Heavy Hitters

stream of elts $a_1, a_2, \ldots, a_t, \ldots$ e.g. IP addresses

each $a_i \in [n]$ ex: $n = 2^{64}$

For any time $t \leq n$

let f_x^t : # times elt x appears in $a_1, a_2, \ldots, a_t$ (frequency)

Goal: return all elts x s.t. f_x large $> \frac{n}{k}$

using $o(n,t)$ space

If $f_x^t > \frac{n}{k}$, add x to list of "heavy hitters"

Provably impossible with sublinear space

Modified goal:

① If $f_x > \frac{n}{k}$ add it to list.

② If x added to list, then w.p. $\geq 1-\delta$, $f_x^t > \frac{n}{k} - \varepsilon n$

Ex Suppose $k = 20$, $\varepsilon = 0.01$, $\delta = \frac{1}{2^{10}}$

① If $f_x > \frac{n}{20}$, x is definitely on list of HH
$= 0.05n$

② if x added to list, then with prob $\geq 1 - \frac{1}{2^{10}}$

$f_x > \frac{n}{k} - \varepsilon n = 0.05n - 0.01n = 0.04n$

Modified goal:

① If $f_x > \frac{n}{k}$ add it to list.

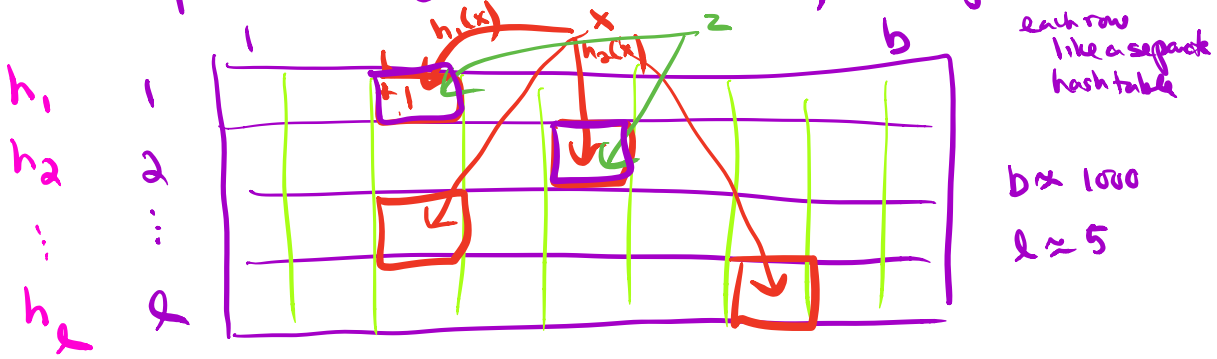
② If x added to list, then w.p. $\geq 1-\delta$, $f_x^t > \frac{n}{k} - \varepsilon n$

use elegant, very simple data structure Count-Min-Sketch.

Count Min Sketch

designer specifies k, δ, ε $\Rightarrow$ b, ℓ.

keep 2D array: ℓ hash tables, each of size b



when elt x shows up

Inc(x): $\forall\ 1 \le j \le \ell$ increment $CMS[j][h_j(x)]$

Observation ①: at any time t, $\forall_j \forall_x$ $CMS[j][h_j(x)] \ge f_x^+$

Count(x): return $\min_{1 \le j \le \ell} CMS[j][h_j(x)]$

by Observation ① $\text{Count}(x) \ge f_x^+$ if $\text{Count}(x) > \frac{1}{k}$ add x to HH list.

Assumptions:

① hash fns behave like random maps in the following sense

$\forall_{x \ne y} x, y \in [n]$ $\forall\ 1 \le j \le \ell$ $\Pr(h_j(x) = h_j(y)) = \frac{1}{b}$

② hash fns $1 \le j \le \ell$ are indep of each other

Analysis Fix time t, elt x

Define $Z_j = CMS[j][h_j(x)]$ $Z_j \ge f_x^+$

$$Z_j = f_x^+ + \sum_{\substack{y \ne x \\ y \in [n]}} f_y^+ \, W_{xy}^+$$

W_{xy}^+ is a r.v.

$W_{xy}^+ = \begin{cases} 1 & \text{if } h_j(x) = h_j(y) \\ \end{cases}$

$$E(Z_j) = f_x^+ + \sum_{y \neq x} f_y^+ \underbrace{E(\omega_{xy}^+)}_{\Pr(h_j(x)=h_j(y)) = \frac{1}{b}} \leq f_x^+ + \frac{t}{b} \leq f_x^+ + \frac{n}{b}$$

$t \leq n$

$$E\left(Z_j - f_x^+\right) \leq \frac{n}{b}$$

By Markov's Ineq ($\alpha = 2$)

$$\Pr\left(Z_j - f_x^+ \geq \frac{2n}{b}\right) \leq \frac{1}{2} \quad (*)$$

Markov's Inequality

X is a nonnegative r.v.

$$\Pr(X > \alpha E(X)) \leq \frac{1}{\alpha}$$

(α: const ≥ 0)

$$\Pr\left(\underbrace{\text{count}(x)}_{\min_{1 \leq j \leq \ell} Z_j} \geq f_x^+ + \frac{2n}{b}\right) \leq \frac{1}{2^\ell}$$

$$= \Pr\left(Z_1 \geq f_x^+ + \frac{2n}{b}, Z_2 \geq f_x^+ + \frac{2n}{b}, \ldots, Z_\ell \geq f_x^+ + \frac{2n}{b}\right)$$

$$= \Pr\left(Z_1 \geq f_x^+ + \frac{2n}{b}\right) \Pr\left(Z_2 \geq f_x^+ + \frac{2n}{b}\right) \cdots \Pr\left(Z_\ell \geq f_x^+ + \frac{2n}{b}\right)$$

$=$ indep of reasoning $=$ indep of Z_j's

by $(*)$ $\leq \frac{1}{2^\ell}$

$$\Pr\left(\underbrace{\text{count}(x)}_{\min_{1 \leq j \leq \ell} Z_j} \geq f_x^+ + \underbrace{\frac{2n}{b}}_{\varepsilon n}\right) \leq \underbrace{\frac{1}{2^\ell}}_{\delta}$$

k, ε, δ

Modified goal:

① If $f_x > \frac{n}{k}$ add it to list.

② If x added to list, then w.p. $\geq 1-\delta$, $f_x^+ > \frac{n}{k} - \varepsilon n$

$\delta = \frac{1}{2^\ell}$ $\quad \varepsilon n = \frac{2n}{b}$ $\quad \left(\varepsilon = \frac{2}{b}\right)$

$\ell = \log_2\left(\frac{1}{\delta}\right)$ $\quad \equiv b = \frac{2}{\varepsilon}$

$b = 200$

$\ell = 5 \Rightarrow \varepsilon = \frac{2}{b} = \frac{2}{200} = 0.01$

$\ell = 5 \Rightarrow$ error prob is $\frac{1}{2^5} = \frac{1}{32}$

$\frac{n}{k} = 0.04n$

✓ if x has $f_x > 0.04n$ × definitely output

if output x, then with prob $\geq \frac{31}{32}$, $f_x \geq \frac{n}{k} - \varepsilon n$

$= 0.04n - 0.01n$

$= 0.03n$

p prime number $> n$

$$\mathcal{H} = \left\{ h(x) = (ex + g) \bmod p \bmod b \;\middle|\; \begin{array}{l} 1 \leq e \leq p-1 \\ 0 \leq g \leq p-1 \end{array} \right\}$$

$|\mathcal{H}| = p(p-1)$ universal family.

$h(x) = (e_i x + g_i) \bmod p \bmod b \rightarrow e_1, g_1$

$\rightarrow e_2, g_2$

$\rightarrow e_3, g_3$

If h is chosen uniformly at random from $\mathcal{H}$

$\forall x \neq y \quad \Pr_{h \in \mathcal{H}}(h(x) = h(y)) = \frac{1}{b}$

Markov's Inequality

X nonnegative r.v. . For any positive const c,

$$\Pr(X \geq cE(X)) \leq \frac{1}{c}$$

Proof:

$$E(X) = \sum_{x < cE(X)} x \Pr(X=x) + \sum_{x \geq cE(X)} x \Pr(X=x)$$

$$\geq 0 + \sum_{x \geq cE(X)} cE(X) \Pr(X=x)$$

$$= cE(X) \Pr(X \geq cE(X))$$

$$\Rightarrow \Pr(X \geq cE(X)) \leq \frac{1}{c}$$