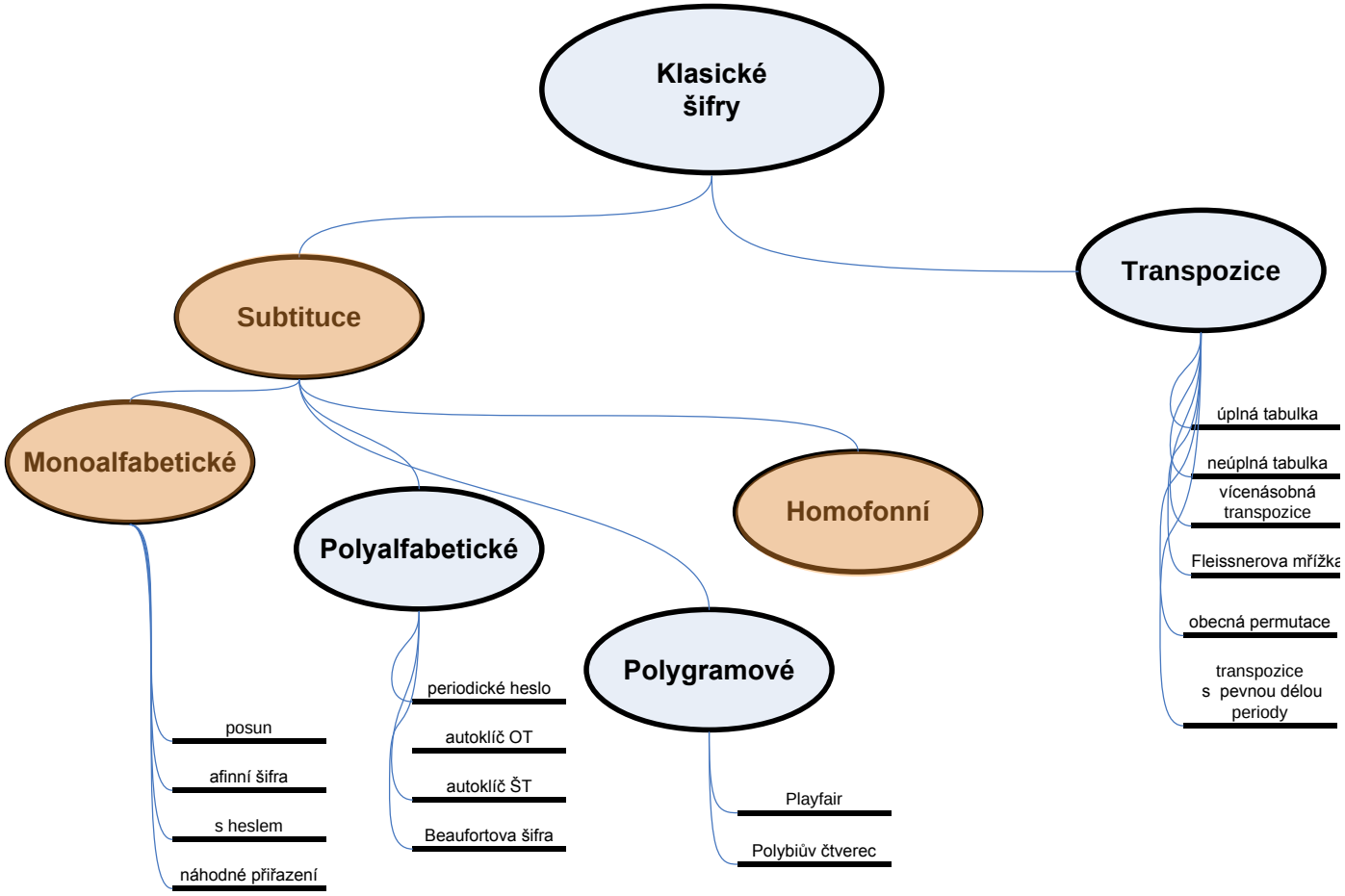
České vysoké učení technické v Praze
Fakulta elektrotechnická
Katedra telekomunikační techniky

2.cvičení – Monoalfabetické substituční šifry, frekvenční analýza, index koincidence

Ing. Tomáš Vaněk, Ph.D. tomas.vanek@fel.cvut.cz

Rozdělení klasických šifer



Rozdělení klasických substitučních šifer

- Jednoduchá substituční šifra (monoalfabetická substituce, jednoduchá záměna) je šifra, ve které se každý znak otevřeného textu nahradí příslušným znakem šifrovaného textu.
- Existuje unikátní mapování znaků abecedy OT na znaky abecedu ŠT ➔ monoalfabetická
- Homofonní substituční šifra se podobá jednoduché substituční šifře, avšak jeden znak otevřeného textu může byt nahrazen jedním z několika možných znaků šifrovaného textu. Znak „A“ by mohl být nahrazen např. 5, 10, 13 nebo 25, „B“ např. 6 nebo 15 atd. Počet znaků zašifrovaného textu pro jeden znak otevřeného textu se může lišit.
- Polygramová substituční šifra je ta, ve které šifrování probíhá mezi skupinami znaků. Skupina „AA“ může být nahrazená skupinou „JH“, „AB“ skupinou „DK“ atd.
- Polyalfabetická substituční šifra se skládá z několika jednoduchých šifer, které se postupně pro jednotlivé znaky otevřeného textu střídají.

Mapování anglické abecedy

- pro potřeby šifrovacích transformací je nejprve nutné převést abecedu na čísla
- používá se tento předpis

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

Monoalfabetická substituce – jednoduchý posun

Šifrování:

$C_i = P_i + A \bmod 26 \quad A \in \langle 0;25 \rangle$

Dešifrování:

$P_i = C_i - A \bmod 26$

- Césarova šifra A=3
- celkem existuje 26 variant jednoduchého posunu (25 použitelných)
- luštění bez znalosti klíče
  - brute force (pouze 25 kombinací)
  - frekvenční analýza

Monoalfabetická substituce – afinní šifra

Šifrování:

$C_i = A \cdot P_i + B \bmod 26 \quad B \in \langle 0;25 \rangle, \quad A \in \langle 1;25 \rangle \quad \gcd(A,26)=1$

Dešifrování:

$P_i = A^{-1}(C_i - B) \bmod 26$; A^{-1} je multiplikativní inverze A mod 26

- existuje $\varphi(26) \cdot 26 = 312$ kombinací
  φ - Eulerova funkce φ
- jednoduchý posun je speciální případ afinní šifry
- luštění bez znalosti klíče
  - brute force
  - frekvenční analýza + zpětný výpočet koeficientů A,B

Monoalfabetická substituce s klíčem

Abeceda OT: 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

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

- pokud se v klíči opakují některá písmena, pak se zaznamená pouze první výskyt: LOKOMOTIVA -> LOKMTIVA
- nejprve se opíše klíč a poté zbývající znaky abecedy (s vynechanými písmeny, která jsou součástí klíče)
- při špatně zvoleném klíči pomůže FA
  - od posledního znaku klíče v abecedě se abeceda OT mapuje 1:1 na abecedu ŠT (viz příklad nahoře, kde od znaku **s** se znaky mapují sami na sebe**)**
- luštění bez znalosti klíče možné s využitím frekvenční analýzy

Monoalfabetická subst. – náhodná transformace

- náhodné přiřazení znaků abecedy OT a ŠT
- počet možných kombinací $26! \approx 4 \times 10^{26}$

Příklad:

Abeceda OT: 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

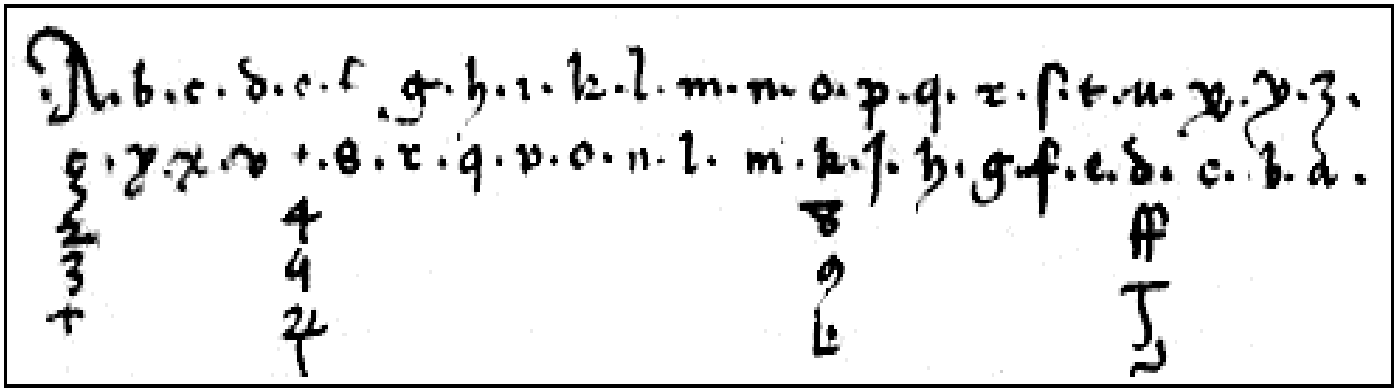
Abeceda ŠT: G = ` E * ~ 2 U D F : I > C - 5 R 7 V S 3 { H X / M

Homofonní šifra

- cca polovina 14. století
- snahy o zdokonalení monoalfabetické substituce
- jeden znak OT je zašifrován na různé znaky ŠT
- počet možných znaků závisí na četnosti výskytu daného znaku
- její použití lze jednoduše rozpoznat - počet různých znaků v ŠT je výrazně větší než počet znaků v dané národní abecedě
- homofonní šifra nepřinesla výrazné vylepšení bezpečnosti
- různé reprezentace jednoho znaku lze odhalit pomocí vyhledání „skoro shodných“ slov
  - stejná slova OT zašifrovaná do ŠT lišící se jen velmi málo např. 50 12 9 87 81 32 21
    50 12 9 6 81 32 21

Homofonní šifra – historické příklady

- Simeon de Crema z Mantovy (1401)

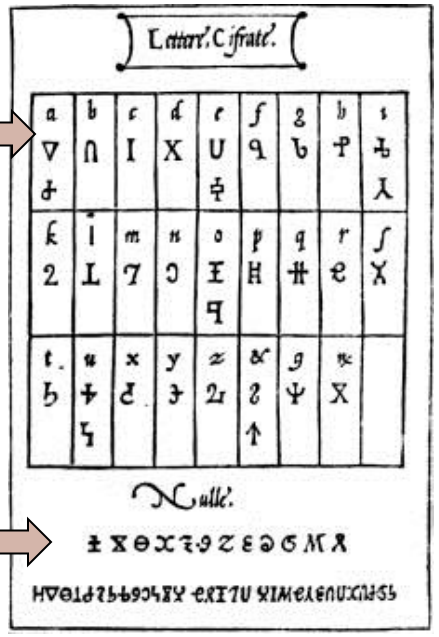


Nomenklátor

- kombinace homofonní monoalfabetické substituce a
  - částečné kódové knihy pro slabiky / jména
  - klamačů
  - symbolů pro zdvojení znaku
- od 14. století
- používán přes 500 let, protože byl
  - jednoduchý
  - rychlý
- kódová kniha měla i stovky až tisíce slov

více reprezentací jednoho znaku

klamače



Homofonní šifra

- I u homofonních šifer lze v ŠT objevit vzory charakteristické pro dvojice/trojice znaků (frekvence digramů/trigramů)

Řešení: 1) polygramové substituce
  - šifrují najednou více znaků OT
  - Playfair, Hillova šifra

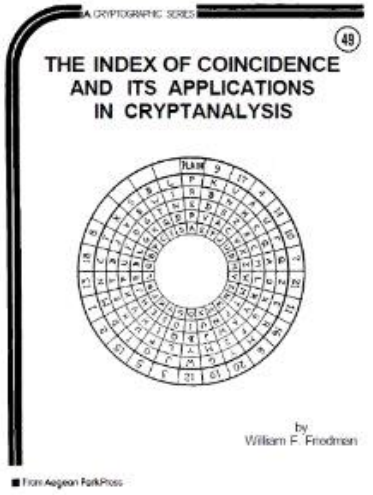
2) polyalfabetické šifry
  - používají více abeced
  - Vigenére, Beaufort

- více na příštím cvičení

Nástroje pro luštění monoalfabetických substitucí

- Frekvenční analýza
  - 9. st. n.l.
  - أبو يوسف يعقوب إبن إسحاق الكندي
    (Abū Yūsuf Ya'qūb ibn Isḥāq al-Ṣabbāḥ al-Kindī)
  - A Manuscript On Deciphering Cryptographic Messages

- Index koincidence
  - 1922
  - William Frederick Friedman
  - The Index of Coincidence And Its Applications in Cryptanalysis



Index koincidence

- IC je míra relativní četnosti znaků v šifrovém textu, neboli odhad pravděpodobnosti, že při náhodném výběru dvou znaků ŠT oba znaky reprezentují stejný znak OT

- Pro abecedu o *c* prvcích platí, že

$$IC = \frac{\sum_{i=1}^{c} n_i(n_i - 1)}{N(N-1)}$$

N délka textu

n_i četnost výskytu znaku *i* v textu délky *N*

c počet znaků abecedy

Index koincidence

- Pro anglickou abecedu (26 znaků) platí,

$$IC = \frac{26\sum_{i=1}^{26} n_i(n_i - 1)}{N(N-1)} \approx \frac{26\sum_{i=1}^{26} n_i^2}{N^2} = \sum_{i=1}^{26} p_i^2 \cong 0,067$$

N délka textu

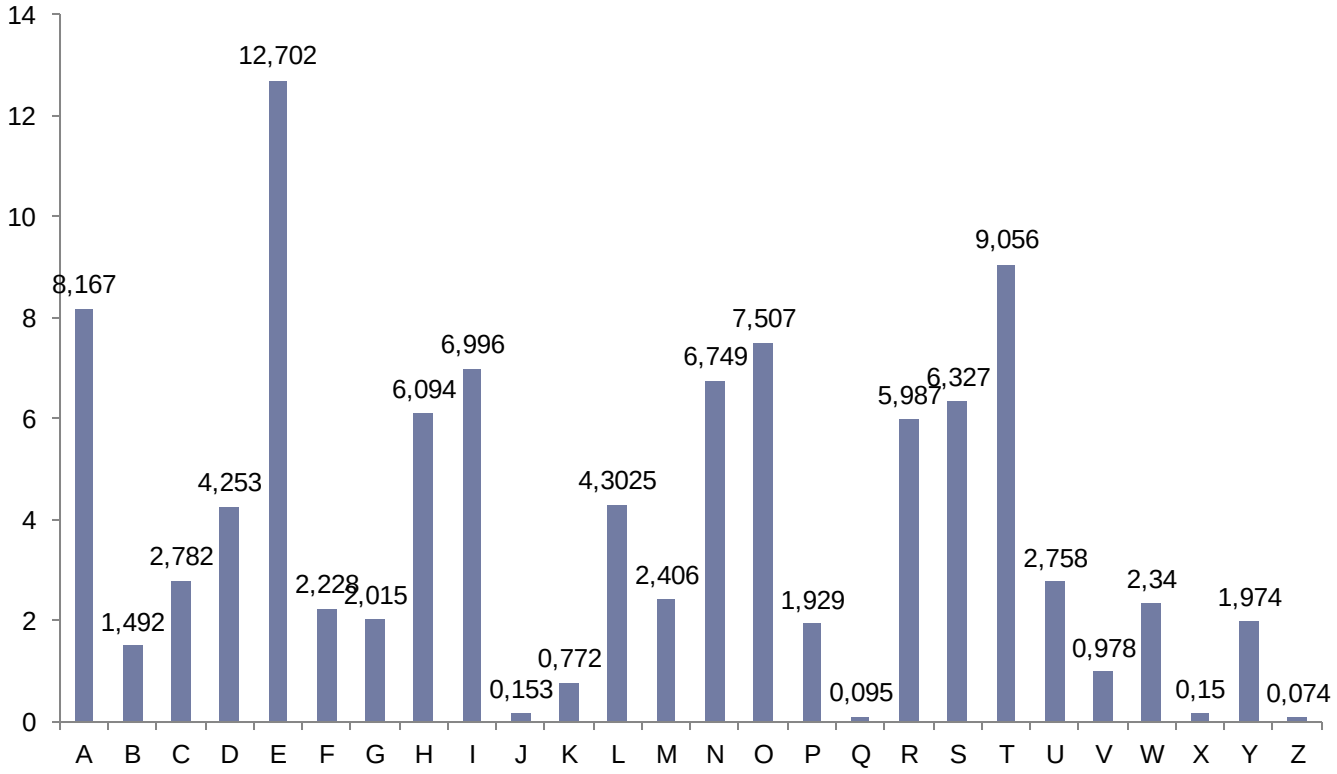
n_i četnost výskytu znaku i v textu délky N

p_i relativní četnost výskytu znaku i v anglické abecedě

- vztah platí pro N >> n
- IC má tuto hodnotu pro OT, ale také ŠT, který vznikl s využitím monoalfabetické substituce a/nebo transpozice
- přesná hodnota IC je mimo způsobu zašifrování ovlivněna také délkou a charakterem analyzovaného textu

Frekvenční analýza

- nástroj pro luštění monoalfabetických substitucí



Rozložení četností výskytu jednotlivých znaků v běžné angličtině

Další frekvenční statistiky jazyka

- kromě analýzy jednotlivých znaků, lze analyzovat i četnost výskytu dvojic (digramy) nebo trojic (trigramy) resp. obecně n-tic (n-gramy)

- nejčastější digramy v AJ
  - TH, HE, IN, ER, AN, RE, ED, ON, ES, ST, EN, AT, TO, NT, HA, ND, OU, EA, NG, AS, OR, TI, IS, ET, IT, AR, TE, SE, HI, OF

- nejčastější trigramy v AJ
  - THE, ING, AND, HER, ERE, ENT, THA, NTH, WAS, ETH, FOR, DTH

Kryptoanalýza monoalfabetických substitučních šifer

Krok č.1 - určit jak byl text zašifrován

1) Pokud IC ~ 0,067 **může** se jednat o monoalfabetickou substituci
2) Frekvenční analýza - pokud jsou v grafickém vyjádření vidět výrazné rozdíly mezi četnostmi výskytu jednotlivých znaků (hodnoty na 10% i pod 2%) a současně rozložení četností neodpovídá otevřenému textu, jedná se o monoalfabetickou substituci

Krok č.2 - dešifrování šifrového textu

1) Prostý posun – pouze 26 kombinací, stačí posunovat ŠT a sledovat odchylky od FA otevřeného anglického textu

Kryptoanalýza monoalfabetických substitučních šifer

2) Afinní šifra - 2 možnosti

a) hrubou silou

existuje pouze 312 možných kombinací → vygenerovat je → nalézt správné řešení (ručně nebo s využitím slovníku)

b) chytrou hlavou – výpočet konstant A,B

1. Z výsledků FA odhadnout, jak se transformovaly dva konkrétní znaky OT (např. E → J a A → X).
2. Sestrojit dvě rovnice $C_i = A \cdot P_i + B \mod 26$ o dvou neznámých (A,B), kde za C_i, P_i se dosadí číselné hodnoty odpovídající hodnotám odhadnutých znaků.
3. Rovnice vyřešit. Pokud rovnice nemají řešení, je nutné se vrátit zpět ke kroku 1 a zvolit jinou transformaci znaků OT na ŠT

Kryptoanalýza monoalfabetických substitučních šifer

3) Monoalfabetická šifra s klíčem

a) napsat nebo si najít program, který text rozluští

- Kvalitní program tento typ šifry na běžném počítači rozluští v řádu jednotek sekund.

b) ručně s využitím FA

- odhad několika znaků s velkými četnostmi výskytu – E,T,A případně ještě I,O,N nebo S
- částečná rekonstrukce textu
- hledání struktur jako je T_E , T_E_E …nalezení H
- postupné odhalování další znaků OT

Dotazy