of the ciphertext. One may observe the pattern of the frequency diagram for both plaintext and ciphertext is the same but just at different letter.

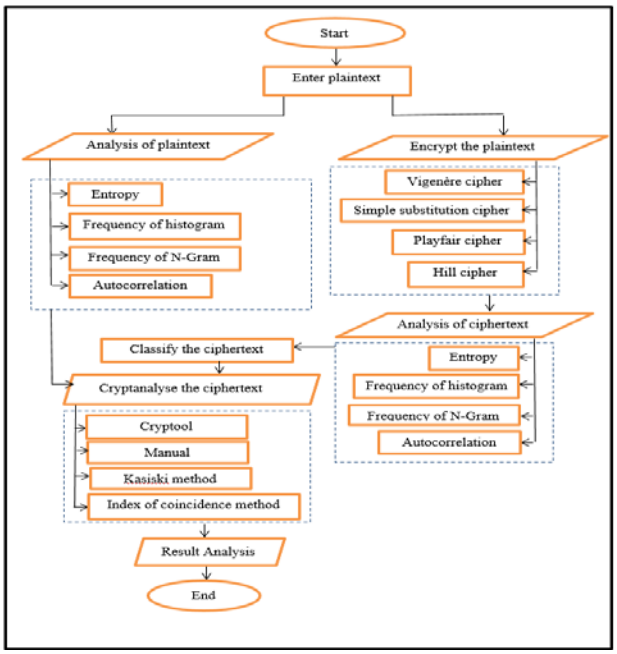

Fig. 1Research Methodology

3.1 Encrypt the plaintext

Two set of plaintexts are chosen. One set of plaintext has 3000 letters (Fig. 2). Another set of plaintext has 300 letters (Fig. 3). These plaintexts are encrypted using Playfair cipher, Hill cipher, Vigenère cipher and Random Substitution cipher.

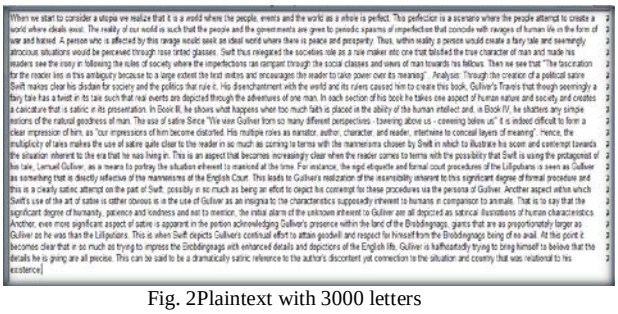

Fig. 2Plaintext with 3000 letters

Fig. 3Plaintext with 300 letters

3.2 Analysis the plaintext

Four analyses are carried out to evaluate the plaintext: frequency of English letter, frequency of N-Gram, autocorrelation and entropy of the ciphertext. These outputs are used to cryptanalyse the ciphertext.

The frequency analysis of the English letter on the Random substitution cipher's ciphertext have some similarity with the frequency analysis of the English letter. By comparing both frequency analysis we can have a rough idea on which letter is substituted. Beside that both the frequency of N-Gram for the ciphertext and the N-Gram for ordinary English alphabets will show the similarity for the Random Substitution cipher. This testing will make us to gain the information of plaintext easily.

The frequency analysis of N-Gram is to guess the word of Hill Cipher and Playfair Cipher. This is because it have the same N-Gram distribution with the ordinary English alphabets N-Gram distribution. Whereas, the auto-correlation is used to find out the distribution of repeated word in the ciphertext. This help us to find out the ciphertext is encrypted using Vigenère cipher. The entropy is find out to make the further comparison for this research. There are special characteristics for each cipher. This characteristic will help us to find out the Cipher.

3.3 Analysis the Ciphertext

Four analyses are carried out to evaluate the ciphertext: frequency of English letter, frequency of N-Gram, autocorrelation and entropy of the ciphertext. N-Gram and frequency of English character are used to guess the ciphertext was encrypted using which historical cipher. The entropy of ciphertext is find to see how much information do the ciphertext have. The autocorrelation is used to find out the key length used for Vigenère cipher.

3.4 Cryptanalyse the Ciphertext

There are four methods are used to identified the keys that are used to encrypt the plaintext: cryptool, manual calculation, kasiski method and index of coincidence method.

3.5 Result Analysis

The cryptanalysed results are analysed and are examine the security levels of these four historical ciphers. The measurements are based on the difficulty to crack the ciphers and to find the correct secret key.

4. Result and Analysis

4.1 Result analysis for plaintext with 3000 characters

Plaintext with 3000 characters is analysed using cryptool. The plaintext has 4.08 entropy with maximum possible entropy 4.70. The plaintext contains 25 different characters compared to the 26 characters of the selected alphabets. The histogram for this plaintext is follow the frequency of English alphabet where letter "E" has around 12% of frequency and "T" has second highest frequency with 11%. The autocorrelation for this plaintext is a scattered graph.

4.2 Cryptanalysis ciphertext from corresponding the plaintext with 3000 characters

Ciphertext 1 is analysed using cryptool. The entropy is 4.62 with maximum possible entropy 4.70. The ciphertext contains 26 different characters of the selected alphabets. The histogram for this ciphertext is not follow the frequency of English alphabet where letter "W" has the highest frequency with around 7%. The lowest frequency is alphabet "U" which is 1%. The autocorrelation for this ciphertext is evenly distributed and the peak of the graph are in a fixed period. The separation between two peaks is at multiplication of 6.

Ciphertext 2 is analysed using cryptool. The entropy is 4.51 with maximum possible entropy 4.70. The ciphertext contains 26 different characters of the selected alphabets. The histogram for this ciphertext, the highest frequency alphabet is "T" and "E", both frequencies are 8%. The frequency diagram for ciphertext 2 are evenly distributed. The highest frequency diagram for ciphertext 2 is "TX" (2%) and followed by "HD" (1.6%). Ciphertext 2 consists of 1283 trigram combinations. "TXE" has the highest frequency with 1.6%. The autocorrelation of ciphertext 2 is not evenly distributed.

Ciphertext 3 is analysed using cryptool. The entropy is 4.40 with maximum possible entropy 4.70. The ciphertext contains 25 different characters compared to the 26 characters of the selected alphabets. The histogram for this ciphertext, the highest frequency alphabet is "D" with 10.5% and followed by alphabet "K" with frequency 9%. The frequency diagram for ciphertext 3 are evenly distributed. The highest frequency diagram for ciphertext 3 is "QK" (4%) and followed by "CX" (2.2%). The autocorrelation of ciphertext 3 do not have any pattern, it looks scramble around.

Ciphertext 4 is analysed using cryptool. The entropy is 4.08 with maximum possible entropy 4.70. The ciphertext contains 26 different characters of the selected alphabets. The histogram for this ciphertext, the highest frequency alphabet is "Y" and "M", the frequencies are 12% and 10.6% respectively. The frequency diagram for ciphertext 4 are evenly distributed. The highest frequency diagram for ciphertext 4 is "ME" (4%) and followed by "EY" (3%). Ciphertext 4 consists of 693 trigram combinations. "MEY" has the highest frequency with 3%. The autocorrelation of ciphertext 4 is scattered.

Table 1 summarized the finding ciphertext 1 – 4. Based on the result, ciphertext 1 is encrypted using Vigenère cipher as the peak of autocorrelation for ciphertext 1 is fixed. Ciphertext 2 is encrypted using Hill cipher as the entropy is higher compare with the plaintext. The ciphertext 2 also has the output of bigram and trigram. Ciphertext 3 is encrypted using Playfair cipher as it has no trigram frequency. Lastly, the ciphertext 4 is encrypted using Random Substitution cipher as the entropy is similar with the plaintext. The frequency of single character is similar with the plaintext, but at different character.

Table 1 Analysis of ciphertexts with 3000 characters

Analysis	**Ciphertext 1**	**Ciphertext 2**	**Ciphertext 3**	**Ciphertext 4**
Entropy	4.62	4.51	4.40	4.08
Max. Entropy	4.70	4.70	4.70	4.70
Frequency of English letter	Do not follows the frequency of English alphabets.	English letter frequency evenly distributed.	Do not follows the frequency of English alphabets.	Follows English letter alphabets frequency.
Bigram	Evenly distributed	Evenly distributed except for "TX".	Have most bigrams. QK have highest occurrence bigram.	Have patterns of English letter bigrams frequency
Trigram	Evenly distributed	Evenly distributed but "TXE" and "AND" have highest value of frequency	Do not have trigram.	Follows the trigram frequencies of English language
Auto-correlation	The peak is in fixed period	There are no fixed correlation and the peak is uncertain.	There are no fixed correlation and the peak is uncertain.	Auto-correlation is scattered and do not have any pattern.
Automation Cryptanalysed (Cryptool)	Yes	No	No	Yes
Cipher	Vigenère Cipher	Hill cipher	Playfair cipher	Random Substitution cipher

Once the ciphertexts have been identified. Then, Known Plaintext Frequency Statistical Attack is performed to find the decryption key.

The ciphertext 1 is encrypted using Vigenère cipher. To retrieve the key, I_c is calculated as shown in table 2. The possible key length for Vigenère cipher is 12 as the I_c is the highest value.

Table 2 Index of coincidence vegenère cipher

Length = 1 $I_C \approx 0.0425 \pm 0.033$	Length = 6 $I_C \approx 0.0699 \pm 0.005$	Length = 11 $I_C \approx 0.0425 \pm 0.033$
Length = 2 $I_C \approx 0.0489 \pm 0.026$	Length = 7 $I_C \approx 0.0422 \pm 0.033$	Length = 12 $I_C \approx 0.0703 \pm 0.005$
Length = 3 $I_C \approx 0.0532 \pm 0.022$	Length = 8 $I_C \approx 0.0493 \pm 0.026$	Length = 13 $I_C \approx 0.0426 \pm 0.032$
Length = 4 $I_C \approx 0.0491 \pm 0.026$	Length = 9 $I_C \approx 0.0532 \pm 0.022$	Length = 14 $I_C \approx 0.0485 \pm 0.027$
Length = 5 $I_C \approx 0.0425 \pm 0.032$	Length = 10 $I_C \approx 0.0489 \pm 0.026$	Length = 15 $I_C \approx 0.0529 \pm 0.022$

Once the key length is identified. The ciphertext is divided into 12 block size. Then, one may refer to the bigram and trigram to find the plaintext. For example, the bigram "EV" is replaced with "TH" or trigram

"EVA" which has high frequency, then replace "EVA" with "THE". Alphabet "A" is replaced with "E", meaning the decryption key is "W". This key is used to decrypt the first word so one may guess the first alphabet. The unknown key to be replaced as "A". Next, the first letter of the plaintext is guessed to be "WHEN" because at the starting of the sentence only possible word use that end with HEN is "WHEN". Hence,"B" is replaced to "W" and the encryption key is"F". Next, the key "FLOWAAAAAAAA" a word "THAX" is found in the paragraph. The "X" is the 5th place in the key. "X" replaced as "T", the decryption key is "E". This key is applied to the system. The 6th letter is suspected to be "E" because two alphabets in 5th position and 6th position is suspected to be "WE". "HDI" in 9th, 10th and 11st positionsare suspected to be "THE" due to it is most repeated and placed different position then "EVA". Therefore,decryption key for these positions are "OWE".The decryption key after replacement is "FLOWERAAOWEA". The rest of the decryption key is to be found by guessing the alphabets. "IY","TS" are placed from position 6th to position 9th. The ciphertext "YT" is suspected to be "T" and "I" because the word should be "IT" and"IS". Therefore, the decryption key for 7th and 8thposition should be "FL". The incomplete decryption key for this cipher will be FLOWERFLOWEA. The remaining decryption key is to be find by using the guessing of the alphabet. The last pair of ciphertext and plaintext should be "K" to "T", then the final decryption key is found as "FLOWERFLOWER".

The ciphertext 2 is encrypted using Hill cipher. Based on the most occurring bigram alphabets for ciphertext 2 are "TH", "HE", "IN" and "ER". With this information the frequency of the bigram is compared to find the decryption key of this cipher. the following bigram most repeated are "NT" and "RE","RG" and "XE" are the in top six most recurring bigram based on frequency of bigram of ciphertext.Therefore, $CP^{-1} = K$, trial and error method is used to find out the key, such that guess "TX" as the "TH", "HO" as "HE", "ET" as "ER" and "IR" as "IN". Lastly, found the decryption key, $K_D = \begin{pmatrix} 1 & 0 \\ 20 & 1 \end{pmatrix}$.

The ciphertext 3 is encrypted using Playfair cipher. Based on the most occurring bigram alphabets for ciphertext 3 is "QK". Therefore, it is suspected the plaintext of "QK" is "TH". Distance between pairs "KH" and "QT" is measured to do the placement. "KH" has two alphabetsin between which is "I" or "J". QT has two alphabets in between which is "R" or "S". With a letter between Q and T is used as a letter in the decryption key.

	H	I/J	K	
	Q	R/S	T	

A letter R/S is placed inside the box to show that either one alphabet is placed in the decryption key. Next suspected pair is "KB" should be "HE". The distance between "KB" is too long. Contradict for "HE", the distance between this characters is too short. But, one may see that the distance between "KH" and "BE" is almost equal. This problem is solved by inserting "B" above "H" based on alphabet order. "E" at the same row with "B" and above "K".

	B	C/D	E	F
G	H	I/J	K	
	Q	R/S	T	

A pair of letter written between "E" and "B", one of this may be is the decryption key. "E" and "H"has two alphabets in between which are "F" and "G".The rest of the word is written inside. With this amount of information, we mayfind the decryption key using known plaintext frequency statistical attack. The decryption key that is used to decrypt certain alphabets to find out a clue on the letter. The remaining letter after "T" is counted which is six alphabets. The boxes are filled with the characters after "T".

	B	C/D	E	F
G	H	I/J	K	
	Q	R/S	T	U
V	W	X	Y	Z

The letter that is not presence in the box is checked. There are eight alphabets missing and two alphabetshave unknown placing. The next step is decrypt remaining ciphertext using this decryption key. First six letters had been tried using the founded decryption key. It only decrypt four alphabets which are "WHCDWE". As per guessing the word should be "WHEN WE". Plaintext "EN" should be decrypted to "CD". Hence,"C" is placed beside "E". "D"is placed above "E"."N" is placed beside "D".

		N	D	
	B	C	E	F
G	H	I/J	K	
	Q	R/S	T	U
V	W	X	Y	Z

"MD" is most occurred and it should be "ER" based on the most frequently used English bigram. The only placed available for "M" to take letter "E" is beside "B". Therefore, the key after replacing"M" is as shown below. Whereas "R" will be same row as "D" and same column of "M".

R		N	D	
M	B	C	E	F
G	H	I/J	K	
	Q	S	T	U
V	W	X	Y	Z

After decrypting the ciphertext, the first part of the plaintext consists of few ciphertexts which are"WH EN WE ST RO TX". It is known that the plaintext should be "when we start". Therefore,"RO" should be "AR". This shows that "A" and "O" shares the same row. "O" is replaced with "R" that means "O" is placed at the end of the row. Where "A" placed at the empty space. The rest of the alphabet is replaced using the unused alphabet for key in accending order. The key to this cipher will be as below.

R	A	N	D	O
M	B	C	E	F

G	H	I/J	K	L
P	Q	S	T	U
V	W	X	Y	Z

The ciphertext 4 is encrypted using Random Substitution cipher. Key recovery for Random Substitution cipher is simple by comparing and mapping the frequency histogram of plaintext and ciphertext. The decryption key is "POIUYTREWQASDFGHJKLMNBVCXZ".

4.3 Result analysis for plaintext with 300 characters

Plaintext with 300 characters is analysed using cryptool. The plaintext has 4.16 entropy with maximum possible entropy 4.70. The plaintext contains 23 different characters compared to the 26 characters of the selected alphabets. The histogram for this plaintext is follow the frequency of English alphabet where letter "E" has around 11% of frequency and "A" has second highest frequency with 10%. The autocorrelation for this plaintext is a scattered graph.

4.4 Cryptanalysis ciphertext from corresponding the plaintext with 300 characters

Ciphertext 5 is analysed using cryptool. The entropy is 4.60 with maximum possible entropy 4.70. The ciphertext contains 26 different characters of the selected alphabets. The highest histogram frequency for this ciphertext "S" has 7%. The lowest frequency is alphabet "R" which is 1%. The autocorrelation for this ciphertext is evenly distributed and the peak of the graph are in a fixed period. The separation between two peaks is at multiplication of 2.

Ciphertext 6 is analysed using cryptool. The entropy is 4.16 with maximum possible entropy 4.70. The ciphertext contains 23 different characters compared to the 26 characters of the selected alphabets. The histogram for this ciphertext, the highest frequency alphabet is "C" about 11%, followed by alphabet "M" with 10%. The frequency diagram for ciphertext 6 are evenly distributed. The highest frequency diagram for ciphertext 6 is "UA" (3%) and followed by "CO" (2.6%). Ciphertext 6 consists of 198 trigram combinations. "IIR", "IRK", "KOV", "LMI", "MII", "RKO" have the highest frequency with 2.9101%. The autocorrelation of ciphertext 6 is not evenly distributed.

Ciphertext 7 is analysed using cryptool. The entropy is 4.54 with maximum possible entropy 4.70. The ciphertext contains 26 different characters of the selected alphabets. The histogram for this ciphertext, the highest frequency alphabet is "N" with 7% and followed by alphabet "A" and "S" with each frequency of 6.9%. The frequency diagram for ciphertext 7 are evenly distributed. The highest frequency diagram for ciphertext 7 is "BU" (1.8%) and followed by "RQ" (1.6%). The autocorrelation of ciphertext 7 do not have any pattern, it looks scramble around.

Ciphertext 8 is analysed using cryptool. The entropy is 4.49 with maximum possible entropy 4.70. The ciphertext contains 25 different characters compared to the 26 characters of the selected alphabets. The histogram for this ciphertext, the highest frequency alphabet is "H" and "Y", the frequencies are 8.5% and 6.9% respectively. The frequency diagram for ciphertext 8 are evenly distributed. The highest frequency diagram for ciphertext 8 is "HN" and "VX", both frequencies are 3.8%. Ciphertext 8 do not have trigram combinations. The autocorrelation of ciphertext 8 is scattered.

Table 3 summarized the finding ciphertext 5 – 8. Based on the result, ciphertext 5 is encrypted using Vigenère cipher as the peak of autocorrelation for ciphertext 5 is fixed. Ciphertext 6 is encrypted using Random Substitution cipher as the entropy is similar with the plaintext. The frequency of single character is similar with the plaintext, but at different character. Ciphertext 7 is encrypted using Hill cipher as the entropy is higher compare with the plaintext. The ciphertext 7 has the output of bigram and trigram. Lastly, the ciphertext 8 is encrypted using Playfair cipher as it has no trigram frequency.

Noted that, the process of finding the decryption key for these ciphertexts is similar with the process described in Section *B: Cryptanalysis ciphertext from corresponding the plaintext with 3000 characters*. Therefore, these processes are discarded in here. After following the steps above, one may know the encryption key or decryption key for these four ciphers. The key for Vigenère cipher is "HOUSE". The key for Random Substitution cipher is "MNBVCXZASDFGHJKLPOIUYTREWQ". The key for Hill cipher is $\begin{pmatrix} 1 & 13 \\ 3 & 14 \end{pmatrix}$. Lastly, the key for Playfair cipher as below:

Y	O	U	T	H
A	B	C	D	E
F	G	I/J	K	L
M	N	P	Q	R
S	T	W	X	Z

Table 3 Analysis of ciphertexts with 300 characters

Analysis	**Ciphertext 5**	**Ciphertext 6**	**Ciphertext 7**	**Ciphertext 8**
Entropy	4.60	4.16	4.54	4.49
Frequency of English letter	Did not follow standard English letter frequency	Follows English letter alphabets frequency.	Do not have the pattern of English letter frequency. Evenly distributed.	Do not follows the frequency of English alphabets.
Bigram	Evenly distributed	Have patterns of English letter bigrams frequency	Evenly distributed except for "BU".	Have most bigrams. QK have highest occurrence bigram.
Trigram	Evenly distributed	Follows the trigram frequencies of English language	Evenly distributed but "ASS" and "SSY" have highest value of frequency	Do not have trigram.
Auto-correlation	The peak is in fixed period	Auto-correlation is scattered and do not have any pattern.	Autocorrelation is did not have fixed key length.	There are no fixed correlation and the peak is uncertain.
Automation	Yes	Yes	No	No

Cryptanalysed (Cryptool)				
Cipher	Vigenère Cipher	Simple Substitution cipher	Hill cipher	Playfair cipher

5. Discussion

Two data sets are used in this research which are the plaintext with 3000 characters and plaintext with 300 characters. The plaintext with 3000 characters and 300 characters are encrypted using Playfair cipher, Hill cipher, Vigenère cipher and Random Substitution cipher. The outputs are eight ciphertexts. Four ciphertexts have approximately 3000 characters each and another four ciphertexts have approximately 300 characters each.

Two plaintexts and eight ciphertexts are analysed. Results shown that all ciphers are breakable. Since, all are breakable, to evaluate which cipher is more secure compare with others. Hence, the number of steps to breaks these ciphers from two data sets are used to determine the security levels. Meaning that, if the cipher needs more step to break it, then it is more secure or wise versa. The summary is shown in Table 4.

Based on the finding, the plaintext and ciphertexts with 3000 characters give more information for cryptanalysing compared with the plaintext and ciphertexts with 300 characters, such as for Hill Cipher, Vigenère cipher and Playfair cipher. The cryptanalyses give no different for Random Substitution cipher for these two datasets. This is because, one may find the frequency of English alphabets for plaintext and ciphertext are same but only at different character. Hence, Random Substitution cipher has the least secure compare with others.

The next cipher is meant to be Vigenère cipher. By using the autocorrelation, the key length of Vigenère cipher is identified. Once the key length is identified, bigram of ciphertext is mapping with the common bigram of plaintext. With this, the decryption key is identified. However, if the plaintext does not provide more information for cryptanalysis, the number of trial and error to guess the decryption key is increase. At the end, the decryption key still be able to recover.

Next, the Playfair cipher is more reliable compared to the Vigenère cipher. The letters of Playfair cipher are arranged in pairs. When a pair of letter in ciphertext is known, most of the character placing for the decryption key will be revealed. This is because the decryption key will follow alphabetic order after the keyword. However, when the ciphertext character decrease the guessing of decryption key need more time in trial and error.

Lastly Hill Cipher be the most reliable among all these ciphers. This is because when the number of character of ciphertext decrease the possibility of the frequency of bigram to be guessedis decreased. Then ciphertext frequency is being distorted and make us to guess more pair of plaintext to identify the decryption key.

Table 4 Comparison of cryptanalysing ciphertexts from two types of plaintexts

Historical cipher	**Plaintext with 3000 characters**	**Plaintext with 300 characters**
Hill Cipher	The bigram provides useful information in guessing the plaintext character. For example, "TX" has been repeated many times. This means by guessing of "TX" to be "TH" may help in cryptanalysing the ciphertext easily.	The bigram did not provide enough information in guessing the plaintext character. This is because got bigram "BU", "RQ", "SS", and "AS" repeated evenly. This had delayed the process of cryptanalysing the ciphertext. Hence, increases the number of trial and error in finding the key.
Random Substitution cipher	The entropy of plaintext is similar with the entropy of ciphertext. The frequency histogram of the characters is same for both plaintext and ciphertext, the different only the alphabet.	
Vigenère cipher	The autocorrelation has the same peak which is used to identify the key length.	
	The repetition of bigram of ciphertext aid in guessing the plaintext, hence, guessing the key	The bigram of ciphertext do not provide any useful information finding the plaintext as the frequencies are similar. Hence, increase the number of trial and error.
Playfair cipher	Both has no frequency of trigram.	
	The frequency of bigram is higher; this provides more information to find the encryption key.	The frequency of bigram is low; it is hard to find the encryption. To find the encryption key, the ciphertext is arranged in 5 by 5 matrix. From that, one may guess the few keywords to identify the key.

6. Conclusions

The comparison between the four ciphers, Random Substitution cipher, Vigenère Cipher, Playfair cipher and Hill cipher are completed using known plaintext frequency statistical attack. The Random Substitution cipher and Vigenère cipher are the weaker ciphers as the steps of cryptanalysed these two ciphers are simplercompare to the Playfair cipher and Hill cipher. Noted that, both ciphers can auto-cryptanalysing by using the Cryptool, subjected the ciphertext length is long.

The security of Playfair cipher is stronger thanRandom Substitution cipher and Vigenère cipher. One main reason is, the Playfair cannot be cryptanalysing using tool. Playfair cipher only can break by manual trial and error. The steps are simpler compare with Hill cipher. Playfair cipher can recover the decryption key by using the ciphertext alone. Therefore, Hill cipher has better security compare with Playfair cipher. Hill cipher is encrypted using modulus, matrix and inverse matrix. The calculation is slightly complex compare with Playfair cipher. Besides that, the cryptanalysing process is slower if the given ciphertext is shorter as the number of trial and error increase. Hence, the Hill cipher is more secure compare with Playfair cipher.

7. Acknowledgment

The authors express appreciation to the University Tun Hussein Onn Malaysia (UTHM). This research is supported by Faculty Computer Science and Information Technology (FSKTM), Research Management Centre (RMC) and Gates IT Solution Sdn. Bhd. under its publication scheme.

References

[1] B. Schneier, *Applied Cryptography Protocols, Algorithms, and Source Code in C*, John Wiley & Sons, 2015.

[2] W. Stallings, *Cryptography and Network Security: Principles and Practice, Global Edition 7th Edision*, Pearson Education Limited, 2016.

[3] T. J. Chua, C. W. Chuah, N. H. Abd Rahman and I. R. A Hamid, *Audio Steganography with Embedded Text*, IOP Conference Series: Materials Science and Engineering, vol. 226, no. 1, pp 1-10, 2017.

[4] A. Chai, C.W. Chuah, *SHEARS INC. Salon Management System*, International Journal on Information Visualization, vol. 1, no. 4-2, pp. 246 – 249, 2017.

[5] K. Y. Lee, C. W. Chuah, *Secret Channel using Video Steganography*, International Journal on Information Visualization, vol. 1, no. 4-2, pp. 240-245, 2017.

[6] L. S. Hill, *Cryptography in an Algebraic Alphabet*, The American Mathematical Monthly, vol. 36, pp. 306 – 312, June-July 1929.

[7] I. IA, A. Momammed and D. Hossam, *How to Repair the Hill Cipher*, Journal of Zhejiang University-Science A, Springer, vol. 7, no. 12, pp. 2022 – 2030, 2006.

[8] S. Singh, *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*, Anchor, 2000.

[9] D. R. Stinson, *Cryptography Theory and Practices 3rd Edition*, Chapman and Hill/CRC, 2006.

[10] D. E. Denning, *Cryptography and Data Security*, Addisson-Wesley Publishing Company, Reading, 1982.

[11] D. Shallcross, *The Playfair Cipher*, Vinculum, vol 45, no. 2, pp. 4-6, 2008.

[12] A. Kahate, *Cryptography and Network Security, 2nd Edition*, Tata McGraw-Hill Publishing Company Limited, New Delhi, 2008.

[13] W. F. Friedman, *The Index of Coincidence and its Applications in Cryptography*, Riverbank Laboratories, 1922.

[14] T. Helleseth, and P. V. Kumar, *Sequences with Low Correlation,* Handbook of Coding Theory, eds. V. S. Press and W. C. Huffman, Elsevier, Amsterdam, 1998.

[15] V. Shoup, *A Computational Introduction to Number Theory and Algebra,* Cambridge University Press, 2008.