# Универсальное семейство хеш-функций

◁ семейство хеш-функций $\mathcal{H}$

$\forall\ h \in \mathcal{H},\ h : K \to \{0, 1, \dots, m-1\}$

$\forall\ k_1, k_2 \quad k_1 \neq k_2 \qquad P_{h \leftarrow \mathcal{H}}\{h(k_1) = h(k_2)\} \leqslant \frac{1}{m}$

$n/m$

Утв 1: Сложность неуспешного поиска $O(1+\alpha)$

▷ $\exists\ X_{ke}$ – случайная величина $h(k) = h(e)$

$E\left[\sum_k X_{ke}\right]$ – длина цепочки

$\sum_k E[X_{ke}] \leqslant \sum_k \frac{1}{m} = \frac{n}{m} = \alpha$ ◁

Утв 2: Сложность удачного поиска $O(1+\alpha)$

▷ $\exists\ X_{ke}$ – — " — $\qquad E[X_{ee}]$

$E\left[\sum_k X_{ke}\right] = \sum_k E[X_{ke}] = 1 + \sum_{k \neq e} E[X_{ke}] = 1 + \frac{n-1}{m} \leqslant 1 + \alpha$

◁

Th. $\mathcal{H} = \{((ak + b) \bmod p) \bmod m\}$

$p$ – простое, $p > m$, $a \in \{1, 2, \dots, p-1\}$, $b \in \{0, 1, \dots, p-1\}$

$\mathcal{H}$ – универсальное семейство хеш-функций

▷ ◁ $k_1 \neq k_2 \bmod p$ $\qquad |\mathcal{H}| = p \cdot (p-1)$

$t_1 = ak_1 + b \bmod p$

$t_2 = ak_2 + b \bmod p$

$k_1 \neq k_2 \bmod p \Rightarrow t_1 \neq t_2$

$t_1 = t_2 \Leftrightarrow a(k_2 - k_1) = 0 \bmod p$, $k_2 - k_1 \neq 0$

$$\#\{(a,b)\} = p\cdot(p-1) \quad \longleftrightarrow \quad \#\{(t_1,t_2)\}$$

$(a,b)$ распределены равномерно на
$$\{1,\dots,p-1\}\times\{0,\dots,p-1\}$$

$\Rightarrow (t_1,t_2)$ распределены равномерно на
$$\{0,\dots,p-1\}^2 \setminus \{(i,i)\}$$

$$P[t_1 = t_2 \bmod m] = \frac{l}{m}$$

Зафиксируем $t_1$

$$\#[t_2,\ t_2 = t_1 \bmod m] \le \left\lceil \frac{p}{m} \right\rceil - 1 \le$$

$$\le \frac{p+m-1}{m} - 1 = \frac{p-1}{m}$$

$$P_{\substack{2\\ t_1\neq t_2}}[t_1 = t_2 \bmod m] = \frac{1}{p}\cdot\frac{p-1}{m} \le \frac{1}{m}$$

◁

## Совершенное хеширование

◁ статическую постановку

▷ у нас дано $n$ ключей

Посчитаем $E\left[\#\text{ коллизий в таблице размером } n^2\right]$

▷ $m = n^2$

Воздействуем УСХФ

$$E[X] = \sum_{k_1\neq k_2} E[h(k_1)=h(k_2)] = \frac{1}{m}\cdot C_n^2 =$$

$$= \frac{1}{n^2}\cdot\frac{n\cdot(n-1)}{2} < \frac{1}{2}$$

## Неравенство Маркова

$P[X \geq \alpha] \leq E[X]/\alpha$

▷ $E[X] = \sum\limits_{x_i} x_i \cdot P[X = x_i] = \sum\limits_{x_i < \alpha} x_i P[X = x_i] +$

$+ \sum\limits_{x_i \geq \alpha} x_i P[X = x_i] \geq \sum\limits_{x_i \geq \alpha} x_i P[X = x_i] \geq \sum\limits_{x_i \geq \alpha} \alpha P[X = x_i] =$

$= \alpha \cdot P[X \geq \alpha]$

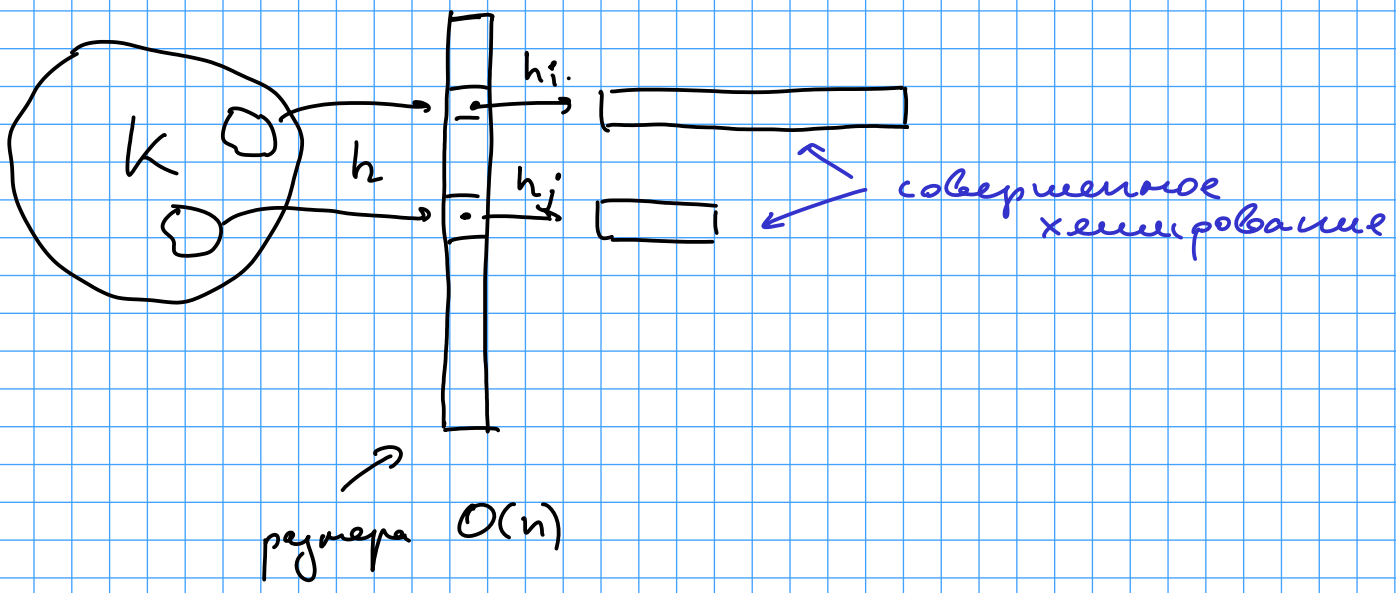
$E[X] \geq \alpha \cdot P[X \geq \alpha]$ ◁

Применим нер-во Маркова для $X$:

$P[X \geq 1] \leq \frac{1}{2}/1 \Rightarrow$ с вер-тью $\leq \frac{1}{2}$ будет коллиз.

Проблема: $O(n^2)$ памяти

Решение: двухуровневое хеширование




$n_i$ – количество эл-ов в ячейке $i$ первой хеш-таблицы

Кол-во памяти: $\sum\limits_i n_i^2 + O(n)$

$\sum n_i^2 \leq \left(\sum n_i\right)^2 = n^2$

Хитрость: $n_i^2 = n_i + 2 \cdot \frac{n_i(n_i-1)}{2} = n_i + 2C_{n_i}^2$

$$E\left[\sum_i n_i^2\right] = E\left[\sum_i n_i + 2C_{n_i}^2\right] =$$

$$= n + 2\underbrace{E\left[\sum_i C_{n_i}^2\right]}_{E[\#\text{коллизий}]} = n + 2\cdot\left(\frac{1}{n}\cdot C_n^2\right) =$$

$$= n + \frac{2}{n}\cdot\frac{n\cdot(n-1)}{2} = 2n - 1 \le 2n$$

$$E[\text{память} \ge 4n] \le \frac{2n}{4n} \le \frac{1}{2}$$

Доступ в таблице за $O(1)$ в худшем

## Фильтр Блюма

Способ хранения множества, если разрешены ошибки вида false positive

Возьмём набор хеш-функций из УСХФ.

$h_1, h_2, \dots, h_k$

Add (K):

Поставим 1 в ячейки $T[h_1(k)]$, $T[h_2(k)]$...

Find (K):

Проверим, что $T[h_i(k)] = 1$ для всех $i$

$\exists$ есть $n$ эл-ов

Вероятность false positive:

$\neq h_1$

Вероятность отсутствия коллизии $ch_i = \left(1 - \frac{1}{m}\right)^n$

Вероятность отсутствия коллизии со всеми $h_i$

$\left(1 - \frac{1}{m}\right)^{nk}$ (в одной ячейке)

По коллизии по всем $l$ ячейкам

$$\left(1 - \left(1 - 1/m\right)^{ne}\right)^l = \left(1 - e^{-ne/m}\right)^l$$