Classical Cryptography

Polyalphabetic cryptanalysis

Karst Koymans

Informatics Institute
University of Amsterdam
(version 22.3, 2023/02/02 11:37:56 UTC)

Friday, February 17, 2023

Table of Contents

The IoC of a polyalphabetic cipher (1)

- We assume that we work with a repeating-key cipher
  - Assume no letters repeat in the key itself
  - Assume the text length is n and the period is p
  - For simplicity suppose p divides n, so $n = p \cdot q$ and $q = \frac{n}{p}$
- Let κ_r be the IoC of random text (≈0.038)
- Let κ_e be the IoC of English plaintext (≈0.066)
- If we split up the cryptogram in p columns
  - then each column of size q is monoalphabetic in itself
  - and letters in different columns seem unrelated

The IoC of a polyalphabetic cipher (2)

So if we pick two different letters from the cryptogram
we expect an index of coincidence of (approximately)

$$\text{IoC} \approx \frac{n(n-q)\kappa_r + n(q-1)\kappa_e}{n(n-1)}$$

or

$$\text{IoC} \approx \frac{n-q}{n-1}\kappa_r + \frac{q-1}{n-1}\kappa_e$$

- For $p = n, q = 1$ this reduces to κ_r (random)
- For $p = 1, q = n$ this reduces to κ_e (monoalphabetic)

Determination of an unknown period (1)

Solving for p and writing κ_i for the IoC we get from the previous estimation

$$p \approx \frac{\kappa_e - \kappa_r}{\kappa_i - \kappa_r + \frac{\kappa_e - \kappa_i}{n}}$$

So if n is large enough this reduces to

$$p \approx \frac{\kappa_e - \kappa_r}{\kappa_i - \kappa_r} \approx \frac{0.028}{\kappa_i - 0.038}$$

Determination of an unknown period (2)

- The **Kasiski test**
- Look for repetitions of groups of letters in the cryptogram
- See how far they are apart and collect these distances
- Probably the repetitions come from a repetition in the plaintext
- In that case the distance d is a multiple of the period p
- A probable p follows from the consideration of all those d's
- **Charles Babbage** (1791 – 1871) probably invented this method years before **Friedrich Kasiski** (1805–1881) did

Babbage

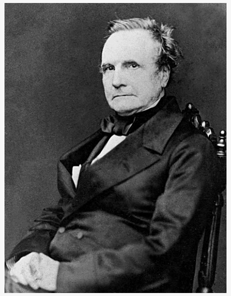


Figure 1: Charles Babbage (1791 – 1871)

Source: `https://en.wikipedia.org/wiki/Charles_Babbage`

The Kasiski method

Adapted from slides by Hans van der Meer

Kasiski method

Until 1863 Vigenère is “le chiffre indéchiffrable”

Then major Friedrich Kasiski publishes
“Die Geheimschriften und die Dechiffrier-kunst”
a method to determine the period

uses repetitions in phase with this period

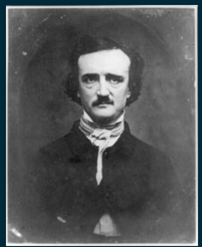
William F. Friedman, Riverbank Publication nr 22, 1920
The Index of Coincidence and its Application in Cryptography

Repetitions

```
pt: EENCURSUSVANHETMATHEMATISCHCENTRUM
 k: STOEIPOESSTOEIPOESSTOEIPOESSTOEIPO   real
ct: WXBGCGGYKNTBLMIAELZXAEBXGGZUXBXZJA

pt: EENCURSUSVANHETMATHEMATISCHCENTRUM
 k: STOEIPOESSTOEIPOESSTOEIPOESSTOEIPO   fake
ct: WXBGCGGYKNTBLMIAELZXAEBXGGZUXBXZJA

pt: EENCURSUSVANHETMATHEMATISCHCENTRUM   coinci
 k: STOEIPOESSTOEIPOESSTOEIPOESSTOEIPO   dental
ct: WXBGCGGYKNTBLMIAELZXAEBXGGZUXBXZJA
```

Kulp message

```
Ge Jeasgdxv,

Zij gl mw, laam, xzy zmlwhfzek
ejlvdxw kwke tx lbr atgh lbmx
aanu bai Vsmukkss pwn vlwk agh
gnumk wdlnzweg jnbxvv oaeg enwb
zwmgy mo mlw wnbx mw al pnfdcfpkh
wzkex hssf xkiyahul. Mk num yexdm
wbxy sbc hv wyx Phwkgnamcuk?
```

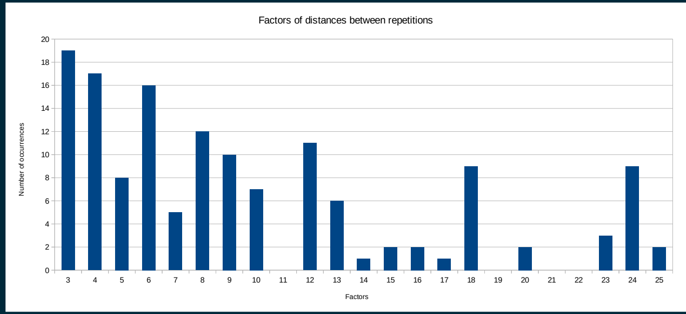


1839 from Kulp, Lewiston, Pennsylvania, USA
to Edgar Allen Poe, ed. Alexander’s Weekly Messenger

Note: jeasgdxv should be ieiasgdxv

Kasiski analysis

```
zij gl mw, laam, xzy zmlwhfzek ejlvdxw kwke tx
lbr atgh lbmx aanu bai vsmukkss pwn vlwk agh
gnumk wdlnzweg jnbxvv oaeg enwb zwmgy mo mlw
wnbx mw al pnfdcfpkh wzkex hssf xkiyahul mk
num yexdm wbxy sbc hv wyx phwkgnamcuk
```



3 letters = THE ?

zij gl mw, laam, xzy zmlwhfzek ejlvdxw kwke tx
lbr atgh lbmx aanu bai vsmukkss pwn vlwk agh
gnumk wdlnzweg jnbxvv oaeg enwb zwmgy mo mlw
wnbx mw al pnfdcfpkh wzkex hssf xkiyahul mk
num yexdm wbxy sbc hv wyx phwkgnamcuk

XYZ = the → key letters

Position on period 12

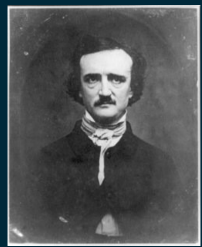
I	J										Z	ZIJ
Y										X	Z	XZY
B	R										L	LBR
		B	A	I								BAI
	P	W	N									PWN
								A	G	H		AGH
									M	L	W	MLW
N	U	M										NUM
S	B	C										SBC
					W	Y	X					WYX

Key letters

B	F										G	ZIJ
U										E	S	XZY
U	N										S	LBR
		I	T	E								BAI
	W	P	J									PWN
								H	Z	D		AGH
									T	E	S	MLW
U	N	I										NUM
Z	U	Y										SBC
					D	R	T					WYX
U	N	I	T	E	D	R	T	A	T	E	S	

Note: The R in DRT should be DST and is one of the many mistakes in the cryptogram

Kulp message decoded



Mr Alexander,
how ys it, that, the messenger
arrives here at the sace time
with the Saturgay cou rier and
other satuzdao paters when
avco rdidg to the cate it is
publishrd three days previous. Is
the fault witg you or tge
Possmastyrs?

Note the many mistakes (introduced by the editor?)

Determination of an unknown period (3)

- The κ **test**
- Friedman's original application of the theory of coincidence
- This time we look at **two** texts
  - that we compare **character by character**
- We expect coincidences κ_r and κ_e for respectively two random and two English texts
- The trick is to compare some cryptogram with a **displaced** (shifted, slid) copy of **itself**
- If the displacement is a multiple of the period coincidences rise

Superimposition

- Knowing the period we can **superimpose** (Dutch: "in diepte leggen") the cryptogram
- Each column is monoalphabetic
- This makes cryptanalysis easy if the cipher is based for instance on a Vigenère with plain alphabet
- Each monoalphabet is then additive and we need only one letter for each column to determine it
- Simple letter frequency counts usually suffice

Repeating-key framework for compositions

- Repeating-key polyalphabetic ciphers
- Each monoalphabetic cipher is either
  - Additive
    - So this is a standard Vigenère
  - Affine
    - The first cipher alphabet is mixed up by a decimation

Keywords of the same length

- Composition gives a similar cipher
- The combined keyword length stays the same
  - Composition of additives stays additive
    - The keyword is the addition of keywords
    - Which makes it somewhat harder-to-guess
  - Composition of affines stays affine
    - The keyword is a linear combination of keywords
    - Also the decimation changes
    - Can you find out the exact formulas?

Keywords of different lengths

- Let the length of the keywords K and L be a and b respectively
- Let lcm(a,b) be the least common multiple of a and b
- Let $a' = \text{lcm}(a, b)/b$ and $b' = \text{lcm}(a, b)/a$
- Reduce this situation to keywords of the same length
  - Consider keywords KK...K (b' times) and LL...L (a' times)
  - This results in two keywords of equal length lcm(a,b)