CS 5602 Lecture 14 Historical Ciphers I

George Markowsky
Computer Science Department
Missouri University of Science and Technology

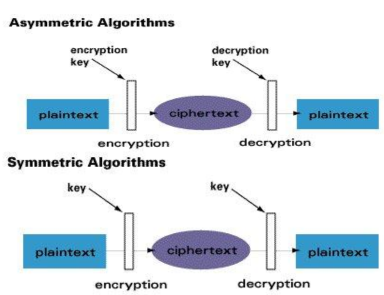
1

Next Steps

- Chapter 2 in the book is on Elliptic Curve Cryptography, which is a very important topic in contemporary cryptography
- The basic idea here is to use ideas from Projective Geometry to create groups that are used for cryptographic purposes
- The topic is presented without any motivation in Chapter 2, so if we went right into it, I would have to give you a series of lectures on pure math so you could understand what is going on
- I think that at this point, it would be best to go into some less mathematical forms of cryptography, and then return to elliptic curves near the end of the course
- We will skip Chapter 2 for now and proceed to Chapter 3 for now
- We will return to Chapter 2 later in the course

2

Types of Algorithms



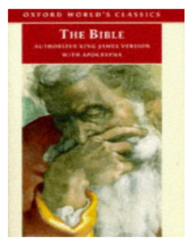
3

Transposition Ciphers

- The idea here is to permute the actual letters of the message in some way
- Some examples are:
  - Reverse pairs of letters
    - HELLO WORLD
    - EHLL OOWLRD
  - Remove every second letter and put it at the end.
    - HELLO WORLD
    - HLOWRDEL OL
  - Reverse the letters
    - HELLO WORLD
    - DLROW OLLEH
  - Pig Latin
- Many trickier transpositions are possible as you will see throughout the course

4

Cryptography in the Bible



- Jewish writers in the Bible used *atbash*, which is a substitution cypher in which the first letter is replaced the last, the second by the next to last, etc.
- Thus, Babylon comes out Sheshach or Sheshech

5

Atbash

- This illustrates how the atbash algorithm works

ABCDEFGHIJKLMNOPQRSTUVWXYZ
ZYXWVUTSRQPONMLKJIHGFEDCBA

XIBKGLTIZKSB

Atbash

6

Atbash and Group Theory

- Let A = { A..Z }. We will assume that blanks (represented by β) map to blanks
- Let A* = { all words of finite length made from elements of A }
- Let S_A be the permutation group on A
- For each $\sigma \in S_A$, $\sigma: A \rightarrow A$
- Let $\sigma^*: A^* \rightarrow A^*$ be given by $\sigma^*(c_1c_2 \ldots c_k) = \sigma(c_1)\,\sigma(c_2) \ldots \sigma(c_k)$
- Thus there are |A|! different permutations on A*
- Consider the permutation $\pi: A \rightarrow A$ given by $\pi(A) = Z$, $\pi(B) = Y$, …, $\pi(Z) = A$, $\pi(\beta) = \beta$
- It is clear that $\pi^2 = 1$ and encryption and decryption are identical
- Because { 1, π } is a subgroup of order 2 of S_A, Atbash is not very secure!
- Once you know the encryption method, the decryption is straightforward!

7

Atbash and Group Theory

- For ease of computation use range(|A|), in this case { 0, …, 25}
- In this scheme the Atbash permuation is $\alpha(k) = 25 - k$ so you can see that $\alpha(\alpha(k)) = 25 - (25 - k) = k$

8

Atbash in Python

```
A = 'ABCDEFGHIJKLMNOPQRSTUVWXYZ'
# print(len(A))

def atbash(word):
    word = word.upper()
    oword = ""
    for c in word:
        oword += A[25-A.index(c)]
    return oword

testWords = "hello crypto smooth decode encode".split()

for w in testWords:
    print (w,atbash(w),atbash(atbash(w)))
```

```
hello SVOOL HELLO
crypto XIBKGL CRYPTO
smooth HNLLGS SMOOTH
decode WVXLWV DECODE
encode VMXLWV ENCODE
```

9

Atbash2 in Python

```
A = 'ABCDEFGHIJKLMNOPQRSTUVWXYZ'
Rev = A[::-1]
ATB = { }

for i in range(len(A)):
   ATB[A[i]] = Rev[i]

def atbash2(word):
    word = word.upper()
    oword = ""
    for c in word:
        oword += ATB[c]
    return oword

for w in testWords:
    print (w,atbash(w),atbash(atbash(w)))
```

```
hello SVOOL HELLO
crypto XIBKGL CRYPTO
smooth HNLLGS SMOOTH
decode WVXLWV DECODE
encode VMXLWV ENCODE
```

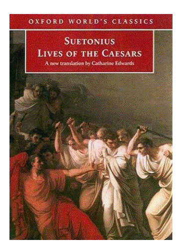
10

Substitution Cyphers

- Atbash is an example of a substitution cypher
- Widely used in various detective stories
  - *The Gold Bug* by Edgar Allan Poe about his detective Legrand
  - *The Adventure of the Dancing Men* by Arthur Conan Doyle about his detective Sherlock Holmes
- Just remember that there are groups underlying all cryptographic methods

11

Julius Caeser and Cryptography



- According to Suetonius, Julius Caeser used some simple substitution ciphers
- Below is a substitution cipher where each letter is replaced by a letter 3 further in the alphabet. Caeser only used shifting by 3

ABCDEFGHIJKLMNOPQRSTUVWXYZ

DEFGHIJKLMNOPQRSTUVWXYZABC

12

Julius Caeser's Ciphers

- This illustrates how a Caeser cipher works

CRYPTO 3

Encrypt Shift Size

13

Group Theory and Substitution Ciphers

- Let A, A*, S_A, σ, and σ* be as before
- There are |A|! different permutations on A*
- For this limited alphabet, this is 26! = 403,291,461,126,605,635,584,000,000
- How do we find the prime factors of 26! intelligently?
- It has no prime factor bigger than 23!
- $26! = 2^a 3^b 5^c 7^d 11^e 13^f 17^g 19^h 23^i$
- Note that g = h = i = 1. Why?
- What is f?
- f = 2. What are e and d?
- e = 2 and d = 3. What are a, b, and c?
- a = 23, b = 10, c = 6, so $26! = 2^{23}3^{10}5^6 7^3 11^2 13^2 17^1 19^1 23^1$

14

The Prime Factorization of n!

- In general, $n! = \prod_{p\ prime \le n} p^{e_p}$ and $e_p = \left\lfloor \frac{n}{p^1} \right\rfloor + \left\lfloor \frac{n}{p^2} \right\rfloor + \left\lfloor \frac{n}{p^3} \right\rfloor + \cdots + \left\lfloor \frac{n}{p^k} \right\rfloor$ where $p^k \le n < p^{k+1}$
- Why is this true?
- If we are looking for subgroups of S_n we need to look at all the possible factors of n!
- In general there are $\prod_p (e_p + 1)$ factors of n! where p ranges over all primes ≤ n
- Why?
- How many potential orders for subgroups of S_{26}?

15

Orders of Subgroups of S_{26}

- We know that $26! = 2^{23}3^{10}5^6 7^3 11^2 13^2 17^1 19^1 23^1$, so the number of potential orders of subgroups of S_{26} is what?
- = 24*11*7*4*3*3*2*2*2 = 532,224
- Clearly, we need better tools than we have now to really take S_{26} apart
- What about real alphabets such standard ASCII with characters ranging from ord(32) to ord(126) which gives us 95 characters to play with!
- Nothing but fun to work with S_{95} !

16

Classical Cryptology

- Invented in the Middle East
  - Interested in riddles and puzzles
  - Described advanced variations of substitutions ciphers
  - Introduced frequency analysis of letters and letter combinations
  - Influenced Europeans

17

Substitution Ciphers

- You have some permutation of letters that permits you to substitute one letter for another. *Often will remove blanks.*

ABCDEFGHIJKLMNOPQRSTUVWXYZ
BADCFEHGJILKNMPORQTSVUXWZY

HELLO WORLD GFKKPXPQKC

→

18

Analyzing Substitution Ciphers

- Frequency Table
- Can use multiple alphabets, etc.

13	9	8	8	7	7	7	6	6	4	4	3	3	3	3	2	2	2	1	1	1	-	-	-	-	-
E	T	A	O	N	I	R	S	H	L	D	C	U	P	F	M	W	Y	B	G	V	K	Q	X	J	Z

19

Classical Cryptology

- Renaissance political intrigues sparked a resurgence of cryptology
- Leon Battista Alberti (ca. 1465) the Father of Western Cryptology
- Giovanni Soro of Venice the first great Western cryptanalyst
- Cryptanalysis by Thomas Phelippes supplied evidence against Mary, Queen of Scots (ca. 1586)
  - *A weak cipher is worse than no cipher at all*
- Many famous names
- Many countries had *Black Chambers*

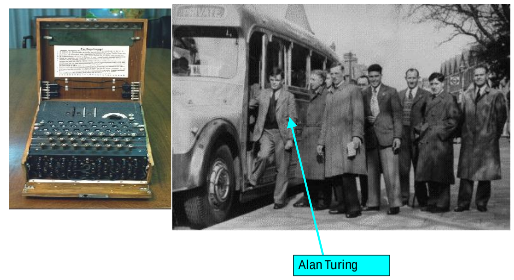
20

Modern Communications

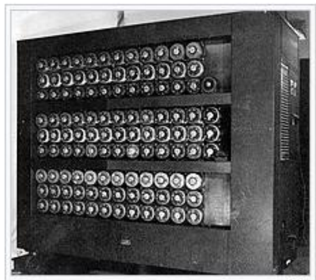
- Telegraphy greatly increased volume of cryptographic messages
  - Interception sporadic and difficult
  - Cables can be taped sometimes
- Radio communication made cryptanalysis come into its own
  - Assumption is that enemy has all the text
- Volume of traffic might limit complexity of cryptographic system
  - High speed computers change this

21

Enigma

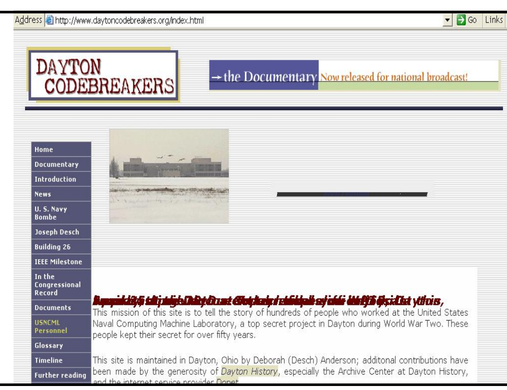


22

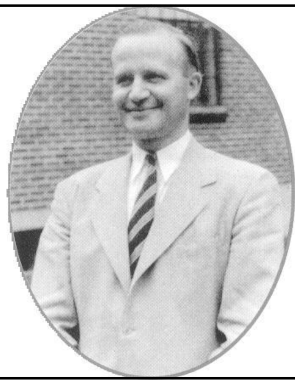


A wartime picture of a Bletchley Park Bombe

23

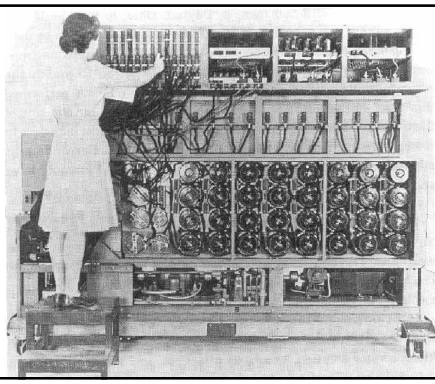


24



Joseph Desch

25



26

Transposition Ciphers

- The idea here is to permute the actual letters of the message in some way
- Some examples are:
  - Reverse pairs of letters
    - HELLO WORLD
    - EHLL OOWLRD
  - Remove every second letter and put it at the end.
    - HELLO WORLD
    - HLOWRDEL OL
  - Reverse the letters
    - HELLO WORLD
    - DLROW OLLEH
- Many trickier transpositions are possible

27

Types of Cryptographic Systems

- *Restricted use systems* -- must keep nature of encoding and decoding secret
- *General use systems* -- nature of encoding and decoding is generally known -- must use a *key* to help safeguard system
  - *Secret-key systems* -- most traditional systems -- same key for encoding and decoding
  - *Public-key systems* -- public key provided for encoding and a private key used for decoding

28

Keys and Codes

- A *key* is a small amount of information needed to use a cryptographic system
- For a Caeser type cipher only 26 keys are possible, which is a ridiculously small number.
- For a general substitution cipher, 26! $\approx$ 4*10^{26} keys are possible
- Substitution ciphers can easily be broken using frequency analysis

29

The One-Time Pad

- There is one classical provably secure cryptographic system called the *one-time pad*
- As the name suggests, you can only use it once and then it must be replaced
- Very secure, but not very handy
- Uses the xor operator $\oplus$

30

The One-Time Pad

- Recall that 0⊕0 = 0, 0⊕1 = 1, 1⊕0 = 1, and 1⊕1 = 0
- Like +, ⊕ is commutative (a ⊕b = b ⊕a) and (a ⊕(b ⊕c) = (a ⊕b) ⊕c).
- In addition, (a ⊕b) ⊕a = b
- Makes ⊕ handy for computer graphics -- i.e., xoring something to itself cancels it out

31

The One-Time Pad

- The idea here is that if you have a one-time pad and a message, then the sender sends Message ⊕ OTP
- The receiver then does (Message ⊕OTP) ⊕OTP = Message ⊕(OTP ⊕OTP) = Message
- If the pad is used twice it is possible to deduce what it is

32

The One-Time Pad

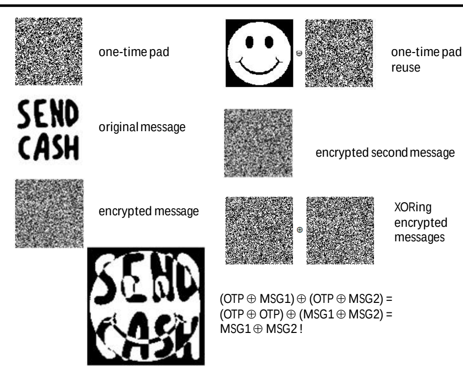
- Why is the one-time pad unbreakable if used only once?
- Because, for any string S and message M of the same length as M, there is a one-time pad Q, such that M ⊕Q = S
  - *Proof:* Let Q = M ⊕S!
- Why don't we just use one-time pads all the time?
- Was used on the hotline between US and USSR

33

One-Time Pad Reuse

- **Don't Do It!**
- Why not?
- The following graphical example comes from
- https://cryptosmith.com/2008/05/31/stream-reuse/

34



35

Problems with Keys

- Distribution
- Updating
- Security
- Distribution
- Updating
- Security
- How can the public use cryptography?
- The next group of slides comes from Smart's book

36

Symmetric Encryption

Encryption of most data is accomplished using fast block and stream ciphers. These are examples of symmetric encryption algorithms. In addition all historical, i.e. pre-1960, ciphers are symmetric in nature and share some design principles with modern ciphers.

The main drawback with symmetric ciphers is that they give rise to a problem of how to distribute the secret keys between users, so we also address this issue.

We also discuss the properties and design of cryptographic hash functions and message authentication codes. Both of which will form basic building blocks of other schemes and protocols within this book.

In the following chapters we explain the theory and practice of modern symmetric ciphers, but first we consider historical ciphers.

37

Stream Cipher (Wikipedia)

- *A stream cipher is a symmetric key cipher where plaintext digits are combined with a pseudorandom cipher digit stream (keystream). In a stream cipher, each plaintext digit is encrypted one at a time with the corresponding digit of the keystream, to give a digit of the ciphertext stream. Since encryption of each digit is dependent on the current state of the cipher, it is also known as state cipher. In practice, a digit is typically a bit and the combining operation an exclusive-or (XOR).*

38

1. Introduction

An encryption algorithm, or cipher, is a means of transforming plaintext into ciphertext under the control of a secret key. This process is called encryption or encipherment. We write

$$c = e_k(m),$$

where

- m is the plaintext,
- e is the cipher function,
- k is the secret key,
- c is the ciphertext.

The reverse process is called decryption or decipherment, and we write

$$m = d_k(c).$$

39

Usually in cryptography the communicating parties are denoted by A and B. However, often one uses the more user-friendly names of Alice and Bob. But you should not assume that the parties are necessarily human, we could be describing a communication being carried out between two autonomous machines. The eavesdropper, bad girl, adversary or attacker is usually given the name Eve.

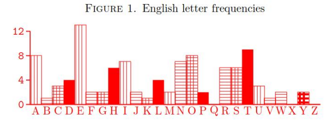
In this chapter we shall present some historical ciphers which were used in the pre-computer age to encrypt data. We shall show that these ciphers are easy to break as soon as one understands the statistics of the underlying language, in our case English. In Chapter 5 we shall study this relationship between how easy the cipher is to break and the statistical distribution of the underlying plaintext.

40

TABLE 1. English letter frequencies

Letter	Percentage	Letter	Percentage
A	8.2	N	6.7
B	1.5	O	7.5
C	2.8	P	1.9
D	4.2	Q	0.1
E	12.7	R	6.0
F	2.2	S	6.3
G	2.0	T	9.0
H	6.1	U	2.8
I	7.0	V	1.0
J	0.1	W	2.4
K	0.8	X	0.1
L	4.0	Y	2.0
M	2.4	Z	0.1

FIGURE 1. English letter frequencies



41

TABLE 2. English bigram frequencies

Bigram	Percentage	Bigram	Percentage
TH	3.15	HE	2.51
AN	1.72	IN	1.69
ER	1.54	RE	1.48
ES	1.45	ON	1.45
EA	1.31	TI	1.28
AT	1.24	ST	1.21
EN	1.20	ND	1.18

42

Trigrams

The most common trigrams are, in decreasing order,
THE, ING, AND, HER, ERE, ENT, THA, NTH, WAS, ETH, FOR.

43

0	A
1	B
2	C
3	D
4	E
5	F
6	G
7	H
8	I
9	J
10	K
11	L
12	M
13	N
14	O
15	P
16	Q
17	R
18	S
19	T
20	U
21	V
22	W
23	X
24	Y
25	Z

2. Shift Cipher

We first present one of the earliest ciphers, called the shift cipher. Encryption is performed by replacing each letter by the letter a certain number of places on in the alphabet. So for example if the key was three, then the plaintext A would be replaced by the ciphertext D, the letter B would be replaced by E and so on. The plaintext word HELLO would be encrypted as the ciphertext KHOOR. When this cipher is used with the key three, it is often called the Caesar cipher, although in many books the name Caesar cipher is sometimes given to the shift cipher with any key. Strictly this is not correct since we only have evidence that Julius Caesar used the cipher with the key three.

There is a more mathematical explanation of the shift cipher which will be instructive for future discussions. First we need to identify each letter of the alphabet with a number. It is usual to identify the letter A with the number 0, the letter B with number 1, the letter C with the number 2 and so on until we identify the letter Z with the number 25. After we convert our plaintext message into a sequence of numbers, the ciphertext in the shift cipher is obtained by adding to each number the secret key k modulo 26, where the key is a number in the range 0 to 25. In this way we can interpret the shift cipher as a *stream cipher*, with key stream given by the repeating sequence

$$k, k, k, k, k, k, \ldots$$

44

This key stream is not very random, which results in it being easy to break the shift cipher. A naive way of breaking the shift cipher is to simply try each of the possible keys in turn, until the correct one is found. There are only 26 possible keys so the time for this exhaustive key search is very small, particularly if it is easy to recognize the underlying plaintext when it is decrypted.

We shall show how to break the shift cipher by using the statistics of the underlying language. Whilst this is not strictly necessary for breaking this cipher, later we shall see a cipher that is made up of a number of shift ciphers applied in turn and then the following statistical technique will be useful. Using a statistical technique on the shift cipher is also instructive as to how statistics of the underlying plaintext can arise in the resulting ciphertext.

Take the following example ciphertext, which since it is public knowledge we represent in blue.

45

GB OR, BE ABG GB OR: GUNG VF GUR DHRFGVBA:
JURGURE 'GVF ABOYRE VA GUR ZVAQ GB FHSSRE
GUR FYVATF NAQ NEEBJF BS BHGENTRBHF SBEGHAR,
BE GB GNXR NEZF NTNVAFG $\boxed{N}$FRN BS GEBHOYRF,
NAQ OL BCCBFVAT RAQ GURZ? GB QVR: GB FYRRC;
AB ZBER; NAQ OL $\boxed{N}$FYRRC GB FNL JR RAQ
GUR URNEG-NPUR NAQ GUR GUBHFNAQ ANGHENY FUBPXF
GUNG SYRFU VF URVE GB, 'GVF N PBAFHZZNGVBA
QRIBHGYL GB OR JVFU'Q. GB QVR, GB FYRRC;
GB FYRRC: CREPUNAPR GB QERNZ: NL, GURER'F GUR EHO;
SBE VA GUNG FYRRC BS QRNGU JUNG QERNZF ZNL PBZR
JURA JR UNIR FUHSSYRQ BSS GUVF ZBEGNY PBVY,
ZHFG TVIR HF CNHFR: GURER'F GUR ERFCRPG
GUNG ZNXRF PNYNZVGL BS FB YBAT YVSR;

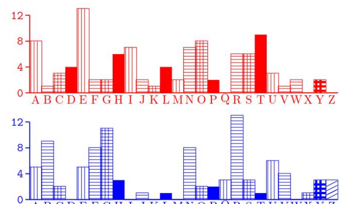
N must be A (shift 13) or I (shift 5)

What simplifies cracking this code?

0	A
1	B
2	C
3	D
4	E
5	F
6	G
7	H
8	I
9	J
10	K
11	L
12	M
13	N
14	O
15	P
16	Q
17	R
18	S
19	T
20	U
21	V
22	W
23	X
24	Y
25	Z

46

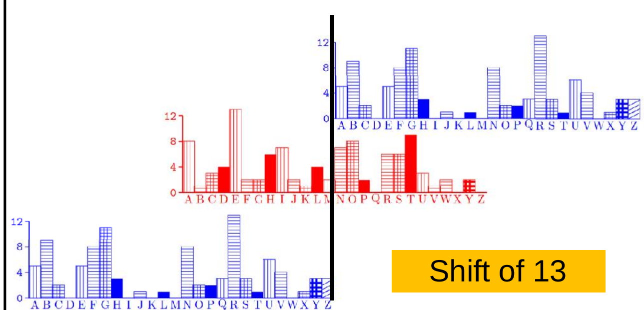
FIGURE 2. Comparison of plaintext and ciphertext frequencies for the shift cipher example



By comparing the two bar graphs in Fig. 2 we can see by how much we think the blue graph has been shifted compared with the red graph. By examining where we think the plaintext letter E may have been shifted, one can hazard a guess that it is shifted by one of
2, 9, 13 or 23.
Then by trying to deduce by how much the plaintext letter A has been shifted we can guess that it has been shifted by one of
1, 6, 13 or 17.
The only shift value which is consistent appears to be the value 13, and we conclude that this is the most likely key value. We can now decrypt the ciphertext, using this key. This reveals, that

0	A
1	B
2	C
3	D
4	E
5	F
6	G
7	H
8	I
9	J
10	K
11	L
12	M
13	N
14	O
15	P
16	Q
17	R
18	S
19	T
20	U
21	V
22	W
23	X
24	Y
25	Z

47



Shift of 13

48

To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them? To die: to sleep;
No more; and by a sleep to say we end
The heart-ache and the thousand natural shocks
That flesh is heir to, 'tis a consummation
Devoutly to be wish'd. To die, to sleep;
To sleep: perchance to dream: ay, there's the rub;
For in that sleep of death what dreams may come
When we have shuffled off this mortal coil,
Must give us pause: there's the respect
That makes calamity of so long life;

49

Tobe,ornottobe:thatisthequestion:
Whether'tisnoblerinthemindtosuffer
Theslingsandarrowsofoutrageousfortune,
Ortotakearmsagainstaseaoftroubles,
Andbyopposingendthem?Todie:tosleep;
Nomore;andbyasleeptosayweend
Theheart-acheandthethousandnaturalshocks
Thatfleshisheirto,'tisaconsummation
Devoutlytobewish'd.Todie,tosleep;
Tosleep:perchancetodream:ay,there'stherub;
Forinthatsleepofdeathwhatdreamsmaycome
Whenwehaveshuffledoffthismortalcoil,
Mustgiveuspause:there'stherespect
Thatmakescalamityofsolonglife;

50

Soyoumayaskifmodernciphersencryptplaintextswithnoredundancy?Theanswerisno,evenifonecompressesthedata,amoderncipheroftenaddssomeredundancytotheplaintextbeforeencryption.Thereasonisthatwehaveonlyconsideredpassiveattacks,i.e.anattackerhasbeenonlyallowedtoexamineciphertextsandfromtheseciphertextstheattacker'sgoalistodeterminethekey.Thereareothertypesofattackcalledactiveattacks,intheseanattackerisallowedtogenerateplaintextsorciphertextsofherchoosingandaskthekeyholdertoencryptordecryptthem,thetwovariantsbeingcalledachosenplaintextattackandachosenciphertextattackrespectively.Inpublickeysystemsthatweshallseelater,chosenplaintextsattackscannotbestoppedsinceanyoneisallowedtoencryptanything.Wewouldhowever,liketostopchosenciphertextattacks.Thecurrentwisdomforpublickeyalgorithmsistomakethecipheraddsomeredundancytotheplaintextbeforeitisencrypted.Inthatwayitishardforanattackertoproduceaciphertextwhichhasavaliddecryption.Thephilosophyisthatitisthenhardforanattackertomountachosenciphertextattack,sinceitwillbehardforanattackertochooseavalidciphertextforadecryptionquery.Weshalldiscussthismoreinlaterchapters.

51

3. Substitution Cipher

The main problem with the shift cipher is that the number of keys is too small, we only have 26 possible keys. To increase the number of keys a *substitution cipher* was invented. To write down a key for the substitution cipher we first write down the alphabet, and then a permutation of the alphabet directly below it. This mapping gives the substitution we make between the plaintext and the ciphertext

Plaintext alphabet	ABCDEFGHIJKLMNOPQRSTUVWXYZ
Ciphertext alphabet	GOYDSIPELUAVCRJWXZNHBQFTMK

Encryption involves replacing each letter in the top row by its value in the bottom row. Decryption involves first looking for the letter in the bottom row and then seeing which letter in the top row maps to it. Hence, the plaintext word HELLO would encrypt to the ciphertext ESVVJ if we used the substitution given above.

The number of possible keys is equal to the total number of permutations on 26 letters, namely the size of the group S_{26}, which is

$$26! \approx 4.03 \cdot 10^{26} \approx 2^{88}.$$

Since, as a rule of thumb, it is feasible to only run a computer on a problem which takes under 2^{80} steps we can deduce that this large key space is far too large to enable a brute force search even using a modern computer. Still we can break substitution ciphers using statistics of the underlying plaintext language, just as we did for the shift cipher.

52

XSO MJIWXVL JODIVA STW VAO VY OZJVCO'W LTJDOWX KVAKOAXJTXIVAW VY SIDS XOKSAVLVDQ IAGZWXJQ. KVUCZXOJW, KVUUZAIKTXIVAW TAG UIKJVOLOKXJVAIKW TJO HOLL JOCJOWOAXOG, TLVADWIGO GIDIXTL UOGIT, KVUCZXOJ DTUOW TAG OLOKXJVAIK KVUUOJKO. TW HOLL TW SVWXIAD UTAQ JOWOTJKS TAG CJVGZKX GONOLVCUOAX KOAXJOW VY UTPVJ DLVMTL KVUCTAIOW, XSO JODIVA STW T JTCIGLQ DJVHIAD AZUMOJ VY IAAVNTXINO AOH KVUCTAIOW. XSO KVUCZXOJ WKIOAKO GOCTJXUOAX STW KLVWO JOLTXIVAWSICW HIXS UTAQ VY XSOWO VJDTAIWTXIVAW NIT KVLLTMVJTXINO CJVPOKXW, WXTYY WOKVAGUOAXW TAG NIWIXIAD IAGZWXJITL WXTYY. IX STW JOKOAXLQ IAXJVGZKOG WONOJTL UOKSTAIWUW YVJ GONOLVCIAD TAG WZCCVJXIAD OAXJOCJOAOZJITL WXZGOAXW TAG WXTYY, TAG TIUW XV CLTQ T WIDAIYIKTAX JVLO IA XSO GONOLVCUOAX VY SIDS-XOKSAVLVDQ IAGZWXJQ IA XSO JODIVA.

XSO GOCTJXUOAX STW T LTJDO CJVDJTUUO VY JOWOTJKS WZCCVJXOG MQ IAGZWXJQ, XSO OZJVCOTA ZAIVA, TAG ZE DVNOJAUOAX JOWOTJKS OWXTMLIWSUOAXW TAG CZMLIK KVJCVJTXIVAW. T EOQ OLOUOAX VY XSIW IW XSO WXJVAD LIAEW XSTX XSO GOCTJXUOAX STW HIXS XSO KVUCZXOJ, KVUUZAIKTXIVAW, UIKJVOLOKXJVAIKW TAG UOGIT IAGZWXJIOW IA XSO MJIWXVL JODIVA . XSO TKTGOUIK JOWOTJKS CJVDJTUUO IW VJDTAIWOG IAXV WONOA DJVZCW, LTADZTDOW TAG TJKSIXOKXZJO, GIDIXTL UOGIT, UVMILO TAG HOTJTMLO KVUCZXIAD, UTKSIAO LOTJAIAD, RZTAXZU KVUCZXIAD, WQWXOU NOJIYIKTXIVA, TAG KJQCXVDJTCSQ TAG IAYVJUTXIVA WOKZJIXQ.

53

We can compute the following frequencies for single letters in the above ciphertext:

Letter	Freq	Letter	Freq	Letter	Freq
A	8.6995	B	0.0000	C	3.0493
D	3.1390	E	0.2690	F	0.0000
G	3.6771	H	0.6278	I	7.8923
J	7.0852	K	4.6636	L	3.5874
M	0.8968	N	1.0762	O	11.479
P	0.1793	Q	1.3452	R	0.0896
S	3.5874	T	8.0717	U	4.1255
V	7.2645	W	6.6367	X	8.0717
Y	1.6143	Z	2.7802		

In addition we determine that the most common bigrams in this piece of ciphertext are

TA, AX, IA, VA, WX, XS, AG, OA, JO, JV,

whilst the most common trigrams are

OAX, TAG, IVA, XSO, KVU, TXI, UOA, AXS.

Since the ciphertext letter O occurs with the greatest frequency, namely 11.479, we can guess that the ciphertext letter O corresponds to the plaintext letter E. We now look at what this means for two of the common trigrams found in the ciphertext

- The ciphertext trigram OAX corresponds to E * *.
- The ciphertext trigram XSO corresponds to * * E.

54

We examine similar common similar trigrams in English, which start or end with the letter E. We find that three common ones are given by ENT, ETH and THE. Since the two trigrams we wish to match have one starting with the same letter as the other finishes with, we can conclude that it is highly likely that we have the correspondence

- X = T,
- S = H,
- A = N.

Even after this small piece of analysis we find that it is much easier to understand what the underlying plaintext should be. If we focus on the first two sentences of the ciphertext we are trying to break, and we change the letters which we think we have found the correct mappings for, then we obtain:

THE MJIWTVL JEDIVN HTW VNE VY EZJVCE'W LTJDEWT
KVNKENTJTTIV NW VY HIDH TEKHNVLVDQ INGZWTJQ.
KVUCZTEJW, KVUUZNIKTTIVNW TNG UIKJVELEKTJVNIKW
TJE HELL JECJEWENTEG, TLVNDWIGE GIDITTL UEGIT,
KVUCZTEJ DTUEW TNG ELEKTJVNIK KVUUEJKE.

55

Recall, this was after the four substitutions

O = E, X = T, S = H, A = N.

We now cheat and use the fact that we have retained the word sizes in the ciphertext. We see that since the letter T occurs as a single ciphertext letter we must have

T = I or T = A.

The ciphertext letter T occurs with a probability of 8.0717, which is the highest probability left, hence we are far more likely to have

T = A.

We have already considered the most popular trigram in the ciphertext so turning our attention to the next most popular trigram we see that it is equal to TAG which we suspect corresponds to the plaintext AN*. Therefore it is highly likely that G = D, since AND is a popular trigram in English.

Our partially decrypted ciphertext is now equal to

THE MJIWTVL JEDIVN HAW VNE VY EZJVCE'W LAJDEWT
KVNKENTJATIV NW VY HIDH TEKHNVLVDQ INDZWTJQ.
KVUCZTEJW, KVUUZNIKATIVNW AND UIKJVELEKTJVNIKW
AJE HELL JECJEWENTED, ALVNDWIDE DIDITAL UEDIA,
KVUCZTEJ DAUEW AND ELEKTJVNIK KVUUEJKE.

This was after the six substitutions

O = E, X = T, S = H,
A = N, T = A, G = D.

56

We now look at two-letter words which occur in the ciphertext:

- IX
  This corresponds to the plaintext *T. Therefore the ciphertext letter I must be one of the plaintext letters A or I, since the only two-letter words in English ending in T are AT and IT. We already have worked out what the plaintext character A corresponds to, hence we must have I = I.
- XV
  This corresponds to the plaintext T*. Hence, we must have V = O.
- VY
  This corresponds to the plaintext O*. Hence, the ciphertext letter Y must correspond to one of F, N or R. We already know the ciphertext letter corresponding to N. In the ciphertext the probability of Y occurring is 1.6, but in English we expect F to occur with probability 2.2 and R to occur with probability 6.0. Hence, it is more likely that Y = F.
- IW
  This corresponds to the plaintext I*. Therefore, the plaintext character W must be one of F, N, S and T. We already have F, N, T, hence W = S.

All these deductions leave the partial ciphertext as

THE MJISTOL JEDION HAS ONE OF EZJOCE'S LAJDEST
KONKENTJATIONS OF HIDH TEKHNOLODQ INDZSTJQ.
KOUCZTEJS, KOUUZNIKATIONS AND UIKJOELEKTJONIKS AJE
HELL JECJESENTED, ALONDSIDE DIDITAL UEDIA,
KOUCZTEJ DAUES AND ELEKTJONIK KOUUEJKE.

Now you can finish this up on your own!

This was after the ten substitutions

O = E, X = T, S = H, A = N, T = A,
G = D, I = I, V = O, Y = F, W = S.

57

4. Vigenère Cipher

The problem with the shift cipher and the substitution cipher was that each plaintext letter always encrypted to the same ciphertext letter. Hence underlying statistics of the language could be used to break the cipher. For example it was easy to determine which ciphertext letter corresponded to the plaintext letter E. From the early 1800s onwards, cipher designers tried to break this link between the plaintext and ciphertext.

The substitution cipher we used above was a mono-alphabetic substitution cipher, in that only one alphabet substitution was used to encrypt the whole alphabet. One way to solve our problem is to take a number of substitution alphabets and then encrypt each letter with a different alphabet. Such a system is called a polyalphabetic substitution cipher.

For example we could take

Plaintext alphabet	`ABCDEFGHIJKLMNOPQRSTUVWXYZ`
Ciphertext alphabet one	`TMKGOYDSIPELUAVCRJWXZNHBQF`
Ciphertext alphabet two	`DCBAHGFEMLKJIZYXWVUTSRQPON`

Then the plaintext letters in an odd position we encrypt using the first ciphertext alphabet, whilst the plaintext letters in even positions we encrypt using the second alphabet. For example the plaintext word HELLO, using the above alphabets would encrypt to SHLJV. Notice that the two occurrences of L in the plaintext encrypt to two different ciphertext characters. Thus we have made it harder to use the underlying statistics of the language. If one now does a naive frequency analysis we no longer get a common ciphertext letter corresponding to the plaintext letter E.

We essentially are encrypting the message two letters at a time, hence we have a block cipher with block length two English characters. In real life one may wish to use around five rather than just two alphabets and the resulting key becomes very large indeed. With five alphabets the total key space is

$$(26!)^5 \approx 2^{441},$$

but the user only needs to remember the key which is a sequence of

$26 \cdot 5 = 130$ letters.

58

However, just to make life hard for the attacker, the number of alphabets in use should also be hidden from his view and form part of the key. But for the average user in the early 1800s this was far too unwieldy a system, since the key was too hard to remember.
Despite its shortcomings the most famous cipher during the 19th-century was based on precisely this principle. The Vigen`ere cipher, invented in 1533 by Giovan Batista Belaso, was a variant on the above theme, but the key was easy to remember. When looked at in one way the Vigen`ere cipher is a polyalphabetic block cipher, but when looked at in another, it is a stream cipher which is a natural generalization of the shift cipher.
The description of the Vigen`ere cipher as a block cipher takes the description of the polyalphabetic cipher above but restricts the possible plaintext alphabets to one of the 26 possible cyclic shifts of the standard alphabet. Suppose five alphabets were used, this reduces the key space down to

$$26^5\ (11{,}881{,}376) \approx 2^{23}\ (8{,}388{,}608) \approx 10^7$$

and the size of the key to be remembered as a sequence of five numbers between 0 and 25.

However, the description of the Vigen`ere cipher as a stream cipher is much more natural. Just like the shift cipher, the Vigen`ere cipher again identifies letters with the numbers 0, . . . , 25. The secret key is a short sequence of letters (e.g. a word) which is repeated again and again to form a keystream. Encryption involves adding the plaintext letter to a key letter. Thus if the key is SESAME, encryption works as follows,

```
THISISATESTMESSAGE
SESAMESESAMESESAME
LLASUWSXWSFQWWKASI
```

Again we notice that A will encrypt to a different letter depending on where it appears in the message.

59

As an example, suppose the ciphertext is given by

UTPDHUG NYH USVKCG MVCE FXL KQIB. WX RKU GI TZN, RLS BBHZLXMSNP
KDKS; CEB IH HKEW IBA, YYM SBR PFR SBS, JV UPL O UVADGR HRRWXF. JV ZTVOOV
YH ZCQU Y UKWGEB, PL UQFB P FOUKCG, TBF RQ VHCF R KPG, OU KFT ZCQU MAW
QKKW ZGSY, FP PGM QKFTK UQFB DER EZRN, MCYE, MG UCTFSVA, WP KFT ZCQU
MAW KQIJS. LCOV NTHDNV JPNUJVB IH GGV RWX ONKCGTHKFL XG VKD, ZJM VG
CCI MVGD JPNUJ, RLS EWVKJT ASGUCS MVGD; DDK VG NYH PWUV CCHIIY RD DBQN
RWTH PFRWBBI VTTK VCGNTGSF FL IAWU XJDUS, HFP VHCF, RR LAWEY QDFS
RVMEES FZB CHH JRTT MVGZP UBZN FD ATIIYRTK WP KFT HIVJCI; TBF BLDPWPX
RWTH ULAW TG VYCHX KQLJS US DCGCW OPPUPR, VG KFDNUJK GI JIKKC PL KGCJ
IAOV KFTR GJFSAW KTZLZES WG RWXWT VWTL WP XPXGG, CJ FPOS VYC BTZCUW
XG ZGJQ PMHTRAIBJG WMGFG. JZQ DPB JVYGM ZCLEWXR: CEB IAOV NYH JIKKC
TGCWXF UHF JZK.
WX VCU LD YITKFTK WPKCGVCWIQT PWVY QEBFKKQ, QNH NZTTW IRFL IAS
VFRPE ODJRXGSPTC EKWPTGEES, GMCG
TTVVPLTFFJ; YCW WV NYH TZYRWH LOKU MU AWO, KFPM VG BLTP VQN RD DSGG
AWKWUKKPL KGCJ, XY OPP KPG ONZTT ICUJCHLSF KFT DBQNJTWUG. DYN MVCK
ZT MFWCW HTWF FD JL, OPU YAE CH LQ! PGR UF, YH MWPP RXF CDJCGOSF, XMS
UZGJQ JL, SXVPN HBG!

60

There is a way of finding the length of the keyword, which is repeated to form the keystream, called the *Kasiski test*. First we need to look for repeated sequences of characters. Recall that English has a large repetition of certain bigrams or trigrams and over a long enough string of text these are likely to match up to the same two or three letters in the key every so often. By examining the distance between two repeated sequences we can guess the length of the keyword. Each of these distances should be a multiple of the keyword, hence taking the greatest common divisor of all distances between the repeated sequences should give a good guess as to the keyword length.

Let us examine the above ciphertext and look for the bigram WX. The gaps between some of the occurrences of this bigram are 9, 21, 66 and 30, some of which may have occurred by chance, whilst some may reveal information about the length of the keyword. We now take the relevant greatest common divisors to find,

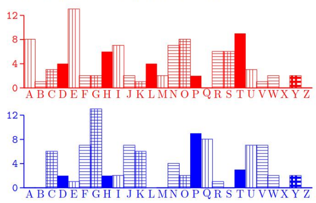
$$\gcd(30,66) = 6,$$
$$\gcd(3,9) = \gcd(9,66) = \gcd(9,30) = \gcd(21,66) = 3.$$

We are unlikely to have a keyword of length three so we conclude that the gaps of 9 and 21 occurred purely by chance. Hence, our best guess for the keyword is that it is of length 6.

61

Now we take every sixth letter and look at the statistics just as we did for a shift cipher to deduce the first letter of the keyword. We can now see the advantage of using the histograms to break the shift cipher earlier. If we used the naive method and tried each of the 26 keys in turn we could still not detect which key is correct, since every sixth letter of an English sentence does not produce an English sentence. Using our earlier histogram based method is more efficient in this case.

FIGURE 3. Comparison of plaintext and ciphertext frequencies for every sixth letter of the Vigenère example, starting with the first letter

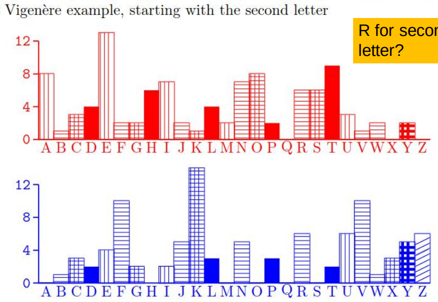


C for first letter?

62

FIGURE 4. Comparison of plaintext and ciphertext frequencies for every sixth letter of the Vigenère example, starting with the second letter

R for second letter?



0 A
1 B
2 C
3 D
4 E
5 F
6 G
7 H
8 I
9 J
10 K
11 L
12 M
13 N
14 O
15 P
16 Q
17 R
18 S
19 T
20 U
21 V
22 W
23 X
24 Y
25 Z

63

Continuing in a similar way for the remaining four letters of the keyword we find the keyword is

CRYPTO.

The underlying plaintext is then found to be:

Scrooge was better than his word. He did it all, and infinitely more; and to Tiny Tim, who did not die, he was a second father. He became as good a friend, as good a master, and as good a man, as the good old city knew, or any other good old city, town, or borough, in the good old world. Some people laughed to see the alteration in him, but he let them laugh, and little heeded them; for he was wise enough to know that nothing ever happened on this globe, for good, at which some people did not have their fill of laughter in the outset; and knowing that such as these would be blind anyway, he thought it quite as well that they should wrinkle up their eyes in grins, as have the malady in less attractive forms. His own heart laughed: and that was quite enough for him.

He had no further intercourse with Spirits, but lived upon the Total Abstinence Principle, ever afterwards; and it was always said of him, that he knew how to keep Christmas well, if any man alive possessed the knowledge. May that be truly said of us, and all of us! And so, as Tiny Tim observed, God bless Us, Every One!

The above text is taken from *A Christmas Carol* by Charles Dickens.

64