Lecture #5 1/26

Eli Goldin

Last time: stream $x_1 \dots x_m \in [n]$, $n = \#IP = 2^{32}$

Goal: estimate # distinct.

Algo: Bottom-k

using $h: [n] \to [q]$.

$k = \Theta(1/\varepsilon^2)$ enough

for $(1 \pm \varepsilon)$ approx $(\delta > \frac{1}{\varepsilon^2})$.

Add. remarks:

1) $\to [q]$.

$h: [n] \to \{0, 1, \dots M-1\}$.

$\to \{0, 1/M, 2/M, \dots \frac{M-1}{M}\}$.

M large enough.

$h(x)$ vs $h(y)$.

Tiny issues $h(x)$ can collide $h(y)$

on different items.

$M = n^3$. $\mathbb{E}[\text{collisions}] \le n^2/M$
$\le 1/n$.

2) fully random?
possible with "limited randomness".

Def: $\mathcal{H}$ is k-wise independent
if $\forall a_1, a_2, \dots a_k \in \{0, 1/M, \dots \frac{M-1}{M}\}$,
$\forall x_1 \dots x_k \in [n]$ distinct.
$\Pr_{h \in \mathcal{H}}[h(x_i) = a_i \ \forall i = 1..k] = \frac{1}{M^k}$.

$\Pr[h(x)=a \wedge h(y)=a]$

$\mathcal{H}$ univ. if
$\Pr[h(x)=h(y)]$
$= \sum_{a \in \{0, 1/M, \dots \frac{M-1}{M}\}} \Pr[h(x)=h(y)=a]$
$= \sum_a \frac{1}{M^2} = \frac{M}{M^2}$
$= 1/M$ ⊗

Rem: if $\mathcal{H}$ is 2-wise indep
$\Rightarrow$ univ.

Rem: $\forall k \ge 2$, can construct
k-wise indep $\mathcal{H}$ of
size $\lg |\mathcal{H}| \le O(k \cdot \lg n)$.

Thm [Thorup '13]: for Bottom-k also,
enough to use 2-wise indep
h.f.

Problem: Stream $x_1, x_2, \ldots x_m \in [n]$.

Most frequent?

Def: freq. vector f_x = how many times x appeared in str.

$f \in \mathbb{R}^n$.

$\|f\|_1 = \sum_x f_x = m$ = len. stream.

→ x s.t. f_x is maximal.

Simple sol: store vector f explicitly

→ space $O(n)$.

Thm: - can't do exactly.

- can't do 2-approx, using random.

Bad case: all items appear once

except 1 item app 2x.

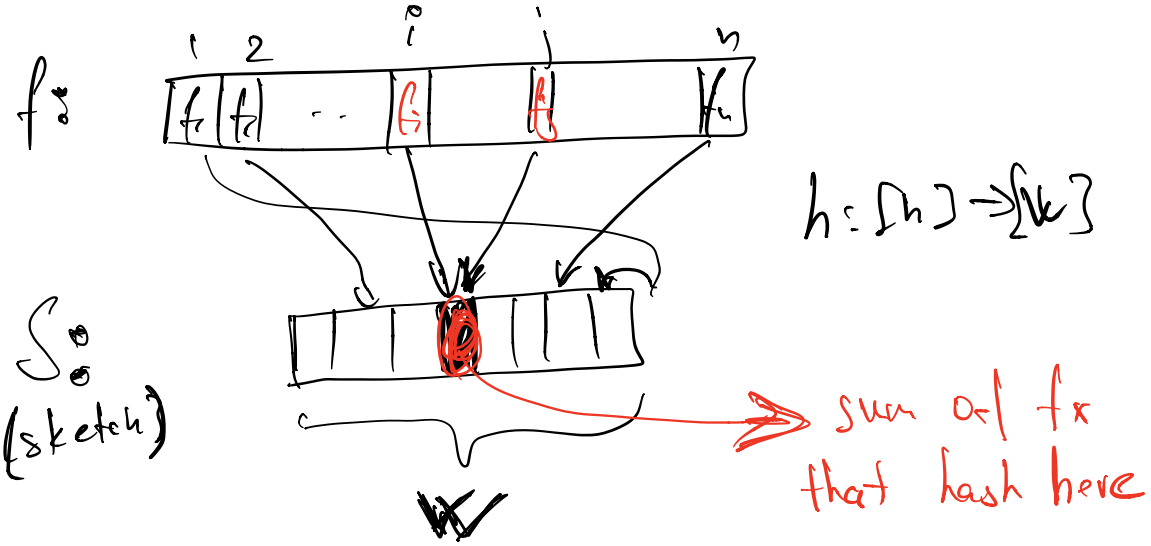
Related Problem: - report items which are sufficiently frequent

Def: $\phi \in (0,1)$, item $x \in [n]$ is ϕ-heavy hitter if $f_x \geq \phi \cdot \sum_{y \in [n]} f_y = \phi \cdot m$.

New goal: find all items that are ϕ-HH. OK space depends on ϕ.

Algorithm CountMin:

Basic idea: to get some estimate $\tilde{f}_x$ for f_x.



Bucket-Algo:

- init: $S[1..k] = 0$

- when we see x in stream $\{f_x := f_x + 1\}$
  $S[h(x)] := S[h(x)] + 1$.
- @ end: to get $\hat{f}_x = S[h(x)]$.

ϕ-HH: $f_x \geq \phi \cdot m$.

Obs: $\hat{f}_x \geq f_x$.

Claim: $\Pr_h[\hat{f}_x > f_x + \varepsilon\phi m] \leq 0.1$,
as long as $w \geq \Omega\left(\frac{1}{\varepsilon\phi}\right)$.

Proof:
$$\mathbb{E}_h[\hat{f}_x] = \mathbb{E}_h\left[\sum_{y \in [n]} \mathbb{1}[h(x)=h(y)] \cdot f_y\right]$$
$$= \mathbb{E}\left[f_x + \sum_{y \neq x} \mathbb{1}[h(x)=h(y)] \cdot f_y\right]$$
$$= f_x + \sum_{y \neq x} f_y \cdot \underbrace{\Pr[h(x)=h(y)]}_{1/w}$$
$$\leq f_x + \frac{\sum f_y}{w} = f_x + \frac{m}{w}.$$

$$\mathbb{E}[\hat{f}_x - f_x] \leq m/w.$$

$\hat{f}_x - f_x$ is ≥ 1

↳ positive r.v.

$\Rightarrow$ By Markov's ineq:

$$\Pr\left[\hat{f}_x - f_x > \underbrace{10 \cdot m/w}\right] \leq 0.1.$$

$$= \varepsilon \cdot \phi \cdot m$$

$\Rightarrow$ set $w = 10/\varepsilon\phi$. ☒

The ϕ-HH algorithm:

- for each $x \in [n]$, compute $\hat{f}_x$ from S,
- if $\hat{f}_x \geq \phi \cdot m$ $\Rightarrow$ report as ϕ-HH.

Corollary: - if $f_x \geq \phi m \Rightarrow$ algo reports it as HH.

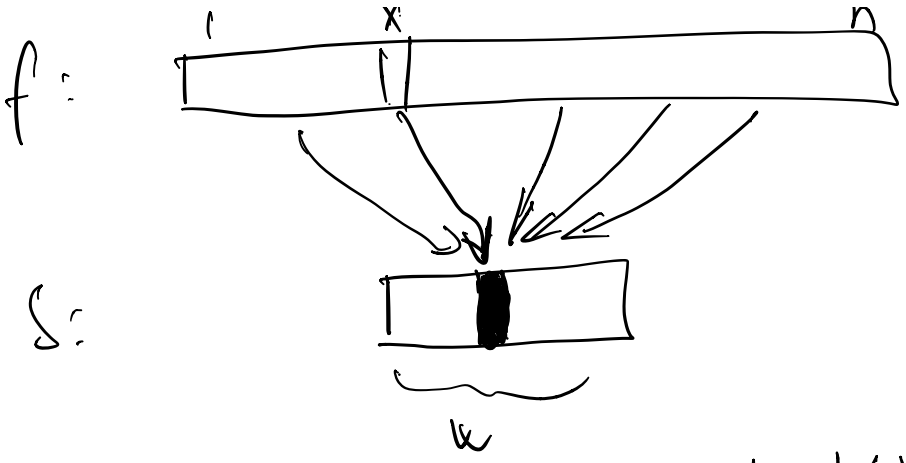
$\hat{f}_x \geq \phi m$ ⟹

- if $f_x < \phi m - \varepsilon\phi m = (1-\varepsilon)\phi m$, then not reported as HH.

$\hat{f}_x < \phi m$ with $\geq 90\%$.

Issue:

↳ some non-HH x's are reported as HH's.



in fact all y's s.t. $h(y) = h(x)$ will be considered heavy.

Full Count Min algo

Repeat "basic idea" for $L = O(\lg n)$ times.

Algo:

- inits $S_i[1..w]$, $h_i : [n] \to w$ $i = 1..L$
- @ item x in stream: $S_i[h_i(x)] := S_i[h_i(x)] + 1$ $i = 1..L$
- estimator: $\hat{f}_x := \min_{i=1..L} S_i[h_i(x)]$

L indep experiments

Claim: for $w = \frac{10}{\varepsilon\phi}$,

$$L = O(\lg n)$$

$$\left[\tilde{f}_x - f_x \le \varepsilon\phi m, \text{ for all } x \in [n]\right]$$

with probability $\ge 1 - 1/n$.

Pf: by Claim above, $\forall$ fixed $i \in [L]$:

$$\Pr_{h_i}\left[S_i[h_i(x)] - f_x > \varepsilon\phi m\right] < \frac{1}{10}.$$

for fixed $x \in [n]$:

$$\Pr_{h_1 \cdots h_L}\left[\tilde{f}_x - f_x > \varepsilon\phi m\right]$$

$$= \prod_{i=1..L} \Pr_{h_i}\left[S_i[h_i(x)] - f_x > \varepsilon\phi m\right]$$

$$\le \left(\tfrac{1}{10}\right)^L = e^{-L \cdot \ln 10} \le \frac{1}{n^2},$$

$= \ln n^2$

for $L = \frac{2 \ln n}{\ln 10} = O(\lg n)$.

union bound

$$\Pr\left[\exists x \text{ s.t. } \tilde{f}_x - f_x > \varepsilon\phi m\right] \le$$

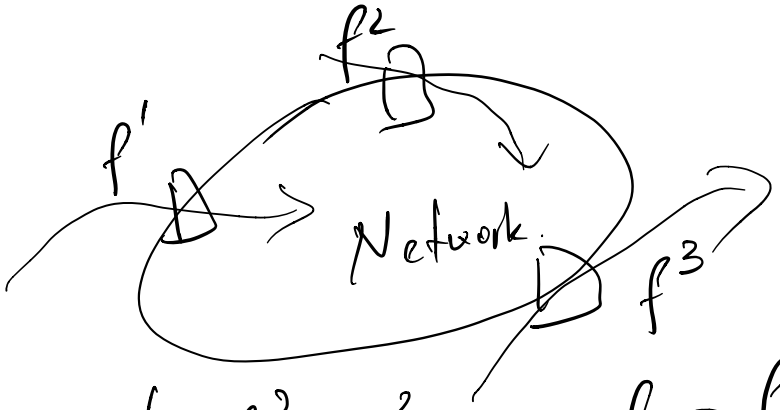
$$\leq \sum_{x \in [n]} \Pr\left[\vec{f}_x - f_x > \varepsilon\phi m\right]$$

$$\leq n \cdot 1/n^2 = 1/n. \quad \boxtimes$$

Space: $O(L \cdot w)$ words (count up to m)

$\Rightarrow O\left(\frac{\lg n}{\varepsilon\phi}\right)$ words for ϕ-HH (up to $1 \pm \varepsilon$ approx.)

Obs: consider situation where:



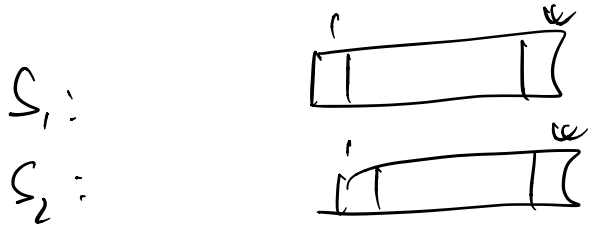
$f = f^1 + f^2 + f^3 \rightarrow f_x = f_x^1 + f_x^2 + f_x^3$

$\forall x \in [n]$.

How to compute sketch that able to estimate $\vec{f}_x$ for all $x \in [n]$.

Solution: each router computes CM sketch of its freq. vect. (using some hash funct.), then sum it all up.

CM sketch(x) $\in \mathbb{R}^{2 \cdot w}$



$= A \cdot f$

matrix of size $2w \times n$.

$A_{(i,j),x} = 1$ iff $h_i(x) = j$

$i \in [1], j \in [w], x \in [n]$.

$$Af_1 + Af_2 + Af_3 = A(f_1 + f_2 + f_3)$$

sketch @ router 1

sketch of freq. vector $f_1 + f_2 + f_3$.

CountMin is a **linear** sketch.