9. CVIČENÍ Z DATOVÝCH STRUKTUR 1, ZS24/25

Hurá, opět hešování!!! Univerzální, k-nezávislé, tabelační a možná i perfektní.

1. *Opakování definic: Modulo univerzálního systému nemusí být univerzální.* 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 $m > c$ 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. (Hint: podívejte se na příklad 3 z minula.)

2. *Tabelační (tabulkové) hešování.* Dokažte, že tabelační hešování je 2-nezávislé a *není* 4-nezávislé, pokud používáme alespoň dvě tabulky.

3. *Hlavní chod.* Dokažte, že tabelace je 3-nezávislá.

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

i) 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.

ii) Začneme s případem, 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 $h(x) = a, h(y) = b, h(z) = c$?

iii) 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?

iv) Uvědomte si, že toto stačí.

4. *Bonus: Perfektní hešování dle Fredmana, Komlóse a Szemerédiho (FKS).* Máme dánu n-prvkovou množinu S jako podmnožinu nějakého (obrovského) univerza $\mathcal{U}$, např. 64-bitové integery. Cílem je navrhnout pro S datovou strukturu velikosti $O(n)$, která zvládne pro zadaný dotaz x zjistit, jestli x náleží v S, v konstantním čase vždy. (V čem klasická hešovací tabulka velikosti $O(n)$ nesplňuje požadavky?)

a) Připomeňte si narozeninový paradox, tedy nejmenší počet lidí s rovnoměrně náhodnými narozeninami takový, aby s pravděpodobností alespoň 50% měli dva stejné narozeniny.
b) Hešování s úplně náhodnou (nebo c-univerzální) hešovací funkcí funguje podobně. Jak zhruba musíme mít velkou tabulku, aby nastala kolize s pravděpodobností méně jak 50%?
c) Perfektní hešování vybudujeme takto: Pořídíme si hešovací funkci $h : \mathcal{U} \to [m]$ pro $m = O(n)$, kterou vybereme náhodně z c-univerzální rodiny. Pro každou přihrádku $i \in [m]$ zavedeme tabulku druhé úrovně, která bude dost velká, aby tam nenastala kolize s pravděpodobností alespoň 50% (pokud by kolize nastala, změníme hešovací funkci).
d) Teď už stačí jen omezit celkovou velikost tabulek druhé úrovně ve střední hodnotě.

Užitečné definice

Definition 1 (c-univerzální systém fcí) *Systém $\mathcal{H}$ funkcí $h : \mathcal{U} \to [m]$ je c-univerzální pro $c > 0$, pokud pro všechna $x \neq y$ platí $\Pr_h[h(x) = h(y)] \leq \frac{c}{m}$. Systém $\mathcal{H}$ je univerzální, pokud je c-univerzální pro nějaké $c > 0$.*

Definition 2 (k-nezávislý systém fcí) *Systém $\mathcal{H}$ funkcí $h : \mathcal{U} \to [m]$ je (k, c)-nezávislý pro nějaká $k \geq 1, c > 0$, pokud $\Pr_{h \in \mathcal{H}}[h(x_1) = a_1 \wedge \ldots \wedge h(x_k) = a_k] \leq \frac{c}{m^k}$ pro libovolná $x_1, \ldots, x_k$ různá, $a_1, \ldots, a_k$ ne nutně různá.*

Systém $\mathcal{H}$ je k-nezávislý, pokud je (k, c)-nezávislý pro nějakou nezávislou konstantu c.

Definition 3 (Tabulkové hashování) *Představme si, že chceme zahashovat w-bitové řetízky do ℓ-bitových řetízků. Řetízek $x \in \{0,1\}^w$ pak rozložíme do d částí délky w/d, které značíme x^i. Můžeme tedy psát $x = x^1 x^2 \ldots x^d$. Pak generování naší hashovací funkce $h : \{0,1\}^w \to \{0,1\}^\ell$ vypadá tak, že vybereme uniformně náhodně d funkcí $T_i : \{0,1\}^{w/d} \to \{0,1\}^\ell$ (tyto reprezentujeme tabulkou, proto tabulkové hashování). Vyhodnocujeme pak $h(x) = \bigoplus_{i=1}^d T_i(x^i) = T_1(x^1) \oplus T_2(x^2) \oplus \cdots \oplus T_d(x^d)$, kde $\oplus$ značí XOR (po jednotlivých bitech).*

Theorem 1 (Markovova nerovnost) *Buď X nezáporná náhodná veličina. Pak $\forall \varepsilon > 0$ platí $P[X \geq \varepsilon] \leq \frac{\mathbb{E}[X]}{\varepsilon}$. Ekvivalentně pro jakékoliv $d \geq 1$, $P[X \geq d \cdot \mathbb{E}[X]] \leq \frac{1}{d}$.*