Classical Cryptology II

Hill Cipher

Use an n×n matrix M which is invertible mod 26 (gcd(det M, 26) = 1). Block the message into blocks of size n, convert to numbers and multiply each block by M to encrypt (reduce results mod 26). To decrypt, repeat procedure using M^{-1}.

Jenna has cute kids

$$M = \begin{pmatrix} 3 & 17 \\ 1 & 6 \end{pmatrix}$$

J E N N AH A S C U T E K I D S
9 4 13 13 0 7 0 18 2 20 19 4 10 8 3 18

5 21 0 13 7 16 18 4 0 24 9 9 12 10 1 3
F V A N H Q S E A Y J J M K B D

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
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

Cryptanalysis of a Monoalphabetic Substitution Cipher

CKPKH GVGCK UGZQA GCKUG CLGPQ FJZIG

PQQAF QQLHG FJZEF QGKEF CCQAG LOULJ

QFRGM OGPQA FUGZO SJBQA GLOTS

MFOKS JZKOQ VKIGE KOGFJ ZKJGI XKJGT

OGMQP LCGJQ CXQKO GPQYD

Cryptanalysis of a Monoalphabetic Substitution Cipher

Cryptogram		English (based on 135 letters)	
G	21	e	17
Q	16	t	13
K	12	a, o	11
F,J,O	9	n, i	10
C	8	s	9
L,P,Z	6	r	8
A	5	h	7
U	4	l, d	5
E,I,M,S	3	c, u	4
H,T,V,X	2	p,f,m,w	3
B,D,R,Y	1	y,b,g	2
		v,k	1

Cryptanalysis of a Monoalphabetic Substitution Cipher

Digrams in Cryptogram	English Digrams
QA 5	th 4
GP,JZ,OG,PQ 4	he 3
KO,FJ,CK,AG,UG 3	an,in,er,re,es 2
GC,GZ,GF,GL,GM,QF	on,ea,ti,at,st,en,nd
QQ,KU,KJ,FQ,JQ,JG	to,nt,ed,is,ar,ou,te
LO,ZK,AF,EF,IG,SJ 2	of,it,ha,se,et,or 1

Trigrams in Cryptogram	English Trigrams (in order of frequency)
GPQ 4	the
QAG, FJZ 3	and
QAF,JZK,OGP,KOG,CKU,AGL	tha
UGZ,GFJ, GLO,KUG,KJG 2	ent
	ion

Cryptanalysis of a Monoalphabetic Substitution Cipher

We start our cryptanalysis by trying to identify some very high frequency letters:

G is the most frequently occurring letter, so we assume that it is an "e".

"the" is the most frequent trigram in English, so we look for frequent trigrams that end in G, i.e., QAG, KOG, KUG and KJG.

"t" should be a high frequency letter and "h" a medium frequency letter. QAG has a slight advantage over KUG in this regard.

The "th" digram is the most frequent, supporting QAG as the correct choice.

Cryptanalysis of a Monoalphabetic Substitution Cipher

Identify "t", "h" and "e" as Q, A and G respectively.

```
      e e    e th e   e   e t     e  tth  tt  e
CKPKH GVGCK UGZQA GCKUG CLGPQ FJZIG PQQAF QQLHG
      te      the       t  e   e th   e      th
FJZEF QGKEF CCQAG LOULJ QFRGM OGPQA FUGZO SJBQA
e               t    e     e      e     e    e t
GLOTS MFOKS JZKOQ VKIGE KOGFJ ZKJGI XKJGT OGMQP
  e t   t   e t
LCGJQ CXQKO GPQYD
```

F (occurring in block 7) must be a vowel and a high frequency one like "a" or "o" (as opposed to u or i). "tha" is a high frequency trigram, and QAF is on our list, so we associate F with a.

"an" is a high frequency digram and "and" a high frequency trigram. FJ and FJZ would seem to fit, so associate J with n and Z with d.

Cryptanalysis of a Monoalphabetic Substitution Cipher

e e ed t h e e e t a nd e t t h a t t e
CKPKH GVGCK UGZQA GCKUG CLGPQ FJZIG PQQAF QQLHG
a nd a t e a t h e n t a e e t h a e d n t h
FJZEF QGKEF CCQAG LOULJ QFRGM OGPQA FUGZO SJBQA
e a nd t e e a n d n e n e e t
GLOTS MFOKS JZKOQ VKIGE KOGFJ ZKJGI XKJGT OGMQP
en t t e t
LCGJQ CXQKO GPQYD

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
F . . ZG . . A . . . . . J . . . . . Q . . . . . .

e,t,a,o,n are the most frequent letters and we have found 4. The high frequency K must surely be our missing letter. Associate K with o.

"he" and "re" are high frequency digrams. OG is on our list so associating O with r is pausible considering that "r" is a medium frequency letter, and so is O.

Cryptanalysis of a ....

o o e e o ed t h e o e e t and e t t h a t t e
CKPKH GVGCK UGZQA GCKUG CLGPQ FJZIG PQQAF QQLHG
and a t e o a t h e r n t a e r e t h a ed r n t h
FJZEF QGKEF CCQAG LOULJ QFRGM OGPQA FUGZO SJBQA
e r a r o nd o r t o e o r e an do ne o ne r e t
GLOTS MFOKS JZKOQ VKIGE KOGFJ ZKJGI XKJGT OGMQP
en t t o r e t
LCGJQ CXQKO GPQYD
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
F . . ZG . . A . . . . . JK . . O .Q . . . . . .

Now looking at ..oreandone..one.. (above blocks 21 and 22), we can see ".. and one .. one..", a common phrase which would indicate that IX should be associated with by.

Examining the keyword substitution list, we clearly see the end of the alphabet in place. This would force us to associate s with P and z with Y. The spacing would require that q is associated with either M or N, but the low frequency of "q" favors the association of q with N.

Cryptanalysis of a ....

o s o e e o e d t h e o e e s t andb e s t t h a t t e
CKPKH GVGCK UGZQA GCKUG CLGPQ FJZIG PQQAF QQLHG
and a t e o a t h e r n t a e r e s t h a e d r n t h
FJZEF QGKEF CCQAG LOULJ QFRGM OGPQA FUGZO SJBQA
e r a r o nd o r t obe or e an do neb yone r e t s
GLOTS MFOKS JZKOQ VKIGE KOGFJ ZKJGI XKJGT OGMQP
ent y t o r e s t z
LCGJQ CXQKO GPQYD

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
F I . ZG . . A . . . . . J K .NOPQ . . . . XY

The remaining high frequency letters are C and L, while the last of the high frequency letters in English is an "i". The occurrence of ..CC.. in block 11 makes the choice of C for i a poor one, so we associate L with i. The order of frequency for doubles in English is "ss", "ee", "tt", "ff", "ll", "mm", and "oo". Thus C is most likely f, l or m. f and m would give poor fits in block 11 so we associate C with l.

Cryptanalysis of a ....

l o s o e e l o e d t h e l o e l i e s t andbe s t t h a t t i e
CKPKH GVGCK UGZQA GCKUG CLGPQ FJZIG PQQAF QQLHG
and a t e o a l l t h e i r i n t a e r e s t h a ed r n t h
FJZEF QGKEF CCQAG LOULJ QFRGM OGPQA FUGZO SJBQA
e i r a r o ndo r t obe o r e an do neb y o ne r e t s
GLOTS MFOKS JZKOQ VKIGE KOGFJ ZKJGI XKJGT OGMQP
i l en t l y t o r es t z
LCGJQ CXQKO GPQYD

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
F I . ZG . . AL . . C . J K . NOPQ . . . .XY

From the keyword list, we can see that p should be associated with M.
Blocks 4 and 5 indicate U associated with v to get the word "lovliest".
Block 8 indicates H associated with m to get "time".
Blocks 10 and 11 indicate E associated with f to get "of all".
Blocks 12 and 13 indicate R associated with g to get "vintage".
Block 24 indicates T associated with c to get "crept".

Cryptanalysis of a ....

l o s om e e l o v e d t h e l o v e l i e s t andbe s t t ha t t i m e
CKPKH GVGCK UGZQA GCKUG CLGPQ FJZIG PQQAF QQLHG
and f a t e o f a l l t h e i r v i n t ag e p r e s t h av e d r n t h
FJZEF QGKEF CCQAG LOULJ QFRGM OGPQA FUGZO SJBQA
e i r c pa r o ndo r t ob ef o r e an don eb y o nec r e p t s
GLOTS MFOKS JZKOQ VKIGE KOGFJ ZKJGI XKJGT OGMQP
i l e n t l y t o r e s t z
LCGJQ CXQKO GPQYD

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
F I T ZGERAL . . CH J KMNOPQ. U . . XY

A final checking of the substitution list gives the remaining associations:
j with D
k with B
u with S
w with V
x with W

Cryptanalysis of a ....

The cryptogram is finally solved (with keyword FITZGERALD).

Lo! some we loved, the lovliest and best
That Time and Fate of all their Vintage prest,
Have drunk their Cup a Round or two before
And one by one crept silently to rest. (zj - nulls)

Substitution Ciphers

Polyalphabetic (Vigenère - 1586)

In this type of substitution cipher, a letter is not
fl inki sinf or itnowflinkis infori t nowfli nk isi
NY BUSA LGCJ CW ANOGPNEDGSWF KVUVVZ T YSPYPZ VC VGB

always replaced by the same letter. This procedure
nforit nowflink is inf orit nowfli nkis inforitno
NQKRGL ESLQLKRN JQ BUJ GRUX YSPYPZ GRQK XETQVLNES

distorts the statistical frequencies of the original
wflinkis inf oritnowflin kisinforitn ow fli nkisinfo
ZNDBBBBK BUJ GKIMVGPNNIY PZWYHJBTQXF FB YSM BBQYQAFZ

message. A keyword or phrase is used, and duplicate
ritnowf l inkisin fo ritnow fl inki nfo ritnowfli
DMLFOCJ L SRJEGZQ TF GPKNGA ND CFOL NSR UCIYWYFEM

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
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

Cryptanalysis of a Vigenère

The first and most important step is to determine the keylength. We give three methods.

1. **Kasiski Test** (1863 - Major F.W. Kasiski, German cryptologist): Length of keyword is a divisor of the gcd of the distances between identical strings of length at least 3.

2. **Sliding Strips**: Write the entire cryptogram on two separate strips of paper. Line up the strips, one above the other and slide the top strip one position at a time. At each step count the number of agreements of letters in the two strips. The slide position that gives the greatest number of agreements is usually the keylength.

Cryptanalysis of a Vigenère

3. **Friedman Test** (1925, Colonel William Frederick Friedman (1891-1969)) also called the *Kappa Test*:

The **index of coincidence** of a message is the probability that a randomly chosen pair of letters in the message are equal. If the message has length n and n_i denotes the number of occurrences of the i^{th} letter then the index, denoted by **I**, is given by:

$$I = \frac{\sum_{i=1}^{26} n_i(n_i-1)}{n(n-1)}.$$

Cryptanalysis of a Vigenère

Now we can also calculate this index for any language source if we know the probabilities of occurrence of each of the letters. Thus, if p_a is the probability of occurrence of the letter a, for example, then we get:

$$I_{\text{Source}} = p_a p_a + p_b p_b + \cdots + p_z p_z = \sum_{i=a}^{z} p_i^2 .$$

Using our knowledge of these probabilities we can easily calculate that $I_{\text{English}} \sim 0.065$ and if we had a random source of English letters then $I_{\text{Random}} \sim 0.038$ (= 1/26).

Cryptanalysis of a Vigenère

This index can give information about a message. For instance, if a ciphered message was either a transposition or a monoalphabetic substitution then one would expect to have $I_{Message} \sim I_{English}$, but if a polyalphabetic substitution was used then this value should decrease (but no lower than 0.038) since the polyalphabetic procedure tends to randomize the occurrences of the letters.

Let us now apply this index to a Vigenère ciphertext. If the ciphertext has length n and the keyword has length k (and $n >> k$) then in the positions corresponding to the same letter of the keyword, the ciphertext has been created with a monoalphabetic substitution, so if one were to calculate the index of just those positions, we should get 0.065. On the other hand if one were to calculate the index using only pairs from different letters of the keyword, the index would be much lower (0.038 if the keyword letters were randomly chosen).

Cryptanalysis of a Vigenère

We may therefore calculate the expected number (A) of pairs of equal letters in the following way:

Pick a letter from the ciphertext (n choices), there are (n/k - 1) remaining letters that have used the same keyword letter [we are neglecting round-off error] and so,

$$\frac{n\left(\frac{n}{k}-1\right)}{2} = \frac{n(n-k)}{2k}$$

pairs of this type. There are (n - n/k) remaining letters that have used a different keyword letter [assuming the keyword letters are all distinct], and so there are

$$\frac{n\left(n-\frac{n}{k}\right)}{2} = \frac{n^2(k-1)}{2k}$$

Cryptanalysis of a Vigenère

pairs of this type. Therefore,

$$A = \frac{n(n-k)}{2k}(0.065) + \frac{n^2(k-1)}{2k}(0.038)$$

and so,

$$I_{\text{Ciphertext}} = \frac{A}{\binom{n}{2}} = \frac{n-k}{k(n-1)}(0.065) + \frac{n(k-1)}{k(n-1)}(0.038) = \frac{1}{k(n-1)}(0.027\,n + 0.065\,k)$$

from which we may solve for k (keyword length):

$$k = \frac{0.027\,n}{(n-1)\,I_{\text{Ciphertext}} - 0.038\,n + 0.065}.$$

Cryptanalysis of a Vigenère

Given the keylength k, calculate the frequencies of letters that are k positions apart. These correspond to the same cyclic shift cipher, so the most frequent letter in these positions should correspond to an "e" - thus giving you the shift for these positions.

A more accurate procedure is as follows:

Let A_0 be the vector of letter probabilities in order for English.

Let A_i be the i^{th} cyclic shift of vector A_0 ($0 \leq i < 26$).

Let W be the frequency vector of the letters a,b,c, ... etc in the positions you are looking at.

Take the dot product of W with each A_i, the largest value obtained will identify the shift i.

Keylength Determination

NYBUS ALGCJ CWANO **GPN**ED GSWFK VUVVZ T**YSPY** **PZ**VCV

GBNQK RGLES LQLKR NJQ**BU** **JG**RUX **YSPYP** **Z**GRQK XETQV LNESZ

NDBBB BK**BUJ** **G**KIMV **GPN**NI YPZWY HJBTQ XFFBY SMBBQ YQAFZ

DMLFO CJLSR JEGZQ TFGPK NGAND CFOLN SRUCI YWYFEM

Distances between YSPYPZ's and BUJG's are both 34, while between the GPN's it is 85. By Kasiski we would expect a keylength which divides gcd(85,34) = 17. I.e., 17 in this case.

There are n = 171 letters and we calculate $I_{crypt} \approx 0.0472$ (which indicates a non-random polyalphabetic substitution). We can then calculate the keylength k = 2.901 (terrible result due to small sample size)

Key Determination

NYBUS ALGCJ CWANO GP**N**ED GSWFK VUVVZ TYSP**Y** PZVCV

GBNQK RGLES L**Q**LKR NJQBU JGRUX YSP**Y**P ZGRQK XETQV LNESZ

NDBBB BKBUJ GKIMV GP**N**NI YPZWY HJBTQ XFFB**Y** SMBBQ YQAFZ

DMLFO C**J**LSR JEGZQ TFGPK NGA**N**D CFOLN SRUCI YWY**F**EM

Simple method says e→N, but this doesn't work !!

$WA_1 = 0.466$ $WA_2 = 0.505$ $WA_3 = 0.235$ .. $\mathbf{WA_5 = 0.537}$

... $WA_9 = 0.437$ .....

So a shift of 5 is indicated.

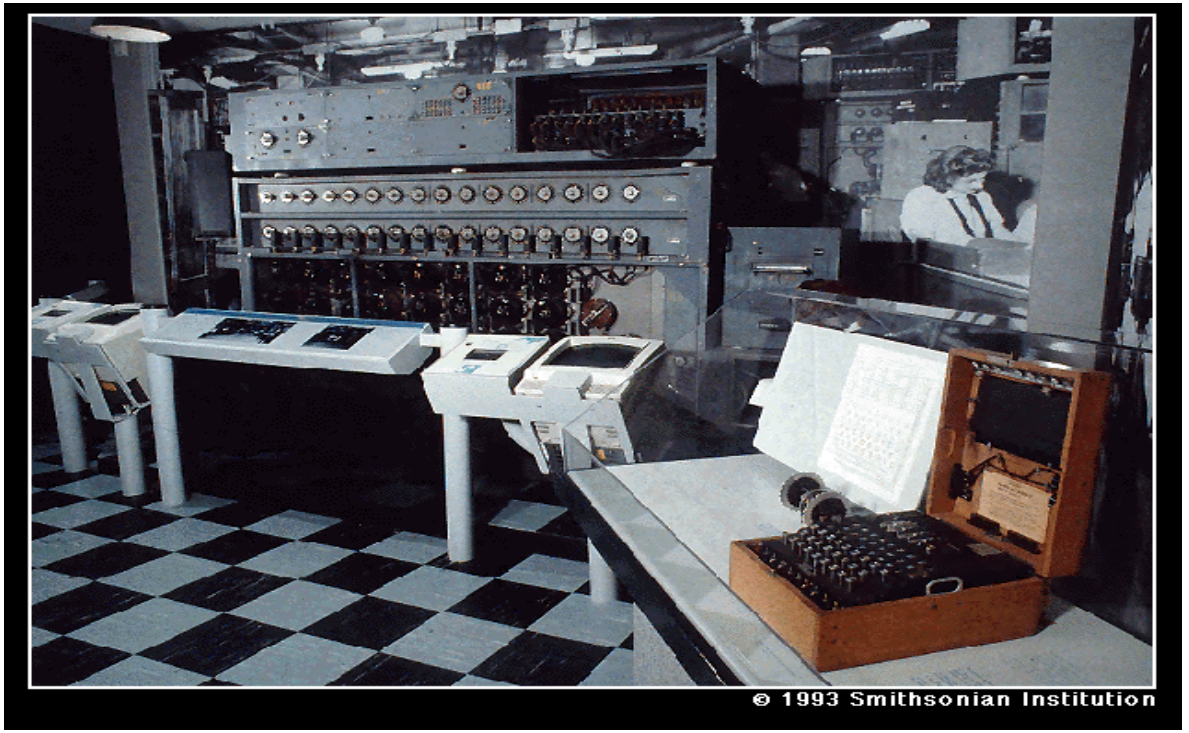
$i \rightarrow N$ $t \rightarrow Y$ $l \rightarrow Q$ $e \rightarrow J$ $a \rightarrow F$

and the first letter of the key is **f.**

Cryptomachines

Machines used to encipher and decipher during wartime (WWII).

- Hagelin (US Army)
- Enigma (German Army and Navy)



- Purple (Japanese Diplomatic)

Enigma

Some references:

David Khan, *The Code Breakers: The story of secret writing*

David Khan, *Seizing the Enigma: The race to break the U-boat codes*

Wladyslaw Kozaczuk, *Enigma: How the german machine cipher was broken, and how it was read by the Allies in WWII*

The One-Time Pad

One-Time Pad (**Vernam Cipher**)

Vignère Cipher with keylength = size of message and the key is formed randomly.

Perfect Security, it can not be broken. Used in the Washington-Moscow Red Phone line.

Disadvantage, both parties must have the same key. (Books of random numbers or letters)

One-Time Pad

Quoted from Neal Stephenson's ***Cryptonomicon:***

These sheets were typed up by a Mrs. Tenney, an aged vicar's wife who works at Bletchley Park. Mrs. Tenney has a peculiar job which consists of the following: she takes two sheets of onionskin paper and puts a sheet of carbon paper between them and rolls them into a typewriter. She types a serial number at the top. Then she turns the crank on a device used in bingo parlors, consisting of a spherical cage containing twenty-five wooden balls, each with a letter printed on it (the letter J is not used). After spinning the cage the exact number of times specified in the procedure manual, she closes her eyes, reaches through a hatch in the cage, and removes a ball at random. She reads the letter off the ball and types it, then replaces the ball, closes the hatch, and repeats the process. From time to time, serious-looking men come into the room, exchange pleasantries with her, and take away the sheets that she has produced. These sheets end up in the possession of men like Waterhouse, and men in infinitely more desperate and dangerous circumstances, all over the world. They are called one-time pads.

Modern Cryptanalysis

Levels of Attack:

Ciphertext only (cryptanalyst sees only the enciphered text)

Plaintext attack (cryptanalyst has both the enciphered text and the clear message that it came from)

Chosen plaintext attack (cryptanalyst can chose the clear message that is enciphered, as often as needed)

Chosen ciphertext attack (cryptanalyst can chose ciphertext to be decrypted, as often as needed ... but is not given the key).

Modern Cryptanalysis

The "philosophy" of modern cryptanalysis is embodied in Kerckhoffs' principle, which was formulated in the book ***La cryptographie militaire*** (1883) by the Dutch philologist Jean Guillaume Hubert Victor François Alexandre Auguste Kerckhoffs von Nieuwenhof, as he is called in all his full glory.

Kerckhoffs' Principle: *The security of a cryptosystem must not depend on keeping secret the crypto-algorithm. The security depends only on keeping secret the key.*