Cryptanalysis of Classical Ciphers

Cryptology (1)

Martin Stanek

2025

KI FMFI UK Bratislava

- cryptoanalysis of some classical ciphers
  - simple substitution cipher, Vigenère cipher
- relevant concepts
  - entropy, redundancy, unicity distance, etc.
- almost any reasonable approach works
  - most of them can be automated
  - see dCode.fr for other ciphers and cryptanalysis automation

Simple substitution cipher (SSC)

- alphabet $A = \{\mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}\}$
- plaintexts and ciphertexts: $P = C = A$
- keys: $K = \{\pi \mid \pi \text{ is a permutation on } A\}$
- encryption: $E_\pi(p) = \pi(p)$
- decryption: $D_\pi(c) = \pi^{-1}(c)$
- $|K| = 26! \approx 2^{88.38} \Rightarrow$ brute force attack is infeasible

Example (spaces preserved):

$$\pi = \frac{\texttt{ABCDEFGHIJKLMNOPQRSTUVWXYZ}}{\texttt{GIHCOPMWKFXLARYZBVQTJNDEUS}}$$

```
HELL IS THE IMPOSSIBILITY OF REASON
                 ↓
WOLL KQ TWO KAZYQQKIKLKTU YP VOGQYR
```

- COA (Ciphertext only attack)
  - only the ciphertext is known
  - not even a partial knowledge of plaintext
  - attacks in other scenarios are trivial
- observation: SSC does not change the frequencies of letters, it renames them
- assumptions:
  - sufficiently long ciphertext
  - probabilities of plaintext letters (individual and conditional) are non-uniform

Non-uniformity of natural languages

- natural language is redundant
  - not all characters are required to reconstruct or understand the original text
  - example (every 5th character removed):
    I sa we ake ff ad uke th sit frm orbt. Is the ony wa to be sue.
- similar observations can be made for other types of plaintext (alphabet is different)
  - source code, audio and video files, images, etc.
- English texts:
  - letters: E (12.7%), T (9.1%), A (8.2%), O (7.5%), I (7.0%), …
  - bigrams: TH, HE, IN, ER, AN, …
  - the first letter of a word: T, A, O, I, S, …

```
GIYU YUTKJL YU K CNOQ UYJRMTKO BJN. YG HBJUYUGU BW TYGGTN NTUN GIKJ GIN UNK UKJL, KJL YU KXBMG
GIONN AYTNU TBJR. YGU XONKLGI KG JB FBYJG NZHNNLU K PMKOGNO BW K AYTN. YG YU UNFKOKGNL WOBA GIN
AKYJ TKJL XQ K UHKOHNTQ FNOHNFGYXTN HONNE, BBSYJR YGU VKQ GIOBMRI K VYTLNOJNUU BW ONNLU KJL UTYAN,
K WKCBOYGN ONUBOG BW GIN AKOUI INJ. GIN CNRNGKGYBJ, KU AYRIG XN UMFFBUNL, YU UHKJG, BO KG TNKUG
LVKOWYUI. JB GONNU BW KJQ AKRJYGMLN KON GB XN UNNJ. JNKO GIN VNUGNOJ NZGONAYGQ, VINON WBOG
ABMTGOYN UGKJLU, KJL VINON KON UBAN AYUNOKXTN WOKAN XMYTLYJRU, GNJKJGNL, LMOYJR UMAANO, XQ GIN
WMRYGYCNU WOBA HIKOTNUGBJ LMUG KJL WNCNO, AKQ XN WBMJL, YJLNNL, GIN XOYUGTQ FKTANGGB; XMG GIN
VIBTN YUTKJL, VYGI GIN NZHNFGYBJ BW GIYU VNUGNOJ FBYJG, KJL K TYJN BW IKOL, VIYGN XNKHI BJ GIN
UNKHBKUG, YU HBCNONL VYGI K LNJUN MJLNOROBVGI BW GIN UVNNG AQOGTN, UB AMHI FOYSNL XQ GIN
IBOGYHMTGMOYUGU BW NJRTKJL. GIN UIOMX INON BWGNJ KGGKYJU GIN INYRIG BW WYWGNNJ BO GVNJGQ WNNG, KJL
WBOAU KJ KTABUG YAFNJNGOKXTN HBFFYHN, XMOGINJYJR GIN KYO VYGI YGU WOKROKJHN.
```

- top 5 letters: N (14.44%), G (10.53%), K (7.71%), U (7.34%), Y (6.98%)
- a “brave” guess: `N` → `e`, `G` → `t`, `K` → `a`

Cryptanalysis example (2)

```
tIYU YUTaJL YU a CeOQ UYJRMTaO BJe. Yt HBJUYUtU BW TYttTe eTUe tIaJ tIe Uea UaJL, aJL YU aXBMt
tIOee AYTeU TBJR. YtU XOeaLtI at JB FBYJt eZHeeLU a PMaOteO BW a AYTe. Yt YU UeFaOateL WOBA tIe
AaYJ TaJL XQ a UHaOHeTQ FeOHeFtYXTe HOeeE, BBSYJR YtU VaQ tIOBMRI a VYTLeOJeUU BW OeeLU aJL UTYAe,
a WaCBOYte OeUBOt BW tIe AaOUI IeJ. tIe CeRetatYBJ, aU AYRIt Xe UMFFBUeL, YU UHaJt, BO at TeaUt
LVaOWYUI. JB tOeeU BW aJQ AaRJYtMLe aOe tB Xe UeeJ. JeaO tIe VeUteOJ eZtOeAYtQ, VIeOe WBOt
ABMTtOYe UtaJLU, aJL VIeOe aOe UBAe AYUeOaXTe WOaAe XMYTLYJRU, teJaJteL, LMOYJR UMAAeO, XQ tIe
WMRYtYCeU WOBA HIaOTeUtBJ LMUt aJL WeCeO, AaQ Xe WBMJL, YJLeeL, tIe XOYUtTQ FaTAettB; XMt tIe
VIBTe YUTaJL, VYtI tIe eZHeFtYBJ BW tIYU VeUteOJ FBYJt, aJL a TYJe BW IaOL, VIYte XeaHI BJ tIe
UeaHBaUt, YU HBCeOeL VYtI a LeJUe MJLeOROBVtI BW tIe UVeet AQOtTe, UB AMHI FOYSeL XQ tIe
IBOtYHMTtMOYUtU BW eJRTaJL. tIe UIOMX IeOe BWteJ attaYJU tIe IeYRIt BW WYWteeJ BO tVeJtQ Weet, aJL
WBOAU aJ aTABUt YAFeJetOaXTe HBFFYHe, XMOtIeJYJR tIe aYO VYtI YtU WOaROaJHe.
```

– multiple tIe words: I → h

Cryptanalysis example (3)

thYU YUTaJL YU a CeOQ UYJRMTaO BJe. Yt HBJUYUtU BW TYttTe eTUe thaJ the Uea UaJL, aJL YU aXBMt
thOee AYTeU TBJR. YtU XOeaLth at JB FBYJt eZHeeLU a PMaOteO BW a AYTe. Yt YU UeFaOateL WOBA the
AaYJ TaJL XQ a UHaOHeTQ FeOHeFtYXTe HOeeE, BBSYJR YtU VaQ thOBMRh a VYTLeOJeUU BW OeeLU aJL UTYAe,
a WaCBOYte OeUBOt BW the AaOUh heJ. the CeRetatYBJ, aU AYRht Xe UMFFBUeL, YU UHaJt, BO at TeaUt
LVaOWYUh. JB tOeeU BW aJQ AaRJYtMLe aOe tB Xe UeeJ. JeaO the VeUteOJ eZtOeAYtQ, VheOe WBOt
ABMTtOYe UtaJLU, aJL VheOe aOe UBAe AYUeOaXTe WOaAe XMYTLYJRU, teJaJteL, LMOYJR UMAAeO, XQ the
WMRYtYCeU WOBA HhaOTeUtBJ LMUt aJL WeCeO, AaQ Xe WBMJL, YJLeeL, the XOYUtTQ FaTAettB; XMt the
VhBTe YUTaJL, VYth the eZHeFtYBJ BW thYU VeUteOJ FBYJt, aJL a TYJe BW haOL, VhYte XeaHh BJ the
UeaHBaUt, YU HBCeOeL VYth a LeJUe MJLeOROBVth BW the UVeet AQOtTe, UB AMHh FOYSeL XQ the
hBOtYHMTtMOYUtU BW eJRTaJL. the UhOMX heOe BWteJ attaYJU the heYRht BW WYWteeJ BO tVeJtQ Weet, aJL
WBOAU aJ aTABUt YAFeJetOaXTe HBFFYHe, XMOtheJYJR the aYO VYth YtU WOaROaJHe.

- thaJ: J → n

Cryptanalysis example (4)

```
thYU YUTanL YU a CeOQ UYnRMTaO Bne. Yt HBnUYUtU BW TYttTe eTUe than the Uea UanL, anL YU aXBMt
thOee AYTeU TBnR. YtU XOeaLth at nB FBYnt eZHeeLU a PMaOteO BW a AYTe. Yt YU UeFaOateL WOBA the
AaYn TanL XQ a UHaOHeTQ FeOHeFtYXTe HOeeE, BBSYnR YtU VaQ thOBMRh a VYTLeOneUU BW OeeLU anL UTYAe,
a WaCBOYte OeUBOt BW the AaOUh hen. the CeRetatYBn, aU AYRht Xe UMFFBUeL, YU UHant, BO at TeaUt
LVaOWYUh. nB tOeeU BW anQ AaRnYtMLe aOe tB Xe Ueen. neaO the VeUteOn eZtOeAYtQ, VheOe WBOt
ABMTtOYe UtanLU, anL VheOe aOe UBAe AYUeOaXTe WOaAe XMYTLYnRU, tenanteL, LMOYnR UMAAeO, XQ the
WMRYtYCeU WOBA HhaOTeUtBn LMUt anL WeCeO, AaQ Xe WBMnL, YnLeeL, the XOYUtTQ FaTAettB; XMt the
VhBTe YUTanL, VYth the eZHeFtYBn BW thYU VeUteOn FBYnt, anL a TYne BW haOL, VhYte XeaHh Bn the
UeaHBaUt, YU HBCeOeL VYth a LenUe MnLeOROBVth BW the UVeet AQOtTe, UB AMHh FOYSeL XQ the
hBOtYHMTtMOYUtU BW enRTanL. the UhOMX heOe BWten attaYnU the heYRht BW WYWteen BO tVentQ Weet, anL
WBOAU an aTABUt YAFenetOaXTe HBFFYHe, XMOthenYnR the aYO VYth YtU WOaROanHe.
```

- multiple anL words: L → d
- thOee: O → r

thYU YUTand YU a CerQ UYnRMTar Bne. Yt HBnUYUtU BW TYttTe eTUe than the Uea Uand, and YU aXBMt
three AYTeU TBnR. YtU Xreadth at nB FBYnt eZHeedU a PMarter BW a AYTe. Yt YU UeFarated WrBA the
AaYn Tand XQ a UHarHeTQ FerHeFtYXTe HreeE, BBSYnR YtU VaQ thrBMRh a VYTderneUU BW reedU and UTYAe,
a WaCBrYte reUBrt BW the AarUh hen. the CeRetatYBn, aU AYRht Xe UMFFBUed, YU UHant, Br at TeaUt
dVarWYUh. nB treeU BW anQ AaRnYtMde are tB Xe Ueen. near the VeUtern eZtreAYtQ, Vhere WBrt
ABMTtrYe UtandU, and Vhere are UBAe AYUeraXTe WraAe XMYTdYnRU, tenanted, dMrYnR UMAAer, XQ the
WMRYtYCeU WrBA HharTeUtBn dMUt and WeCer, AaQ Xe WBMnd, Yndeed, the XrYUtTQ FaTAettB; XMt the
VhBTe YUTand, VYth the eZHeFtYBn BW thYU VeUtern FBYnt, and a TYne BW hard, VhYte XeaHh Bn the
UeaHBaUt, YU HBCered VYth a denUe MnderRrBVth BW the UVeet AQrtTe, UB AMHh FrYSed XQ the
hBrtYHMTtMrYUtU BW enRTand. the UhrMX here BWten attaYnU the heYRht BW WYWteen Br tVentQ Weet, and
WBrAU an aTABUt YAFenetraXTe HBFFYHe, XMrthenYnR the aYr VYth YtU WraRranHe.

– the rest is easy, continue with decryption: Yndeed, thYU, multiple YU, …

Plaintext:

THIS ISLAND IS A VERY SINGULAR ONE. IT CONSISTS OF LITTLE ELSE THAN THE SEA SAND, AND IS ABOUT THREE MILES LONG. ITS BREADTH AT NO POINT EXCEEDS A QUARTER OF A MILE. IT IS SEPARATED FROM THE MAIN LAND BY A SCARCELY PERCEPTIBLE CREEK, OOZING ITS WAY THROUGH A WILDERNESS OF REEDS AND SLIME, A FAVORITE RESORT OF THE MARSH HEN. THE VEGETATION, AS MIGHT BE SUPPOSED, IS SCANT, OR AT LEAST DWARFISH. NO TREES OF ANY MAGNITUDE ARE TO BE SEEN. NEAR THE WESTERN EXTREMITY, WHERE FORT MOULTRIE STANDS, AND WHERE ARE SOME MISERABLE FRAME BUILDINGS, TENANTED, DURING SUMMER, BY THE FUGITIVES FROM CHARLESTON DUST AND FEVER, MAY BE FOUND, INDEED, THE BRISTLY PALMETTO; BUT THE WHOLE ISLAND, WITH THE EXCEPTION OF THIS WESTERN POINT, AND A LINE OF HARD, WHITE BEACH ON THE SEACOAST, IS COVERED WITH A DENSE UNDERGROWTH OF THE SWEET MYRTLE, SO MUCH PRIZED BY THE HORTICULTURISTS OF ENGLAND. THE SHRUB HERE OFTEN ATTAINS THE HEIGHT OF FIFTEEN OR TWENTY FEET, AND FORMS AN ALMOST IMPENETRABLE COPPICE, BURTHENING THE AIR WITH ITS FRAGRANCE.

Remarks on cryptanalysis of SSC

- several possibilities to improve and automate the cryptanalysis
- language dictionary
  - searching for patterns (known, duplicate and distinct characters)
  - in any phase of cryptanalysis
- better statistical models of plaintext lanaguage based on bigrams, trigrams, …
- optimization techniques, for example hill climbing:
  - start with random/default key
  - test all two-letters swaps in the key, and score the resulting decrypted text
  - do the swap that maximizes the score
  - repeat until no further “improving” swaps are possible (example)

When is the plaintext unique?

- ciphertext `QWERT` $\mapsto$ possible plaintexts: `today`, `terms`, `index`, `jacob`, `delay`, …
- intuitively, longer ciphertexts $\Rightarrow$ less freedom for meaningful plaintexts
- let's estimate the expected length of the ciphertext such that only one meaningful plaintext is possible
  - equivalently, when there is only one "meaningful" key?

Entropy

- let X be a discrete random variable (countable number of distinct values)
  - in our case X will be finite: values: $x_1, \dots, x_n$; probabilities $p_1, \dots, p_n$
- Entropy of X (lg denotes $\log_2$):

$$H(X) = -\sum_{i=1}^{n} p_i \cdot \lg p_i$$

- entropy is a expected information (in bits) when given a value of X
  - how many bits are needed to represent X
- examples:
  - coin tossing: $H\left(\frac{1}{2}, \frac{1}{2}\right) = -2 \cdot \frac{1}{2} \cdot \lg \frac{1}{2} = 1$
  - uniformly random letters from English alphabet: $H_{\mathrm{rnd}}\left(\frac{1}{26}, \dots, \frac{1}{26}\right) = -\lg \frac{1}{26} \approx 4.70$

Entropy of a language

- using probability distribution of individual letters:

$$H_{\mathrm{EN},1}(0.127;\ 0.091;\ 0.082;\ \ldots) \approx 4.15 \text{ bits/letter}$$

- we can analyze bigrams, trigrams, … and estimate entropy of n-grams (for $n \to \infty$):

$$\begin{aligned} H_{\mathrm{EN},2}(\text{digrams}) &\approx 3.65 \text{ bits/letter} \\ H_{\mathrm{EN},3}(\text{trigrams}) &\approx 3.22 \text{ bits/letter} \\ &\ldots \\ H_{\mathrm{EN}} &\approx 1.50 \text{ bits/letter} \end{aligned}$$

- other languages can be analyzed similarly

Plaintext redundancy

- let q be an alphabet size
- let Y be a random variable of length n
- redundancy of Y: $D_n = n \lg q - H(Y)$
- how much longer is the plaintext than the bitstring needed for its representation

English (26 letters, sufficiently large n):
$D_n = (4.7 - 1.5) \cdot n = 3.2 \cdot n$

Unicity distance

- scenario: COA
- idea: redundant bits in a plaintext allow to recognize the correct key
- unicity distance of the plaintext: $\min\{n \in \mathbb{N} \mid D_n \geq H(K)\}$

SSC (keys uniformly distributed):
- $H(K) = \lg 26! \approx 88.38$

$$3.2 \cdot n \geq 88.38 \quad \Rightarrow \quad n \geq 28$$

- polyalphabetic substitution
  - key length n
  - concatenation of n independent shift ciphers
- $K = (k_1, \ldots, k_n) = \mathbb{Z}_{26}^n$
- $P = C = \{\mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}\}^n \leftrightarrow \mathbb{Z}_{26}^n$
- $E_k(p_1, \ldots, p_n) = (p_i + k_i \bmod 26)_{i=1,\ldots,n}$
- $D_k(c_1, \ldots, c_n) = (c_i - k_i \bmod 26)_{i=1,\ldots,n}$
- long text encrypted as a sequence of independent blocks

Example:

k = `BLUE`

plaintext: `HOUSTONWEHAVEAPROBLEM`

```
HOUS|TONW|EHAV|EAPR|OBLE|M
BLUE|BLUE|BLUE|BLUE|BLUE|B
          ↓
IZOW|UZHA|FSUZ|FLJV|PMFI|N
```

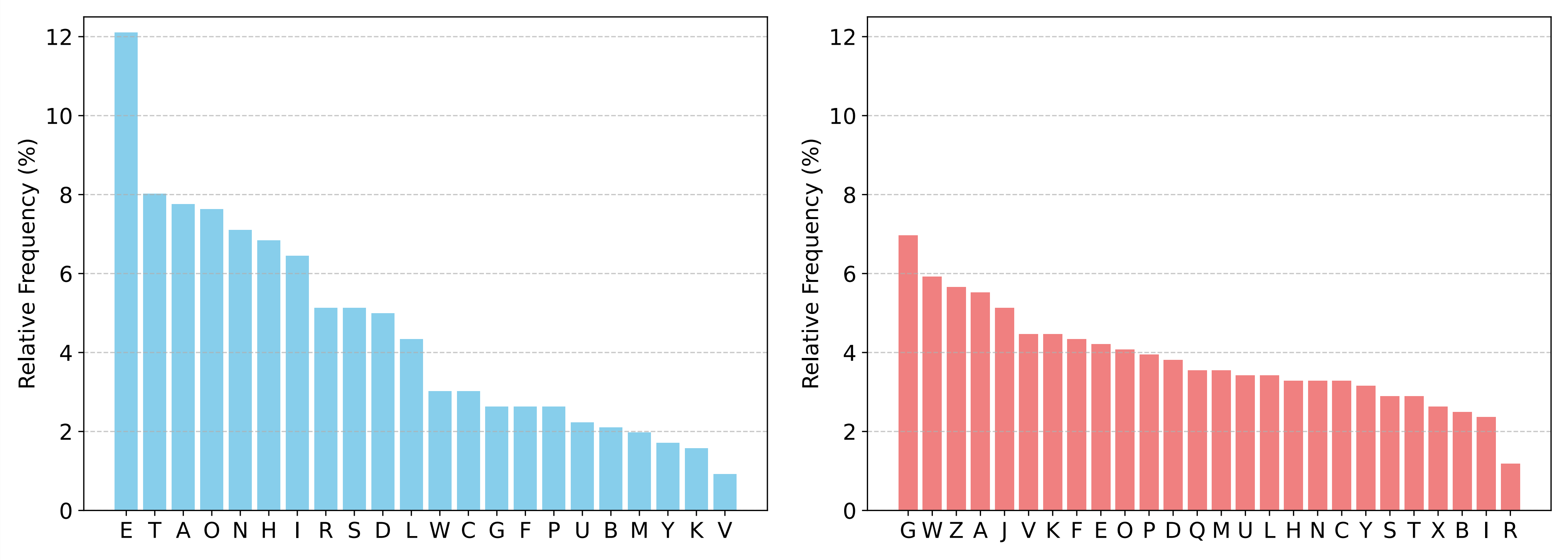
- no more one-to-one correspondence between plaintext and ciphertext letters
- “flat” probability distribution

- we again assume COA scenario
  - otherwise the cryptanalysis is trivial
- the first step: find n (if the value is unknown)
  - ignore: run cryptanalysis for $n = 1, 2, \ldots$ until we succeed
  - Kasiski's method:
    - equal substrings are encrypted equally, if their distance is a multiple of n
    - find multiple pairs of equal substrings, and compute the gcd of their distances
    - substrings should be sufficiently long (for example 4 or 5 letters)
  - index of coincidence: computation of specific probabilities

Cryptanalysis example – the ciphertext

```
UPVLDQKZYSOZHBTWWFWDHXEONGVQEZMCGXEDLOULWSXJHEUMWQGFPGKZWHLQKNYSJPTHNSJQBNCNE
XXVNDAKEPTCEZZWGZEZWOJWDUZCYZWDOZUODUVCNAJQMCGFKDMCUAZQHAVZADBQGJPAMCGWWEMJHD
KZWJPTNUWBGTAFPZGFWEEJRKDAIVPVWSBIUZKBTKRJKMVCGVXKTNVWABYGKYDFHAULABLGGSZUGBF
GSZMJCTHMVFISLXBFGLDQFJWLDAYVESRQBAQMJPMCGVAZHAYZEOADYSOUGNGSNOAJTVADBIIEUOTW
VGSMBOKHWELZFVKIGOJWOFXKUOKDGCQDHAPDPLDQVZPLNQUTVZAOXVUWHQLNVJAMWJHVNGGFGFBQX
OCFZNROJWHUZCVGBMYGKUGQKDPYKUEGCELMUJXWPTXYQGNUYJWFZFAZNSPOAVPVIMWZOQSMRDPLKM
EJPYHAPMQGIFADECWZWCGSRKPDVZPTXWTGSZHKKMIEFJMWWZWOGJNMVZFOEFARQGZQGWGJPTLGKCA
FAZHGNQVVULHQHACFAYBBTSJFLCKHPTKJWYDFAZIDKAFJPWYANGFVEYETESPOAVIDEYINGGBNHYKW
OXRDPYEZLOTSJSXACFPMLOKULALZUTKIXYUZKGEYGJONXIVCJQXNJWWPLOJJKIGWCUGMGYEZEZLKQ
AJFBIIMLITMFOEFACGJAMGYVZADXVFSNWEVECHGLOTWAKXOWJJQWPRGJFAZPWSOHHGJ
```

Relative frequencies of letters (plaintext vs. ciphertext)



Kasiski’s method

```
UPVLDQKZYSOZHBTWWFWDHXEONGVQEZMCGXEDLOULWSXJHEUMWQGFPGKZWHLQKNYSJPTHNSJQBNCNE
XXVNDAKEPTCEZZWGZEZWOJWDUZCYZWDOZUODUVCNAJQMCGFKDMCUAZQHAVZADBQGJPAMCGWWEMJHD
KZWJPTNUWBGTAFPZGFWEEJRKDAIVPVWSBIUZKBTKRJKMVCGVXKTNVWABYGKYDFHAULABLGGSZUGBF
GSZMJCTHMVFISLXBFGLDQFJWLDAYVESRQBAQMJPMCGVAZHAYZEOADYSOUGNGSNOAJTVADBIIEUOTW
VGSMBOKHWELZFVKIGOJWOFXKUOKDGCQDHAPDPLDQVZPLNQUTVZAOXVUWHQLNVJAMWJHVNGGFGFBQX
OCFZNROJWHUZCVGBMYGKUGQKDPYKUEGCELMUJXWPTXYQGNUYJWFZFAZNSPOAVPVIMWZOQSMRDPLKM
EJPYHAPMQGIFADECWZWCGSRKPDVZPTXWTGSZHKKMIEFJMWWZWOGJNMVZFOEFARQGZQGWGJPTLGKCA
FAZHGNQVVULHQHACFAYBBTSJFLCKHPTKJWYDFAZIDKAFJPWYANGFVEYETESPOAVIDEYINGGBNHYKW
OXRDPYEZLOTSJSXACFPMLOKULALZUTKIXYUZKGEYGJONXIVCJQXNJWWPLOJJKIGWCUGMGYEZEZLKQ
AJFBIIMLITMFOEFACGJAMGYVZADXVFSNWEVECHGLOTWAKXOWJJQWPRGJFAZPWSOHHGJ
```

- `SPOAV` distance: 156, `FOEFA` distance: 186 $\longrightarrow \gcd(156, 186) = 6$

Index of coincidence

- probability, that two randomly drawn letters from a given string are equal

$$\mathrm{IC}(x) = \frac{\sum_i f_i(f_i - 1)}{t(t-1)}$$

  where $t = |x|$, and f_i are counters for individual letters in x

- expectations for an English alphabet (26 letters), and sufficiently long string:
  - random string: $\mathrm{IC}_{\mathrm{rnd}} = 26 \cdot \frac{1}{26^2} = \frac{1}{26} \approx 0.0385$
  - English text: $\mathrm{IC}_{\mathrm{EN}} \approx 0.127^2 + 0.091^2 + \ldots \approx 0.6555$

- test hypotheses for $n = 1, 2, \ldots$
- given n a "trace" of string $x = x_1, x_2, \ldots$ (for example a ciphertext) is a string containing every n-th letter of x:
  - 1st trace: $x_1, x_{n+1}, x_{2n+1}, \ldots$
  - 2nd trace: $x_2, x_{n+2}, x_{2n+2}, \ldots$, etc.
- observation: shift cipher does not change the IC
- expectation:
  - for correct n all traces will have IC closer to $\mathrm{IC}_{\mathrm{EN}}$
  - for incorrect n, the traces are a mixture of several shift ciphers $\Rightarrow$ IC closer to $\mathrm{IC}_{\mathrm{rnd}}$

IC calculated for our ciphertext (the first trace):

n	IC
1	0.0411
2	0.0456
3	0.0479
4	0.0481
5	0.0395
6	0.0559

n	IC
7	0.0421
8	0.0459
9	0.0412
10	0.0428
11	0.0373
12	0.0551

- decrypt all traces separately
  - individual letters with statistical properties of the plaintext
  - try all shifts and find the best one(s)
  - combine shifted traces into the plaintext and verify its correctness
- mutual index of coincidence, $\mathrm{MIC}(x, y)$ is the probability, that randomly drawn letters from x and y are equal
  - $\mathrm{MIC}(x, y) = \frac{\sum_i f_i \cdot f_i'}{t \cdot t'}$, where $t = |x|$, $t' = |y|$, and f_i (f_i') are counters for individual letters in x (y)
- find mutual relations between 1st trace and all other traces
  - shift other traces until MIC(1st tarce, shited trace) is close to $\mathrm{IC}_{\mathrm{EN}}$
  - test all shifts of the 1st trace (other relative shifts are already known)

Cryptanalysis example – finding correct shifts (separate traces)

- scoring an error of shifted trace
  - how "far" are frequencies of `E`, `T`, `A`, and `O` from their expected values:
    $$\text{error} = \left((\text{fr}_\text{E} - 0.127)^2 + (\text{fr}_\text{T} - 0.091)^2 + (\text{fr}_\text{A} - 0.082)^2 + (\text{fr}_\text{O} - 0.075)^2\right)^{1/2}$$
  - nothing special, anything reasonbale will work (trace length is important)
  - you can include more letters, the least frequent letters, etc.
  - large gaps between smallest errors; decrypt and analyze digrams/words, ...

- top three shifts with the smallest errors for each trace (bold – the correct shift)

shift	error
H	0.047
T	**0.062**
W	0.080

shift	error
V	**0.045**
K	0.072
F	0.095

shift	error
C	**0.055**
R	0.081
Q	0.086

shift	error
S	**0.039**
V	0.079
F	0.085

shift	error
W	**0.032**
L	0.079
Z	0.087

shift	error
M	**0.035**
B	0.086
K	0.102

Original plaintext

```
But there was no great difficulty in the first stage of my adventure.
Upper Swandam Lane is a vile alley lurking behind the high wharves
which line the north side of the river to the east of London Bridge.
Between a slop-shop and a gin-shop, approached by a steep flight of
steps leading down to a black gap like the mouth of a cave, I found the
den of which I was in search. Ordering my cab to wait, I passed down
the steps, worn hollow in the centre by the ceaseless tread of drunken
feet; and by the light of a flickering oil-lamp above the door I found
the latch and made my way into a long, low room, thick and heavy with
the brown opium smoke, and terraced with wooden berths, like the
forecastle of an emigrant ship.
Through the gloom one could dimly catch a glimpse of bodies lying in
strange fantastic poses, bowed shoulders, bent knees, heads thrown
back, and chins pointing upward, with here and there a dark,
lack-lustre eye turned upon the newcomer.
```

1. *Find the top 10 digrams for C or C++ programming language. Estimate the entropy of the language in bits/character.*
2. *Calculate the unicity distance for Vigenère cipher with key length 6 (English plaintext).*
3. *Autokey is a Vigenère cipher variant, where the key is concatenated with the plaintext for encryption and decryption.*
   a) *Why and how does the decryption work?*
   b) *Propose a method how to attack this cipher.*

```
HOUSTONWEHAVEAPROBLEM
BLUEHOUSTONWEHAVEAPRO
          ↓
IZOWACHOXVNRIHPMSBAVA
```