TJMM
14 (2022), No. 1, 43-51

THE VIGENÈRE CIPHER

ANA-MARIA DOBRIŢOIU

Abstract. This paper contains an overview to a well-known encryption method, Vigenère's cipher, which although easy to understand and implement, seems impossible for beginners to break, which is why it has been described as "le chiffre indéchiffrable".

1. Introduction

Cryptography is the study of mathematical techniques used to encrypt and decrypt data. The main purpose is to ensure secure communication, i.e. sensible information must remain private during the communication process.

Cryptanalysis is the science of breaking cryptosystems ([5], p.3).

From a functional perspective, cryptography and cryptanalysis mirror each other in function ([1], p.435).

1.1. Terminology

Encryption scheme or cryptosystem = a system that essentially transforms plaintext into ciphertext and conversely ([2]).

Plaintext (*cleartext*) = the message whose meaning must be hidden.

Encryption = the process of transforming the plaintext in such a way that its meaning is hidden (except for the sender and the intended recipient of the message).

Ciphertext = the converted data resulted.

Decryption = the process of converting the ciphertext into plaintext.

Cryptographic algorithm (*cipher*) = a mathematical function used in the encryption and decryption process. Generally, there are two related functions: one for encryption and one for decryption.

This algorithm works in combination with a key to encrypt the plaintext. Different keys lead to the plaintext being encrypted to different ciphertexts.

The security of the encrypted information depends on the strength of the cryptographic algorithm and the secrecy of the key.

1.2. Encryption methods

1.2.1. Singlet/Secret/Symmetric key cryptography

This encryption method involves using a shared secret key between the sender and the receiver.

The involved parties use functions dependent on the same predetermined key. Usually, the key is randomly generated.

Weak point: The strength of the symmetric key algorithm lies primarily in maintaining the secrecy of the key. Hence the need for a proper exchange of private keys. If

2010 *Mathematics Subject Classification.* 68W01, 68W32, 68P99.

Key words and phrases. Cryptography, Cryptanalysis, Encryption method, Vigenère cipher.

the number of participant to a transaction increases, the number of potential weakness points increases.

1.2.2. Asymmetric/Public-key cryptography

In public-key cryptography, a user possesses a secret key as in symmetric cryptography but also a public key.

Asymmetric algorithms can be used for applications such as digital signatures and key establishment, and also for classical data encryption.

Weak points: Public-keys cryptography is more computationally costly, due to the unique nature of the keys. Compared to the keys used in the symmetric key cryptography, public keys are also more vulnerable to bruce force and man-in-the-middle attacks.

2. The Vigenère cipher

2.1. Historical context

The Vigenère cipher is an adaptation of the **Trithemius cipher**, which was first introduced by Johannes Tritemius in his book, ”Polygraphiae libri sex, loannis Trithemii abbatis Peapolitani, quondam Spanheimensis, ad Maximilianum Caesarem” (”Six books of polygraphy”), which was printed and published in 1518 ([1], p.133).

The Tritemius cipher is a variation of the **Caesar cipher** (mono-alphabetic cipher), steganographic cipher in which each letter was represented as a word taken from a succession of columns. For this cipher, the author used a square matrix (or tableau), ”which is the elemental form of polyalphabetic substitution” ([1], p.133).

Unlike the Tritemius cipher, the Vigenère cipher used a passphrase as the key for a repeated polyalphabetic cipher (an encryption key).

The Vigenère cipher was wrongfully attributed to Blaise de Vigenère, who was the author of the **Autokey cipher**, which he introduced in his book, ”Traicte des Chiffres”, printed in 1586.

The Vigenère cipher was first presented by Giovan Battista Bellaso in 1553, in his book, ”La cifra del Sig. Giovan Battista Bellaso”.

2.2. Mechanism

The standard Vigenère cipher is a polyalphabetic cipher which uses latin alphabets and a short repeating keyword. Its tabula recta contains 26×26 characters (Table 1).

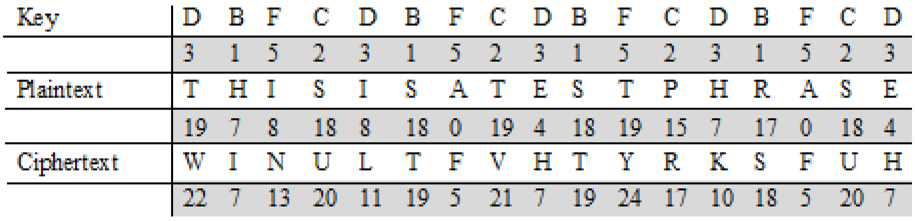
The Vigenère cipher that uses a key of length m is an example of a symmetric variable-length scheme over the alphabet of letters.

In order to apply the Vigenère cipher, one can use:

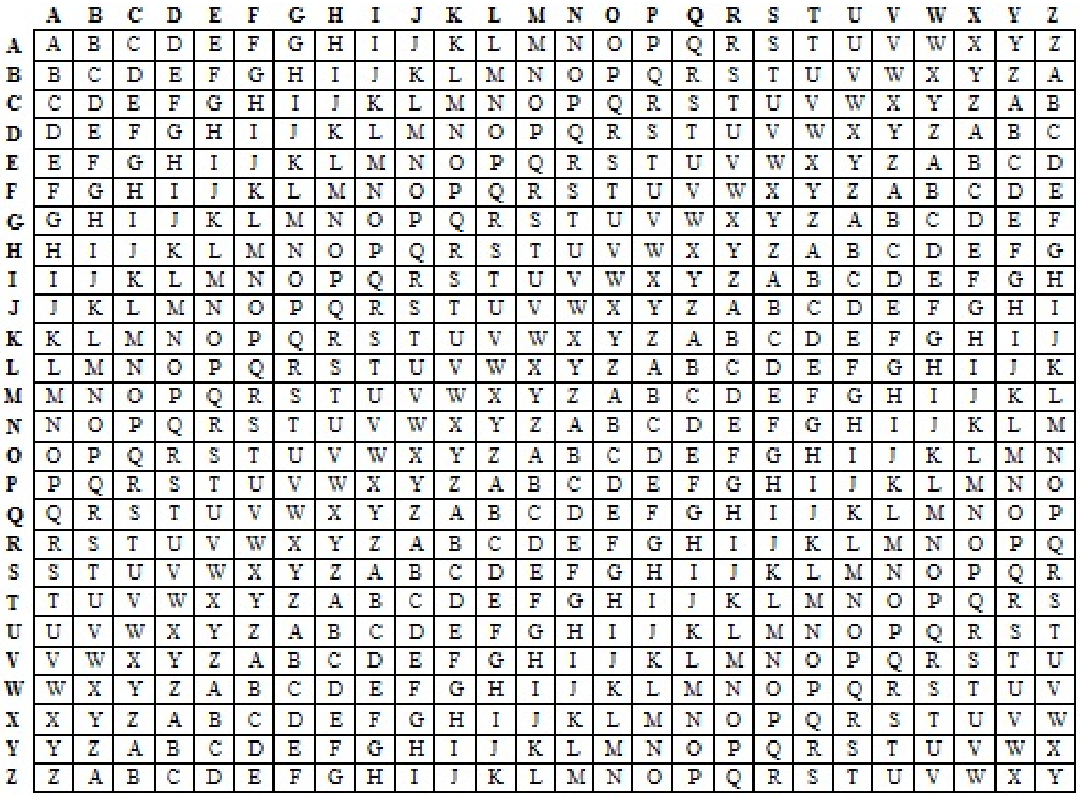
a. a group of characters as a key.

Steps:

Let’s consider the key $k = (n_1, n_2, \cdots, n_m)$, where $n, m \in \mathbb{N}$, $0 \leq n \leq 25$ and m is the length of the key. Each character of the ciphertext is obtained using the formula $c_i = (k_i + t_i)\%26$, $i \leq m$.



Key	D	B	F	C	D	B	F	C	D	B	F	C	D	B	F	C	D
	3	1	5	2	3	1	5	2	3	1	5	2	3	1	5	2	3
Plaintext	T	H	I	S	I	S	A	T	E	S	T	P	H	R	A	S	E
	19	7	8	18	8	18	0	19	4	18	19	15	7	17	0	18	4
Ciphertext	W	I	N	U	L	T	F	V	H	T	Y	R	K	S	F	U	H
	22	7	13	20	11	19	5	21	7	19	24	17	10	18	5	20	7



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

Table 1. Tabula recta (26 × 26) used by the Vigenère Cipher

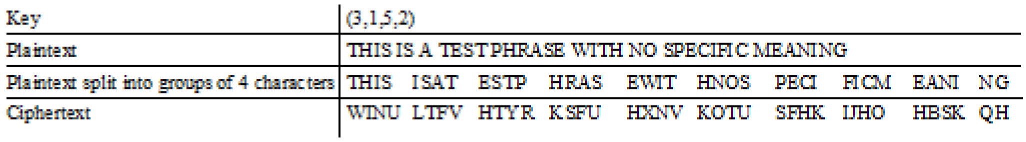
b. a group digits as a key, in order to indicate the number of positions shifted by a character.

Steps:

Let's consider the key $k = (n_1, n_2, \cdots, n_m)$, where $n, m \in \mathbb{N}$,$0 \leq n \leq 25$, m is the length of the key.

Split the message to be encrypted into a number of groups, most of them having the length of the key.

Shift the each letter of every group with the corresponding number of positions, n_i.



Key	(3,1,5,2)									
Plaintext	THIS IS A TESTPHRASE WITH NO SPECIFIC MEANING									
Plaintext split into groups of 4 characters	THIS	ISAT	ESTP	HRAS	EWIT	HNOS	PECI	FICM	EANI	NG
Ciphertext	WINU	LTFV	HTYR	KSFU	HXNV	KOTU	SFHK	IJHO	HBSK	QH

The main element is the group of characters. The encryption function that maps a group of characters of length k to another group of characters of the same length:

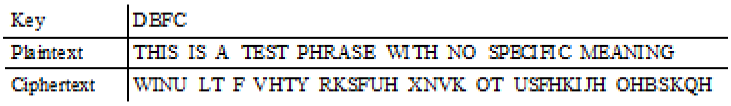
$$C_{n_1,n_2,\cdots,n_k} : \mathbb{Z}/26\mathbb{Z} \times \cdots \times \mathbb{Z}/26\mathbb{Z} \to \mathbb{Z}/26\mathbb{Z} \times \cdots \times \mathbb{Z}/26\mathbb{Z} \ ,$$

$$C_{n_1,n_2,\cdots,n_k}(x_1, x_2, \cdots, x_k) = (x_1 + n_1, x_2 + n_2, \cdots, x_k + n_k).$$

Each component is a Caesar cipher. For a key that has the length m, there are $26^k = 4 \cdot 10^{26}$ possible choices.

c. use the tabula recta to encrypt the plaintext

Steps:
Select the column corresponding to the character n_i from the key, $i \leq m$.
Select the row corresponding to the character t_i from the plaintext, $i \leq m$.
Their intersection corresponds to the character c_i from the ciphertext, $i \leq m$



Key	DBFC
Plaintext	THIS IS A TEST PHRASE WITH NO SPECIFIC MEANING
Ciphertext	WINU LT F VHTY RKSFUH XNVK OT USFHKIJH OHBSKQH

In the context of the English alphabet, the strength of the Vigenère cipher consists of the fact that it is not susceptible to frequency analysis because the cipher rotates through different shifts. Therefore, the same plaintext letter will not always be encrypted using the same ciphertext letter ([8]).

The real weakness of the Vigenère cipher lies in its periodicity ([9]).

Let's consider the key $k = (k_1, k_2, \cdots, k_m)$, where m is the length of the key, $m \in \mathbb{N}$, $V : \{A, B, \cdots, Z\} \to \{A, B, \cdots, Z\}$, the Vigenère encryption function; the plaintext $t_1, t_2, \cdots, t_n$ into the ciphertext $C_1, C_2, \cdots, C_n$, where $n \in \mathbb{N}$. Then, the Vigenère cipher with key k can be expressed as:

$$C = (C_1, C_2, \cdots, C_n) = V_k(t_1, t_2, \cdots, t_n) =$$
$$= ((t_1 + k_1)\%26, (t_2 + k_2)\%26, \cdots, (t_r + k_{r\%m})\%26, \cdots, (t_n + k_{n\%m})\%26).$$

The periodicity of the Vigenère cipher can be noticed in the following substring of the ciphertext:

$$\tilde{C_i} = \Big(C_1, C_{1+m}, \cdots, C_{1+rm}, \cdots, C_{1+[\frac{n}{m}]m}\Big) =$$
$$= ((t_1 + k_1)\%26, (t_{1+m} + k_1)\%26, \cdots, (t_{1+rm} + k_1)\%26, \cdots, (t_{1+[\frac{n}{m}]m} + k_1)\%26)$$
$$= T_{k_1}\big(t_1, t_{1+m}, \cdots, t_{1+rm}, \cdots, t_{1+[\frac{n}{m}]m}\big).$$

Therefore, if the plaintext is encrypted with a shift cipher with shift k_i, $\forall\ 1 \leq i \leq m$, it follows:

$$\Big(C_i, C_{i+m}, \cdots, C_{i+rm}, \cdots, C_{i+[\frac{n}{m}]m}\Big) = \Big(t_i, t_{i+m}, \cdots, t_{i+rm}, \cdots, t_{i+[\frac{n}{m}]m}\Big).$$

The ciphertext $C = \big(C_1, C_2, \cdots, C_n\big)$ will be split into m separate substrings of the ciphertext:

$$\tilde{C_i} = \Big(C_1, C_{1+m}, \cdots, C_{1+rm}, \cdots, C_{1+[\frac{n}{m}]m}\Big).$$

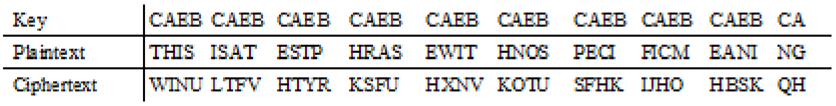
3. The Kasiski Method

Since its publication, the Vigenère cipher was considered to be unbreakable. But in 1863, Friedrich W. Kasiski published "Die Geheimschriften und die Dechif-frir-kunst" (Cryptography and the art of decryption) about a general solution for polyalphabetic ciphers with repeating keywords ([1]).

Kasiski exploits the periodicity of the Vigenère cipher (the repetition of the key). One may look for repeated fragments in the ciphertext and compile a list of the distances that separate the occurences. Then, it is possible (but not certain) that the length of the keyword is the greatest common divisor of these values.

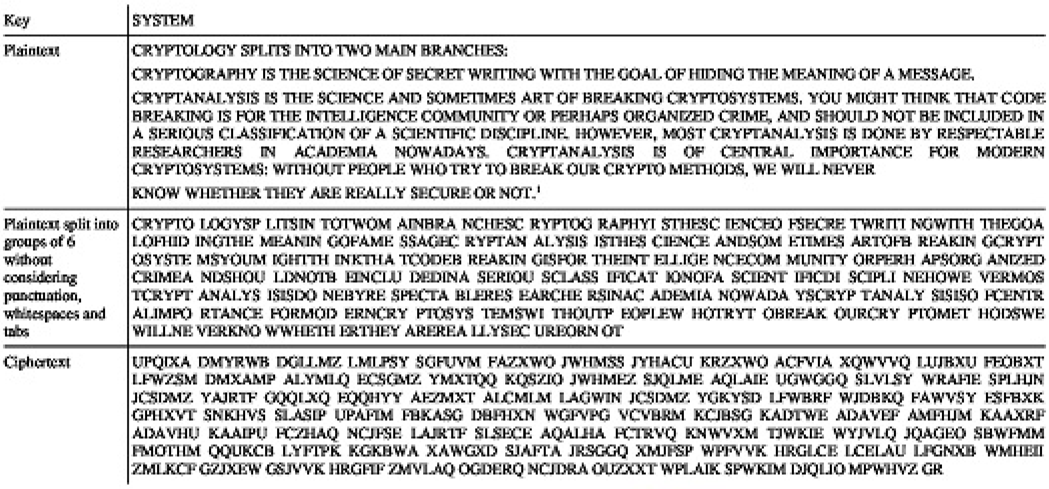
If a repeated substring in a plaintext is encrypted by the same substring in the keyword, then the ciphertext contains a repeated substring and the distance of the two occurences is a multiple of the keyword length.

Not every repeated string in the ciphertext arises in this way; but, the probability of a repetition by chance is noticeably smaller ([6]).

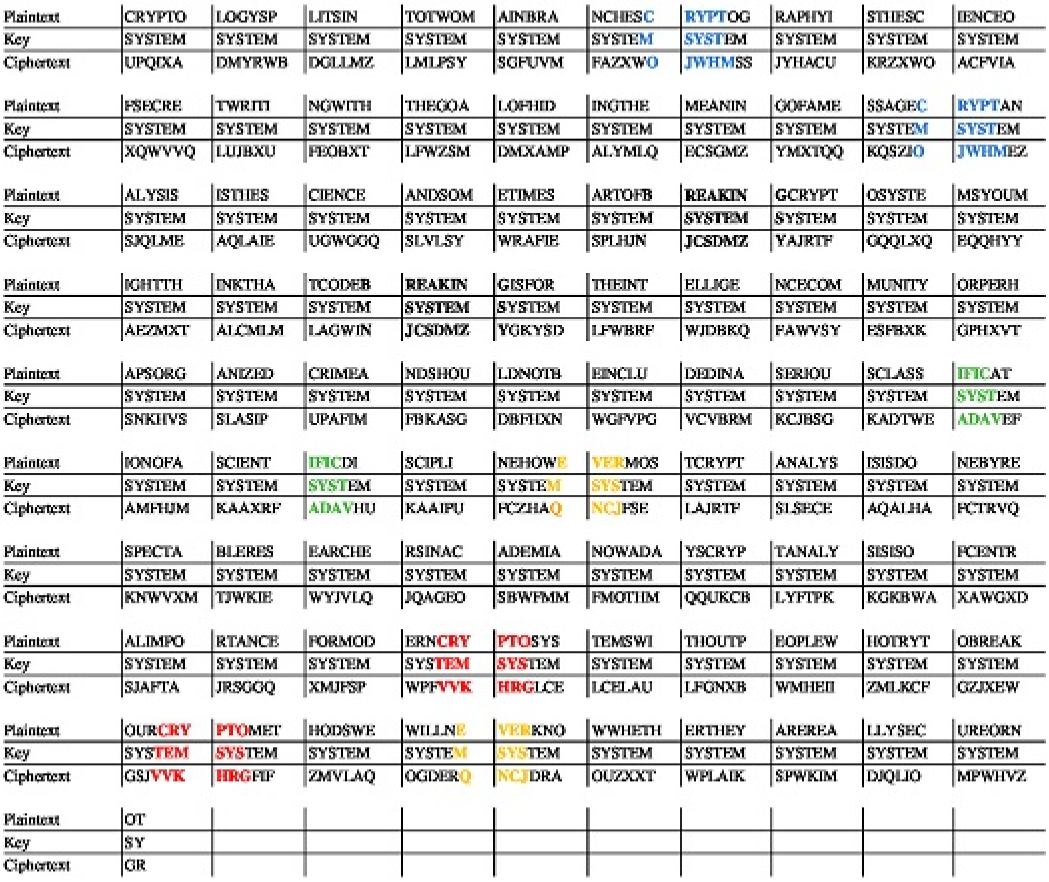


Key	CAEB	CAEB	CAEB	CAEB	CAEB	CAEB	CAEB	CAEB	CAEB	CA
Plaintext	THIS	ISAT	ESTP	HRAS	EWIT	HNOS	PECI	FICM	EANI	NG
Ciphertext	WINU	LTFV	HTYR	KSFU	HXNV	KOTU	SFHK	IJHO	HBSK	QH

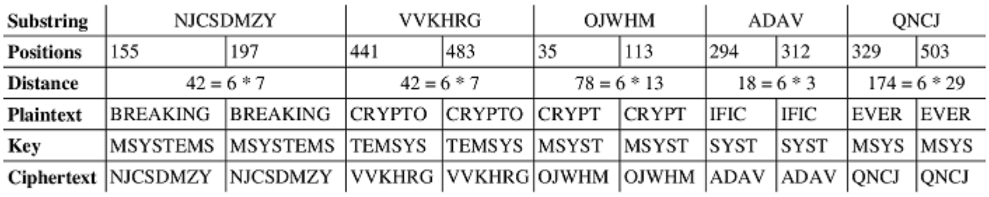
In this case, there is no repeated substring of length at least 2. Therefore, Kasiski's method will fail.



Key	SYSTEM
Plaintext	CRYPTOLOGY SPLITS INTO TWO MAIN BRANCHES: CRYPTOGRAPHY IS THE SCIENCE OF SECRET WRITING WITH THE GOAL OF HIDING THE MEANING OF A MESSAGE. CRYPTANALYSIS IS THE SCIENCE AND SOMETIMES ART OF BREAKING CRYPTOSYSTEMS. YOU MIGHT THINK THAT CODE BREAKING IS FOR THE INTELLIGENCE COMMUNITY OR PERHAPS ORGANIZED CRIME, AND SHOULD NOT BE INCLUDED IN A SERIOUS CLASSIFICATION OF A SCIENTIFIC DISCIPLINE. HOWEVER, MOST CRYPTANALYSIS IS DONE BY RESPECTABLE RESEARCHERS IN ACADEMIA NOWADAYS. CRYPTANALYSIS IS OF CENTRAL IMPORTANCE FOR MODERN CRYPTOSYSTEMS: WITHOUT PEOPLE WHO TRY TO BREAK OUR CRYPTO METHODS, WE WILL NEVER KNOW WHETHER THEY ARE REALLY SECURE OR NOT.[1]
Plaintext split into groups of 6 without considering punctuation, whitespaces and tabs	CRYPTO LOGYSP LITSIN TOTWOM AINBRA NCHESC RYPTOG RAPHYI STHESC IENCEO FSECRE TWRITI NGWITH THEGOA LOFHID INGTHE MEANIN GOFAME SSAGEC RYPTAN ALYSIS ISTHES CIENCE ANDSOM ETIMES ARTOFB REAKIN GCRYPT OSYSTE MSYOUM IGHTTH INKTHA TCODEB REAKIN GISFOR THEINT ELLIGE NCECOM MUNITY ORPERH APSORG ANIZED CRIMEA NDSHOU LDNOTB EINCLU DEDINA SERIOU SCLASS IFICAT IONOFA SCIENT IFICDI SCIPLI NEHOWE VERMOS TCRYPT ANALYS ISISDO NEBYRE SPECTA BLERES EARCHE RSINAC ADEMIA NOWADA YSCRYP TANALY SISISO FCENTR ALIMPO RTANCE FORMOD ERNCRY PTOSYS TEMSWI THOUTP EOPLEW HOTRYT OBREAK OURCRY PTOMET HODSWE WILLNE VERKNO WWHETH ERTHEY AREREA LLYSEC UREORN OT
Ciphertext	UPQIXA DMYRWB DGLLMZ LMLPSY SGFUVM FAZXWO JWHMSS JYHACU KRZXWO ACFVIA XQWVVQ LUJBXU FEOBXT LFWZSM DMXAMP ALYMLQ ECSGMZ YMXTQQ KQSZIO JWHMEZ SJQLME AQLAIE UGWGGQ SLVLSY WRAFIE SPLHJN JCSDMZ YAJRTF GQQLXQ EQQHYY AEZMXT ALCMLM LAGWIN JCSDMZ YGKYSD LFWBRF WJDBKQ FAWVSY ESFBXK GPHXVT SNKHVS SLASIP UPAFIM FBKASG DBFHXN WGFVPG VCVBRM KCJBSG KADTWE ADAVEF AMFHJM KAAXRF ADAVHU KAAIPU FCZHAQ NCJFSE LAJRTF SLSECE AQALHA FCTRVQ KNWVXM TJWKIE WYJVLQ JQAGEO SBWFMM FMOTHM QQUKCB LYFTPK KGKBWA XAWGXD SJAFTA JRSGGQ XMJFSP WPFVVK HRGLCE LCELAU LFGNXB WMHEII ZMLKCF GZJXEW GSJVVK HRGFIF ZMVLAQ OGDERQ NCJDRA OUZXXT WPLAIK SPWKIM DJQLIO MPWHVZ GR



Plaintext	CRYPTO	LOGYSP	LITSIN	TOTWOM	AINBRA	NCHESC	RYPTOG	RAPHYI	STHESC	IENCEO
Key	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM
Ciphertext	UPQIXA	DMYRWB	DGLLMZ	LMLPSY	SGFUVM	FAZXWO	JWHMSS	JYHACU	KRZXWO	ACFVIA
Plaintext	FSECRE	TWRITI	NGWITH	THEGOA	LOFHID	INGTHE	MEANIN	GOFAME	SSAGEC	RYPTAN
Key	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM
Ciphertext	XQWVVQ	LUJBXU	FEOBXT	LFWZSM	DMXAMP	ALYMLQ	ECSGMZ	YMXTQQ	KQSZIO	JWHMEZ
Plaintext	ALYSIS	ISTHES	CIENCE	ANDSOM	ETIMES	ARTOFB	REAKIN	GCRYPT	OSYSTE	MSYOUM
Key	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM
Ciphertext	SJQLME	AQLAIE	UGWGGQ	SLVLSY	WRAFIE	SPLHJN	JCSDMZ	YAJRTF	GQQLXQ	EQQHYY
Plaintext	IGHTTH	INKTHA	TCODEB	REAKIN	GISFOR	THEINT	ELLIGE	NCECOM	MUNITY	ORPERH
Key	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM
Ciphertext	AEZMXT	ALCMLM	LAGWIN	JCSDMZ	YGKYSD	LFWBRF	WJDBKQ	FAWVSY	ESFBXK	GPHXVT
Plaintext	APSORG	ANIZED	CRIMEA	NDSHOU	LDNOTB	EINCLU	DEDINA	SERIOU	SCLASS	IFICAT
Key	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM
Ciphertext	SNKHVS	SLASIP	UPAFIM	FBKASG	DBFHXN	WGFVPG	VCVBRM	KCJBSG	KADTWE	ADAVEF
Plaintext	IONOFA	SCIENT	IFICDI	SCIPLI	NEHOWE	VERMOS	TCRYPT	ANALYS	ISISDO	NEBYRE
Key	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM
Ciphertext	AMFHJM	KAAXRF	ADAVHU	KAAIPU	FCZHAQ	NCJFSE	LAJRTF	SLSECE	AQALHA	FCTRVQ
Plaintext	SPECTA	BLERES	EARCHE	RSINAC	ADEMIA	NOWADA	YSCRYP	TANALY	SISISO	FCENTR
Key	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM
Ciphertext	KNWVXM	TJWKIE	WYJVLQ	JQAGEO	SBWFMM	FMOTHM	QQUKCB	LYFTPK	KGKBWA	XAWGXD
Plaintext	ALIMPO	RTANCE	FORMOD	ERNCRY	PTOSYS	TEMSWI	THOUTP	EOPLEW	HOTRYT	OBREAK
Key	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM
Ciphertext	SJAFTA	JRSGGQ	XMJFSP	WPFVVK	HRGLCE	LCELAU	LFGNXB	WMHEII	ZMLKCF	GZJXEW
Plaintext	OURCRY	PTOMET	HODSWE	WILLNE	VERKNO	WWHETH	ERTHEY	AREREA	LLYSEC	UREORN
Key	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM	SYSTEM
Ciphertext	GSJVVK	HRGFIF	ZMVLAQ	OGDERQ	NCJDRA	OUZXXT	WPLAIK	SPWKIM	DJQLIO	MPWHVZ
Plaintext	OT									
Key	SY									
Ciphertext	GR									



Substring	NJCSDMZY		VVKHRG		OJWHM		ADAV		QNCJ	
Positions	155	197	441	483	35	113	294	312	329	503
Distance	42 = 6 * 7		42 = 6 * 7		78 = 6 * 13		18 = 6 * 3		174 = 6 * 29	
Plaintext	BREAKING	BREAKING	CRYPTO	CRYPTO	CRYPT	CRYPT	IFIC	IFIC	EVER	EVER
Key	MSYSTEMS	MSYSTEMS	TEMSYS	TEMSYS	MSYST	MSYST	SYST	SYST	MSYS	MSYS
Ciphertext	NJCSDMZY	NJCSDMZY	VVKHRG	VVKHRG	OJWHM	OJWHM	ADAV	ADAV	QNCJ	QNCJ

The distance between repeating substrings can be a multiple of the keyword length. If a match is purely coincidental, then the factors of the distance might not be multiples of the keyword length.

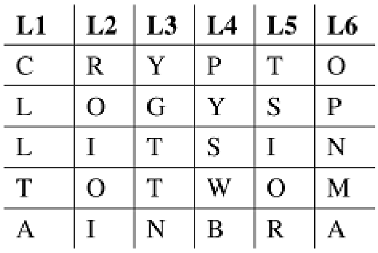
By decomposing into prime factors, we see that the greatest common divisor is 6. Therefore, the length of the key is 6.

A short plaintext with a relatively long keyword may produce a ciphertext in which no repetition can be found.

Long repeated substrings in a ciphertext are not likely to be by chance, whereas short repeated substrings may appear more often and some of which may be purely by chance.

In order to find out the key letters, one must use frequency analysis on the ciphertext split into sequences of characters having the length m (the length of the key).

For every letter that we consider to be part of the key, we have a standard English frequency distribution (red) and get a shifted English frequency distribution (blue).



L1	L2	L3	L4	L5	L6
C	R	Y	P	T	O
L	O	G	Y	S	P
L	I	T	S	I	N
T	O	T	W	O	M
A	I	N	B	R	A

For every position of the key, we consider each letter, until we get the shifted English frequency distribution to match the standard English frequency distribution.

Let's consider letter "a" as the first character for the key. Shift the frequencies of the column by 0.

Shift by 0. The standard English frequency distribution vs the shifted English frequency distribution.

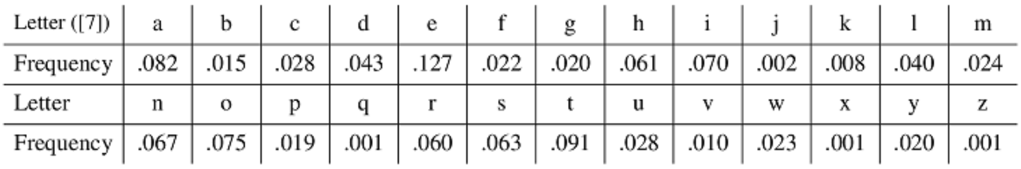
Consider letter "s" as the first character for the key. Shift the frequencies of the column by 18.

Shift by 18. The standard English frequency distribution vs the shifted English frequency distribution.

We continue shifting until the blue graph matches the red graph.

3.1. Frequency analysis

Random parts of the English language have a standard frequency distribution of the English letters.



Letter ([7])	a	b	c	d	e	f	g	h	i	j	k	l	m
Frequency	.082	.015	.028	.043	.127	.022	.020	.061	.070	.002	.008	.040	.024
Letter	n	o	p	q	r	s	t	u	v	w	x	y	z
Frequency	.067	.075	.019	.001	.060	.063	.091	.028	.010	.023	.001	.020	.001

Steps:

Count the occurrences of each of the letters in the ciphertext (e.g., 17 X's, 12 B's, 9 C's, 7 P's, etc.).

Using a standard English frequency table, guess the identity of the letters based on the table. In most of the English texts, the nine most frequent letters will be E, T, A, I, O, N, S, R, H.

Consider how letters relate to one another. One can often tell which letters are vowels because most letters appear before and/or after them. Thus, if there is a letter in the ciphertext that appears in pairs with many different letters, it may be a vowel. To find out which one it is, one must take into consideration the fact that ”a” is almost never doubled, ”e” is incredibly frequent, and ”u” is relatively rare.

3.2. The Friedman test

In order for the Kasiski attack to work, the keyword must be repeated. The ideal situation consists of having a long plaintext and a short key.

In 1922, William Friedman published ”The Index of Coincidence and Its Applications in Cryptography”, a statistical test based upon frequency that can be used to determine whether a cipher is polyalphabetic or monoalphabetic (only one ciphertext alphabet is used) and for polyalphabet ciphers can estimate the number of alphabets (the length of the keyword for the Vigenère cipher).

The index of coincidence (i.e., the repeat rate) for a ciphertext is the probability that two letters selected at random from it are identical ([10]). It is used to estimate the length of the unknown keyword.

Let's consider:

N - the length of the text be N;

n - the size of the alphabet;

a_i - the i^{th} letter in the alphabet.

Suppose a_i appears in the given text F_i times.

The number of a_i occurences in the text is F_i. There are F_i different choices to pick the first a_i, and to pick the second a_i we have $F_i - 1$ different choices (since one a_i has already been selected), etc.

Since there are $N(N-1)$ different ways of picking two characters from the text, the probability of having two a_i is $\frac{F_i(F_i-1)}{N(N-1)}$.

Since the alphabet hasndifferent letters and the above formula applies to each of them, the probability of having two identical letters from the text is:

$$I = \sum_{i=1}^{n} \frac{F_i(F_i-1)}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i=1}^{n} F_i(F_i-1).$$

If the percentage of letter a_i is p_i (see the standard frequency distribution of the English letters table above), the number of occurences of the i^{th} letter is:

$$F_i = p_i \cdot N \implies I = \sum_{i=1}^{n} p_i \frac{p_i N - 1}{N(N-1)}.$$

If $N \to \infty$ then $I \approx \sum_{i=1}^{n} p_i^2$.

If the text is randomly generated, the frequency of each letter is $p_i = \frac{1}{n}$ and $I = \frac{1}{n}$, where n is the size of the alphabet.

Then, if the plaintext is written in English, the probability of selecting two identical letters is:

$$\begin{array}{lllllllll} aa & + & bb & + & cc & + & \cdots & + & zz \\ 0.082 \times 0.082 & & 0.015 \times 0.015 & & 0.028 \times 0.028 & & & & 0.001 \times 0.001 \end{array}$$

For a monoalphabetic cipher (i.e., a permutation of the letters of a single alphabet), the frequencies of the letters is $I \approx 0.0656010$.

For a polialphabetic cipher, the frequencies of the letters should be closer to uniformity. The probability of selecting two identical letters is: $I \approx (\frac{1}{26} \cdot \frac{1}{26}) + (\frac{1}{26} \cdot \frac{1}{26}) + \cdots + (\frac{1}{26} \cdot \frac{1}{26}) = \frac{1}{26} \approx 0.038$.

Steps:

Knowing l, the length of the key, one can arrange the ciphertext into l columns. Each column corresponds to a Caesar cipher.

The columns might not all have the same length. However, one can assume that the number of letters in the ciphertext is large enough so that the length of each column can be estimated to be $\approx \frac{n}{l}$.

Case 1: Choose a letter from the ciphertext. By selecting a letter, one also selects a column. The probability that the next letter chosen comes from the same column is $\frac{\frac{n}{l}-1}{n-1}$.

Because both letters are selected from the same Caesar cipher alphabet, $I \approx 0.065$.

Therefore, the probability that both letters are identical and selected from the same column $\approx \frac{\frac{n}{l}-1}{n-1} \cdot 0.065$.

Case 2: Choose two identical letters from two different columns of the ciphertext. Calculate the probability of this event.

Select a letter from the ciphertext. The probability that the next letter comes from a different column is $\frac{n-\frac{n}{l}}{n-1}$.

Because the two letters are selected from different Caesar cipher alphabets, the probability that both are the same is approximately the same as for a random alphabet, 0.038.

So, the probability that both letters are selected from different columns and are identical is $\approx \frac{n-\frac{n}{l}}{n-1} \cdot 0.038$.

To get an approximation of the index of coincidence I (the probability that the two letters selected are identical), we add the probabilities resulted from both cases:

$$I \approx \frac{\frac{n}{l} - 1}{n-1} \cdot 0.065 + \frac{n - \frac{n}{l}}{n-1} \cdot 0.038$$

$$\implies (n-1)I + 0.065 - 0.038n \approx 0.027\frac{n}{l} \implies l \approx \frac{0.027n}{(n-1)I + 0.065 - 0.038n}.$$

References

[1] Kahn, David, *The Codebreakers*, Macmillan co., 1996.

[2] Knospe, Heiko, *A Course in Cryptography*, American Mathematical Soc., 2019.

[3] Lewand, Robert, *Cryptological Mathematics*, Cambridge University Press, 2000.

[4] Martin, Keith, *Everyday cryptography: Fundamental principles and applications*, Oxford University Press, 2012.

[5] Paar, Christof, Pelzl, Jan, *Understanding Cryptography : a Textbook for Students and Practitioners*, Springer, 2010.

[6] Pommerening, Klaus, *Kasiski's Test: Couldn't the Repetitions be by Accident?*, Cryptologia, vol. 30, no. 4, 2006, pp. 346–352.

[7] Trappe, Wade, Washington, Lawrence C., *Introduction to Cryptography with Coding Theory*, Pearson, 2nd ed., 2006.

[8] https://crypto.interactive-maths.com/kasiski-analysis-breaking-the-code.html

[9] https://web.stanford.edu/class/stats47n/coursework/notes/lecture4.pdf

[10] https://pages.mtu.edu/ shene/NSF-4/Tutorial/VIG/Vig-IOC.html

Technical University of Cluj-Napoca
master student, Software Engineering Degree Program
master's degree thesis, advisor: Prof. Dr. Eng. Alin-Dumitru Suciu
Cluj-Napoca, Romania
E-mail address: `dobritoiu.ana@gmail.com`