The Vigenere Cipher

CMPUT 331

Overview

- The idea of the Vigenere Cipher is to use a different key for each letter of the message.
- Unlike substitution cipher, the Vigenere cipher cannot be easily broken by frequency analysis.
- Invented in 1562, it was called “le chiffre indechiffrable” (“the indecipherable cipher”).
- It was finally broken in 1854 by Charles Babbage, “the father of computers”.

Caesar’s cipher wheel

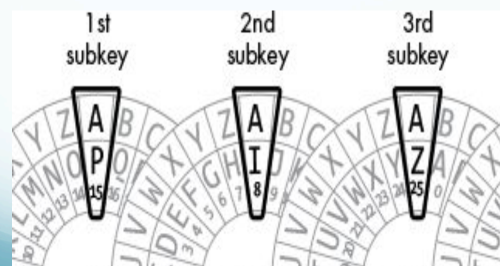
Each ciphertext letter is “the sum” of the plaintext letter and the shift value:

$C_i = M_i + K \bmod 26$

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

I	W	T		C	T	L		E	P	H	H	L	D	G	S		X	H		H	L	D	G	S	U	X	H	W
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
T	H	E		N	E	W		P	A	S	S	W	O	R	D		I	S		S	W	O	R	D	F	I	S	H

Multiple Caesar ciphers



Vigenere Cipher

- The Vigenere cipher is like Caesar cipher, but with multiple keys/shifts.
- The keyword is aligned with the message:

  Message: **thesunandthemoon**

  Key: **KINGKINGKINGKING**

  Cipher: **DPRYEVNTXBUKWWBT**
- Each ciphertext letter is “the sum” of the keyword letter and the plaintext letter:

  $C_i = (M_i + K_i) \bmod 26$

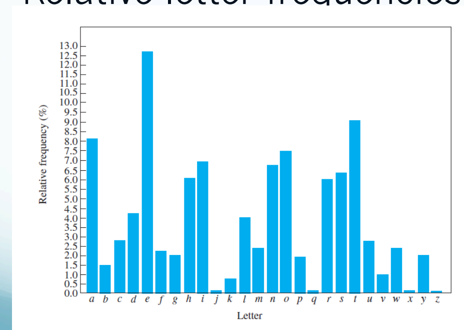
The Vigenere square

	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

Number of possible keys

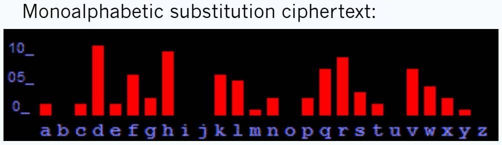
- Key length = K
- Possible keys for Vigenere: 26^K
- K = 10 => $26^{10} > 10^{14}$
- K = 20 => $26^{20} > 10^{28}$
- K = 30 => $26^{30} > 10^{42}$
- For substitution: $26! > 10^{26}$

Relative letter frequencies



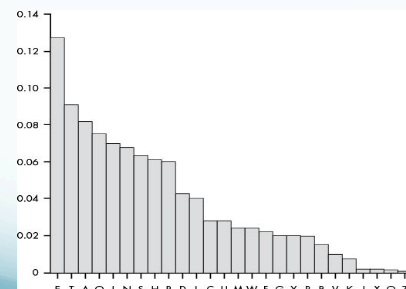
Frequency distributions

Monoalphabetic substitution ciphertext:



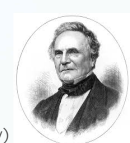
Polyalphabetic Vigenere ciphertext:

Letters by frequency

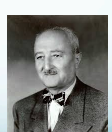


Cracking Vigenere

- Babbage/Kasiski/Sweigart
  - find key length with repeated n-gram offsets
  - for each substring, find shift by calculating the match score (or visually)



- William Friedman
  - find key length with Index of Coincidence
  - for each substring, find shift with Index of Mutual Coincidence (IMC)



Calculating the match score

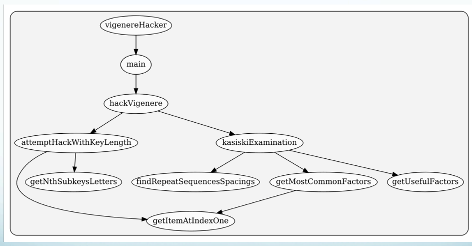
ASRXJILPWMCYOUEQNTHBFZGKVD ← 5 matches

EISNTHAOCLRFDGWVMUYBPZXQJK ← 9 matches

ETAOINSHRDLCUMWFGYPBVKJXQZ

Most frequent | Ignore middle 14 | Least frequent

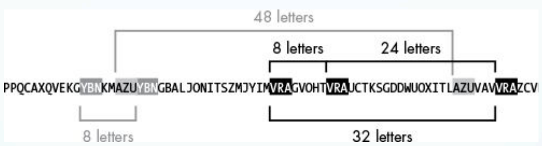
vigenereHacker call graph



Repeated n-grams

- `THEDOGANDTHECAT (plaintext message)`
- `ABCDEFGHIABCDEF (key #1)`
- `TIGGSLGULTIGFEY (ciphertext #1)`
- `XYZXYZXYZXYZXYZ (key #2)`
- `QFDAMFXLCQFDZYS (ciphertext #2)`

Kasiski Examination



Index of Coincidence (IC)

- The probability that two randomly selected letters from a ciphertext will be the same.

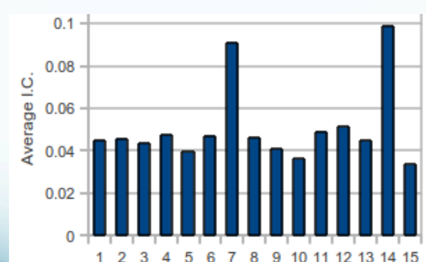
$$IC = \frac{\sum_{i=A}^{i=Z} c_i(c_i - 1)}{N(N-1)}$$

- It measures the "flatness" of the frequency distribution
- The approximate value for plain English text is 0.065
- The approximate value for random text is 0.039
- IC does not change if you apply a substitution cipher!

IC examples

- Let's compute IC on some multi-sets of characters
- IC = (c_1 * (c_1 – 1) + … c_n * (c_n – 1) / (n * (n – 1))
- X = {a,a,a,a,b,b,c,d,d,d,e,e,e,e,e}
- IC(X) = (4*3 + 2*1 + 1*0 + 3*2 + 5*4) / (15*14) = (12 + 2 + 0 + 6 + 20) / 210 ≈ 0.19
- Y = {a,a,a,b,b,b,c,c,c,d,d,d,e,e,e}
- IC(Y) = (3*2) * 5 / (15*14) = 30 / 210 ≈ 0.14
- If n is large, (c_i – 1)/(n – 1) ≈ c_i/n = p_i, so IC ≈ $\sum_{i=1}^{i=N} p_i^2$

Averaging IC for key lengths



Index of Mutual Coincidence

- The probability that two randomly selected letters from two texts *x* and *y* will be the same.

$$IMC = \frac{\sum_{i=A}^{i=Z} c_i^x \cdot c_i^y}{N_x \cdot N_y} = \sum_{i=A}^{i=Z} f_i^x \cdot f_i^y$$

- N_x – length of text *x*
- c_i^x – count of letter *i* in text *x*
- $f_i^x = c_i^x / N_x$ – frequency of letter *i* in text *x*

IMC examples

- Let's compute IC on some multi-sets of characters
- IMC = ($x_1 * y_1 + \ldots x_n * y$) / ($N_x * N_y$)
- X = {a,a,a,a,b,b,c,d,d,d,e,e,e,e,e}
- Y = {a,a,a,a,a,b,b,b,b,c,c,d,e,e,e}
- IMC(X,Y) = (4*5 + 2*4 + 1*2 + 3*1 + 5*3) / (15*15) = (20 + 8 + 2 + 3 + 15) / 225 ≈ 0.21
- IMC(X,X) = 55 / 225 ≈ 0.24
- higher IMC indicates a better distribution match

Caesar shifts via IMC

- Assume language with 5-character alphabet:
- f(a) = .27 f(b) = .13 f(c) = .07 f(d) = .20 f(e) = .33
- Ciphertext X = {a,b,b,b,c,c,c,c,c,d,d,d,d,e,e}
- Which shift gives best IMC?
- e.g. shift a → c: X(2) = {c,d,d,d,e,e,e,e,e,a,a,a,a,b,b}
- IMC values: 0.165, 0.213, **0.244**, 0.213, 0.165
- shift 2 is the best match
- this technique allows us to guess each key letter

Running-Key Cipher

- The weakness of Vigenere is its cyclical nature
  - It's easy to break, once you guess the key length
- What if the key is as long as the message?
  - The key could be a book text or a list of words
- Then, Kasiski's method does not work
  - But it can be broken by exploiting n-gram patterns
  - Because both the plaintext and the key are English
- To break a running-key ciphertext, you can either:
  - alternate partial decryption of the key and plaintext
  - or, use n-gram language models like in Asn 6

The One-Time Pad Cipher

- A Vigenere cipher is unbreakable if the key is:
  - as long as the message ("running key")
  - truly random
  - used only once
- All possible keys/decipherments are equally likely
- It is not practical to use for everyday encryption.

Plaintext	IFYOUWANTTOSURVIVEOUTHEREYOUVEGOTTOKNOW
Key	KCQYZHEPXAUTIQEKXEJMORETZHZTRWWQDYLBTTV
Ciphertext	SHOMTDECQTILCHZSSIXGHYIKDFNNMACEWRZLGHR

Evolution of Shift Ciphers

Each key letter encodes a unique shift:

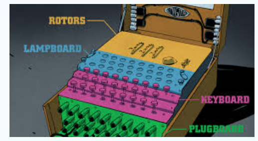
- DDDDDDDDDDDDDDDDDDDD (Caesar Cipher)
- ABCDEFGHIJKLMNOPQRS (Trithemius Cipher)
- SECRETSECRETSECRETS (Vigenere Cipher)
- THEDESIREFORSECRECY (Running-Key Cipher)
- QTXPLKGGUREMPDWXBSR (One-Time Pad Cipher)

Cracking N-Time Pad

- Even if a one-time pad is truly random, reusing it makes it breakable
- It can be broken by using the same techniques as those for breaking a running-key ciphertext
- The method is explained in Appendix G
  - $C_1 = M_1 + K$ and $C_2 = M_2 + K$
  - So, $K = C_1 - M_1 = C_2 - M_2$
  - And, $C_1 - C_2 = M_1 - M_2$
- N-Time Pad can also be viewed as a Vigenere cipher with the key of length $|M| \div N$ (see Chapter 21)

The Enigma Machine

- An encryption device used by the Germans in WWII
- Components:
  - keyboard
  - plugboard (P)
  - rotors (L,M,R)
  - reflector (U)
  - lampboard

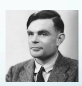


- Encryption formula: $E = PRMLUL^{-1}M^{-1}R^{-1}P^{-1}$
- Number of keys: $(3!)(26^3)(26!/(6!)(10!)(2^{10})) \approx 10^{17}$

Cracking the Enigma

- Polish breakthroughs: Rejewski's *bombe*
- Bletchley Park: "the geese that never cackled"
  - 49 *bombes*
  - predictable messages and keys (*cribs*)
  - cryptic crosswords

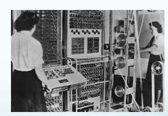
- Alan Turing (1912 – 1954)
  - Turing machine (1936) – model of computer
  - Turing Test of Artificial Intelligence (1950)
  - convicted of homosexuality (1952)



- Anonymous cryptanalysts: no credit in lifetime

Enter Computers

- The German Lorenz cipher was broken with Colossus – the precursor of the modern computer.
- After WWII, cryptographers started using computers for both encoding and cryptanalysis.
- The advantages of computers:
  - speed
  - flexibility
  - binary representation



- Every encipherment algorithm is still a combination of substitution and transposition.