Tikimybių teorijos ir statistikos metodai kriptografijoje

R. Čiegis

Vilniaus Gedimino technikos universitetas
e-mail: rc@vgtu.lt

Spalio 20 d., 2020, Vilnius

Pakeitimų šifras

Slaptųjų dokumentų ir informacijos slėpimo poreikis yra toks pat senas, kaip ir mūsų istorija. Šis poreikis apima labai didelę dalį žmonių veiklos sričių (teisingiau, sunku įvardinti tokią sritį, kur informacijos slėpimas nebūtų naudojamas).

Šiandien jau niekuo nepasitikime – slaptažodžiai, raktai, šifravimo algoritmai supa mus kiekviename žingsnyje, kiekviename įrenginyje, kiekviename dokumente. Košmariškame sapne sapnuojame, kad užmiršome prisijungimo prie banko sąskaitos ar telefono slaptažodį.

Tvari tik kova tarp informacijos slėpimo (šifravimo) ir jos "nulaužimo" algoritmų. Čia niekada nebuvo galutinio nugalėtojo, tik trumpalaikės pergalės.

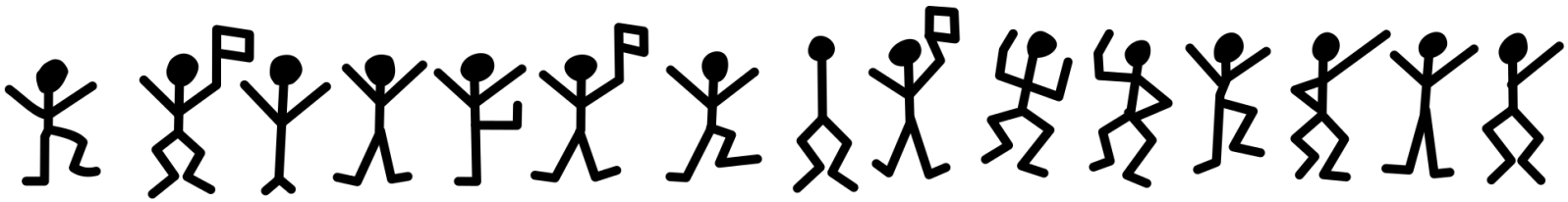
Šiandien apžvelgsime du klasikinius šifravimo algoritmus – jie puikūs, išradingi ir tikrai nelengvai iššifruojami. Daug vilčių turėta, kad jie absoliučiai saugūs (ypač antrojo algoritmo atveju).

Tačiau baziniai tikimybių teorijos ir statistikos rezultatai, kuriuos pateikiame mūsų studentams matematikos studijų pirmuose semestruose, sugriovė šias viltis. Neblogas pavyzdys, galintis pareklamuoti šių kursų naudą VGTU studentų išsilavinime.

Klasikiniame pakeitimo šifre pradinio teksto raidės pakeičiamos kitais simboliais (bijekcijos atvaizdavimas), bet jų išdėstymo eiliškumas nesikeičia.

Šis šifravimo metodas aprašytas daugelyje labai populiarių knygų (ten, beje pateikiami ir slaptojo teksto dešifravimo žingsniai, pastarieji yra bene įdomiausia šių pasakojimų dalis).

Arthur Conan Doyle "The Adventure of the Dancing Men"



Edgar Allan Poe "The Gold-Bug"

53??+305))6*;4826)4?.)4?);80
6*;48+8@60))85;1?(;:?*8+83(88)
5*+;46(;88*96*?;8)*?(;485);5*+
2:*?(;4956*2(5*-4)8@8*;40692
85);)6+8)4??;1(?9;48081;8:8?1
;48+85;4)485+528806*81(?9;48
;(88;4(??34;48)4?;161;:188;??;

A good glass in the bishop's hostel in the devil's seat
forty-one degrees and thirteen minutes northeast and by north
main branch seventh limb east side
shoot from the left eye of the death's-head
a bee line from the tree through the shot fifty feet out.

1. Kadangi alfabete turime 26 raides, tai šifruodami pranešimą galime pasirinkti vieną iš

$$26! \approx 10^{26}$$

pakeitimų.

2. Toks variantų skaičius yra daug didesnis už DES raktų skaičių

$$2^{56} \approx 10^{18}$$

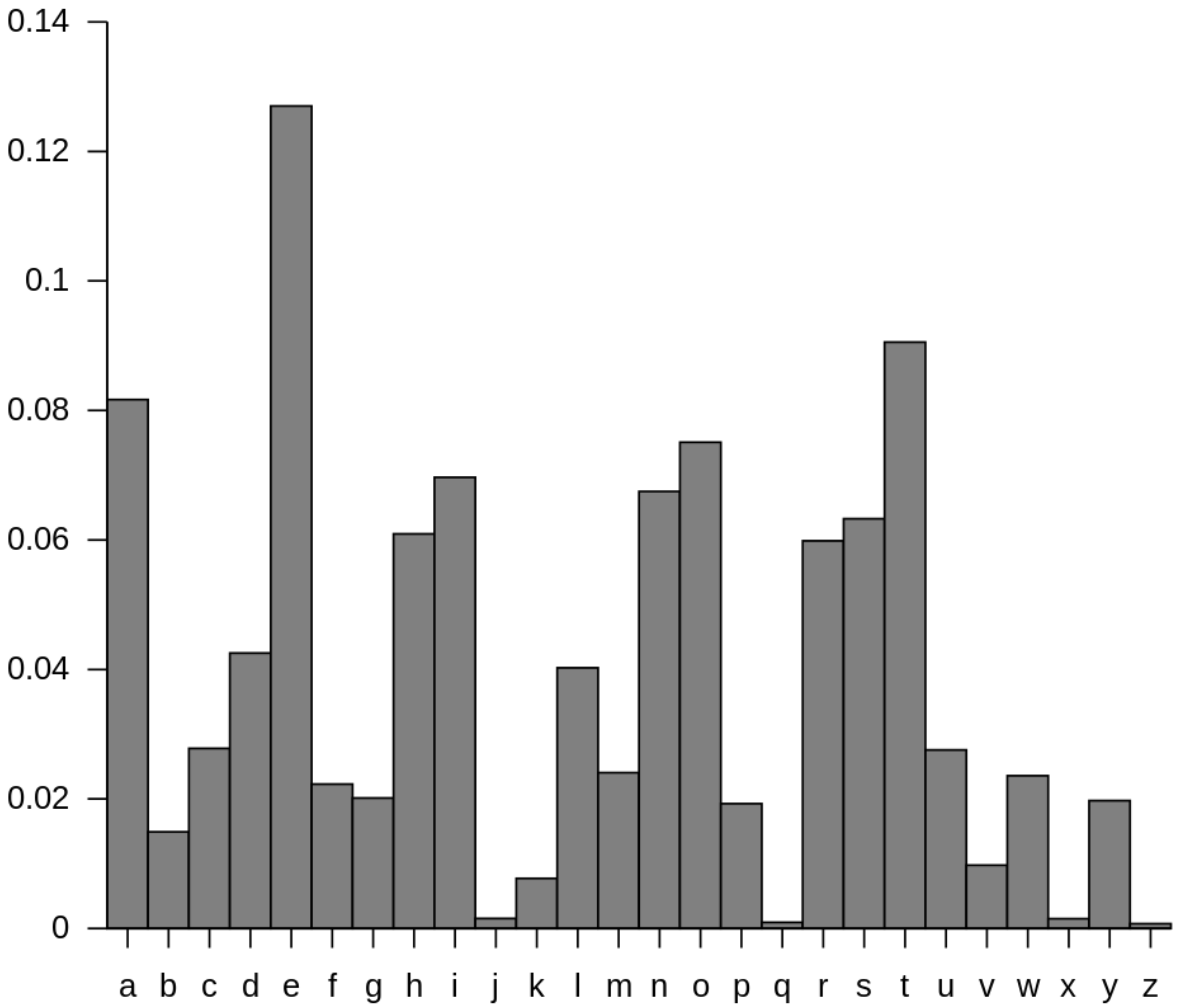
Taigi, atrodytų, kad klasikinis pakeitimų šifravimo algoritmas yra visiškai saugus.

Savo pasuose saugome biometrinius duomenis, kurie vienareikšmiškai identifikuoja mus pasienio punktuose, bankuose, policijoje.

Gerai žinomas pirštų antspaudų metodas kriminalistikoje.

Pasirodo, kad kiekvienos kalbos raidžių bei jų sekų dažniai yra unikalūs ir stabilūs kalbos požymiai (ši išvada grindžiama klasikine statistine dažnių analize).

Anglų kalbos abėcėlės raidžių dažniai



1. E, T, A ir O yra populiariausios raidės.
2. Z, Q, X ir J yra rečiausiai naudojamos raidės.

3. TH, ER, ON ir AN yra populiariausios raidžių poros
4. SS, EE, TT ir FF yra populiariausios pasikartojančių raidžių poros.

Iš dažnių analizės seka ir daugiau svarbių išvadų apie balsių ir priebalsių tarpusavio išsidėstymą.

Taigi po šio statistinės dažnių analizės verdikto klasikinis pakeitimo šifras nebegali būti laikomas patikimu – raidėms nepavyks pasislėpti tik "persirengus".

Visgi net ir žinant šią informaciją, pakeitimo šifro "nulaužimas" yra nelengva užduotis, reikalaujanti sprendėjo kruopštumo, loginio mąstymo ir kantrybės. Patirtis ir profesionalumas atsiranda tik išsprendus didesnį kiekį uždavinių.

Yra sukurta daug pakeitimo algoritmo modikacijų, kurios paslepia raidžių dažnių skirtumus. Tarp jų yra net ir tokių, kuriems parodyta, jog naujieji šifravimo algoritmai yra "nenulaužiami".

Visgi pastarieji nėra tinkami platesniam naudojimui – o žmonija norėjo, nori ir norės šifruoti vis daugiau informacijos .

Vigenere šifras

Šifrą 1553 metais sukūrė Giovan Batista Belasso. Tačiau jį išpopuliarino B. Vigenere apie 1586 metus, jo vardu algoritmas dažniausiai ir vadinamas.

Algoritme panaudotos dvi idėjos:

1. Bazinis alfabeto pakeitimo algoritmas – Cezario (arba rot m) algoritmas. Naujas alfabetas gaunamas, atlikus raidžių postūmį per m pozicijų (mod 26 liekanų aritmetikoje).

2. Pasirenkamas atsitiktinis raktas (dažniausiai prasmingas žodis $R_1\,R_2\,\ldots\,R_n$) ir šifruojant $(j+nk)$-ąją teksto raidę naudojamas postūmis per R_j pozicijų.

Tokiu būdu ta pati alfabeto raidė, priklausomai nuo jos vietos tekste, gali būti užšifruota n skirtingais simboliais ir visų simbolių statistiniai dažniai yra panašūs.

Vigenere algoritmo dešifravimas nežinant rakto yra daug sudėtingesnis, nei klasikinio pakeitimo šifro. Tik po trijų šimtų metų buvo sukurtas bendras šio šifro "nulaužimo" metodas (1863, Friedrich Kasiski). Ir vėl svarbiausiu veikėju tapo statitiniai metodai.

Vigenere šifravimo lentelė

	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

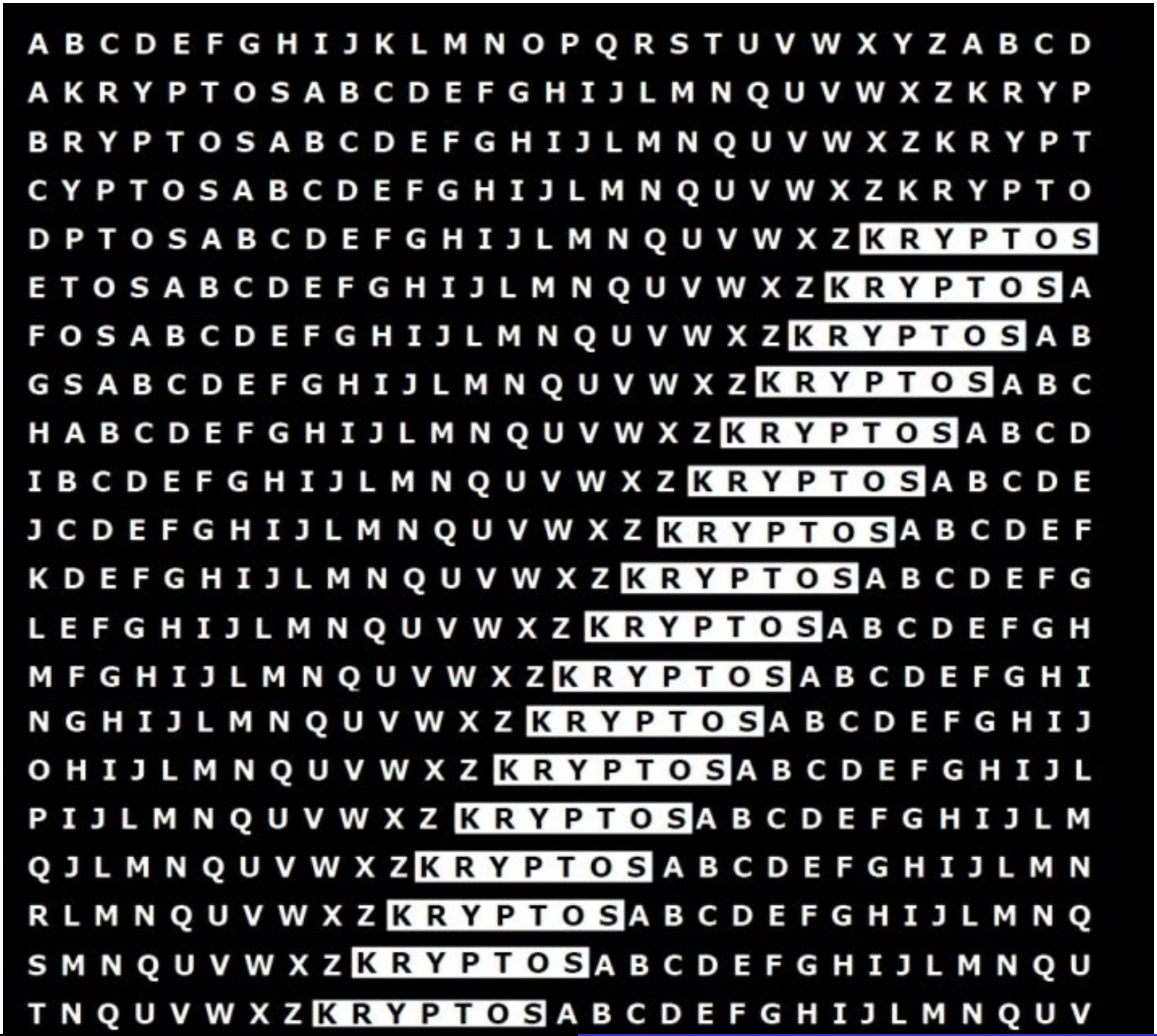
Raktas(5): LEMON, pranesimas(12): ATTACKATDAWN

A T T A C K A T D A W N
L E M O N L E M O N L E
L X F O P V E F R N H R

$$C_i = E_K(M_i) = (M_i + K_i) \mod 26$$
$$M_i = D_K(C_i) = (C_i - K_i + 26) \mod 26$$

Nagrinėkime garsųjį Kryptos uždavinį. Tai JAV menininko Jim Sanborn sukurta skulptūra, 1990 metų lapkričio mėnesį ji pastatyta Centrinės Žvalgybos Valdybos (CŽV) teritorijoje Virdžinijos valstijoje Lengli mieste.

Jos stenduose užšifruoti keturi pranešimai, kurie ir praėjus 30 metų vis dar nėra pilnai iššifruoti. Du pirmieji pranešimai užšifruoti modifikuotu Vigenere algoritmu.


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 B C D
A K R Y P T O S A B C D E F G H I J L M N Q U V W X Z K R Y P
B R Y P T O S A B C D E F G H I J L M N Q U V W X Z K R Y P T
C Y P T O S A B C D E F G H I J L M N Q U V W X Z K R Y P T O
D P T O S A B C D E F G H I J L M N Q U V W X Z K R Y P T O S
E T O S A B C D E F G H I J L M N Q U V W X Z K R Y P T O S A
F O S A B C D E F G H I J L M N Q U V W X Z K R Y P T O S A B
G S A B C D E F G H I J L M N Q U V W X Z K R Y P T O S A B C
H A B C D E F G H I J L M N Q U V W X Z K R Y P T O S A B C D
I B C D E F G H I J L M N Q U V W X Z K R Y P T O S A B C D E
J C D E F G H I J L M N Q U V W X Z K R Y P T O S A B C D E F
K D E F G H I J L M N Q U V W X Z K R Y P T O S A B C D E F G
L E F G H I J L M N Q U V W X Z K R Y P T O S A B C D E F G H
M F G H I J L M N Q U V W X Z K R Y P T O S A B C D E F G H I
N G H I J L M N Q U V W X Z K R Y P T O S A B C D E F G H I J
O H I J L M N Q U V W X Z K R Y P T O S A B C D E F G H I J L
P I J L M N Q U V W X Z K R Y P T O S A B C D E F G H I J L M
Q J L M N Q U V W X Z K R Y P T O S A B C D E F G H I J L M N
R L M N Q U V W X Z K R Y P T O S A B C D E F G H I J L M N Q
S M N Q U V W X Z K R Y P T O S A B C D E F G H I J L M N Q U
T N Q U V W X Z K R Y P T O S A B C D E F G H I J L M N Q U V

VFPJUDEEHZWETZYVGWHKKQETGFQJNCE
GGWHKK?DQMCPFQZDQMMIAGPFXHQRLG
TIMVMZJANQLVKQEDAGDVFRPJUNGEUNA
QZGZLECGYUXUEENJTBJLBQCRTBJDFHRR
YIZETKZEMVDUFKSJHKFWHKUWQLSZFTI
HHDDDUVH?DWKBFUFPWNTDFIYCUQZERE
EVLDKFEZMOQQJLTTUGSYQPFEUNLAVIDX
FLGGTEZ?FKZBSFDQVGOGIPUFXHHDRKF
FHQNTGPUAECNUVPDJMQCLQUMUNEDFQ
ELZZVRRGKFFVOEEXBDMVPNFQXEZLGRE
DNQFMPNZGLFLPMRJQYALMGNUVPDXVKP
DQUMEBEDMHDAFMJGZNUPLGEWJLLAETG

Jeigu žinotume rakto ilgį n (dar ne patį raktą), tai galėtume visą užšifruotą tekstą padalinti į n dalių. Kiekviename poaibyje jau naudojamas paprastas Cezario šifras, su tuo pačiu postūmiu m.

Taigi pirmasis dešifravimo uždavinys yra nustatyti Vigenerės šifro rakto ilgį n.

Jeigu rakto ilgis $n = 7$, tai kiekviena septintoji pranešimo raidė yra užšifruota naudojant tokį patį poslinkį.

Taigi visą šifruotą tekstą padaliname į n dalių.

Imkime kurį nors poaibį, kurio ilgis M raidžių. Pažymėkime i-tosios abėcėlės raidės dažnį šioje aibėje f_i, $i = 0, \ldots, 25$.

Spręskime klasikinį statistikos uždavinį: raskime tikimybę, kad atsitiktinai pasirinktos dvi raidės sutampa.

$$P = \sum_{i=0}^{25} \frac{f_i(f_i - 1)}{M(M-1)}.$$

Jeigu visų raidžių dažniai yra vienodi, tai

$$P_0 \approx \frac{1}{26} = 0.038.$$

Jeigu turime anglų kalba parašytą tekstą, tai panaudojame statistinių dažnių lentelę ir gauname tikimybę

$$P_E \approx \sum_{i=0}^{25} p_i^2 \approx 0.065.$$

Atliekame statistinę analizę – įvairiems n skaičiuojame atskirų šifruoto teksto dalių eksperimentinių sutapimo indeksų aritmetinį vidurkį.

Darbinė hipotezė, kad ieškomas rakto ilgis n atitinka tą atvejį, kai sutapimo indeksas artimas 0.065.

VFPJUDEEHZWETZYVGWHKKQETGFQJNCEGGWH
KKDQMCPFQZDQMMIAGPFXHQRLGTIMVMZJANQ
LVKQEDAGDVFRPJUNGEUNAQZGZLECGYUXUEE
NJTBJLBQCRTBJDFHRRYIZETKZEMVDUFKSJH
KFWHKUWQLSZFTIHHDDDUVHDWKBFUFPWNTDF
IYCUQZEREEVLDKFEZMOQQJLTTUGSYQPFEUN
LAVIDXFLGGTEZFKZBSFDQVGOGIPUFXHHDRK
FFHQNTGPUAECNUVPDJMQCLQUMUNEDFQELZZ
VRRGKFFVOEEXBDMVPNFQXEZLGREDNQFMPNZ
GLFLPMRJQYALMGNUVPDXVKPDQUMEBEDMHDA
FMJGZNUPLGESWJLLAETG

Keyword length 1: 0.0453
Keyword length 2: 0.0468
Keyword length 3: 0.0448
Keyword length 4: 0.0553
Keyword length 5: 0.0486
Keyword length 6: 0.0466
Keyword length 7: 0.0450
Keyword length 8: 0.0680
Keyword length 9: 0.0426
Keyword length 10: 0.0495
Keyword length 11: 0.0452
Keyword length 12: 0.0546
Keyword length 13: 0.0480
Keyword length 14: 0.0457
Keyword length 15: 0.0482

Taigi jau nustatėme rakto ilgį: $n = 8$.

Antrasis uždavinys yra surasti postūmius tarp dviejų gretimų rakto raidžių (vis dar rasime ne patį raktą, bet svarbią informaciją).

Pavyzdys: tarkime kad postūmis tarp pirmosios ir antrosios rakto raidžių yra lygus 2.

Kokia yra antroji rakto raidė, jei pirmoji yra R?

Pasinaudosime, kad kiekvieno poaibio abėcėlės pakeitimas atliekamas Cezario algoritmu.

Nagrinėkime dvi aibes X ir Y, sudarytas iš M_x ir M_y raidžių. Pažymėkime raidžių dažnius f_i^x, f_i^y, $i = 0, \ldots, 25$.

Statistikos uždavinys: iš kiekvienos aibės atsitiktinai pasirenkame po vieną raidę, raskite tikimybę, kad raidės sutampa.

$$P_s(X,Y) = \sum_{i=0}^{25} p_i^x p_i^y = \sum_{i=0}^{25} \frac{f_i^x}{M_x} \frac{f_i^y}{M_y}.$$

Kadangi žinome, jog visų aibės X (atitinkamai Y) elementų postūmiai yra vienodi, tai skaičiuojame kiekvieno galimo atvejo tikimybę:

$$P_s(X,Y,r) = \sum_{i=0}^{25} p_i^x p_{i+r}^y, \quad r = 0, \ldots, 25.$$

Indeksų reikšmes skaičiuojame liekanų $\mod 26$ aritmetikoje.

Darbinė hipotezė: atstumas tarp dviejų gretimų rakto raidžių yra toks, kurį atitinka didžiausia tikimybė.

Pastaroji turėtų būti artima 0.065.

i= 0 j= 1 shift= 1
i= 1 j= 2 shift= 24
i= 2 j= 3 shift= 3
i= 3 j= 4 shift= 6
i= 4 j= 5 shift= 17
i= 5 j= 6 shift= 0
i= 6 j= 7 shift= 1

Dabar jau liko paskutinis žingsnis – surasti, kokia yra pirmoji rakto raidė.

Šį kartą teks pasinaudoti nuobodžiu, bet patikimu įrankiu: pilnuoju variantų perrinkimu. Patikriname visas abėcėlės raides.

1)KRZYBZZK:
VETHGEFEHXXCVKPVGVIXMUFTGEUHEDFGGVIXMEUM
2)RYKPCKKR:
UDPGFDEDGWWBUZYUFUHWLQEPFDQGDCEFFUHWLDQL
3)YPRTDRRY:
QCYFECDCFVVAQXRQEQGVJNDYECNFCBDEEQGVJCNJ
4)PTYOEYYP:
NBREDBCBEUUSNWKNDNFUIMCRDBMEBACDDNFUIBMI
5)TOPSFPPT:
MAKDCABADQQOMVZMCMEQHLBKCALDASBCCMEQHALH
6)OSTAGTTO:
LSZCBSASCNNTLUXLBLDNGJAZBSJCSOABBLDNGSJG
7)SAOBHOOS:
JOXBAOSOBMMPJQWJAJCMFISXAOIBOTSAAJCMFOIF

8)ABSCISSA: ITWASTOTALLYINVISIBLEHOWSTHATPOSSIBLETHE
9)BCADJAAB:
HPVSOPTPSJJRHMUHOHAJDGTVOPGSPYTOOHAJDPGD
10)CDBELBBC:
GYUOTYPYOIIKGLQGTGSICFPUTYFOYRPTTGSICYFC
11)DECFMCCD:
FRQTPRYRTHHZFJNFPFOHBEYQPRETRKYPPFOHBREB
12)EFDGNDDE:
EKNPYKRKPGGXEIMEYETGADRNYKDPKZRYYETGAKDA

ITWASTOTALLYINVISIBLEHOWSTHATPOSSIBLE
THEYUSEDTHEEARTHSMAGNETICFIELDXTHEINF
ORMATIONWASGATHEREDANDTRANSMITTEDUNDE
RGRUUNDTOANUNKNOWNLOCATIONXDOESLANGLE
YKNOWABOUTTHISTHEYSHOULDITSBURIEDOUTT
HERESOMEWHEREXWHOKNOWSTHEEXACTLOCATIO
NONLYWWTHISWASHISLASTMESSAGEXTHIRTYEI
GHTDEGREESFIFTYSEVENMINUTESSIXPOINTFI
VESECONDSNORTHSEVENTYSEVENDEGREESEIGH
TMINUTESFORTYFOURSECONDSWESTXLAYERTWO

Matome, kad net ir bazinės statistikos žinios yra puikus pagalbininkas – iššifravome įdomų Vigenerės algoritmu paruoštą dokumentą.

Tiems, kurie nori patys išbandyti šią metodiką, rekomenduoju perskaityti skulptūros pirmąją teksto dalį. Šis tekstas beje yra trumpesnis, taigi jo dešifravimas yra sudėtingesnis.

Kriptografijoje mažiau, visada yra sudėtingiau.

O kaip iššifruosite tekstą, kai nežinome abiejų raktų?

....YTBMKAQOUWJLIZRFGXDVCSHPEN

Z ABCDEFGHIJKLMNOPQRSTUVWXYZ
R BCDEFGHIJKLMNOPQRSTUVWXYZA
E CDEFGHIJKLMNOPQRSTUVWXYZAB
A DEFGHIJKLMNOPQRSTUVWXYZABC
H EFGHIJKLMNOPQRSTUVWXYZABCD
X FGHIJKLMNOPQRSTUVWXYZABCDE
N GHIJKLMNOPQRSTUVWXYZABCDEF
P HIJKLMNOPQRSTUVWXYZABCDEFG
C IJKLMNOPQRSTUVWXYZABCDEFGH