Heavy Hitters (HH)

There is no alg to solve HH in one pass with sublinear space.

We relax our requirements $\Rightarrow$ ε-HH

Suppose at time t $\quad \forall i \in U$

x_i^t = # times seen i up to time t

Will require estimates

$$\tilde{x}_i^t \in x_i^t \pm \varepsilon t \qquad (*)$$

that uses space $<<$ $\min(t, n)$

(# distinct elts)

Idea: counting Bloom filter

$h: U \to [m]$ random universal A array with m elts

when elt i arrives

$A[h(i)]++;$

$\tilde{x}_i^+ = \underbrace{A[h(i)]}_{\text{r.v.}} = x_i^+ + \sum_{j \neq i} x_j^+ 1_{h(j)=h(i)}$

$E[A[h(i)]] = x_i^+ + \sum_{j \neq i} \frac{x_j^+}{m} \leq x_i^+ + \frac{n}{m}$

with $m = \frac{1}{\varepsilon}$ satisfy (*) in expectation

Want low error w/ high probability

$\Rightarrow$ amplify success probability w/ indep repetitions

Do same thing independently ℓ times w/ ℓ independently selected hash fns

$h_1, h_2, \ldots h_\ell$

$\downarrow \quad \downarrow \quad \ldots \quad \downarrow$

$A_1 \quad A_2 \quad \ldots \quad A_\ell$

Called CountMin sketch

How should we combine entries? We have overestimates

$$\Rightarrow \tilde{z}_i^+ = \min_{1 \le j \le \ell} A_j[h_j(i)]$$

$$E(x_i^+ - \tilde{x}_i^+) \le \frac{n}{m}$$

$\Rightarrow$ Markov Inequality

$$Pr\left(x_i^+ - \tilde{x}_i^+ > \frac{2n}{m}\right) \le \frac{1}{2}$$

$$Pr\left(\tilde{z}_i^+ > \frac{2n}{m}\right) = Pr(\text{all estimates too high}) \le \frac{1}{2^\ell}$$

How to choose parameters?

$\tilde{z}_i^+ \in [x_i^+, x_i^+ + \frac{2n}{m}]$ with prob $\geq 1 - \frac{1}{2^\ell}$

$\frac{2}{m} = \varepsilon$ $\quad$ $\ell = \log_2(\frac{1}{\delta})$

$\tilde{z}_i^+ \in [x_i^+, x_i^+ + \varepsilon n]$ with prob $\geq 1 - \delta$

useful when $x_i^+ > 2\varepsilon n$

Space usage: ℓm counters of $\log_2 T$ bits each

$\Rightarrow O(\frac{\log \frac{1}{\delta}}{\varepsilon} \log T)$

also space for hash fns: $O(\ell \log U)$

Example: want to output all items that constitute $\geq 4\%$ of flow and output no items that constitute $< 2\%$