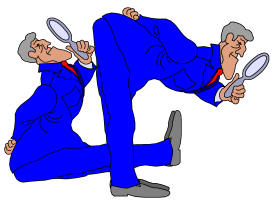
3. Dopřejme si jistoty neboli trochu teorie

Jan Paseka

Ústav matematiky a statistiky
Masarykova univerzita

23. listopadu 2023

O čem to bude



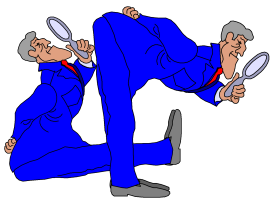
1 Úvod do problematiky
2 Šifrovací systémy
3 Perfektní bezpečnost
4 Redundance přirozeného jazyka a bod unicity

Úvod do problematiky

Mnozí lidé používají svoji inteligenci k zjednodušení, mnozí k zesložitění.
(Erich Kästner)

FI V této kapitole se pokusíme vytvořit teoretický základ pro naše dřívější úvahy. Zejména vymezíme pojem "perfektní bezpečnosti" šifrovacího systému.

O čem to bude



1 Úvod do problematiky

2 Šifrovací systémy
- Teoretické základy

3 Perfektní bezpečnost

4 Redundance přirozeného jazyka a bod unicity

Teoretické základy I

Podle našich předchozích představ si dohodnou odesílatel a příjemce ***nějaký*** klíč a zašifrují s ním zprávu. Správný pohled na věc se od této představy jemně odlišuje.

Budeme nyní uvažovat ***systémy***, které sestávají z nějaké ***množiny*** zpráv, příslušných kryptogramů a klíčů. V případě, že bychom následující myšlenky chtěli provést zcela precizně, museli bychom se držet axiomatiky; pro naše účely však bude lepší vysvětlení pojmů pomocí typických příkladů.

Takovýmto typickým příkladem množiny zpráv je sbírka matematických knih v knihovně sekce matematika týkajících se kódování.

Teoretické základy II

Získáme pak šifrovací systém, pokud budeme navíc uvažovat všech 312 afinních šifer s příslušnými kryptogramy.

Pro jiný případ stačí vzít všechna slova cizího původu vyskytující se v tomto textu, všech 26 aditivních posouvacích šifrování a výsledné kryptogramy.

Teoretické základy III

Zaveďme následující označení. Pomocí písmene **M** (message) označíme množinu všech zpráv, **C** (cryptogram) množinu všech kryptogramů (obvykle se jedná o řetězce nad konečnými abecedami Σ_1 a Σ_2) a **K** (key) množinu všech klíčů.

Šifrovacím systémem (***kryptosystémem***) pak nazýváme trojici $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ spolu s dodatečným předpokladem, že existují funkce (neboli algoritmy) e a d takové, že

$$e : \mathbf{M} \times \mathbf{K} \to \mathbf{C} \qquad \text{a} \qquad d : \mathbf{C} \times \mathbf{K} \to \mathbf{M}$$

a že pro všechna $(M, K) \in \mathbf{M} \times \mathbf{K}$ platí:

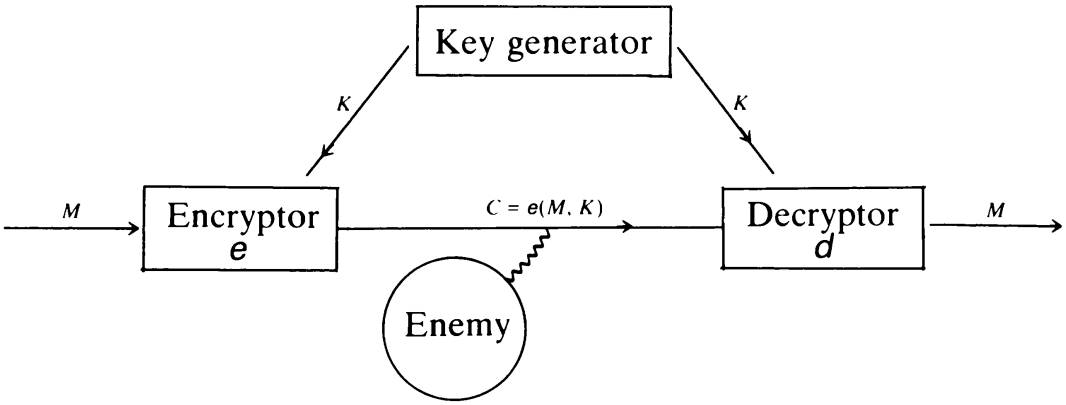
$$d(e(M, K), K) = M.$$

Teoretické základy IV

Zejména tedy pro každý klíč K máme invertibilní funkci (transformaci) $f_K : \mathbf{M} \to \mathbf{C}$ tak, že

$$f_K(M) = e(M, K) \quad \text{a} \quad {f_K}^{-1}(f_K(M)) = M.$$

Systém $(f_K)_{K \in \mathbf{K}}$ je nazýván ***šifrovací algoritmus***.



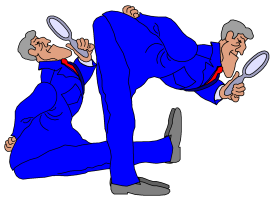
Teoretické základy V

Výše uvedená definice byla formulována praotcem moderní kryptografie Claudem E. **Shannonem**.

Uveďme dvě triviální pozorování:

1. Je možné, že dvě různé transformace převádí tutéž zprávu na jeden kryptogram.
2. Skutečnost, že transformace je invertibilní, implikuje $|\mathbf{M}| \leq |\mathbf{C}|$.

O čem to bude



1 Úvod do problematiky

2 Šifrovací systémy

3 Perfektní bezpečnost
- Bezpečný systém
- Komunikační kanál
- Ekvivokace
- Příklady
- Vlastnosti
- Skládání kryptosystémů

4 Redundance přirozeného jazyka a bod unicity

Bezpečný systém I

Nyní víme, co je šifrovací systém. Těžištěm tohoto odstavce je podání definice a popisu bezpečného šifrovacího systému.

Intuivně řečeno znamená "perfektní bezpečnost", že Mr. X nemá žádnou šanci zvětšit své znalosti o systému, i kdyby měl k dispozici všechno vědění a všechnu počítačovou kapacitu světa.
Předpokládejme nyní, že máme šifrovací systém $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ a že
(a) p_i je pravděpodobnost, že je odeslána zpráva M_i, $1 \leq i \leq n = |M|$; tyto pravděpodobnosti se nazývají **a priori** (nebo **teoretické**) **pravděpodobnosti** a jsou přirozeně každému dobrému kryptoanalytikovi známy.
(b) pravděpodobnost, že je použit klíč K_j je k_j a výběr klíče ***nezávisí na zprávě***, která je přenášena.

Bezpečný systém II

Tato dvě rozdělení pravděpodobností indukují rozdělení pravděpodobností na množině možných kryptogramů, kde pro jistý kryptogram C, řekněme C_u, je pravděpodobnost, že ***"náhodný" kryptogram*** C je roven kryptogramu C_u, určena vztahem

$$P(C = C_u) = \sum p_i k_j,$$

kde v sumě na pravé straně sčítáme přes všechny dvojice zpráva-klíč (M_i, K_j) takové, že $e(M_i, K_j) = C_u$.

Výše uvedené lze přeformulovat následovně pomocí pojmu náhodné veličiny.

Bezpečný systém III

Mějme tři náhodné veličiny M, K, C tak, že

(a) jev $M = M_i$ znamená, že ***byla odeslány zpráva*** $M_i \in \mathbf{M}$,

(b) jev $K = K_j$ znamená, že ***byl pro šifrování vybrán klíč*** $K_j \in \mathbf{K}$ a

(c) jev $C = C_u$ znamená, že ***byl zachycen kryptogram*** $C_u \in \mathbf{C}$.

Zejména tedy

(a) $P(M = M_i) = p_i$ a

(b) $P(K = K_j) = P(K = K_j | M = M_i) = k_j$

pro všechna i a všechna j.

Komunikační kanál I

Připomeňme, že výše uvedená situace je analogická situaci v teorii kódování pro případ zdroje bez paměti.

Přitom považujeme zdroj za proud symbolů jisté konečné abecedy. Zdroj má obvykle nějaký náhodný mechanismus, který je založen na statistice situace, která je modelovaná.

Tento náhodný mechanismus může být poměrně dost komplikovaný, ale my se budeme pro okamžik soustředit na následující opravdu speciální a jednoduchý příklad.

Značí-li X_i i-tý symbol vytvořený zdrojem, dohodneme se pak, že, pro každý symbol a_j, pravděpodobnost

$$P(X_i = a_j) = p_j$$

je nezávislá na i a tedy je nezávislá na všech minulých nebo v budoucnosti vyslaných symbolech.

Komunikační kanál II

Jinak řečeno, $X_1, X_2, \ldots$ je právě posloupnost identicky distribuovaných, nezávislých náhodných veličin. Takovýto zdroj nazveme ***zdrojem s nulovou pamětí*** nebo ***zdrojem bez paměti*** a jeho entropie H je definována jako

$$H = -\sum p_j \log p_j,$$

kde sčítáme přes množinu j takových, že $p_j > 0$.

Připomeňme, že jsou-li $X_1, \ldots, X_m$ náhodné proměnné takové, že každá z nich nabývá pouze konečně mnoha hodnot, lze pak považovat $\mathbf{X} = (X_1, \ldots, X_m)$ za náhodný vektor a definovat souhrnou entropii $X_1, \ldots, X_m$ jako

$$H(\mathbf{X}) = -\sum_k p(x_1, \ldots, x_m) \cdot \log_2 p(x_1, \ldots, x_m), \qquad (3.1)$$

kde $p(x_1, \ldots, x_m) = P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m)$.

Komunikační kanál III

Předpokládejme dále, že X je náhodná proměnná na pravděpodobnostním prostoru Ω a A je událost z Ω. Nabývá-li X konečné množiny hodnot $\{a_i : 1 \leq i \leq m\}$, je přirozené definovat ***podmíněnou entropii*** náhodné proměnné X určenou událostí A jako

$$H(X|A) = -\sum_{k=1}^{m} P(X = a_k|A) \log P(X = a_k|A).$$

Úplně stejně, je-li Y jiná náhodná proměnná nabývající hodnot b_k $(1 \leq k \leq m)$, definujeme ***podmíněnou entropii*** náhodné proměnné X určenou náhodnou proměnnou Y jako

$$H(X|Y) = \sum_j H(X|Y = b_j) P(Y = b_j).$$

Komunikační kanál IV

Považujeme $H(X|Y)$ za entropii náhodné proměnné X určenou jistou hodnotou Y zprůměrovanou přes všechny hodnoty, jichž může Y nabývat.

Diskrétní kanál bez paměti je charakterizován vstupní abecedou $\Sigma_1 = \{a_1, \ldots, a_m\}$ vstupních znaků, výstupní abecedou $\Sigma_2 = \{b_1, \ldots, b_n\}$ výstupních znaků a ***maticí* P *kanálu***

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \ldots & \ldots & p_{1n-1} & p_{1n} \\ p_{21} & p_{22} & \ldots & \ldots & p_{2n-1} & p_{2n} \\ \vdots & \vdots & \ldots & \ldots & \vdots & \vdots \\ p_{m-11} & p_{m-12} & \ldots & \ldots & p_{m-1n-1} & p_{m-1n} \\ p_{m1} & p_{m2} & \ldots & \ldots & p_{mn-1} & p_{mn} \end{pmatrix}.$$

Komunikační kanál V

Způsob používání kanálu je následující: každá posloupnost $(u_1, u_2, \ldots, u_N)$ symbolů ze vstupní abecedy Σ_1 na vstupu se převede na posloupnost $(v_1, v_2, \ldots, v_N)$ téže délky symbolů z výstupní abecedy Σ_2 na výstup tak, že

$$P(v_k = b_j | u_k = a_i) = p_{ij} \quad (1 \leq i \leq m, 1 \leq j \leq n),$$

a to nezávisle pro každé $k, 1 \leq k \leq N$.

Implicitně je ve výše uvedeném obsaženo, že pro každé i, $1 \leq i \leq m$ platí

$$\sum_j p_{ij} = 1.$$

Komunikační kanál VI

Matice **P** s nezápornými hodnotami taková, že součet prvků v každém řádku je roven 1, se nazývá ***stochastická matice***; v teorii náhodných procesů mluvíme o ***matici přechodu markovského řetězce***.

Kapacita komunikačního kanálu je míra jeho schopnosti přenášet informaci. Formální definice je motivována níže uvedeným:
Předpokládejme, že máme diskrétní kanál bez paměti se vstupní abecedou $\Sigma_1 = \{a_1, \ldots, a_m\}$, výstupní abecedou $\Sigma_2 = \{b_1, \ldots, b_n\}$ a maticí P kanálu

$$P = \left[p_{ij}\right] = P(\mathbf{b}_j \text{ obdrženo} | \mathbf{a}_i \text{ odesláno}).$$

Komunikační kanál VII

Přidáme-li k tomuto kanálu zdroj $\mathcal{S}$ bez paměti, který vysílá symboly $a_1, \ldots, a_m$ s pravděpodobnostmi $p_1, \ldots, p_m$, pak výstup kanálu můžeme považovat za zdroj $\mathcal{J}$ bez paměti, který vysílá symboly $b_1, \ldots, b_n$ s pravděpodobnostmi $q_1, \ldots, q_n$, kde

$$\begin{aligned} q_j &= \sum_{i=1}^{m} P(\mathbf{b}_j \text{ obdrženo} | \mathbf{a}_i \text{ odesláno}) P(\mathbf{a}_i \text{ odesláno}) \\ &= \sum_{i=1}^{m} p_i p_{ij}. \end{aligned}$$

Jsou-li **U** a **V** dva náhodné vektory, definujeme ***informaci o* U *poskytnutou* V** jako číslo

$$I(\mathbf{U}|\mathbf{V}) = H(\mathbf{U}) - H(\mathbf{U}|\mathbf{V}).$$

Komunikační kanál VII

Jinak řečeno, $I(\mathbf{U}|\mathbf{V})$ vyjadřuje množství nejistoty o **U** odstraněné **V**. Totiž, množství informace, průměrně obsažené v jednom znaku zprávy, je entropie vstupního rozdělení $H(\mathcal{S}) = -\sum_i p_i \log p_i$. Při přenosu diskrétním kanálem se ztratí informace

$$H(\mathcal{S}/\mathcal{J}) = -\sum_{i,j} p_{i,j} \log p_{i,j}$$

průměrně na jeden znak.

Zbývá pak $H(\mathcal{S}) - H(\mathcal{S}/\mathcal{J})$ přenesené informace. Jestliže entropii počítáme v bitech a známe průměrnou dobu τ, kterou kanál spotřebuje na přenos jednoho znaku, je rychlost přenosu

$$I(\mathcal{S}|\mathcal{J}) = \frac{H(\mathcal{S}) - H(\mathcal{S}/\mathcal{J})}{\tau} \text{ bitů za sekundu.}$$

Komunikační kanál VIII

Často se za jednotku času volí jeden přenos znaku a potom

$$I(\mathcal{S}|\mathcal{J}) = H(\mathcal{S}) - H(\mathcal{S}/\mathcal{J}) \text{ bitů za jednotku času.}$$

Informace o $\mathcal{S}$ podaná pomocí $\mathcal{J}$ je pak rovna

$$I(\mathcal{S}|\mathcal{J}) = H(\mathcal{S}) - H(\mathcal{S}|\mathcal{J}) = H(\mathcal{S}) + H(\mathcal{J}) - H(\mathcal{S},\mathcal{J})$$

a je to funkce, která závisí pouze na pravděpodobnostním rozdělení $q_1, \ldots, q_n$, a maticí kanálu **P**.

Komunikační kanál IX

Proto je přirozené definovat ***kapacitu*** C kanálu jako maximální rychlost přenosu, tedy

$$C = \sup I(\mathcal{S}|\mathcal{J}), \tag{3.2}$$

kde supremum je bráno přes všechny zdroje bez paměti $\mathcal{S}$, nebo, ještě přesněji, nad všemi možnými rozděleními pravděpodobností $(p_1, \ldots, p_n)$.

V dalším tedy můžeme považovat **M** za zdroj bez paměti s šifrovací funkcí e, přičemž klíče slouží jako komunikační kanál.

Ekvivokace I

Základním pojmem je pojem ***klíčové ekvivokace*** zavedený Shannonem $H(\mathbf{K}|\mathbf{C})$. Ten nám měří průměrnou nejistotu, která nám zůstává po zachycení kryptogramu **C**.

Podobně budeme definovat ***ekvivokaci zpráv*** jakožto $H(\mathbf{M}|\mathbf{C})$. Občas budeme psát $\mathrm{S} = \langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ a budeme značit $H(\mathrm{S})$ klíčovou ekvivokaci $H(\mathbf{K}|\mathbf{C})$.

Náhodná proměnná zpráv má pak ***entropii zpráv***

$$H(\mathbf{M}) = -\sum p_i \log p_i.$$

Ekvivokace II

Jsou-li po řadě **K** a **C** náhodné proměnné klíčů a kryptogramů, jsou pak ***klíčová entropie*** $H(\mathbf{K})$ a ***entropie kryptogramů*** definovány jakožto

$$H(\mathbf{K}) = -\sum P(\mathbf{K} = K_i) \log P(\mathbf{K} = K_i),$$
$$H(\mathbf{C}) = -\sum P(\mathbf{C} = C_j) \log P(\mathbf{C} = C_j),$$

kde sumace je prováděna přes všechny možné klíče K_i a všechny možné kryptogramy C_j.

Následující vlastnost ekvivokace vyjadřuje tu skutečnost, že daleko více nejistoty je spjato s klíčem než se zprávou.

Věta 3.1

Klíčová ekvivokace je určena ekvivokací zprávy vztahem

$$H(\mathbf{K}|\mathbf{C}) = H(\mathbf{M}|\mathbf{C}) + H(\mathbf{K}|\mathbf{M}, \mathbf{C}).$$

Ekvivokace III

Důkaz.

Připomeňme základní identitu pro entropii

$$H(X|Y) = H(X, Y) - H(Y).$$

Můžeme tedy psát

$$\begin{aligned} H(\mathbf{M}|\mathbf{C}) &= H(\mathbf{M}, \mathbf{C}) - H(\mathbf{C}) \\ &= H(\mathbf{M}, \mathbf{K}, \mathbf{C}) - H(\mathbf{K}|\mathbf{M}, \mathbf{C}) - H(\mathbf{C}). \end{aligned}$$

Nyní tedy i

$$\begin{aligned} H(\mathbf{K}|\mathbf{C}) &= H(\mathbf{K}, \mathbf{C}) - H(\mathbf{C}) \\ &= H(\mathbf{M}, \mathbf{K}, \mathbf{C}) - H(\mathbf{M}|\mathbf{K}, \mathbf{C}) - H(\mathbf{C}). \end{aligned}$$

Ale

$$H(\mathbf{M}|\mathbf{K}, \mathbf{C}) = 0.$$

Ekvivokace IV

Pokračování důkazu.

Totiž, jakmile je znám kryptogram C a klíč K, je jednoznačně určena i zpráva M a tedy míra neurčitosti je nulová. Tedy

$$H(\mathbf{K}|\mathbf{C}) = H(\mathbf{M},\mathbf{K},\mathbf{C}) - H(\mathbf{C}),$$

což, porovnáno s výše uvedeným, nám dává dokazovanou identitu. ∎

Důsledek 3.2

Klíčová ekvivokace je alespoň tak velká jako ekvivokace zprávy.

Ekvivokace V

Lemma 3.3

Pro každý kryptosystém $\langle \mathbf{M},\mathbf{K},\mathbf{C}\rangle$ *platí*

$$H(\mathbf{K},\mathbf{C}) = H(\mathbf{M}) + H(\mathbf{K}).$$

Důkaz.

Protože náhodné veličiny **K** a **M** jsou nezávislé, víme, že $H(\mathbf{M}) + H(\mathbf{K}) = H(\mathbf{M},\mathbf{K})$. Stačí tedy ověřit, že $H(\mathbf{C},\mathbf{K}) = H(\mathbf{M},\mathbf{K})$.

Protože $H(\mathbf{M}|\mathbf{K},\mathbf{C}) = 0$ a $H(\mathbf{C}|\mathbf{K},\mathbf{M}) = 0$, máme

$$\begin{aligned} H(\mathbf{M}|\mathbf{K},\mathbf{C}) &= H(\mathbf{M},\mathbf{K},\mathbf{C}) - H(\mathbf{K},\mathbf{C}) = 0 \\ H(\mathbf{C}|\mathbf{K},\mathbf{M}) &= H(\mathbf{C},\mathbf{K},\mathbf{M}) - H(\mathbf{K},\mathbf{M}) = 0. \end{aligned}$$

Tedy i $H(\mathbf{C},\mathbf{K}) = H(\mathbf{M},\mathbf{K})$. ∎

Ekvivokace VI

Důsledek 3.4

Pro každý kryptosystém $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ *platí*

$$H(\mathbf{C}|\mathbf{K}) = H(\mathbf{M}|\mathbf{K}),\ H(\mathbf{C},\mathbf{K}) = H(\mathbf{M},\mathbf{K})\ a\ H(\mathbf{M}) \leq H(\mathbf{C}).$$

Důkaz.

Stačí ověřit poslední nerovnost.
$H(\mathbf{M}) + H(\mathbf{K}) = H(\mathbf{M},\mathbf{K}) = H(\mathbf{C},\mathbf{K}) \leq H(\mathbf{C}) + H(\mathbf{K})$ tj.
$H(\mathbf{M}) \leq H(\mathbf{C})$. ∎

Příklady I

Příklad 3.5

(a) *Předpokládejme, že "zprávy" v* **M** *jsou písmena slov nějaké (německé) knihy; pravděpodobnost zprávy je pak četnost odpovídajícího písmene; v případě písmene* **e** *je pak* $p(\mathbf{e}) \approx 0.174$.

(b) *Nyní předpokládejme, že "zprávy" v* **M** *jsou dvojice za sebou následujících písmen slov nějaké (německé) knihy; pravděpodobnost zprávy je pak četnost odpovídajícího bigramu.*

Představme si nyní, že Mr. X zachytil kryptogram *C*.
Aby ho byl schopen analyzovat, může (***alespoň teoreticky***) vyzkoušet všechny zprávy a vždy určit pravděpodobnost toho, že kryptogram *C* vznikl zašifrováním zprávy *M*.

Příklady II

Označme pak tyto pravděpodobnosti $p_C(M) = P(M|C)$; mluvíme pak o ***a posteriori*** (nebo ***pozorovaných***) ***pravděpodobnostech***.

Příklad 3.6

(c) *Buď M stejné jako ve výše uvedeném příkladu (a); jako algoritmus budeme uvažovat posouvací šifry se všemi 26 možnými klíči.*

Uvažme, že každé písmeno kryptogramu C má tutéž šanci, že odpovídá určitému písmenu zprávy; např. v 17,4% případů vznikne C z **e**, *v 9,8% případů vznikne z* **n**, *atd.*

Jinak řečeno, **pro každý kryptogram** *C* **platí** $p_C(M) = p(M)$ **pro každou zprávu** *M*.

Úvod
Šifrovací systémy
Perfektní bezpečnost
Redundance
Bezpečný systém
Komunikační kanál
Ekvivokace
Příklady
Vlastnosti
Skládání

Příklady III

Příklad 3.7

(d) *Nyní předpokládejme, že každá zpráva v* **M** *sestává z prvních 100 písmen každé strany prvního dílu slovníku* **das große Brockhaus**. *Algoritmus nechť opět sestává z posouvacích šifer.*

Pro každou zprávu M je její pravděpodobnost $\frac{1}{|M|}$, *malé, ale stále ještě kladné číslo.*

Protože je relativně snadné prověřit, zda určitý kryptogram pochází z určité zprávy (rozdělení písmen v kryptogramu musí přesně odpovídat rozdělení písmen ve zprávě), je $p_C(M)$ *rovno buď jedné nebo nule.*

To znamená obzvlášť, že pro každý kryptogram je $p_C(M) \neq p(M)$.

Příklady III

Příklad 3.8 (Pokračování)

Proberme tuto skutečnost podrobněji:
Předpokládejme, že kryptoanalytik Mr. X zjistí. že pro jistou zprávu M je $p_C(M) > p(M)$. Pak by věděl, že kryptogram C vznikl s vysokou pravděpodobností ze zprávy M.

Tzn., že by se analýzou něco nového naučil. To ale nesmí při perfektním systému nastat.

V případě, že by bylo $p_C(M) < p(M)$, pak by Mr. X věděl, že kryptogram C vznikne s velmi malou pravděpodobností ze zprávy M. I v tomto případě by si Mr. X rozšířil svoje znalosti.

Vlastnosti I

Můžeme tedy definovat:
Šifrovací systém $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ poskytuje **perfektní bezpečnost**, *jestliže pro každý kryptogram C platí*

$$p_C(M) = p(M)$$

pro každou zprávu M.

Jinak řečeno, šifrovacího systém $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ je perfektní, pokud jsou a priori pravděpodobnosti rovny pravděpodobnostem a posteriori (viz příklad (c)).

Posouvací šifry jsou perfektní, jestliže operují nad jednotlivými písmeny. Mr. X se pak může namáhat jak chce; písmena kryptogramu jsou totiž zcela náhodně rozdělena.

Vlastnosti II

Chceme-li perfektnost šifrovacího systému vyjádřit pomocí náhodných proměnných **M** a **C**, je systém perfektní právě tehdy, když **M** ***a*** **C** ***jsou nezávislé***.

Z toho bezprostředně plyne, že šifrovací systém $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ poskytuje **perfektní bezpečnost**, jestliže pro každou zprávu M platí

$$p_M(C) = p(C)$$

pro každý kryptogram C.

Úvod
Šifrovací systémy
Perfektní bezpečnost
Redundance
Bezpečný systém
Komunikační kanál
Ekvivokace
Příklady
Vlastnosti
Skládání

Vlastnosti III

Věta 3.9

Kryptosystém $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ *je perfektní právě tehdy, když*

$$H(\mathbf{M}|\mathbf{C}) = H(\mathbf{M}).$$

Důkaz.

Z teorie informace je známo, že $H(\mathbf{M}|\mathbf{C}) = H(\mathbf{M})$ právě tehdy, když **M** a **C** jsou nezávislé náhodné proměnné tj. to je právě tehdy, když se jedná o perfektní kryptosystém. ∎

Ptejme se nyní, jak můžeme rozpoznat, kdy je šifrovací systém perfektní či nikoliv. K tomu si dokážeme několik jednoduchých kritérií.

Vlastnosti IV

1. Kritérium *Je-li šifrovací systém* $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ *perfektní, pak každá zpráva s odpovídajícím klíčem může být zobrazena na libovolný kryptogram.*

Proč platí toto kritérium?

Uvažme zprávu M a kryptogram C. Protože $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ je perfektní, platí $p_C(M) = p(M)$. V každém šifrovacím systému je $p(M) > 0$, protože každá zpráva se vyskytuje s kladnou pravděpodobností. Dohromady obdržíme $p_C(M) > 0$.

To znamená, že existuje klíč, pomocí kterého se zašifruje M do C. (Kdyby žádný takový klíč neexistoval, nutně bychom měli, že $p_C(M) = 0$.)
Tím je dokázáno první kritérium.

Toto kritérium je velmi užitečné - v "***negativním smyslu***":
Umožní nám rozhodnout, že jisté systémy ***nejsou*** perfektní.

Vlastnosti V

2. Kritérium *Je-li šifrovací systém* $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ *perfektní, pak platí:*

$$|\mathbf{M}| \leq |\mathbf{C}| \leq |\mathbf{K}|.$$

Důkaz.

Zřejmě $|\mathbf{M}| \leq |\mathbf{C}|$.
Proč platí $|\mathbf{C}| \leq |\mathbf{K}|$? Uvažme libovolnou, pevně zvolenou zprávu M a zašifrujme ji pomocí všech možných klíčů z $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$.

Podle prvního kritéria lze M převést do každého možného kryptogramu. Pro každý kryptogram C potřebujeme alespoň jeden klíč (totiž pomocí jednoho klíče nemůžeme M zobrazit na dva různé kryptogramy).

Potřebujeme tedy alespoň tolik klíčů, kolik je kryptogramů. Máme tedy $|\mathbf{C}| \leq |\mathbf{K}|$. ∎

Vlastnosti VI

3. Kritérium *Buď* $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ *šifrovací systém tak, že*

$$|\mathbf{M}| = |\mathbf{C}| = |\mathbf{K}|,$$

ve kterém se všechny klíče vyskytují s toutéž pravděpodobností.

Dále předpokládejme, že ke každé zprávě M *a ke každému kryptogramu* C *existuje právě jeden klíč* K *z* $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ *tak, že* $e(M, K) = C$.

Pak je $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$ *perfektní.*

Vlastnosti VII

Důkaz - 3. Kritérium.

Stačí zřejmě ověřit, že pro každou zprávu M_i platí $P(\mathbf{M} = M_i | \mathbf{C} = C_j) = P(\mathbf{M} = M_i)$ pro každý kryptogram C_j.

Z Bayesova vzorce máme

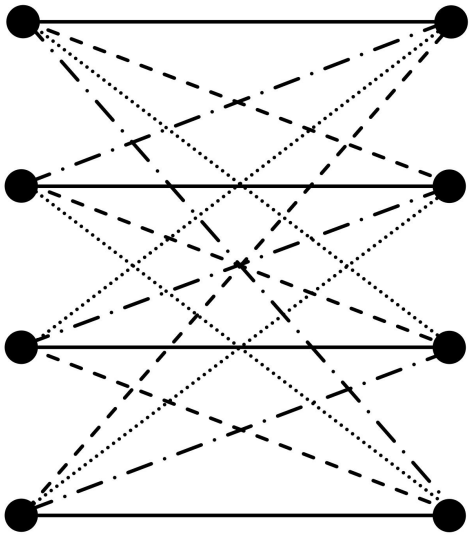
$$P(\mathbf{M} = M_i | \mathbf{C} = C_j) = \frac{P(\mathbf{C} = C_j | \mathbf{M} = M_i) \cdot P(\mathbf{M} = M_i)}{\sum_{k=1}^{|K|} P(\mathbf{C} = C_j | \mathbf{M} = M_k) \cdot P(\mathbf{M} = M_k)}$$

$$= \frac{P(\mathbf{M} = M_i)}{\sum_{k=1}^{|K|} P(\mathbf{M} = M_k)} = P(\mathbf{M} = M_i),$$

neboť $P(\mathbf{C} = C_j | \mathbf{M} = M_k) = \frac{1}{|K|}$ nezávisle na j a k. ∎

Vlastnosti VIII

Na základě třetího kritéria lze přirozeným způsobem konstruovat perfektní šifrovací systémy (čtyřem klíčům a jím příslušným transformacím odpovídají různě šrafované šipky).



Skládání kryptosystémů I

Přirozeným způsobem, jak zvýšit bezpečnost šifrování, je vzít různé systémy a kombinovat je.

Dvě takovéto metody navržené Shannonem jsou stále základem mnoha praktických kryptosystémů.

Jedná se o ***vážený součet*** a ***součin*** kryptosystémů.

Skládání kryptosystémů II

- ***Vážený součet***: Jsou-li $\mathbf{S}_1$ a $\mathbf{S}_2$ dva kryptosystémy se stejným prostorem zpráv $M = M_1 = M_2$ a $0 < p < 1$, je pak jejich ***vážený součet*** $p\mathbf{S}_1 + (1-p)\mathbf{S}_2$ kryptosystém určený následným výběrem: použijeme $\mathbf{S}_1$ s pravděpodobností p a $\mathbf{S}_2$ s pravděpodobností $1-p$.

  Má-li tedy $\mathbf{S}_1$ klíče $K_1, \ldots, K_m$ s pravděpodobnostmi použití p_i pro klíč K_i a $\mathbf{S}_2$ má klíče $K'_1, \ldots, K'_n$ s pravděpodobnostmi použití p'_j pro klíč K'_j, má pak kryptosystém $p\mathbf{S}_1 + (1-p)\mathbf{S}_2$ $m+n$ klíčů $K_1, \ldots, K_m, K'_1, \ldots, K'_n$ s pravděpodobnostmi použití pp_i pro klíč K_i a s pravděpodobnostmi použití $(1-p)p'_j$ pro klíč K'_j.

  Tento postup lze přirozeně rozšířit na více než dva systémy.

Skládání kryptosystémů III

- ***Součin***: Druhý způsob kombinování kryptosystémů $\mathbf{S}_1$ a $\mathbf{S}_2$ je to, že nejprve použijeme na naši zprávu kryptosystém $\mathbf{S}_1$ a potom aplikujeme $\mathbf{S}_2$ na výsledný kryptogram.

  Abychom toto mohli provést, musí být nutně $\mathbf{C}_1 \subseteq \mathbf{M}_2$. Pak můžeme definovat součin jako $\mathbf{S}_1 * \mathbf{S}_2$.

  Jsou-li klíče $K_1, \ldots, K_m$ s pravděpodobnostmi použití p_i pro klíč K_i v kryptosystému $\mathbf{S}_1$ a $\mathbf{S}_2$ má klíče $K'_1, \ldots, K'_n$ s pravděpodobnostmi použití p'_j pro klíč K'_j, má pak kryptosystém $\mathbf{S}_1 * \mathbf{S}_2$ $m.n$ klíčů (K_i, K'_j) s pravděpodobnostmi použití $p_i p'_j$.

  Poznamenejme, že skutečně efektivních klíčů může být méně, protože některé se složených transformací mohou splývat.

Skládání kryptosystémů IV

Poznamenejme, že evidentně platí následující:

Jsou-li $\mathbf{S}_1$, $\mathbf{S}_2$ a $\mathbf{S}_3$ kryptosystémy tak, že níže uvedené operace jsou definovány, $0 < p < 1$, $q = 1 - p$, pak

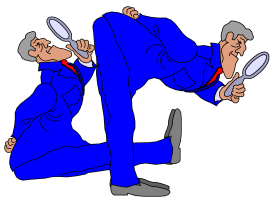
$$\mathbf{S}_3 * (p\mathbf{S}_1 + q\mathbf{S}_2) = p\mathbf{S}_3 * \mathbf{S}_1 + q\mathbf{S}_3 * \mathbf{S}_2,$$

$$(p\mathbf{S}_1 + q\mathbf{S}_2) * \mathbf{S}_3 = p\mathbf{S}_1 * \mathbf{S}_3 + q\mathbf{S}_2 * \mathbf{S}_3,$$

$$\mathbf{S}_1 * (\mathbf{S}_2 * \mathbf{S}_3) = (\mathbf{S}_1 * \mathbf{S}_2) * \mathbf{S}_3,$$

$\mathbf{S}_1 * \mathbf{S}_2$ není obecně rovno $\mathbf{S}_2 * \mathbf{S}_1$.

O čem to bude



1 Úvod do problematiky

2 Šifrovací systémy

3 Perfektní bezpečnost

4 Redundance přirozeného jazyka a bod unicity
- Aproximace jazyka
- Jazyk jako zdroj
- Entropie a redundance
- Redundance a šifrování I

Aproximace jazyka I - aproximace 0. řádu

Věnujme se nyní chvíli zkoumání přirozeného jazyka jakým je například angličtina. Budeme v dalším považovat angličtinu za jazyk skládající se z abecedy o 27 písmenech, z toho je 26 římských písmen a 1 mezera.

První, a velmi špatná aproximace angličtiny je, že vezmeme ***aproximaci 0. řádu***.

V tomto případě mají všechny symboly stejnou pravděpodobnost: každý se tedy vyskytne s pravděpodobností $\frac{1}{27}$ a následující text nám ukáže typickou sekvenci symbolů vytvořenou takovýmto zdrojem:

DM QASCJDGFOZYNX ZSDZLXIKUD.

Úvod Perfektní bezpečnost Aproximace jazyka Odhad redundance
Šifrovací systémy Redundance Jazyk jako zdroj Bod unicity

Aproximace jazyka II - aproximace 1. řádu

Tato aproximace vůbec nevyužívá relativní četnosti symbolů použitých v anglickém jazyce.

Použijeme-li odhady těchto četností, můžeme vytvořit ***aproximaci 1. řádu***, jejímž typickým příkladem je

OR L RW NILI E NNSBATEI.

Ačkoliv je tento přístup zřejmější než aproximace 0. řádu, stále zde není žádná informace o vzájemné závislosti sousedních písmen.

Aproximace jazyka III - Markovův zdroj 1. řádu

Tomuto lze vyhovět například ***Markovovým zdrojem 1. řádu***, kde můžeme použít podmíněné pravděpodobnosti založené na četnostech dvojic písmen tj. ***digramů***:

$$P(i|j) = p(i,j)/p(j),$$

kde $p(i,j)$ je pravděpodobnost výskytu digramu (i,j) a $p(i|j)$ je podmíněná pravděpodobnost výskytu písmene i za předpokladu, že předcházející písmeno je j.

To je však velmi časově náročné a Shannon místo toho navrhl použít metodu Monte Carlo, která má stejný efekt.

Aproximace jazyka IV - metoda Monte Carlo

Vyberme ***náhodně*** text či texty. ***Náhodně*** z textu vyberme první písmeno jakožto první symbol X_1.

Předpokládejme bez újmy na obecnosti, že je to např. B. Opět ***náhodně*** nalistujme nějakou stránku textu a pokračujme na ní dále, až narazíme na první výskyt B. Vezměme za X_2 písmeno textu bezprostředně za B.

Použijeme-li výše uvedenou metodu, lze obdržet následující ***Markovovu aproximaci 1. řádu pro angličtinu***:

OUCTIE IN ARE AMYST TE TUSE SOBE CTUSE.

Aproximace jazyka V - Markovova aproximace 2. řádu

Shannonovu metodu lze použít na to, abychom získali lepší aproximaci tak, že vybereme písmena z textu vzhledem k dvěma předchozím písmenům.

Např., ***Markovovou aproximací druhého řádu*** je posloupnost

HE AREAT BEIS THAT WISHBOUT SEED DAY OFTE, AND HE IS FOR THAT MINUMB LOOTS WILL AND GIIRLS, A DOLL WILL IS FRIECE ABOARICE STRED SAYS.

Úvod Perfektní bezpečnost Aproximace jazyka Odhad redundance
Šifrovací systémy Redundance Jazyk jako zdroj Bod unicity

Aproximace jazyka V - Markovova aproximace latiny

Použijeme-li Shannonovu metodu s Cicerovým dílem ***de Senectute***, obdržíme velmi zřetelnou Markovovu aproximaci latiny:

IENEC FES VIMONILLITUM M ST ER PEM ENIM PTAUL

(Markovova aproximace 1. řádu)

SENECTOR VCI QUAEMODOMIS SE NON FRATURDIGNAVIT SINE VELIUS

(Markovova aproximace 2. řádu).

Aproximace jazyka VI - zdroj slov

Teoreticky může být tato metoda použita pro aproximace libovolně vysokého řádu.

Je však více než namáhavé provádět už aproximace třetího řádu. Lze však akceptovat to, že už aproximace druhého řádu je přijatelná.

Alternativní přístup navržený Shannonem bylo modelování angličtiny nikoliv jako zdroje písmen, nýbrž jako zdroje s množinou anglických ***slov***, jakožto základní abecedou.

Shannon dává přednost náhodnému výběru z textů před metodou četnosti anglických slov.

Aproximace jazyka VII - zdroj slov

Uveďme následující aproximace:

> REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF THE EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE
>
> (slovní aproximace 1. řádu)

> THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHOEVER TOLD THE PROBLEM FOR AN UNEXPECTED.
>
> (slovní aproximace 2. řádu).

Jazyk jako zdroj I

Budeme tedy v dalším považovat přirozené jazyky za zdroje s entropií. Pokusíme se podat jisté odhady a interpretace této entropie, kterou budeme v případě angličtiny značit jako H_E. Je známo, že lze H_E interpretovat pomocí přibližné formule

$$2^{nH_E} \simeq T(n) \qquad (n \text{ dostatečně velké}),$$

kde $T(n)$ označuje počet typických (=smysluplných) posloupností délky n anglického jazyka.

Tento vztah nám však bezprostředně nepomůže s odhadem H_E, protože není znám žádný způsob odhadu $T(n)$.

Úvod
Šifrovací systémy
Perfektní bezpečnost
Redundance
Aproximace jazyka
Jazyk jako zdroj
Odhad redundance
Bod unicity

Jazyk jako zdroj II

Víme však, že existuje 27^n možných posloupností délky n z anglické abecedy a protože $\log 27 = 4.76$, máme

$$H_E \leq 4.76 \text{ bitů na symbol.}$$

Lepší odhad pro H_E můžeme obdržet z aproximace 1. řádu, ve které můžeme použít informaci o rozdílných pravděpodobnostech výskytu písmen.

Například nejpravděpodobnějším symbolem je mezera s pravděpodobností $P(\text{mezera}) = 0.18\ldots$, $P(E) = 0.13\ldots$ atd.

Použijeme-li základní identitu $H(X, Y) \leq H(X) + H(Y)$, dostaneme horní závoru

$$H_E \leq H_E^1 \leq -\sum_{p_i} \log p_i,$$

kde p_i je pravděpodobnost výskytu i-tého symbolu.

Jazyk jako zdroj III

Podobně obdržíme

$$H_E \leq H_E^2 = -\frac{1}{2}\sum_i \sum_j p(i,j)\log p(i,j),$$

kde $p(i,j)$ jsou odhadnuté výskyty symbolů (i,j) a my ignorujeme dvojice symbolů s nulovou pravděpodobností (např. Qq).

Následující tabulka nám ukazuje přehled odhadů entropií pro 26- a 27-písmennou anglickou abecedu.

	26-ti písmenná abeceda	27-ti písmenná abeceda
H_E^0	4.70	4.76
H_E^1	4.14	4.03
H_E^2	3.56	3.32
H_E^3	3.30	3.10

Jazyk jako zdroj IV

Jiný přístup nalezení odhadu entropie je založený na četnosti slov.

Považujme angličtinu za konečný jazyk skládající se ze slov $w_1, \ldots, w_N$, jež se vyskytují nezávisle s pravděpodobnostmi $p_1, \ldots, p_N$.

Pak ***slovní entropie*** H_W je určena vztahem

$$H_W = -\sum_{i=1}^{N} p_i \log p_i.$$

Shannon navrhl, že pak entropii symbolů H_E lze aproximovat jakožto

$$H_E = H_W/\overline{w},$$

kde $\overline{w}$ je průměrná délka slova v angličtině.

Jazyk jako zdroj V

K tomuto přístupu lze mít následující výhrady:

- Slova použitá v angličtině nejsou nezávislá a slovní entropie je spíše odhad slovní entropie prvního řádu.
- Podíl slovní entropie a průměrné délky slova je velmi hrubá aproximace a je nejlépe ji nahradit vhodnou nerovností.

Abychom vyčíslili slovní entropii, použijme pravidlo navržené lingvistou ***G. K. Zipfem*** (1935).

To tvrdí, že, pokud jsou slova přirozeného jazyka uspořádána v klesajícím uspořádání podle jejich pravděpodobností výskytu (p_n pak označuje pravděpodobnost $n-$tého nejvýše pravděpodobného slova), ***dobrá aproximace*** těchto pravděpodobností je určena formulí

$$p_n = A/n,$$

kde A je nějaká konstanta závisející na daném jazyce.

Jazyk jako zdroj VI

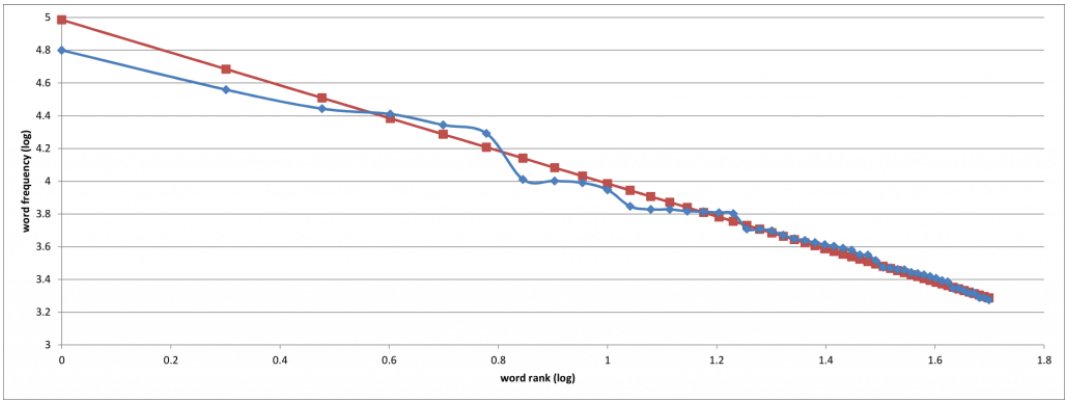
Ačkoliv Zipfův přístup byl kritizován, jeho pravidlo dobře funguje pro tak různé jazyky jako je hebrejština, starogermánština, křováčtina a norština.

Shannon použil Zipfovo pravidlo jakožto aproximaci pro angličtinu s konstantou $A = 0.1$ a počtem slov $M = 12366$. Platí

$$\sum_{n=1}^{12366} p_n = 0.1 \sum_{n=1}^{12366} \frac{1}{n} = 1,$$

a pak je slovní entropie $H_w = 9.72$ bitů na slovo. Protože střední délka $\overline{w}$ anglických slov je 4.5, obdržíme odhad pro $H_{4.5} \simeq 9.72/4.5 = 2.16$ bitů na písmeno.

Jazyk jako zdroj VII



V každém anglickém textu jsou nejčastěji se vyskytujícími slovy the, and, of, to, be, a, in, I, a that. V textu je spousta dalších slov, která se nevyskytují tak často.

Entropie a redundance I

Proveďme nyní následující výpočet:

$$H_W = \sum_{k=1}^{\infty} H(W : |W| = k)P(|W| = k),$$

kde W je "***náhodný***" slovní výstup a $|W|$ označuje délku (nebo počet písmen) výstupu W. Tedy

$$\begin{aligned} H_W &= \sum_{k=1}^{\infty} H(X_1 X_2 \dots X_k)P(|W| = k) \\ &\leq \sum_{k=1}^{\infty} kH(X)P(|W| = k), \end{aligned}$$

kde $H(X)$ je entropie symbolů H_E a nerovnost bezprostředně vyplývá ze základní identity

$$H(U, V) \leq H(U) + H(V).$$

Máme pak $H_W \leq H_{4.5} \sum_{k=1}^{\infty} kP(|W| = k)$, tj.

$$H_W \leq H_{4.5}\overline{w}.$$

Entropie a redundance II

Je známo, že zdroj s entropií H má v abecedě $|\Sigma|$ jednoznačně dešifrovatelné zakódování s minimální průměrnou délkou slova $l(n)$ typického řetězce o n symbolech, přičemž

$$l(n) \approx nH/\log|\Sigma|.$$

Představíme-li si redundanci jako míru zbytečných symbolů (v procentech), je přirozené ji ***definovat*** následujícím přirozeným způsobem:

$$l(n) \simeq n(1 - R/100).$$

Z výše uvedeného pak obdržíme

$$R = 1 - H/\log_2|\Sigma|,$$

uvažujeme-li redundanci jako číslo mezi 0 a 1.

Entropie a redundance III

Přesný odhad redundance je obtížný; odhady zřejmě závisí na vybraném textu. Pokud je ale text zcela náhodný, bude jeho redundance rovna 0.

Uveďme následující příklad

	Bible	Měsíčník
H_1	4,086	4,152
H_{12}	2,397	2,824
R	0,413	0,285
$\overline{w}$	4,060	4,653

Entropie a redundance IV

Zajímavější je studium identických pasáží Bible přeložených do různých jazyků.

Zatímco samojština je jazyk s pouze 16 písmeny, z nichž 60% tvoří samohlásky, ruština před rokem 1917 používala abecedu o 35 písmenech.

Srovnání viz v následující tabulce:

	Angličtina	Ruština	Samojština
H_1	4,114	4,612	3,370
H_{12}	2,397	2,395	2,136
R	0,413	0,474	0,372
$\overline{w}$	4,060	5,296	3,174

Entropie a redundance V

Shannon odhadl, že entropii angličtiny lze redukovat na jeden bit na písmeno, což by nám dávalo řádově redundanci asi 75%. Tento odhad je však nutno interpretovat s jistou opatrností. Například výše uvedené neznamená, že můžeme rekonstruovat zprávu, ve které jsou písmena smazána s pravděpodobností $\frac{3}{4}$. Přesný způsob mazání je také důležitý. Jsou-li písmena smazána s pravděpodobností 0,5, pak například zpráva

MATHEMATICS IS BEAUTIFUL

může být obdržena ve tvaru

MTMASSBUFL;

a tedy by bylo opravdu obtížné získat zprávu pouze z narušeného textu.

Entropie a redundance VI

Bylo dokázáno, že kritická hodnota je $p \simeq 0,25$ a pro vyšší hodnotu je obdržení původní zprávy nemožné.

Jinak řečeno, ačkoliv je teoreticky možné zkrátit vytištěný text na čtvrtinu jeho současné délky, náhodné zkrácení není vhodný způsob, jak toho dosáhnout. Je nám ale jasné, že velkou redukci lze obdržet smysluplným zakódováním.

Např., lze bez obtíží akceptovat pravdivost následujících tvrzení:

- Vynecháme-li nějaké písmeno z textu, lze ho zpětně zrekonstruovat.
- Vynecháme-li všechny samohlásky z textu, lze text zpětně zrekonstruovat.

Entropie a redundance VII

Oba tyto případy jsou příklady suboptimálního zakódování a vezmeme-li redundanci angličtiny mezi 75% a 50%, dostaneme, že entropie H_E splňuje

$$1.19 \leq H_E \leq 2.38.$$

Přes svou dosti mlhavou a nepřesnou povahu má koncept redundance v kryptografii zásadní význam.

Redundance a šifrování I

Dále buď dán kryptosystém $\langle \mathbf{M}, \mathbf{K}, \mathbf{C} \rangle$, položíme $\mathbf{M}_N$ a $\mathbf{C}_N$ pro "náhodné" části zdrojového textu a odpovídajícího kryptogramu délky N. Nyní pak zřejmě

$$\begin{aligned} H(\mathbf{K}|\mathbf{C}_N) &= H(\mathbf{K}, \mathbf{C}_N) - H(\mathbf{C}_N) \\ &= H(\mathbf{M}_N, \mathbf{K}, \mathbf{C}_N) - H(\mathbf{C}_N) \\ &= H(\mathbf{M}_N, \mathbf{K}) - H(\mathbf{C}_N) \\ &= H(\mathbf{M}_N) + H(\mathbf{K}) - H(\mathbf{C}_N). \end{aligned}$$

Definujeme pak ***bod unicity*** U jakožto

$$U := \min\{N > 0 : H(\mathbf{K}|\mathbf{C}_N) = 0\}, \text{ tj.}$$

$$H(\mathbf{M}_U) + H(\mathbf{K}) - H(\mathbf{C}_U) = 0.$$

Redundance a šifrování II

Předpokládejme, že platí následující:

1. Přirozený jazyk, ve kterém šifrujeme, má tu vlastnost, že je dán vhodný odhad $H(M_N)$ jako

   $$H(M_N) \simeq NH,$$

   kde H je entropie jednoho symbolu jazyka;
2. kryptosystém má tu vlastnost, že všechny sekvence délky N symbolů mají stejnou pravděpodobnost jakožto kryptogram; jinak řečeno

   $$H(C_N) \simeq N \log|\Sigma|.$$

Redundance a šifrování III

To není nevhodný požadavek: každý dobrý kryptosystém by měl mít tuto vlastnost. Z výše uvedeného obdržíme

$$UH + H(K) - U\log|\Sigma| = 0,$$

tj.

$$U = \frac{H(K)}{\log|\Sigma| - H}.$$

Obvykle se rovněž předpokládá, že každý klíč můžeme vybrat se stejnou pravděpodobností a to znamená, že

$$U = \frac{\log|\mathbf{K}|}{\log|\Sigma| - H},$$

kde H je entropie symbolu zdroje.

Redundance a šifrování IV

Připomeňme, že existuje těsný vztah mezi bodem unicity kryptosystému a redundancí jazyka, ve kterém se přenáší zpráva.

Přitom redundance R jazyka s entropií H je určena vztahem

$$R = 1 - \frac{H}{\log|\Sigma|},$$

a tedy

$$U = \frac{\log|\mathbf{K}|}{\log|\Sigma| - H} = \frac{\log|\mathbf{K}|}{R\log|\Sigma|}.$$

Zejména pak má-li jazyk ***nulovou redundanci***, je pro každý kryptosystém splňující 1 a 2 bod unicity ***nekonečno***.

Redundance a šifrování V

Příklad 4.1

Předpokládejme, že šifrujeme pomocí jednoduché substituce tak, že máme právě 26! *klíčů.*

Uvažujeme-li $\log 26 = 4,7$ *a entropii anglického jazyka* H_E *rovnu 2 bitům na symbol, obdržíme*

$$U = \frac{\log 26!}{4,7 - 2} = \frac{88,4}{2,7} \simeq 32.$$

Jinak řečeno, při výše uvedené entropii anglického jazyka jsme obdrželi hodnotu bodu unicity rovnu 32 symbolům.

Redundance a šifrování VI

To pak celkem souhlasí s empirickým pozorováním Shannona (1949), který tvrdí, že pro bod unicity

"lze ukázat, že leží mezi krajními body 20 a 30. S 30 písmeny existuje téměř vždy jediné řešení pro kryptogram tohoto typu a s 20 můžeme najít nějaký počet řešení. "

Podobným způsobem Friedman (1973) tvrdí, že

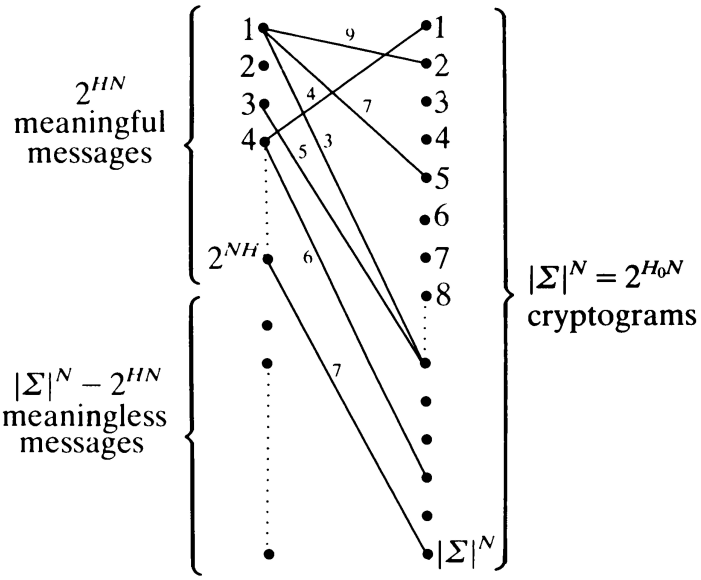
"Prakticky každý příklad 25 nebo více charakterů reprezentujících monoabecední zašifrování smysluplné zprávy v angličtině lze snadno vyřešit".

Redundance a šifrování VII

V roce 1977 M.E. Hellman navrhl alternativní a přítažlivé rozšíření výše zkoumaného přístupu.

V tomto modelu je prostor zpráv rozdělen do dvou disjunktních podmnožin.

První podmnožina obsahuje 2^{HN} smysluplných či typických zpráv, přičemž každá z těchto zpráv má a priori pravděpodobnost 2^{-HN}.



Redundance a šifrování VIII

Zbývající zprávy nemají v našem jazyku smysl a mají pravděpodobnost 0. Zároveň budeme předpokládat, že klíče jsou použity nezávisle na zprávě a se stejnou pravděpodobností.

Je-li C kryptogram, označme $Z(C)$ počet dvojic (M, K) takových, že zpráva M je smysluplná a

$$e(M_i, K_j) = C,$$

pak nepřítel, který zachytí C, bude v pochybách o použitém klíči.

Je-li $|Z(C)| > 1$, je kryptogram C zašifrován pomocí šifrování s ***falešným klíčem***.

Položme

$$s(C) = \max\{[Z(C) - 1], 0\}.$$

Redundance a šifrování IX

Očekávaná hodnota $s(C)$, totiž

$$\overline{s} = \sum_{C \in \mathbf{C}} s(C)P(C),$$

nám odhaduje očekávaný počet šifrování s falešným klíčem. Ale je okamžitě vidět, že

$$\overline{s} = \overline{z} - 1,$$

kde

$$\overline{z} = \sum_{C \in \mathbf{C}} Z(C)P(C) = \sum_{C \in \mathbf{C}} Z^2(C)/2^{NH}|\mathbf{K}|,$$

protože z definice modelu pro každý kryptogram C platí

$$P(C) = Z(C)/2^{NH}|\mathbf{K}|.$$

Redundance a šifrování X

Přitom evidentně

$$\sum_{C \in \mathbf{C}} Z(C) = 2^{NH}|\mathbf{K}|.$$

Aplikujeme-li na výše uvedené jednoduché lemma tvrdící, že pro všechna x_i splňující

$$\sum_{i=1}^{n} x_i = a,$$

máme

$$\sum_{i=1}^{n} x_i^2 \geq a^2/n,$$

a tedy obdržíme

$$\overline{Z} \geq (2^{NH}|\mathbf{K}|)^2/(|\mathbf{C}|2^{NH}|\mathbf{K}|) = 2^{NH}|\mathbf{K}|/|\mathbf{C}|.$$

Redundance a šifrování XI

Můžeme pak vyslovit následující

Věta 4.2

Za předpokladu platnosti výše uvedeného je očekávaný počet šifrování s falešným klíčem odhadnut jako

$$\overline{s} \geq (2^{NH}|\mathbf{K}|/|\mathbf{C}|) - 1.$$

Píšeme-li nyní $\mathbf{K} = 2^{H(K)}$, $\mathbf{C} = 2^{NH_0} = |\Sigma|^N$, kde H_0 je entropie jazyka, lze výše uvedenou větu přepsat jakožto

$$\overline{s} \geq 2^{NH+H(K)-NH_0} - 1,$$

přičemž pravá strana je rovna nule přesně v bodu unicity.

Redundance a šifrování XII

Příklad 4.3

Předpokládejme, že šifrujeme pomocí Vigenerova šifrování otevřený text délky 100 klíčem délky 80 tak, že máme

Víme, že $H_0 = 4,7 = \log 26$ *a* $H = H_E \simeq 1,5$ *bitů,*
$H(K) = 80 \cdot \log 26 = 376$.

Obdržíme pak průměrně alespoň

$$2^{376+100\cdot(1,5-4,7)} = 2^{376-320} \simeq 2^{56}$$

různých šifrování s falešným klíčem pro kryptogram o 100 písmenech.