Kryptografie a počítačová bezpečnost

Kryptoanalytické útoky - kryptoanalýza klasických šifer

Kryptoanalýza (1)

- Kryptoanalytické útoky - útočník zná zpravidla šifrovací algoritmus e a šifrový text c - se liší podle informací, které má útočník k dispozici:
- Ciphertext only attack (COA, útok ze známého šifrového textu nebo také útok se znalostí šifrového textu)
  - útočník zná $c_1=e_k(m_1)$, $c_2=e_k(m_2)$, ..., chce zjistit klíč k, některé m_i (nebo k zvolenému m najít $e_k(m)$).
  - Obvykle se používají různé statistické testy, např. frekvenční analýza.
  - Výhodou útočníka je, pokud má obecnou představu o typu otevřeného textu.
  - Všechny moderní šifrovací algoritmy jsou navrhovány s ohledem na odolnost proti tomuto útoku.
- Další typy útoků později
  - Ciphertext Only
  - Known Plaintext
  - Chosen Plaintext
  - Chosen Ciphertext
  - Chosen Text

Kryptoanalýza (2)

Type of Attack	Known to Cryptanalyst
Ciphertext Only	■ Encryption algorithm ■ Ciphertext
Known Plaintext	■ Encryption algorithm ■ Ciphertext ■ One or more plaintext–ciphertext pairs formed with the secret key
Chosen Plaintext	■ Encryption algorithm ■ Ciphertext ■ Plaintext message chosen by cryptanalyst, together with its corresponding ciphertext generated with the secret key
Chosen Ciphertext	■ Encryption algorithm ■ Ciphertext ■ Ciphertext chosen by cryptanalyst, together with its corresponding decrypted plaintext generated with the secret key
Chosen Text	■ Encryption algorithm ■ Ciphertext ■ Plaintext message chosen by cryptanalyst, together with its corresponding ciphertext generated with the secret key ■ Ciphertext chosen by cryptanalyst, together with its corresponding decrypted plaintext generated with the secret key

Jak určit neznámou šifru? (1)

- Jak určit neznámou šifru, tedy algoritmus pravděpodobně použitý pro šifrování?
- Potřebujeme dostatečně dlouhý šifrový text, alespoň 1000 znaků - velmi krátké ŠT mohou být neprolomitelné, jestliže jejich délka je menší než vzdálenost jednoznačnosti (unicity distance) použité šifry
- Kolik různých znaků ŠT obsahuje?
  a) Pokud pouze 2 různé znaky, je pravděpodobné, že se jedná o Baconovu šifru https://en.wikipedia.org/wiki/Bacon%27s_cipher)
  b) Pokud existuje 5 nebo 6 různých znaků, může to být Polybius nebo obdobná čtvercová šifra, šifra ADFGX nebo ADFGVX.
  c) Pokud ŠT obsahuje více než 26 znaků, pravděpodobně to bude nějaký kód nebo nomenklátor nějakého druhu, nebo homofonní substituční šifra.
  d) Pokud je v ŠT 26 různých znaků, může to vylučovat šifry na základě mřížky 5 x 5, např. Playfair. Pokud je ŠT poměrně dlouhý a má pouze 25 znaků, může to naopak znamenat, že byla použita např. šifra tohoto typu.

Jak určit neznámou šifru? (2)

- Rozlišení transpozičních a ostatních šifer:
  a) Pomocí monogramových frekvencí (distribuce četností znaků)
     a) Nezmění se pomocí transpozice, všechny ostatní šifry mění četnosti znaků.
     b) Pokud distribuce frekvence vypadá přesně jako pro kus např. anglického textu, ale je stále nečitelná, můžeme konstatovat, že je pravděpodobně transpoziční šifra, jinak se přesuneme na další krok.
  b) Dalším krokem je zjistit, zda je šifra substituční šifrou nějakého druhu. Zde vypočteme **index koincidence** (IC - pokud je text podobný angličtině, bude mít IC přibližně 0,06, pokud jsou znaky rozloženy rovnoměrně, bude IC blíže k 0,03 - 0,04.).
     a) Pokud je IC přibližně 0,06, je šifra pravděpodobně substituční šifra.
     b) Je-li nižší, je s největší pravděpodobností nějaká polyalfabetická, polygramová nebo složitější šifra.
  c) Je-li to polyalafabetická šifra, např. Vigenerova, musíme spočítat IC pro různé délky klíče, pro násobky délky klíče obdržíme vysoké hodnoty IC. Žádné další šifry nemají tuto vlastnost.
  d) Je-li šifra polygramová, délka ŠT musí být násobkem délky skupiny znaků, kterou používáme pro substituci. Např. pokud má ŠT lichý počet znaků, nemohla být použita bigramová šifra jako např. Playfair. Pokud např. délka ŠT není násobkem 3, nemůže to být 3x3 Hillova šifra atd.

Index konicidence

- Index koincidence – k určení v jakém přirozeném jazyce je text napsaný nebo k odhadu zda-li byla použita monoalfabetická nebo polyalfabetická substituce.

Náhodně generovaná slova	1/26 = 0.03846 (pro anglickou abecedu)
Čeština/slovenština	0.06027
Angličtina	0.06689
Francouzština	0.07460
Holandština	0.07981
Němčina	0.07667
Italština	0.07329
Ruština	0.05607
Španělština	0.07661

Index konicidence

- Index koincidence vyjadřuje celkovou pravděpodobnost toho, že při náhodném výběru dvou znaků z celého textu budou tyto dva znaky stejné.
- Pokud je v našem šifrovém textu N znaků, existuje celkem N(N-1)/2 způsobů výběru dvojice znaků. Šance na to, že tyto dva znaky budou stejné == celkový počet způsobů výběru 2 stejných znaků, dělený celkovým počtem způsobů výběru dvou znaků.

$$I.C. = \frac{\sum_{i=A}^{i=Z} f_i(f_i - 1)}{N(N-1)}$$

where f_i is the count of letter i (where i = A,B,...,Z) in the ciphertext, and N is the total number of letters in the ciphertext.

- Viz např. https://www.matweb.cz/friedmanuv-test/ , http://practicalcryptography.com/cryptanalysis/text-characterisation/index-coincidence/

Luštění Vigenerovy šifry (1)

- Otestovat, zda by mohla být použita polyalfabetická substituce – pomocí I.C. - je-li IC nižší, než IC předpokládaného přirozeného jazyka, pravděpodobně byla použita polyalfabetická, polygramová nebo složitější šifra.
- Hypotéza – Vigenere → určíme pravděpodobnou délku klíče
  - Kasiského test - Určíme počet znaků ŠT. Najdeme opakované n-gramy (čím větší n, tím je Vig. Š. bezpečnější), určit vzdálenost mezi jejich jednotlivými výskyty. Je-li tato vzdálenost ve většině případů násobkem jednoho celého čísla N, budeme o tomto čísle N uvažovat jako o délce klíče
    - viz VigenerPr1.pdf, VigenerPr2.pdf, https://www.matweb.cz/kasiskeho-test/
    - Nebo použít IC (Friedmanův test). Pro předpokládané délky klíčů |K| rozdělit text do |K| sloupců. Spočítat IC pro každý z |K| sloupců a zprůměrovat. Pro násobky „správné“ délky klíče obdržíme vyšší hodnoty IC
    - Viz http://practicalcryptography.com/cryptanalysis/stochastic-searching/cryptanalysis-vigenere-cipher/

period	avg I.C.
1 :	0.0449443523561
2 :	0.0457833618884
3 :	0.0435885364312
4 :	0.0474962292609
5 :	0.0393612078978
6 :	0.0471437059672
7 :	0.0909922589726
8 :	0.0461858974359
9 :	0.0407804755631
10:	0.0361152882206
11:	0.0491603339901
12:	0.0512663398693
13:	0.0446886446886
14:	0.0988487702773
15:	0.0334554334554

Luštění Vigenerovy šifry (2)

- Rozdělit ŠT do N (resp. |K|) sloupců a rozhodnout, jaká posunutí abecedy byla použita v jednotlivých sloupcích.
- Provést frekvenční analýzu pro každý z N sloupců, resp. kryptoanalýza Shift šifry (několika Shift šifer), např. viz VigenerPr1.pdf, VigenerPr2.pdf
- Nebo např. použít Chí-kvadrát test, který je měřítkem toho, jak podobná jsou dvě diskrétní rozdělení pravděpodobnosti. Pokud jsou obě distribuce shodné, je chí-kvadrát 0, jestliže jsou distribuce odlišné, bude výsledkem vyšší číslo.

$$\chi^2(C,E) = \sum_{i=A}^{i=Z} \frac{(C_i - E_i)^2}{E_i}$$

kde např. C_A je počet (ne pravděpodobnost!) písmene A a E_A je očekávaný počet písmene A (vzhledem k délce šifrového textu a četnosti znaku v přirozeném jazyce).

- Viz http://practicalcryptography.com/cryptanalysis/text-characterisation/chi-squared-statistic/

Kryptoanalýza sloupcové transpozice (1)

- Určení rozměru tabulky:
  - Spočteme délku šifrového textu a snažíme se určit pravděpodobný rozměr tabulky. Ten zjistíme tak, že délku šifrového textu rozložíme na součin prvočísel (prvočíselných dělitelů) a z nich kombinujeme pravděpodobnou velikost tabulky. Máme-li např. šifrový text délky 120 (120= 2*2*2*3*5), pak jsou možné následující velikosti tabulek :
- Tabulky (počet sloupců * počet řádků)
- málo pravděpodobné (bylo by příliš lehké k řešení): 1*120, 2*60 , 3*40, 4*30, 6*20
- složité: 120*1, 60* 2, 60*2
- tabulky: 8*15, 15*8, 12*10, 10*12, 20*6, 30*4, 40*3

Kryptoanalýza sloupcové transpozice (2)

- Příklad
  - `OTSEC NCNUX ATONO TOUTO KXUJU AILBX UVPTD HSEOL KYREN EPSUK ZELID RZPAU` (60 znaků)
  - Určení velikosti tabulky
  - ne : 1*60 , 2*30, 3*20, 4*15,
  - možné tabulky : 15*4, 20*3, 10*6, 6*10
- Poměr samohlásek a souhlásek v češtině je 40:60

Kryptoanalýza sloupcové transpozice (3)

- rozměr 20*3, očekávaný poměr 8:12

```
OECXOTTXULUTSLREUEDP 8:12
TCNANOOUABVDEKEPKLRA 8:12
SNUTOUKJIXPHOYNSZIZU 8:12
```

- rozměr 15*4, očekávaný poměr 6:9

```
OCUOOKUBPSKNULZ 6:9
TNXNUXAXTEYEKIP 6:9
SCAOTUIUDORPZDA 7:8
ENTTOJLVHLESERU 5:10
```

Kryptoanalýza sloupcové transpozice (4)

- rozměr 10*6, očekávaný poměr 4:6

```
OCOTUUSRUD 5:5
TNNOAVEEKR 4:6
SUOKIPONZZ 4:6
EXTXLTLEEP 3:7
CAOUBDKPLA 4:6
NTUJXHYSIU 4:6
```

- rozměr 6*10, očekávaný poměr 2,4 : 3,6

```
OAKUKZ 3:3
TTXVYE 2:4
SOUPRL 2:4
ENJTEI 3:3
COUDND 2:4
NTAHER 2:4
COISPZ 2:4
NULESP 2:4
UTBOUA 4:2
XOXLKU 2:4
```

Kryptoanalýza sloupcové transpozice (5)

- Nejpravděpodobnějšími tabulkami jsou tabulky rozměrů 20*3 a 6*10, následují rozměry 10*6 a nejhůře z testu vyšel rozměr 15*4.
- Správný rozměr je 6*10. Vzhledem k malému počtu sloupců již není problém je správně seřadit a dostaneme příslušný otevřený text.

OAKUKZ		UKAZKO
TTXVYE		VYTEXT
SOUPRL		PROLUS
ENJTEI		TENIJE
COUDND	→	DNODUC
NTAHER		HETRAN
COISPZ		SPOZIC
NULESP		ESUPLN
UTBOUA		OUTABU
XOXLKU		LKOUXX

- Hledaným textem je :
  UKAZKOVY TEXT PRO LUSTENI JEDNODUCHE TRANSPOZICE S UPLNOU TABULKOU XX

Teoretické základy

- Koncept moderní kryptografie navrhli C. Shannon a H. Feistel.
  - Claude Shannon:
    - „Communication Theory of Secrecy Systems“, Bell System Technical Journal, Oct 1949,
    - „Prediction and Entropy of printed English“, Bell System Technical Journal, Jan 1951.
- Koncept:
  - entropie, redundance jazyka (redundance ve zprávě je dostačující k jejímu prolomení),
  - teorie o tom kolik informace je třeba pro zlomení šifrového textu - stanovil teoretickou míru bezpečnosti šifry pomocí neurčitosti otevřeného textu, když je dán šifrový text.
  - https://en.wikipedia.org/wiki/Entropy_(information_theory)
  - http://home.zcu.cz/~vais/TIaK.pdf (strana 5-7)
  - https://en.wikipedia.org/wiki/Redundancy_(information_theory)

Entropie (1)

- Entropie (také Shannonova entropie) je míra neurčitosti zprávy a vyjadřuje **průměrnou** informační hodnotu jedné zprávy X z daného zdroje (zhruba řečeno vyjadřuje, kolik informace je v nějaké zprávě).
- Nechť X je diskrétní náhodná proměnná s možnými hodnotami $x_1, x_2, \ldots, x_n$, které se vyskytují s pravděpodobností $p_1, p_2, \ldots, p_n$, pak je entropie náhodné proměnné X definována jako:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Př.: Mějme čtyři zprávy červená, žlutá, zelená a bílá (barvy odjezdového návěstidla na nádraží), ty mají pravděpodobnost: $\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{4}$, potom entropie je rovna:

$$H(X) = -(\tfrac{1}{4}\log_2\tfrac{1}{4} + \tfrac{1}{2}\log_2\tfrac{1}{2} + \tfrac{1}{8}\log_2\tfrac{1}{8} + \tfrac{1}{4}\log_2\tfrac{1}{4}) =$$
$$-(\tfrac{1}{4}\cdot(-2) + \tfrac{1}{2}\cdot(-1) + \tfrac{1}{8}\cdot(-3) + \tfrac{1}{4}\cdot(-2)) = 1,75$$

Entropie (2)

- Entropie H(M) - **množství informace ve zprávě** – min počet bitů potřebných k „zakódování“, uložení všech možných významů (hodnot) zprávy **za předpokladu, že všechny významy jsou stejně pravděpodobné**.
- Entropie se měří **průměrným** počtem bitů, nezbytných k zakódování zprávy (při optimálním kódování, tj. optimální kódování používá co nejméně bitů k zakódování zpráv)
  - Tj. je to množství informace ve zprávě M, měřené v bitech (proto $\log_2 n$, logaritmus o základu 2), kde n je počet možných významů, tj.

    $$H(M) = \log_2(n)$$

  - Maximální entropie nastává, pokud mají všechny zprávy stejnou pravděpodobnost: $H(M) = \log_2(n)$
  - Minimální nastává, pokud přijde pouze jedna zpráva a má pravděpodobnost 1, potom H(M) = 0
- Příklad - reprezentace dnů v týdnu ve 3 bitech:
  - 000 pondělí, 001 úterý, …110 neděle, 111 nevyužito → 8 významů
  - tj. $H(M) = \log_2 8 = 3$ → entropie je < 3 bity,
  - Kdybychom repr. dny v týdnu 2 znaky (Po, Ut, St, …) * 8 bitů, spotřebujeme 16 bitů, ale jsou v nich obsaženy jen méně než 3 bity informace.

Obsažnost jazyka

- **Rate** r jazyka (obsažnost jazyka vzhledem k 1 znaku) je průměrná entropie (v bitech) na znak ve zprávě

  $r = H(M) / N$, kde N je délka zprávy

- Absolutní rate (obsažnost) R je maximální možná obsažnost jazyka o L stejně pravděpodobných znacích , tedy průměrná entropie znaku pokud by byly všechny zprávy a znaky stejně pravděpodobné

  $R = \log_2 L$, kde L je počet znaků v abecedě (např 26 pro anglickou abecedu bez mezery).

- Přirozený jazyk jí nedosahuje!
- S rostoucím N obsažnost přirozeného jazyka pro zprávy délky N klesá.
- V limitě se blíží nějaké konstantě r - obsažnost jazyka vzhledem k jednomu znaku.

Redundance

- **Redundance D** (nadbytečnost) - nadbytečnost jazyka vzhledem k jednomu písmenu jazyka (to, co **není** nezbytně nutné k přenosu informace) :

$$D = R - r$$

- Angličtina:
  - rate r je cca 1<r<1.5 bitů/znak pro velká N
  - Absolutní rate R je $R = \log_2 26 \cong 4.7004$ bitů/znak
  - Redundance D = R - r = 4.7 -1.5 = 3.2 b/znak pro r=1.5
- Příklad:
  - zpráva, ASCII znaky – z 8 bitů jen 1.5 nese užitečnou informaci a 6.5 b je nadbytečných
  - redundance na bit ASCII textu je $D = 6.5 / 8 \cong 0.82$ b a entropie je 0.18 b informace na každý ASCII bit

Vzdálenost jednoznačnosti (1)

- Všechny klasické šifry (s výjimkou Vernamovy šifry) mají tu vlastnost, že čím delší je šifrový text, tím snazší je šifru rozluštit. Shannonova teorie vysvětluje, proč tomu tak je.
- Intuitivně můžeme posoudit, proč delší šifrový text poskytuje více informací o otevřeném textu.
  - Použijeme-li monoalfabetickou substituci, pak ze šifrového textu o délce jednoho písmene např. ‘A’ nemůžeme usoudit vůbec nic o příslušném otevřeném textu.
  - Ze šifrového textu o dvou písmenech ‘AB’ už můžeme alespoň usoudit, že příslušný otevřený text není složený ze dvou stejných písmen.
  - Má-li smysluplný otevřený text v přirozeném jazyce délku např. 500 písmen, pak si lze těžko představit, že by mohla existovat nějaká permutace písmen, která by jej opět proměnila v **jiný smysluplný** otevřený text v přirozeném jazyce.
  - Proto k šifrovému textu o 500 písmenech může existovat nejvýše jeden smysluplný text v přirozeném jazyce, který nějakou jednoduchou záměnou vede k tomuto šifrovému textu.

Vzdálenost jednoznačnosti (2)

- Nejmenší délka šifrového textu, ke kterému existuje jednoznačně určený smysluplný otevřený text, se nazývá **vzdálenost jednoznačnosti** (Unicity Distance U)
- U je pro každou šifru (každý algoritmus) jiná
- Vzdálenost jednoznačnosti

  $U = H(K) / D$, kde $H(K)$ je entropie klíče a D redundance jazyka
- Pokud otevřený text neobsahuje žádnou redundanci, pak je vzdálenost jednoznačnosti nekonečná; to znamená, že systém je teoreticky nezlomitelný útokem ze známého šifrového textu.

Příklad

- Mějme algoritmus AES (budeme diskutovat později) s délkou klíče 256 bitů. Víme, že otevřený text je v angličtině, v kódování ASCII 8-bit.
- Kolik znaků šifrového textu musíme získat, aby tomu odpovídal jeden smysluplný otevřený text?
- $H(K) = \log_2(2^{256}) = 256$ je entropie klíče, pokud má klíč délku 256 bitů
- Nechť redundance je D = 6.5 bitu na bajt
- U = H(K) /D = 256/6.5 = 39.3 bajtů
- Stačí nám cca 39 (pro 8 bit ASCII) znaků šifrového textu, abychom věděli, že získáme jediný smysluplný otevřený text.