

## 10. cvičení

Datové struktury I, 6. 12. 2022

<https://iuuk.mff.cuni.cz/~chmel/2223/ds1/>

**Věta.** Tabulkové hashování je 3-nezávislé.

**Úloha 1** (Tuhle větu si dokážeme)

Dokažte předcházející větu s následujícím postupem. Mějme  $a, b, c \in \mathbb{Z}_2^{\ell}$ ,  $x \neq y \neq z \neq x \in \mathbb{Z}_2^w$ , a použijeme tabulkové hashování s  $d$  částmi. Pak chceme ukázat, že  $\Pr_{h \in \mathcal{H}}[h(x) = a \wedge h(y) = b \wedge h(z) = c] \leq \frac{1}{m^3}$ .

(a) Prvně si uvědomme, že pokud máme jen jednu část, a tedy jednu tabulku, tvrzení je triviální.

Dále mějme alespoň dvě části. Protože  $x, y, z$  jsou různé, musí se (po dvou) lišit alespoň v jedné části.

(b) Prvně prozkoumejme případ, kdy existuje část  $i$ , že  $x^i, y^i, z^i$  jsou všechny různé. Mějme jakkoliv zvolené ostatní tabulky, kromě tabulky  $T_i$ . S jakou pravděpodobností můžeme zvolit funkci pro tabulku  $T_i$  tak, že  $h(x) = a, h(y) = b, h(z) = c$ ?

(c) Jinak existují (BÚNO) části  $i, j$  takové, že  $z^i = x^i \neq y^i$  a  $y^j = x^j \neq z^j$ . Potom máme následující soustavu rovnic, kde  $v_x, v_y, v_z$  jsou vyXORované výsledky z ostatních tabulek:

$$T_i[x^i] \oplus T_j[x^j] \oplus v_x = a$$

$$T_i[y^i] \oplus T_j[y^j] \oplus v_y = b$$

$$T_i[z^i] \oplus T_j[z^j] \oplus v_z = c$$

Opět si představme, že  $v_x, v_y, v_z$  už známe. S jakou pravděpodobností budou náhodně volené tabulky  $T_i, T_j$  splňovat tuto soustavu rovnic?

(d) Uvědomte si, že toto stačí.

**Lemma.** Necht'  $\mathcal{F}$  je  $c$ -univerzální rodina funkcí  $f: \mathcal{U} \rightarrow [r]$ ,  $\mathcal{G}$  je  $(2, d)$ -nezávislá rodina funkcí  $g: [r] \rightarrow [m]$ . Pak  $\mathcal{H} = \mathcal{F} \circ \mathcal{G} = \{f \circ g: f \in \mathcal{F}, g \in \mathcal{G}\}$  je  $(2, c')$ -nezávislá, kde  $c' = d \cdot (\frac{cm}{r} + 1)$ .

**Tvrzení.** Pro prvočíslo  $p$  a délku vektoru  $d$  definujeme třídu hashovacích funkcí  $\mathcal{R} = \{h_a: a \in \mathbb{Z}_p\}$ , kde  $h_a(x) = \sum_{i=0}^{d-1} x_{i+1} a^i$ . (Bereme  $x \in \mathbb{Z}_p^d$ , indexujeme od jedničky.) Pak  $\mathcal{R}$  je  $d$ -univerzální, a  $\mathcal{R} \circ \mathcal{L}$  je  $(2, 5/2)$ -nezávislý systém, máme-li  $p \geq 4dm$ , kde  $\mathcal{L} = \{h_{a,b}(x) = (ax + b \pmod p) \pmod m: a, b \in \mathbb{Z}_p\}$ .

**Úloha 2** (I tohle si dokážeme)

Dokažte předcházející tvrzení. Začneme s univerzalitou. Chceme tedy dokázat, že pro každá dvě slova  $x \neq y \in \mathbb{Z}_p^d$  je  $\Pr_a[\sum_{i=0}^{d-1} x_{i+1} a^i = \sum_{i=0}^{d-1} y_{i+1} a^i] \leq \frac{d}{p}$ .

Hint: *ηλβν ηγλαζου ωωτ εγςκλζ - ηυεζζογ ρ εζηαζου ρυ ρ εζηαζου σουηιουδ ηρζζεγ εζ 'ακζ ηυεζζογ ηεζ ηυαηηκζ*

Dále použijeme lemma, které jsme si dokázali společně na začátku hodiny, a fakt z přednášky, že  $\mathcal{L}$  je  $(2, 2)$ -nezávislý systém pro  $p \geq 4m$ . Tím jsme už vše dokázali.

**Úloha 3** (Univerzální modulo může rozbít univerzalitu)

Ukažte, že pokud máme univerzální systém hashovacích funkcí  $\mathcal{H}$ , pak systém  $\mathcal{H}'$ , kde ke každé fci navíc přidáme modulo  $m$ , už nemusí být univerzální. Formálně: Dokažte, že pro každé  $c$  a  $m > 1$  existuje univerzum  $\mathcal{U}$  a systém  $\mathcal{H}$  z  $\mathcal{U}$  do  $\mathcal{U}$  tak, že  $\mathcal{H}$  je univerzální, ale  $\mathcal{H}'$  už není  $c$ -univerzální.

**Úloha 4** (Vyhledávání jehly v textu)

Vymyslete algoritmus na nalezení všech výskytů podřetězce  $x$  délky  $n$  v textu  $T$  délky  $m$  pomocí hashování, který běží v průměrném čase (tj. ve střední hodnotě)  $\mathcal{O}(n + m + k \cdot n)$ , kde  $k$  je počet výskytů  $x$  v  $T$ .