

České vysoké učení technické v Praze
Fakulta elektrotechnická
Katedra telekomunikační techniky

A7B32KBE - Teorie informace

Ing. Tomáš Vaněk, Ph.D.





Claude E. Shannon - Teorie informace

1948

A Mathematical Theory of Communication, The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948

<http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>

- 1949

Claude E. Shannon, *Communication Theory of Secrecy Systems*, Bell System Technical Journal, vol.28-4, pp. 656 - 715, 1949,

<http://www.cs.ucla.edu/~jkong/research/security/shannon1949.pdf>

- zásadní práce pro moderní kryptoanalýzu
 - důkaz, že Vernamova šifra je zcela bezpečný kryptosystém (i teoreticky neprolomitelný)
 - zavedl pojmy konfúze a difúze
-



Shannova věta

„The enemy knows the system.“

- jiná formulace Kerchoffova principu
- dnes (doufejme) standardní předpoklad dobrých kryptosystémů
- opačný pól k přístupu „security through obscurity“



Shannonův zjednodušený model šifrovacího systému

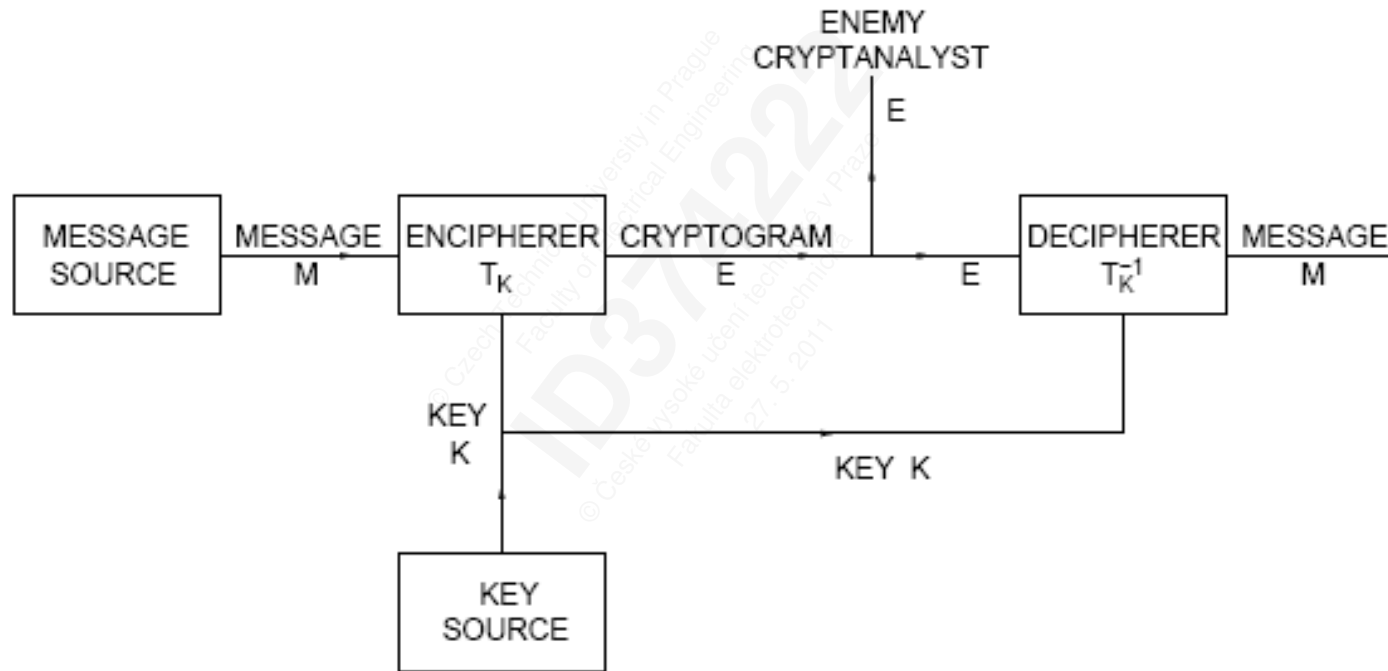


Fig. 1. Schematic of a general secrecy system



Teorie informace

- **Konfuze** - zavádí složité vztahy mezi jednotlivými znaky OT, ŠT a klíče.
Realizuje se pomocí substituce
- **Difúze** – každý bit OT ovlivňuje každý bit ŠT; rozprostírá statistické charakteristiky jednotlivých znaků (dvojic, trojic i větších skupin) napříč celým textem.
Realizuje se pomocí transpozice.
- **Iterace** - důkladné promíchání bitů; V moderních algoritmech jsou konfuze a difúze používány opakovaně – jedna aplikace nestačí (z hlediska bezpečnosti)



Teorie informace

Součinnové šifry (Product cipher)

- Shannon navrhl metodu vytváření šifer opakovaným skládáním (součinem) několika šifer různých typů
- kombinace různé konfúze a difúze

Iterované šifry

- v moderních algoritmech jedna aplikace konfúze a difúze nestačí (z hlediska bezpečnosti)
- řešení - opakovaná aplikace konfúze a difúze
- cíl - důkladné promíchání bitů
- iterované šifry
- **runda** (stejně dílčí šifry v jednotlivých iteracích)



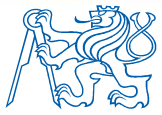
Teorie informace

Teorie informace definuje jak změřit množství informace, obsažené ve zprávě.

Definice „množství informace“ ve zprávě:

Množství informace ve zprávě je dáno minimálním počtem dvojkových signálových prvků (bitů) potřebných k zakódování všech možných stavů této zprávy, za předpokladu, že výskyt každého stavu má stejnou pravděpodobnost.

Množství informace je vyjádřeno pomocí **entropie** zprávy.

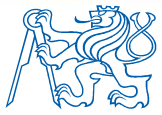


Entropie

- Mějme nějaký zdroj zpráv M , který generující různé zprávy m_1, m_2, \dots, x_m .
- Pravděpodobnost, že zdroj M vydá zprávu m_i si označíme p_i .
- Entropii zprávy ze zdroje X definujeme jako číslo
$$H(M) = \sum_{i=1, \dots, n} -p_i \log_2(p_i).$$
- Číslo $\log_2(p_i)$ je počet bitů, který je třeba k zakódování zprávy m_i , které potřebuje nejmenší možný počet bitů k zakódování všech zpráv ze zdroje M .
- Optimální zakódování bere v úvahu pravděpodobnosti p_i jednotlivých zpráv generovaných zdrojem M .

Entropie zprávy - příklad

- Jakou entropii má zdroj generující jednu zprávu x s pravděpodobností 1 ?
- Jakou entropii má zdroj, který vysílá dvě zprávy x_1, x_2 , každou s pravděpodobností $\frac{1}{2}$?
- Zdroj generující dvě zprávy x_1, x_2 jednu s pravděpodobností 0,1, druhou s pravděpodobností 0,9 ?



Entropie jazyka

- Mějme zdroj zpráv, který generuje zprávy délky k bitů, všechny se stejnou pravděpodobností.
- Zpráv je tedy celkem 2^k a každá má pravděpodobnost 2^{-k}
- Entropie tohoto zdroje je potom
$$\sum_{i=1, \dots, 2^k} (-p_i) \log_2(p_i) = 2^k (-2^{-k}) \log_2(2^{-k}) = k.$$
- Každá zpráva z takového zdroje tak nese informaci k bitů.
- Obsažnost jazyka pro zprávy délky N (=množství informace ve zprávě) definujeme jako průměrnou entropii na jeden znak zprávy, tj. jako číslo

$$r_N = \frac{H(M)}{N}$$



Maximální entropie jazyka

- Má-li jazyk zpráv L stejně pravděpodobných znaků a všechny zprávy jsou stejně pravděpodobné, pak entropie zdroje zpráv délky N v tomto jazyce je

$$L^N (-L^{-N}) \log_2 (L^{-N}) = N \log_2 L$$

a obsažnost jazyka zpráv pro zprávy délky N je tedy $\log_2 L$.

- Toto je **teoretická maximální** možná obsažnost jazyka o L stejně pravděpodobných znacích.
- Maximální entropie jazyka se označuje R .
Přirozený jazyk jí nikdy nedosahuje.
- Max. entropie angličtiny ($L=26$) pro zprávy délky N je

$$R = \log_2 L = \log_2 26 = 4,7 \text{ bit/symbol}$$



Obsažnost a nadbytečnost jazyka

- Proč se nedosahuje maximální obsažnosti (množství informace ve zprávě):
 - jednotlivá písmena nejsou stejně pravděpodobná
 - různé posloupnosti písmen délky N také nejsou stejně pravděpodobné
- Když vygeneruji zprávy:
 - a) *Dnes je úter*
 - b) *Dnes je úterý*tj. prodloužení o jeden znak, nezvýšíme entropii této zprávy o $\log_2 L$ bitů, protože množství informace uložené ve zprávě se nezmění !



Redundance

- Pro rostoucí N obsažnost přirozeného jazyka pro zprávy délky N klesá.
- Pro $N \rightarrow \infty$ se limitně blíží konstantě r .
- Tuto konstantu r nazýváme ***průměrná obsažnost (entropie) jazyka*** vzhledem k jednomu znaku.
- Udává průměrný počet bitů informace obsaženém v jednom znaku jazyka.
- **Průměrná entropie angličtiny je 1,3-1,5 bit/symbol**
- Rozdíl $D = R - r$ pak nazýváme ***nadbytečnost (redundance) jazyka*** vzhledem k jednomu znaku(symbolu).



Entropie a kryptologie

Pro angličtinu tedy platí: $L = 26$, $R = 4,7$ $r = 1,5$ $D = 3,2$.

Poměr D / R udává v procentech, kolik bitů jazyka je nadbytečných.

Entropie kryptosystému $H(K) = \log_2 K$ K - počet klíčů

Jednotkový odstup (Unity distance) $U = H(K)/D$

Odhad minimálního množství šifrovaného textu nutného k tomu, aby výsledkem luštění byl metodou totálních zkoušek byl pravděpodobně pouze jedná smysluplný otevřený text.



Teorie informace

Delší klíč → větší entropie kryptosystému →
větší jednotkový odstup → bezpečnější kryptosystém.

Délka klíče [b]	Jednotkový odstup [B]
40	5,9
56	8,3
64	9,5
128	19,1
256	38,2

Př.: $U = H(K)/D = \log_2(2^{128})/(8-1,3) = 128/6.7 = 19.7$

Právní doložka (licence) k tomuto Dílu (elektronický materiál)

České vysoké učení technické v Praze (dále jen ČVUT) je ve smyslu autorského zákona vykonavatelem majetkových práv k Dílu či držitelem licence k užití Díla. Užívat Dílo smí pouze student nebo zaměstnanec ČVUT (dále jen Uživatel), a to za podmínek dále uvedených.

ČVUT poskytuje podle autorského zákona, v platném znění, oprávnění k užití tohoto Díla pouze Uživateli a pouze ke studijním nebo pedagogickým účelům na ČVUT. Toto Dílo ani jeho část nesmí být dále šířena (elektronicky, tiskově, vizuálně, audiem a jiným způsobem), rozmnožována (elektronicky, tiskově, vizuálně, audiem a jiným způsobem), využívána na školení, a to ani jako doplňkový materiál. Dílo nebo jeho část nesmí být bez souhlasu ČVUT využívána ke komerčním účelům. Uživateli je povoleno ponechat si Dílo i po skončení studia či pedagogické činnosti na ČVUT, výhradně pro vlastní osobní potřebu. Tím není dotčeno právo zákazu výše zmíněného užití Díla bez souhlasu ČVUT. Současně není dovoleno jakýmkoliv způsobem manipulovat s obsahem materiálu, zejména měnit jeho obsah včetně elektronických popisných dat, odstraňovat nebo měnit zabezpečení včetně vodoznaku a odstraňovat nebo měnit tyto licenční podmínky.

V případě, že Uživatel nebo jiná osoba, která drží toto Dílo (Držitel díla), nesouhlasí s touto licencí, nebo je touto licencí vyloučena z užití Díla, je jeho povinností zdržet se užívání Díla a je povinen toto Dílo trvale odstranit včetně veškerých kopií (elektronické, tiskové, vizuální, audio a zhotovených jiným způsobem) z elektronického zařízení a všech záznamových zařízení, na které jej Držitel díla umístil.