

# Kryptoanalýza jednoduchých šifíř

Martin Stanek

Katedra informatiky  
FMFI UK, Bratislava  
`stanek@dcs.fmph.uniba.sk`

Kryptológia 1

# Jednoduchá substitučná šifra (JSŠ)

- ▶  $P = C = A$  (abeceda)
- ▶  $K$  – množina všetkých permutácií na  $A$
- ▶ šifrovanie:  $E_k(p) = k(p)$ ; dešifrovanie:  $D_k(p) = k^{-1}(p)$
- ▶ útočiť hrubou silou nie je efektívne možné,  $\approx 2^{88,4}$  kľúčov (pri 26 znakovej abecede)

# Kryptoanalýza JSŠ

- ▶ útok len so znalosťou šifrového textu (pri silnejších útokoch je kryptoanalýza triviálna)
- ▶ útočník má k dispozícii dostatočne dlhý šifrový text
- ▶ kryptoanalýza sa opiera o frekvenčnú analýzu znakov
- ▶ JSŠ nemení frekvencie znakov, len ich „označenie“

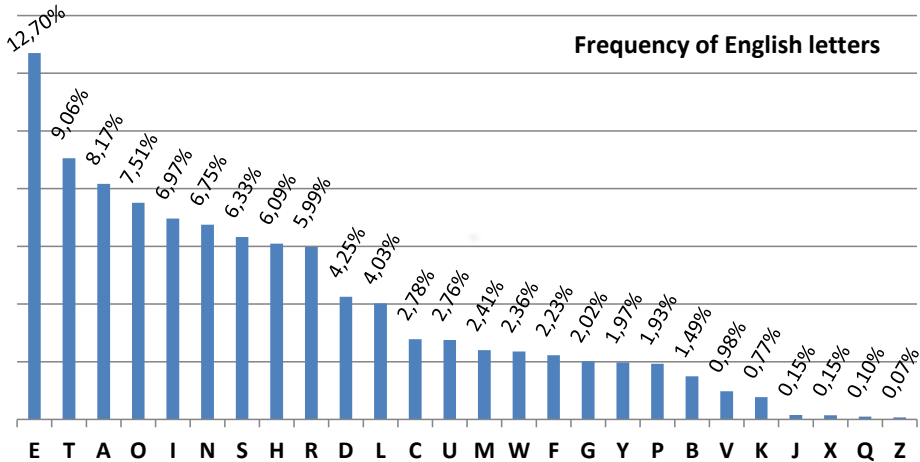
# Prirodzený jazyk (1)

- ▶ redundantný – k pochopeniu/reprezentácii textu nie sú všetky rovnako potrebné
  - ▶ kompresia, vynechanie znakov

Tu v este áme ln jedu stuňu, zktore všeti pijme, ae námtá voa drao pade, leo v jdnej iere a mesom býa dra s dvnástii hlaami, torém každ deň ednu aničk musi dať ožrať leboak bymu neali, ikohoby natú stdňu npusti a mueli b sme d smäu pohnúť. ž z dm na om poávalimešťaia svje divčatá a teaz jerad n kráľvej dére. ato rzkáza kráľmestočiernm súkom obiahnu, — ae aj o dalohlášť, žekto b sa tký naiel, torý y toh drak zabi, že u tú voju céru pol ráľovtvom á za enu apo krľovejsmrtiže dotane elú kajinu  
vynechané každé piate písmeno, (Dobšinský)

## Prirodzený jazyk (2)

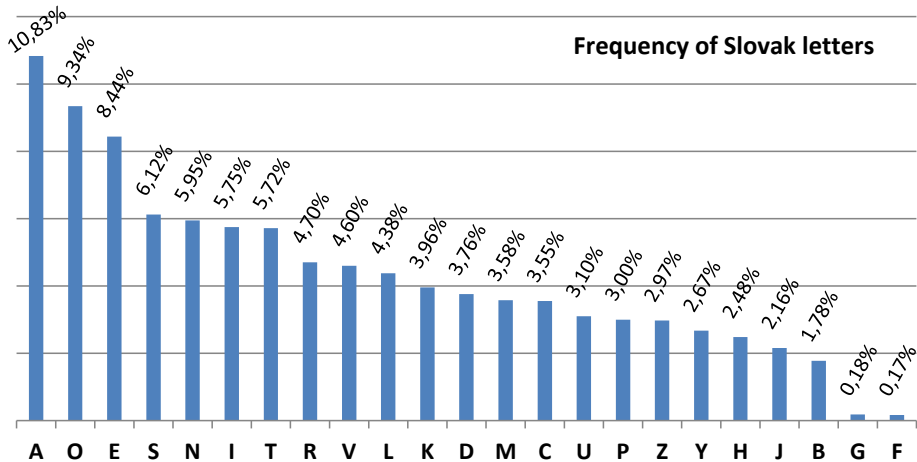
- ▶ rôzne znaky abecedy sa v texte vyskytujú s rôznou pravdepodobnosťou
- ▶ podobne pre dvojice („SA“ vs. „GM“), trojice znakov („ANO“ vs. „RHF“), ...
- ▶ príklad najfrekvencovanejších bigramov v angličtine (bez medzery):  
TH, HE, IN, ER, AN, RE, ND, ON, EN, AT
- ▶ príklad najfrekvencovanejších trigramov v angličtine (bez medzery):  
THE, AND, ING, HER, HAT, HIS, THA, ERE, FOR, ENT
- ▶ príklad najfrekvencovanejších 4-gramov v angličtine (bez medzery):  
THAT, THER, WITH, TION, HERE, OULD, IGH, HAVE
- ▶ príklad najfrekvencovanejších bigramov v slovenčine (bez medzery):  
PR, OV, PO, NA, NE, KO, ST, VA, RE



## Prirodzený jazyk (3)

iné užitočné štatistiky pri kryptoanalýze (angličtina):

Začiatky slov		Konce slov	
T	15,94%	E	19,17%
A	15,50%	S	14,35%
I	8,23%	D	9,23%
S	7,75%	T	8,64%
O	7,12%	N	7,86%
C	5,97%	Y	7,30%
M	4,26%	R	6,93%
F	4,08%	O	4,67%
P	4,00%	L	4,56%
W	3,82%	F	4,08%

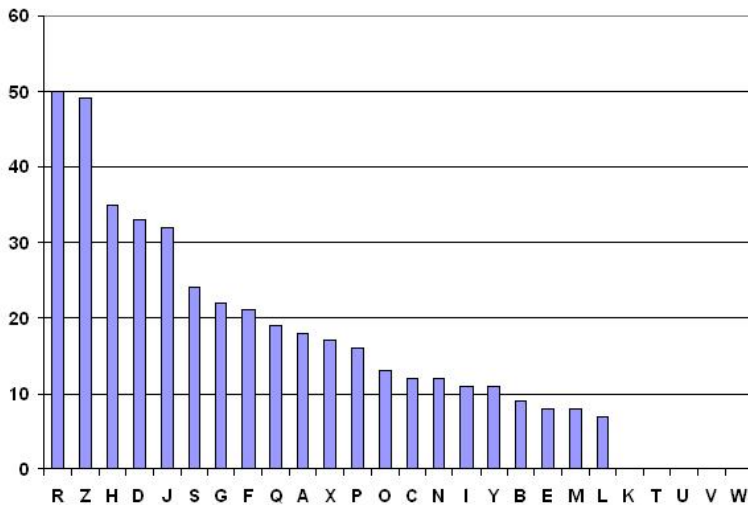




# Šifrový text

SR HJLHZ HPRHD XZQD FHJON JGHJ YGJAN SR AHZPNIE FQRA CPJMRFRQ DXR  
OZIRGSJ. B ZADJS FQRAY XZQZ IJGHZY FDORH SR MSZENIE MDJGHRIE  
PZXZHSDAZF AHZPD OZASIZFRQD GHRFXN HYSJQZF RQJXZ FDROYAHZF RQJXZ  
ZCPRFZFRQD CZGAZOJSJ YGJAN HPRHJ. RA PZXZHSDID SJCPRIZFRQD XQDBAZ  
SJLRAJL ZGRON AHZPJ GR HY FHJON MSZBDQD RAZ EYXN CZ ORBOD  
ZXNIRLSJ GD GHRFRQD HRXZPN R HZ RL G ZCJFSJSJM. XZQZ HZ CZHPJXSJ  
QJXZ HYSRLGD DSODRSD CZAQRORQD GHRFXN HPRHJ BR BRGRE OZ GFZLDIE  
CPRF R CPJHZ GR GSRBDQD FGJHANMD GCZGZXMD CPRIY ERHDH R  
BSJMZBSDH.

## Frekvenčná charakteristika šifrovaného textu



# Kryptoanalýza ŠT

- ▶ ... slides
- ▶ veľa ďalších možných vylepšení kryptoanalýzy
- ▶ automatizovaná kryptoanalýza
- ▶ použitie slovníka príslušného jazyka, vyhľadávanie podľa vzoriek (rovnaké/rôzne znaky), modelovanie jazyka (Markovov zdroj), ...
- ▶ hill climbing (štart z náhodného kľúča, štart z kľúča utriedeného podľa frekvencií) ... ukážka

# Jednoznačnost kryptoanalýzy

- ▶ šifrový text: QWERT
- ▶ otevřený text: kniha, obraz, lucia, ...
- ▶ otevřený text: today, terms, index, ...
  
- ▶ kedy je ŠT jednoznačne dešifrovateľný?
- ▶ (ekvivalentne) kedy ŠT prislúcha práve jeden „zmysluplný“ kľúč?

# Entropia

- ▶  $X$  je diskrétna náhodná premenná
- ▶  $X$  nadobúda hodnoty  $x_1, \dots, x_n$  s pravd.  $p_1, \dots, p_n$
- ▶ entropia n.p.  $X$ :

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

- ▶ entropia  $\approx$  očakávaný počet bitov potrebných na reprezentáciu výsledku  $X$
- ▶ hod mincou:  $H(\frac{1}{2}, \frac{1}{2}) = -2 \cdot (\frac{1}{2} \lg \frac{1}{2}) = 1$  (bit)

# Entropia jazyka

- ▶ na základe frekvenčnej tabuľky individuálnych znakov:

$$H_{\text{rnd}}(\dots) \approx 4,70$$

$$H_{\text{SK}}(\dots) \approx 4,23$$

$$H_{\text{EN}}(\dots) \approx 4,15$$

- ▶ na základe frekvencií  $n$ -tíc znakov (prepočítané na 1 znak):

$$H_{\text{EN}}(\text{dvojice}) \approx 3,65 \text{ bitov/znak}$$

$$H_{\text{EN}}(\text{trojice}) \approx 3,22 \text{ bitov/znak}$$

...

$$H_{\text{EN}} \approx 1,50 \text{ bitov/znak}$$

- ▶ entropia jazyka je entropia pre  $n \rightarrow \infty$  (kompresia)

# Redundancia textu

- ▶ nech abeceda jazyka má  $q$  znakov
- ▶  $Y$  – n.p. označuje otvorený text dĺžky  $n$
- ▶ redundancia textu  $Y$ :

$$D_n = n \lg q - H(Y)$$

- ▶ redundancia  $\approx$  o koľko bitov je OT dlhší ako reťazec potrebný na jeho zápis

# Vzdialenosť jednoznačnosti OT

- ▶ pre prípad útoku len so znalosťou ŠT (COA)
- ▶ vzdialenosť jednoznačnosti OT:

$$\min\{n \in \mathbb{N} \mid D_n \geq H(K)\}$$

- ▶ „nadbytočné“ bity OT umožňujú jednoznačne určiť kľúč
- ▶ Príklad: jednoduchá substitučná šifra (EN)
  - ▶  $H(K) = \lg(26!) \approx 88,38$  (kľúče sú rovnako pravd.)
  - ▶  $D_n = (4,7 - 1,5) \cdot n = 3,2 \cdot n$  (aproximujeme  $H(Y)$  pomocou entropie jazyka)
  - ▶  $3,2 \cdot n \geq 88,38 \Rightarrow n \geq 28$



# Vzdialenosť jednoznačnosti OT

- ▶ pre prípad útoku len so znalosťou ŠT (COA)
- ▶ vzdialenosť jednoznačnosti OT:

$$\min\{n \in \mathbb{N} \mid D_n \geq H(K)\}$$

- ▶ „nadbytočné“ bity OT umožňujú jednoznačne určiť kľúč
- ▶ Príklad: jednoduchá substitučná šifra (EN)
  - ▶  $H(K) = \lg(26!) \approx 88,38$  (kľúče sú rovnako pravd.)
  - ▶  $D_n = (4,7 - 1,5) \cdot n = 3,2 \cdot n$  (aproximujeme  $H(Y)$  pomocou entropie jazyka)
  - ▶  $3,2 \cdot n \geq 88,38 \Rightarrow n \geq 28$

# Vigenereova šifra

- ▶ polyalfabetická substitúcia
- ▶ zreťazenie  $n$  nezávislých posuvných šifier (pripočítavame kľúč), napr.:

OTVORENYTEXT...

KLUCKLUCKLUC...

YEPQBPHADPRV ...

- ▶ „zarovnaná“ frekvenčná charakteristika ŠT

# Kryptoanalýza

- ▶ útok len so znalosťou šifrového textu (COA)
- ▶ prvý krok – určenie dĺžky kľúča  $n$ :
  - ▶ Kasiskiho metóda – rovnaké časti OT sú šifrované rovnako, ak ich vzájomná vzdialenosť je násobkom  $n$  . . . nsd offsetov rovnakých častí ŠT
  - ▶ Index koincidencie – vhodnosť charakteristiky jednotlivých „stôp“ ŠT

# Index koincidence

- ▶  $x = x_1 x_2 \dots x_t$  – reťazec znakov
- ▶ index koincidence reťazca  $x$ ,  $I_c(x)$ , je pravdepodobnosť, že dva náhodne zvolené znaky z  $x$  sú rovnaké

$$I_c(x) = \frac{\sum_i f_i(f_i - 1)}{t(t - 1)},$$

kde  $f_i$  sú početnosti jednotlivých znakov.

- ▶ pre dostatočne dlhý reťazec je  $I_c(x)$  možné aproximovať z pravdepodobností znakov jazyka:

$$I_c^{\text{rnd}}(\dots) \approx 0,0385 \quad (\text{pre 26 znakov})$$

$$I_c^{\text{EN}}(\dots) \approx 0,0655$$

# Vzájomný index koincidencie

- ▶ vzájomný index koincidencie reťazcov  $x$  a  $y$ ,  $MI_c(x, y)$ , je pravdepodobnosť, že náhodne zvolený znak z  $x$  a náhodne zvolený znak z  $y$  sú rovnaké

$$MI_c(x, y) = \frac{\sum_i f_i \cdot f'_i}{t \cdot t'},$$

kde  $f_i$  ( $f'_i$ ) sú početnosti jednotlivých znakov v  $x$  (resp.  $y$ ) a  $t$  ( $t'$ ) je dĺžka  $x$  (resp.  $y$ ).

# Kryptoanalýza ŠT

► ...slides

# Autoklúč

- ▶ využíva otvorený text ako kľúč, napr.:

TOTOJEOTVORENYTEXT...

KLUCTOTOJEOTVORENY...

DZNQCSHHESFXIMKIKR

- ▶ lúštenie nie je zložitejšie ako pri Vigenereovej šifre, napr.:
  - ▶ hypotéza o dĺžke kľúča (skúšame všetky)
  - ▶ skúmanie „stôp“ (reťazec OT dešifrovateľný jedným znakom kľúča)
  - ▶ testovanie správnosti stopy, 26 možností (je z jazyka?)
  - ▶ fungujú aj triviálne metriky typu (pre SK):  
zahodíme všetky stopy, ktoré obsahujú viac neprípustných znakov (q,w,x)  
maximalizujeme súčet štvorcov 7 najčastejších znakov (a,o,e,s,n,i,t)