Last time:

- matching
- intro hashing
- universal hashing

Key idea:

- put the randomness in hash function instead of assuming data is random!
- gives all the nice properties:
  - data distributed randomly throughout table
  - hash fn efficient to store
    efficient to compute.

Today:

3 applications of hashing & lossy compression

- Bloom filters
- Heavy hitters & count-min sketch
- Distinct elts

with short review of variance & tail bounds

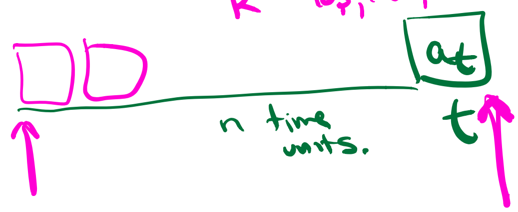
Heavy hitters

stream of elts $\quad$ 3 5 7 3 4
$a_1, a_2, a_3, \ldots,$

at any time t,
let f_x^t : # times seen element x in
$a_1, a_2, \ldots, a_t$

Goal: when elt shows up, output that element
if $f_x^t > \frac{n}{k}$ $\quad$ $\frac{n}{k}$ $\quad$ millions, billions
10's, 100's, 1000's

Such an element is a heavy hitter.



space used proportional to # unique elts

not possible to solve this problem exactly with sublinear space.

Modified goal (ε, δ)

① If $f_x^t > \frac{n}{k}$ $\quad$ output x.

② If x is output, then with prob at least $1-\delta$ it is the case that $f_x^t \geq \frac{n}{k} - \varepsilon n$

For example: suppose $k = 25$, $\varepsilon = 0.01$, $\delta = \frac{1}{2^{10}}$

① If $f_x^t > \frac{n}{25} = 0.04n$, then output x

② If x output then w.prob $\geq \boxed{1 - \frac{1}{2^{10}}}$

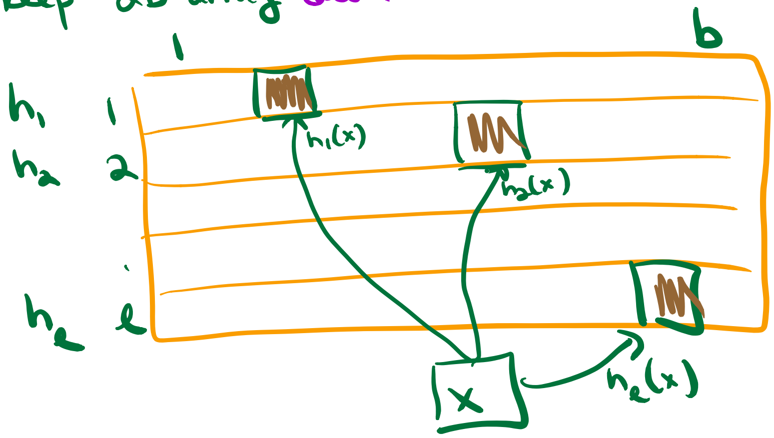
$$f_x^t \geq \underbrace{0.04n}_{\frac{n}{k}} - \underbrace{0.01}_{\varepsilon} n = 0.03n$$

Count-min sketch:

Designer specifies $n, k, \delta, \varepsilon$

$\Rightarrow b, \ell$

keep 2D array called CMS



each row is hash table of size b.

typical values for b & ℓ might be
$b \approx 1000$
$\ell \approx 5$

when element x shows up

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

Observe: $\forall$ time t, $\forall j$, $\forall x$ $\quad CMS[j][h_j(x)] \ge f_x^t$

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

if this value $\ge \frac{n}{k}$ output x as HH.

by observation: $\text{Count}(x) \ge f_x^t$

Construction

① hash fns behave randomly

$\forall\, x, y,\ x \ne y \quad \forall\, 1 \le j \le \ell \quad \Pr\left(h_j(x) = h_j(y)\right) = \frac{1}{b}$

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

$\mathcal{H}$ universal class of hash fns.

Analysis. Fixing $t \le n$, elt x that arrives at time t.

$Z_j = CMS[j][h_j(x)]$ random variable.

$$Z_j = f_x^t + \sum_{y \neq x} f_y^t W_{xy}$$

$$\boxed{W_{xy}^j} = \begin{cases} 1 & h_j(x) = h_j(y) \\ 0 & \text{o.w.} \end{cases}$$

$$E(Z_j) = f_x^t + \sum_{y \neq x} f_y^t \underbrace{E[W_{xy}^j]}_{\frac{1}{b}}$$ by linearity of expectation

$$\le f_x + \frac{t}{b} \le f_x^t + \frac{n}{b}$$

$$E(Z_j - f_x^t) \le \frac{n}{b}$$

$$\Pr\left(Z_j - f_x^t > \frac{2n}{b}\right) \le \frac{1}{2}$$

$$\boxed{\begin{array}{l} \text{Markov's Inequality} \\ X \text{ is nonnegative r.v.} \\ \Pr(X \ge c \cdot E(X)) \le \frac{1}{c} \end{array}}$$

$$\Pr\left(Count(x) - f_x^t > \frac{2n}{b}\right) \le \frac{1}{2^\ell} \Leftarrow \text{bad event}$$

$Z_1 - f_x^t > \frac{2n}{b}$
$Z_2 - f_x^t > \frac{2n}{b}$
$\vdots$
$Z_\ell - f_x^t > \frac{2n}{b}$

Conclusion: $$\Pr\left(Count(x) \ge f_x^t + \frac{2n}{b}\right) \le \frac{1}{2^\ell} *$$

εn $\quad$ δ

Modified goal (ε, δ)

⇒ ① If $f_x^t > \frac{n}{k}$ output x.

⇒ ② If x is output, then with prob at least $1-\delta$ it is the case that $\boxed{f_x^t > \frac{n}{k} - \varepsilon n}$

$\frac{n}{2k}$

choose b & ℓ

so that ⓐ $\frac{2n}{b} = \varepsilon n$ ⓑ $\frac{1}{2^\ell} = \delta$

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

$\varepsilon = \frac{1}{2k}$

$b = 4k$

$\delta = \frac{1}{2^{100}}$

$\ell = 100$

Suppose $\boxed{f_x^t < \frac{n}{k} - \varepsilon n}$

$$\Pr\left(Count(x) \ge \frac{n}{k}\right) \le \Pr\left(Count(x) > \underbrace{f_x^t + \varepsilon n}_{< \frac{n}{k} - \varepsilon n}\right) \le \delta$$

Universal hash family $\mathcal{H}$

$U = \{0, 1, \ldots, u-1\}$

each $h \in \mathcal{H}$ $\quad h : \underline{U} \to \{0, 1, \ldots, b-1\}$

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

$$\forall x \neq y \quad \Pr_{h \in \mathcal{H}}\left(h(x) = h(y)\right) \leq \frac{2}{b}$$

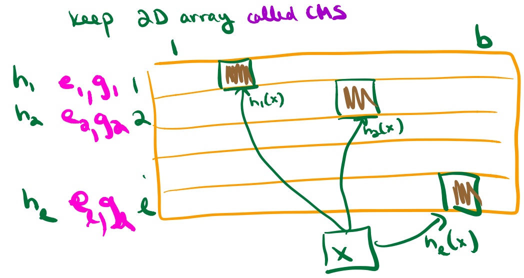
Choose any prime # $p > u$

$$\mathcal{H} = \left\{ h(x) = \overline{(e x + g) \bmod p \bmod b}, \right.$$

where $\quad 1 \leq e \leq p-1$

$\qquad\quad 0 \leq g \leq p-1 \}$

keep 2D array called CMS



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