CAB340 – ASSESSMENT 1

Historical Cryptanalysis, Probability, Information Theory

Callum McNeilage – n10482652 Folder: 152-Student

1 Cryptanalysis of Historical Ciphers

a)

Ciphertext 0

10 Most Common Single-character Frequencies
Histogram Analysis of <0.txt>. File size 1008 bytes.

Descending sorted on frequency.

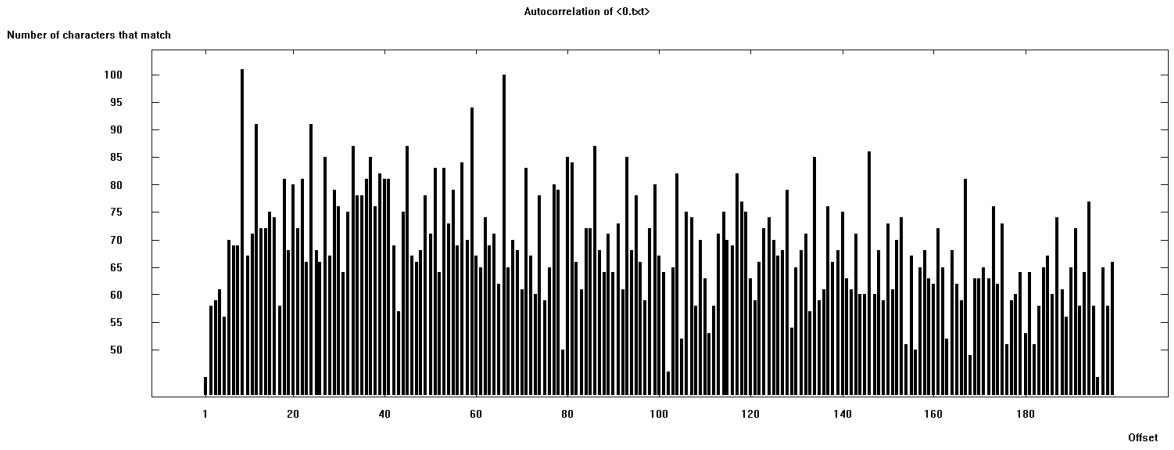
No.	Substring	Frequency (in %)	Frequency
1	Δ	18.0556	182
2	t	9.8214	99
3	e	8.4325	85
4	i	7.3413	74
5	o	6.2500	63
6	a	5.8532	59
7	s	5.5556	56
8	h	5.3571	54
9	n	5.1587	52
10	r	4.5635	46

10 Most Common Digram Frequencies
Digram Analysis of <0.txt>. File size 1008 bytes.

Descending sorted on frequency.

No.	Substring	Frequency (in %)	Frequency
1	Δt	2.5819	26
2	aΔ	1.9861	20
3	tΔ	1.9861	20
4	iΔ	1.6882	17
5	Δe	1.5889	16
6	oΔ	1.3903	14
7	Δa	1.2910	13
8	Δs	1.2910	13
9	ΔΔ	1.1917	12
10	iΔ	1.1917	12

Autocorrelation (As a function of the shift up to 10)



Ciphertext 1

10 Most Common Single-character Frequencies
Histogram Analysis of <1.txt>. File size 1001 bytes.

Descending sorted on frequency.

No.	Substring	Frequency (in %)	Frequency
1	Δ	17.0234	167
2	u	11.3150	111
3	r	8.2569	81
4	s	6.6259	65
5	m	6.5240	64
6	b	5.5046	54
7	q	5.3007	52
8	l	4.4852	44
9	a	4.2813	42
10	h	3.9755	39

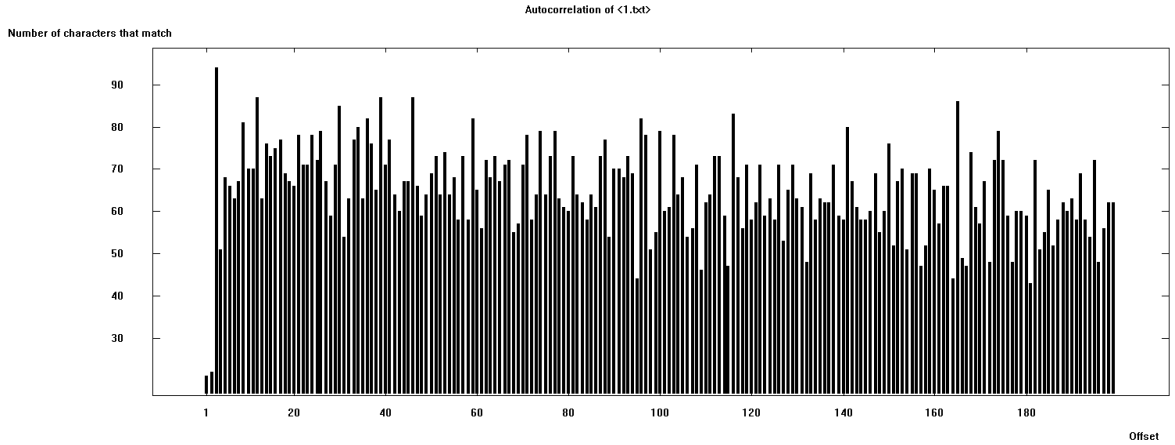
10 Most Common Digram Frequencies

Digram Analysis of <1.txt>. File size 1001 bytes.

Descending sorted on frequency.

No.	Substring	Frequency (in %)	Frequency
1	uΔ	3.9419	38
2	Δr	3.0083	29
3	ra	2.3859	23
4	rΔ	2.1784	21
5	Δs	2.0747	20
6	qΔ	1.9710	19
7	Δc	1.7635	17
8	Δw	1.6598	16
9	uh	1.6598	16
10	Δm	1.5560	15

Autocorrelation (as a function of the shift up to 10)



Ciphertext 2

10 Most Common Single-character Frequencies
Histogram Analysis of <2.txt>. File size 1001 bytes.

Descending sorted on frequency.

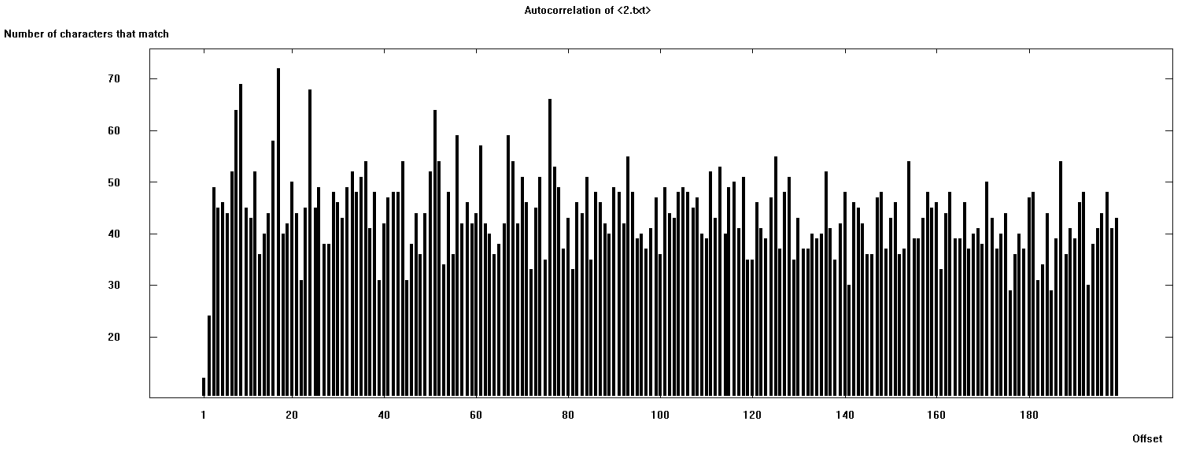
No.	Substring	Frequency (in %)	Frequency
1	Δ	17.0431	166
2	w	4.9281	48
3	D	4.6201	45
4	C	4.1068	40
5	H	4.0041	39
6	s	4.0041	39
7	o	3.3881	33
8	A	3.2854	32
9	r	3.0801	30
10	z	3.0801	30

10 Most Common Digram Frequencies
Digram Analysis of <2.txt>. File size 1001 bytes.

Descending sorted on frequency.

No.	Substring	Frequency (in %)	Frequency
1	sΔ	1.5690	15
2	ΔH	1.4644	14
3	tΔ	1.3598	13
4	lΔ	1.2552	12
5	CΔ	1.1506	11
6	ΔB	0.9414	9
7	eΔ	0.9414	9
8	DΔ	0.8368	8
9	FΔ	0.8368	8
10	wΔ	0.8368	8

Autocorrelation (As a function of the shift up to 10)



Ciphertext 3

10 Most Common Single-character Frequencies
Histogram Analysis of <3.txt>. File size 1001 bytes.

Descending sorted on frequency.

No.	Substring	Frequency (in %)	Frequency
1	Δ	17.0431	166
2	k	4.6201	45
3	z	4.4148	43
4	h	3.9014	38
5	u	3.9014	38
6	b	3.7988	37
7	s	3.7988	37
8	y	3.6961	36
9	e	3.5934	35
10	a	3.4908	34

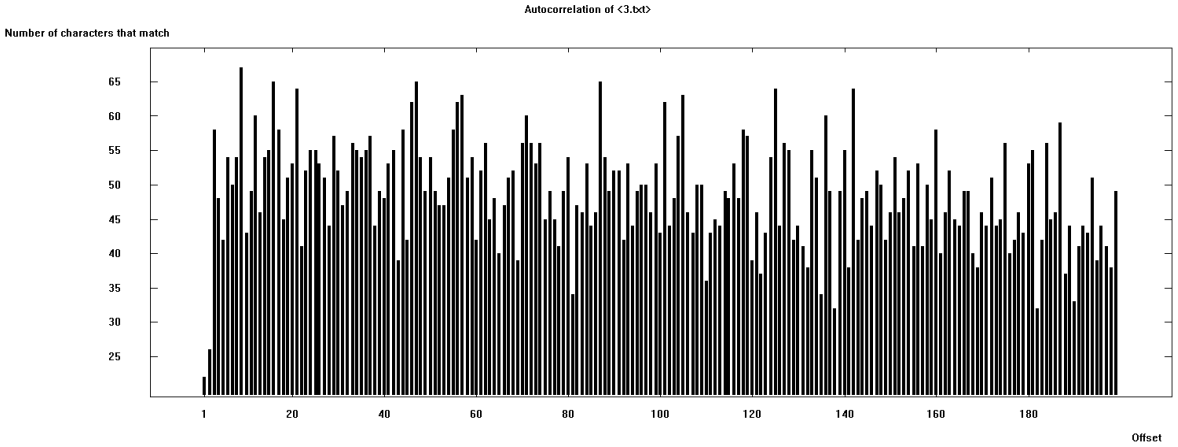
10 Most Common Digram Frequencies

Digram Analysis of <3.txt>. File size 1001 bytes.

Descending sorted on frequency.

No.	Substring	Frequency (in %)	Frequency
1	kΔ	1.7782	17
2	Δq	1.4644	14
3	zΔ	1.2552	12
4	Δe	1.1506	11
5	hy	1.1506	11
6	Δa	1.0460	10
7	Δh	1.0460	10
8	Δs	1.0460	10
9	Δv	1.0460	10
10	bf	1.0460	10

Autocorrelation (As a function of the shift up to 10)



b)

Random Simple Substitution Cipher

By definition, the Random Simple Substitution Cipher assigns a fixed chosen character for each character of the alphabet at random and without collisions. However, this means that the most common letters from English language will correspond to the most common letters in the ciphertext. So, a ciphertext where the frequencies of the 10 most common single-characters are similar to the frequencies of the 10 most common single-characters in English is most likely using the Random Simple Substitution Cipher.

As such, it is most likely that ciphertext 1 is a Random Simple Substitution Cipher. This was determined by comparing the measured values in the 10 Most Common Single-Character Frequencies table with the values in Figure 1 and discounting the space (Δ) character.

English Letter Frequency (based on a sample of 40,000 words)

Letter	Count	Letter	Frequency
E	21912	E	12.02
T	16587	T	9.10
A	14810	A	8.12
O	14003	O	7.68
I	13318	I	7.31
N	12666	N	6.95
S	11450	S	6.28
R	10977	R	6.02
H	10795	H	5.92
D	7874	D	4.32
L	7253	L	3.98
U	5246	U	2.88
C	4943	C	2.71
M	4761	M	2.61
F	4200	F	2.30
Y	3853	Y	2.11
W	3819	W	2.09
G	3693	G	2.03
P	3316	P	1.82
B	2715	B	1.49
V	2019	V	1.11
K	1257	K	0.69
X	315	X	0.17
Q	205	Q	0.11
J	188	J	0.10
Z	128	Z	0.07

Figure 1: http://pi.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html

Vigenère Cipher

The autocorrelation of a given ciphertext is used to decrypt a Vigenère cipher. This is able to be achieved because we expect to see peaks in the autocorrelation data at multiples of the period *d*, which can be observed by plotting autocorrelation.

As such, it is most likely that ciphertext 2 is a Vigenère Cipher. This was determined by observing the Autocorrelation plot and determining that a peak in the data occurs at a multiple of 3 characters, hence a 3-character period was used.

Transposition Cipher

A transposition cipher encrypts plaintext by re-ordering a block of *d* characters by a permutation *f*. As such, it will is expected that the frequency of characters in the ciphertext will be similar to the frequency of characters in the plaintext as the characters are not being replaced with other characters but rather, being re-ordered.

As such, by comparing the '10 Most Common Single-character' tables against Figure 1 as in Random Substitution Cipher, it can be observed that Ciphertext 0 may be using a Transposition Cipher as all characters that appear in the 10 Most common single-characters of ciphertext are among the top letter frequencies in Figure 1.

2 x 2 Hill Cipher

The 2x2 Hill Cipher encrypts plaintext by using a simple substitution cipher on digrams (every first and second character) in the plaintext. This means that frequencies of common digrams in English should appear in the Most Common Digrams tables.

Bigram Frequencies

A.k.a digraphs. We can't list all of the bigram frequencies here, the top 30 are the following (in percent %):

TH : 2.71	EN : 1.13	NG : 0.89
HE : 2.33	AT : 1.12	AL : 0.88
IN : 2.03	ED : 1.08	IT : 0.88
ER : 1.78	ND : 1.07	AS : 0.87
AN : 1.61	TO : 1.07	IS : 0.86
RE : 1.41	OR : 1.06	HA : 0.83
ES : 1.32	EA : 1.00	ET : 0.76
ON : 1.32	TI : 0.99	SE : 0.73
ST : 1.25	AR : 0.98	OU : 0.72
NT : 1.17	TE : 0.98	OF : 0.71

Figure 2: http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/english-letter-frequencies/

As such, by comparing the frequencies of the '10 Most Common Digram' tables against the diagram frequencies outlined in Figure 2, it was determined that Ciphertext 3 was the closest match.

c)

Random Simple Substitution Cipher

For a random simple substitution cipher, it is possible to assign the most commonly occurring characters in English alphabet with the most commonly occurring characters in ciphertext 1's alphabet to effectively perform cryptanalysis.

$$E \rightarrow U,$$

$$T \rightarrow R$$

Vigenère Cipher

By analyzing the Autocorrelation graph of ciphertext 2, it can be identified that the cipher has a period length of 7 characters as each peak is at a multiple of 7 (i.e. 7, 14, 21, etc.). Then, it is possible to attack each one of the three substitution tables of *d* separately to effectively perform cryptanalysis.

Transposition Cipher

In order to effectively decrypt the transposition cipher in ciphertext 0, simply split the ciphertext into blocks of increasing period and use knowledge of anagramming to search for words.

Δi/hO/ru/st/nΔ/ue/ra/Δi/Δa/Δs/sh/um/…

Δih/Our/stn/Δue/raΔ/iΔa/Δss/hum/…

ΔihO/rust/nΔue/raΔi/ΔaΔs/shum/…

Etc…

2x2 Hill Cipher

In order to perform cryptanalysis on Hill Cipher, it is necessary to first recover the key.

1. Let $C = [C_0 C_1 \dots C_{d-1}]$ and $P = [P_0 P_1 \dots P_{d-1}]$
2. Solve $C = KP$ for key matrix K
3. Verify $P = K^{-1}C$ and decrypt any block C_j as $P_j = K^{-1}C_j$

Then decrypt using matrix multiplication

d)

Ciphertext 0 – Transposition Cipher

his Our nature is as much a fact of the existing world as anything and there can be no certainty that it will remain constant It might happen if Kant is right that tomorrow our nature would so change as to make two and two become five This possibility seems never to have occurred to him yet it is one which utterly destroys the certainty and universality which he is anxious to vindicate for arithmetical propositions It is true that this possibility formally is inconsistent with the Kantian view that time itself is a form imposed by the subject upon phenomena so that our real Self is not in time and has no tomorrow But he will still have to suppose that the timeorder of phenomena is determined by characteristics of what is behind phenomena and this suffices for the substance of our argument Reflection moreover seems to make it clear that if there is any truth in our arithmetical beliefs they must apply to things equally whether we think of them or not Two physical objects and two other WJlShGJ

Ciphertext 2 – Vigenère Cipher

MAYBE THAT THE WHOLE OUTER WORLD IS NOTHING BUT A DREAM AND THAT WE ALONE EXIST THIS IS AN UNCOMFORTABLE POSSIBILITY BUTAL THOUGH IT CAN NOT BE STRICTLY PROVED TO BE FALSE THERE IS NOT THE SLIGHTEST REASON TO SUPPOSE THAT IT IS TRUE IN THIS CHAPTER WE HAVE TO SEE WHY THIS IS THE CASE BEFORE WE EMBARK UPON DOUBTFUL MATTERS LET US TRY TO FIND SOME MORE OR LESS FIXED POINT FROM WHICH TO START ALTHOUGH WE ARE DOUBTING THE PHYSICAL EXISTENCE OF THE TABLE WE ARE NOT DOUBTING THE EXISTENCE OF THE SENSED AT A WHICH MADE US THINK THERE WAS A TABLE WE ARE NOT DOUBTING THAT WHILE WE LOOK A CERTAIN COLOUR AND SHAPE APPEAR TO US AND WHILE WE PRESS A CERTAIN SENSATION OF HARDNESS IS EXPERIENCED BY US ALL THIS WHICH IS PSYCHOLOGICAL WE ARE NOT CALLING IN QUESTION IN FACT WHATEVER ELSE MAY BE DOUBTFUL SOME AT LEAST OF OUR IMMEDIATE EXPERIENCES SEEM ABSOLUTELY CERTAIN DESCARTES THE FOUNDER OF MODERN PHILOSOPHY INVENTED A METHOD WHICH MAY STILL BE USED WITH PROFIT THE METHOD OF SYSTEMATIC DOUBT

e)

Plaintext		Ciphertext
$P_0 = (TI) = \begin{pmatrix}19\\8\end{pmatrix}$	$\rightarrow$	$C_0 = (HY) = \begin{pmatrix}7\\24\end{pmatrix}$
$P_1 = (ON) = \begin{pmatrix}14\\13\end{pmatrix}$	$\rightarrow$	$C_1 = (BF) = \begin{pmatrix}2\\5\end{pmatrix}$
$P = [P_0 P_1] = \begin{pmatrix}19 & 14\\8 & 13\end{pmatrix}$	$\rightarrow$	$C = [C_0 C_1] = \begin{pmatrix}7 & 2\\24 & 5\end{pmatrix}$

$$C = KP$$

$$\begin{pmatrix}7 & 2\\24 & 5\end{pmatrix} = \begin{pmatrix}a & b\\c & d\end{pmatrix}\begin{pmatrix}19 & 14\\8 & 13\end{pmatrix}$$

$$\begin{pmatrix}a & b\\c & d\end{pmatrix} = \begin{pmatrix}7 & 2\\24 & 5\end{pmatrix}\begin{pmatrix}19 & 14\\8 & 13\end{pmatrix}^{-1} = \begin{pmatrix}\frac{5}{9} & \frac{-4}{9}\\ \frac{272}{135} & \frac{-241}{135}\end{pmatrix}$$

Plaintext		Ciphertext
$P_0 = (TH) = \begin{pmatrix}19\\7\end{pmatrix}$	$\rightarrow$	$C_0 = (RA) = \begin{pmatrix}17\\1\end{pmatrix}$
$P_1 = (HE) = \begin{pmatrix}7\\4\end{pmatrix}$	$\rightarrow$	$C_1 = (AU) = \begin{pmatrix}1\\20\end{pmatrix}$
$P = [P_0 P_1] = \begin{pmatrix}19 & 7\\7 & 4\end{pmatrix}$	$\rightarrow$	$C = [C_0 C_1] = \begin{pmatrix}17 & 1\\1 & 20\end{pmatrix}$

$$C = KP$$

$$\begin{pmatrix}17 & 1\\1 & 20\end{pmatrix} = \begin{pmatrix}a & b\\c & d\end{pmatrix}\begin{pmatrix}19 & 7\\7 & 4\end{pmatrix}$$

$$\begin{pmatrix}a & b\\c & d\end{pmatrix} = \begin{pmatrix}17 & 1\\1 & 20\end{pmatrix}\begin{pmatrix}19 & 7\\7 & 4\end{pmatrix}^{-1} = \begin{pmatrix}\frac{-29}{5} & \frac{293}{20}\\ \frac{12}{5} & \frac{-27}{10}\end{pmatrix}$$

2 Probability

a)

3x3 joint probability table for Pr(D,S)

P(D,S)	S = 1	S = 2	S = 3
$D = p_1$	$\frac{p}{3}$	$\frac{p}{3}$	$\frac{p}{3}$
$D = p_2$	$\frac{p}{3}$	$\frac{p}{3}$	$\frac{p}{3}$
$D = p_3$	$\frac{p}{3}$	$\frac{p}{3}$	$\frac{p}{3}$

Because there is no information given about which door the diamond is behind, $p_1 = p_2 = p_3$. Therefore, the probability of selecting the correct door is $\frac{1}{3}$. Because each door is equally likely, there is no strategy to picking the correct door.

b)

P(D,S)	$S = \frac{1}{3}$	$T = \frac{2}{3}$
$D = P_1$	$\frac{p}{3}$	$\frac{2p}{3}$
$D = P_2$	$\frac{p}{3}$	$\frac{2p}{3}$
$D = P_3$	$\frac{p}{3}$	$\frac{2p}{3}$

Because there is a combined total of $\frac{1}{3}$ probability that the diamond is behind each of the two doors you do not choose and you are shown that there is 0 probability that the diamond is behind one of them, there is now a $\frac{2}{3}$ probability that the diamond is behind the other door. Therefore, you should always accept the offer to switch as it has you are twice as likely to be correct from a switch as you are from staying on your original choice.

c)

Since the game host is adversarial, you should always do the opposite to what he says, as the host will always want you to pick the bad door. Therefore, if the host opens a bad door and says the prize is behind the door that you chose, you should switch to the door you did not choose and if the host opens a door and says the prize is not behind your door, you should stay on your original choice.

3 Information Theory

a)

$K = [a, b] where$: a = generator gets stuck, b = generator does not get stuck

$$H(K) = \frac{1}{4} log_2(1) + \frac{3}{4} log_2(2^{128})$$
$$H(K) = 96$$

b)

$let\ f(a) = 1, f(b) = 0$:

$$Pr\big(f(a)\big) = \frac{1}{4}, Pr\big(f(b)\big) = \frac{3}{4}$$
$$H(F|K) = \frac{1}{4} \log_2 \left(\frac{1}{\frac{1}{4}}\right) + \frac{3}{4} log_2 \left(\frac{1}{\frac{3}{4}}\right) - H(K)$$
$$H(F|K) = (0.81127) - 96$$
$$H(F|K) = -95.18872$$

Given that conditional entropy measures the left-over uncertainty, since the conditional entropy is negative, there is certainty that the key can be obtained.

c)

This key generator is not secure because the bug happens $\frac{1}{4}$ of the time. This means that roughly a quarter of all keys can be guessed easily by an attacker in a ciphertext-only attack.

d)

As displayed above, generating keys of sufficiently high entropy is **necessary** for the cryptographic security of a system based on them.

e)

A possibly better metric for capturing the security of a non-uniform key distribution, but ideally with the same additive properties as entropy when combining independent variable is the amount of information gain per event.