Vzdálenost jednoznačnosti a absolutně bezpečné šifry

Andrew Kozlík

KA MFF UK

Značení

Pracujeme s šifrou $(\mathcal{P}, \mathcal{C}, \mathcal{K}, E, D)$, kde

- $\mathcal{P}$ je množina otevřených textů,
- $\mathcal{C}$ je množina šifrových textů,
- $\mathcal{K}$ je množina klíčů,
- $E : \mathcal{K} \times \mathcal{P} \to \mathcal{C}$ je šifrovací algoritmus,
- $D : \mathcal{K} \times \mathcal{C} \to \mathcal{P}$ je dešifrovací algoritmus.

Množinám $\mathcal{K}$, $\mathcal{P}$ a $\mathcal{C}$ odpovídají náhodné veličiny:

- $\mathbf{K} : \Omega \to \mathcal{K}$ (pravděpodobnostní rozdělení klíčů),
- $\mathbf{P} : \Omega \to \mathcal{P}$ (pravděpodobnostní rozdělení otevřených textů),
- $\mathbf{C} : \Omega \to \mathcal{C}$ (pravděpodobnostní rozdělení šifrových textů).

Předpokládáme, že veličiny $\mathbf{K}$ a $\mathbf{P}$ jsou nezávislé.

Tvrzení

$$\mathrm{H}(\mathbf{K} \mid \mathbf{C}) = \mathrm{H}(\mathbf{K}) + \mathrm{H}(\mathbf{P}) - \mathrm{H}(\mathbf{C}).$$

Důkaz.

Výraz $\mathrm{H}(\mathbf{K}, \mathbf{P}, \mathbf{C})$ rozepíšeme dvěma způsoby:

$$\begin{aligned}\mathrm{H}(\mathbf{K}, \mathbf{P}, \mathbf{C}) &= \mathrm{H}(\mathbf{P}, (\mathbf{K}, \mathbf{C})) = \overbrace{\mathrm{H}(\mathbf{P} \mid (\mathbf{K}, \mathbf{C}))}^{0} + \mathrm{H}(\mathbf{K}, \mathbf{C}) \\ &= \mathrm{H}(\mathbf{K} \mid \mathbf{C}) + \mathrm{H}(\mathbf{C}),\end{aligned}$$

$$\begin{aligned}\mathrm{H}(\mathbf{K}, \mathbf{P}, \mathbf{C}) &= \mathrm{H}(\mathbf{C}, (\mathbf{K}, \mathbf{P})) = \overbrace{\mathrm{H}(\mathbf{C} \mid (\mathbf{K}, \mathbf{P}))}^{0} + \mathrm{H}(\mathbf{K}, \mathbf{P}) \\ &= \mathrm{H}(\mathbf{K}) + \mathrm{H}(\mathbf{P}).\end{aligned}$$

Podle předpokladu jsou $\mathbf{K}$ a $\mathbf{P}$ nezávislé. □

Příklad

- Máme šifrový text „WNAJW“.
- Vznikl zašifrováním anglického slova Caesarovou šifrou.
- Průchodem všech 26 klíčů najdeme jen 2 možnosti:
  - „river“ ($klíč = +5$),
  - „arena“ ($klíč = +22$).
- Využíváme redundanci lidského jazyka:
  Písmena otevřeného textu nejsou nezávislé náhodné veličiny.
- Čím víc šifrového textu máme, tím víc klíčů umíme vyloučit.
- Kolik šifrového textu je v průměru třeba, abychom mohli klíč určit jednoznačně?

Definice

Nechť $(\mathcal{P}, \mathcal{C}, \mathcal{K}, E, D)$ je šifra.

- Definujeme náhodnou veličinu s pravděpodobnostním rozdělením, které odpovídá n po sobě jdoucím blokům otevřeného textu:

$$\mathbf{P}^n : \Omega \to \mathcal{P}^n.$$

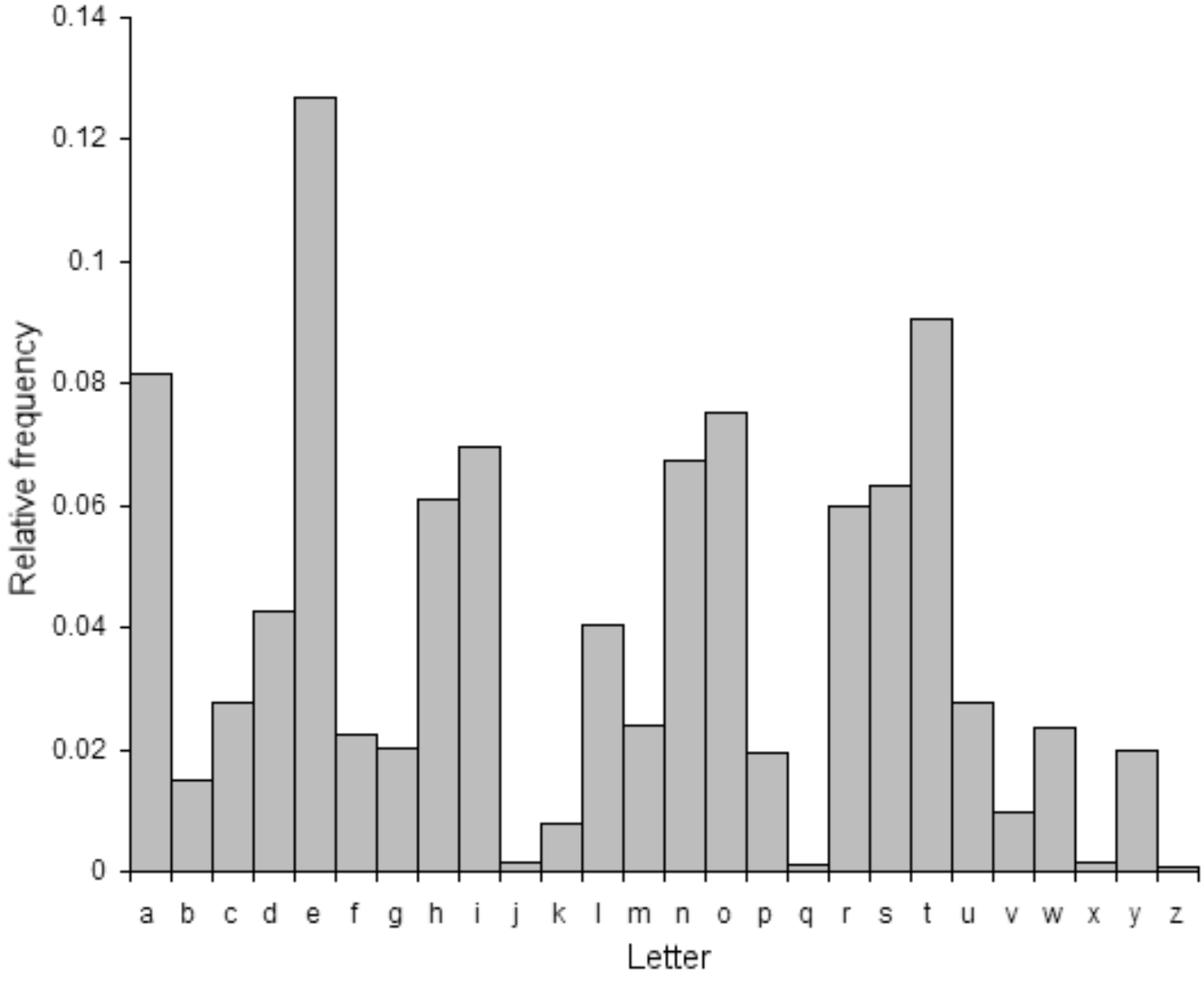
- Pro každé $y \in \mathcal{C}^n$ definujeme množinu všech možných klíčů

$$\mathcal{K}(y) := \left\{ k \in \mathcal{K} : \Pr\left(\mathbf{P}^n{=}D(k, y)\right) > 0 \right\}.$$

Příklad

- V případě Caesarovy šifry máme $\mathcal{K}(\text{„WNAJW“}) = \{5, 22\}$.
- Bloky otevřeného textu mohou být například písmena.

Četnost písmen v anglickém jazyce



Četnost nejčastějších bigramů v anglickém jazyce v %

TH	2,71	RE	1,41	EN	1,13	OR	1,06	NG	0,89
HE	2,33	ES	1,32	AT	1,12	EA	1,00	AL	0,88
IN	2,03	ON	1,32	ED	1,08	TI	0,99	IT	0,88
ER	1,78	ST	1,25	ND	1,07	AR	0,98	AS	0,87
AN	1,61	NT	1,17	TO	1,07	TE	0,98	IS	0,86

Četnost nejčastějších trigramů v anglickém jazyce v %

THE	1,81	HER	0,36	ERE	0,31	ATI	0,26
AND	0,73	FOR	0,34	TIO	0,31	HAT	0,26
ING	0,72	THA	0,33	TER	0,30	ATE	0,25
ENT	0,42	NTH	0,33	EST	0,28	ALL	0,25
ION	0,42	INT	0,32	ERS	0,28	ETH	0,24

Redundance jazyka

Kolik informace je v jednom písmeně jazyka?

- Jazyk s nulovou redundancí: $\log_2 26 \approx 4{,}70$.
- Anglický jazyk v 1. přiblížení:
  Podle pravděpodobnostního rozdělení písmen

$$\mathrm{H}(\mathbf{P}) \approx 4{,}19.$$

- Anglický jazyk ve 2. přiblížení:
  Podle pravděpodobnostního rozdělení bigramů

$$\tfrac{1}{2}\,\mathrm{H}(\mathbf{P}^2) \approx 3{,}90.$$

  Průměrné množství informace v jednom písmeni bigramu.
- Hodnota $\frac{1}{n}\,\mathrm{H}(\mathbf{P}^n)$ klesá s rostoucím n.
  (S rozšiřujícím se kontextem klesá informační hodnota písmene.)

Definice

Nechť

- L je jazyk,
- $\mathbf{P}^n$ je náhodná veličina, jejíž rozdělení odpovídá n-gramům v jazyce L.

Entropii jazyka L definujeme jako

$$\mathrm{H}_L := \lim_{n\to\infty} \frac{\mathrm{H}(\mathbf{P}^n)}{n}.$$

Redundanci jazyka L definujeme jako

$$R_L := 1 - \frac{\mathrm{H}_L}{\log_2|\mathcal{P}|}.$$

Definice

Nechť

- L je jazyk,
- $\mathbf{P}^n$ je náhodná veličina, jejíž rozdělení odpovídá n-gramům v jazyce L.

Entropii jazyka L definujeme jako

$$\mathrm{H}_L := \lim_{n\to\infty} \frac{\mathrm{H}(\mathbf{P}^n)}{n}.$$

Redundanci jazyka L definujeme jako

$$R_L := 1 - \frac{\mathrm{H}_L}{\log_2|\mathcal{P}|}.$$

počet bitů v jednom znaku

maximální možný počet bitů v jednom znaku

% nevyužitého prostoru

Statistiky pro anglický jazyk

$1{,}0 \leq H_L \leq 1{,}5$

- Každé písmeno nese pouze cca 1,25 bitu informace.

$0{,}68 \leq R_L \leq 0{,}79$

- V anglickém textu je 75 % znaků informačně zbytečných.
- Přesněji, pro dostatečně velké n existuje kódování n-gramů, kterým lze v průměru docílit o 75 % kratšího textu.

Věta

Nechť

- $(\mathcal{P}, \mathcal{C}, \mathcal{K}, E, D)$ *je šifra,*
- R_L *je redundance jazyka otevřených textů,*
- $\mathbf{K}$ *má rovnoměrné rozdělení.*

Pak střední hodnota počtu možných klíčů pro šifrový text délky n bloků je

$$\mathrm{E}|\mathcal{K}(\mathbf{C}^n)| \geq \frac{|\mathcal{K}| \cdot |\mathcal{P}|^{n(1-R_L)}}{|\mathcal{C}|^n}.$$

Důkaz.

- Podle věty o maximální entropii je

$$\mathrm{H}(\mathbf{C}^n) \leq \log_2|\mathcal{C}^n| = \log_2|\mathcal{C}|^n.$$

- Posloupnost $\frac{1}{n}\mathrm{H}(\mathbf{P}^n)$ klesá k hodnotě H_L, čili

$$\tfrac{1}{n}\mathrm{H}(\mathbf{P}^n) \geq \mathrm{H}_L = (1-R_L)\log_2|\mathcal{P}|.$$

- Veličina $\mathbf{K}$ má rovnoměrné rozdělení:

$$\mathrm{H}(\mathbf{K}) = \log_2|\mathcal{K}|.$$

- Už jsme dokázali

$$\begin{aligned}\mathrm{H}(\mathbf{K} \mid \mathbf{C}^n) &= \mathrm{H}(\mathbf{K}) + \mathrm{H}(\mathbf{P}^n) - \mathrm{H}(\mathbf{C}^n)\\ &\geq \log_2|\mathcal{K}| + n(1-R_L)\log_2|\mathcal{P}| - \log_2|\mathcal{C}|^n\\ &= \log_2 \frac{|\mathcal{K}|\cdot|\mathcal{P}|^{n(1-R_L)}}{|\mathcal{C}|^n}.\end{aligned}$$

Pokračování důkazu

Máme tedy spodní odhad:

$$\mathrm{H}(\mathbf{K} \mid \mathbf{C}^n) \geq \log_2 \frac{|\mathcal{K}| \cdot |\mathcal{P}|^{n(1-R_L)}}{|\mathcal{C}|^n}.$$

Uděláme ještě horní odhad:

definice podmíněné entropie

$$\mathrm{H}(\mathbf{K} \mid \mathbf{C}^n) = \sum_{y \in \mathcal{C}^n} \Pr(\mathbf{C}^n{=}y)\, \mathrm{H}(\mathbf{K} \mid \mathbf{C}^n{=}y)$$

věta o maximální entropii

$$\leq \sum_{y \in \mathcal{C}^n} \Pr(\mathbf{C}^n{=}y) \log_2 |\mathcal{K}(y)|$$

Jensenova nerovnost

$$\leq \log_2 \sum_{y \in \mathcal{C}^n} \Pr(\mathbf{C}^n{=}y) |\mathcal{K}(y)| = \log_2 \mathrm{E}|\mathcal{K}(\mathbf{C}^n)|.$$

Složením odhadů dostaneme:

$$\log_2 \mathrm{E}|\mathcal{K}(\mathbf{C}^n)| \geq \log_2 \frac{|\mathcal{K}| \cdot |\mathcal{P}|^{n(1-R_L)}}{|\mathcal{C}|^n}.$$

Vzdálenost jednoznačnosti

- Kolik šifrového textu potřebujeme pro known-ciphertext attack?
  - Neřešíme časovou náročnost.
  - Řešíme možnost teoreticky uspět, např. hrubou silou.
- Pro šifrový text y lze útok úspěšně provést, jestliže $\mathcal{K}(y)$ je jednoprvková.
- Dosadíme-li $\mathrm{E}|\mathcal{K}(\mathbf{C}^n)| = 1$, dostaneme podmínku:

$$\frac{|\mathcal{K}| \cdot |\mathcal{P}|^{n(1-R_L)}}{|\mathcal{C}|^n} \leq 1.$$

Upravíme ji, abychom vyjádřili n:

$$n \geq \frac{\log_2|\mathcal{K}|}{\log_2|\mathcal{C}| - (1-R_L)\log_2|\mathcal{P}|}$$

vzdálenost jednoznačnosti

Vzdálenost jednoznačnosti

Pokud $|\mathcal{C}| = |\mathcal{P}|$, nerovnost se zjednoduší:

$$n \geq \frac{\log|\mathcal{K}|}{R_L \log|\mathcal{P}|}.$$

Příklad

- Mějme substituční šifru na anglickém textu:
  - $|\mathcal{K}| = 26!$
  - $|\mathcal{P}| = |\mathcal{C}| = 26$
  - $R_L = 0{,}75$
- Po dosazení dostaneme podmínku $n \geq 25{,}07$.
- K jednoznačnému dešifrování může stačit pouhých 26 znaků šifrového textu!

Vzdálenost jednoznačnosti

Příklad

- Mějme šifru AES na anglickém textu se 128 bitovým klíčem.
- AES pracuje s bloky otevřeného textu o délce 16 bajtů.
- Kódujme jeden znak otevřeného textu do každého bajtu.
- $\log_2|\mathcal{P}| = \log_2 26^{16} \approx 75$
- $\log_2|\mathcal{C}| = 128$
- $\log_2|\mathcal{K}| = 128$
- $R_L = 0{,}75$
- Po dosazení dostaneme podmínku $n \geq 1{,}17$ bloků.
- K jednoznačnému dešifrování mohou stačit pouhé dva bloky, tj. 32 znaků šifrového textu!

Absolutní bezpečnost

Definice

Říkáme, že šifra $(\mathcal{P}, \mathcal{C}, \mathcal{K}, E, D)$ je *absolutně bezpečná*, jestliže

$$\Pr(\mathbf{P}{=}x \mid \mathbf{C}{=}y) = \Pr(\mathbf{P}{=}x)$$

pro všechna $x \in \mathcal{P}$ a $y \in \mathcal{C}$ taková, že $\Pr(\mathbf{C}{=}y) \neq 0$.

Tvrzení

Šifra je absolutně bezpečná, právě když $\mathbf{P}$ *a* $\mathbf{C}$ *jsou nezávislé.*

Důkaz.

Následující podmínky jsou ekvivalentní:

- Šifra je absolutně bezpečná.
- $\Pr(\mathbf{P}{=}x) = \Pr(\mathbf{P}{=}x \mid \mathbf{C}{=}y) = \dfrac{\Pr(\mathbf{P}{=}x, \mathbf{C}{=}y)}{\Pr(\mathbf{C}{=}y)}$

  pro všechna $x \in \mathcal{P}$ a $y \in \mathcal{C}$ taková, že $\Pr(\mathbf{C}{=}y) \neq 0$.
- $\Pr(\mathbf{P}{=}x)\Pr(\mathbf{C}{=}y) = \Pr(\mathbf{P}{=}x, \mathbf{C}{=}y)$

  pro všechna $x \in \mathcal{P}$ a $y \in \mathcal{C}$.

  platí triviálně, pro případ $\Pr(\mathbf{C}{=}y) = 0$
- $\mathbf{P}$ a $\mathbf{C}$ jsou nezávislé.

□

Důsledek

Následující podmínky jsou ekvivalentní:

1. *Šifra je absolutně bezpečná.*
2. $\mathrm{H}(\mathbf{P} \mid \mathbf{C}) = \mathrm{H}(\mathbf{P})$.
3. $\mathrm{H}(\mathbf{C} \mid \mathbf{P}) = \mathrm{H}(\mathbf{C})$.

Věta

Jestliže klíč je volen náhodně s rovnoměrným rozdělením, pak Vernamova šifra je absolutně bezpečná.

Důkaz.

Pro všechna $x \in \mathcal{P}$ a $y \in \mathcal{C}$ platí

$$\begin{aligned}\Pr(\mathbf{P}{=}x, \mathbf{C}{=}y) &= \Pr(\mathbf{P}{=}x, \mathbf{K}{=}x \oplus y) \\ &= \Pr(\mathbf{P}{=}x)\Pr(\mathbf{K}{=}x \oplus y) \\ &= \Pr(\mathbf{P}{=}x)\,|\mathcal{K}|^{-1}.\end{aligned}$$

Podle věty o úplné pravděpodobnosti, pro všechna $y \in \mathcal{C}$:

$$\Pr(\mathbf{C}{=}y) = \sum_{x\in\mathcal{P}} \Pr(\mathbf{P}{=}x, \mathbf{C}{=}y) = \sum_{x\in\mathcal{P}} |\mathcal{K}|^{-1}\Pr(\mathbf{P}{=}x) = |\mathcal{K}|^{-1}.$$

Spojením obou výsledků zjišťujeme, že

$$\Pr(\mathbf{P}{=}x, \mathbf{C}{=}y) = \Pr(\mathbf{P}{=}x)\Pr(\mathbf{C}{=}y) \quad \forall x \in \mathcal{P}, y \in \mathcal{C}.$$

Lemma

$$\mathrm{H}(\mathbf{C}) \geq \mathrm{H}(\mathbf{P})$$

Důkaz.

$$\mathrm{H}(\mathbf{C}) \geq \mathrm{H}(\mathbf{C} \mid \mathbf{K}) = \mathrm{H}(\mathbf{P} \mid \mathbf{K}) = \mathrm{H}(\mathbf{P}).$$

obecná vlastnost entropie

znalost klíče dává jednoznačnou korespondenci mezi otevřenými a šifrovými texty

P a **K** jsou nezávislé

□

Věta (Shannon)

Jestliže je šifra absolutně bezpečná, pak $\mathrm{H}(\mathbf{K}) \geq \mathrm{H}(\mathbf{P})$.

Důkaz.

$$\mathrm{H}(\mathbf{K}) + \mathrm{H}(\mathbf{P}) = \mathrm{H}(\mathbf{K}, \mathbf{P}) = \mathrm{H}(\mathbf{K}, \mathbf{P}, \mathbf{C}) \geq \mathrm{H}(\mathbf{P}, \mathbf{C}).$$

P a **K** jsou nezávislé

známe-li otevřený text a klíč,
pak šifrový text je určen jednoznačně

obecná vlastnost entropie

$$\mathrm{H}(\mathbf{K}) \geq \mathrm{H}(\mathbf{P}, \mathbf{C}) - \mathrm{H}(\mathbf{P}) = \mathrm{H}(\mathbf{C} \mid \mathbf{P}) = \mathrm{H}(\mathbf{C}) \geq \mathrm{H}(\mathbf{P}).$$

právě jsme dokázali

obecná vlastnost entropie

absolutní bezpečnost

předchozí lemma

□

Závěr

- Absolutní bezpečnost je nepraktická, protože k zašifrování zprávy potřebujeme nejdříve utajeně přenést klíč s entropií alespoň tak velkou, jako je entropie samotné zprávy.

- Má-li $\mathbf{P}$ rovnoměrné rozdělení, pak pro Vernamovu šifru máme $\mathrm{H}(\mathbf{K}) = \mathrm{H}(\mathbf{P})$.

- Z Shannonovy věty tedy plyne:
  Má-li $\mathbf{P}$ rovnoměrné rozdělení, pak žádná absolutně bezpečná šifra není z hlediska délky klíče efektivnější než Vernamova šifra.

Další typy bezpečnosti

Výpočetní bezpečnost

- Kryptografický systém je *výpočetně bezpečný*, jestliže nejlepší známý útok, kterým ho lze prolomit, vyžaduje alespoň N operací, kde N je velké číslo převyšující výpočetní možnosti útočníka.
- V současné době alespoň $N \geq 2^{85}$.

Dokazatelná bezpečnost

- Říkáme, že kryptografický systém je *dokazatelně bezpečný*, jestliže jeho prolomením by bylo možné současně řešit nějaký dobře zkoumaný matematický problém, pro který není známa efektivní metoda řešení.
- Například problém faktorizace nebo diskrétního logaritmu.