Lecture 21

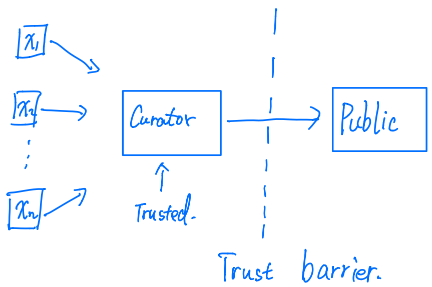
Local Model for DP

- Frequency Estimator
- Heavy Hitters

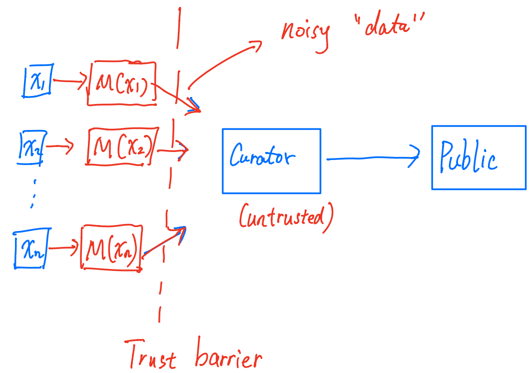
Techniques:

- Count Sketch
- Tree Histogram

Central Model



Local Model



Local Randomizer $M: \mathcal{X} \mapsto \mathcal{Y}$

is (ε, δ) - locally differentially private (LDP) if

$\forall x, x' \in \mathcal{X}, \quad E \subseteq \mathcal{Y}$

$$\mathbb{P}[M(x) \in E] \leq e^{\varepsilon} \mathbb{P}[M(x') \in E] + \delta.$$

Last Lecture:

- Randomized Response
- 1-bit Mean Estimation

Error in Local Model $\approx$ Error in Central Model $\cdot \sqrt{n}$

$\Longleftrightarrow$

quadratic.

more samples

Frequency Estimation in Local Model

$x_1, \dots, x_n \in [d]$ $\quad \{1, \dots, d\}$

data universe.

$\forall x \in [d]$

$f(x) = \sum_{i=1}^{n} \mathbb{1}[x_i = x]$

x_i -coordinate

$x_i \in [d]$ $\quad X_i = \underbrace{\boxed{0 \;\; 0\,0 \dots 1 \;\; 0 \dots 0}}_{\text{length } d.}$

Communication

$\Omega(d)$ bits

Runtime $\begin{cases} \text{Server} \\ \text{User} \end{cases}$

$\Omega(d)$

Local Randomizer

$b_i = \boxed{1 \;\; -1 \;\; \dots\dots \;\; -1 \;\; 1}$

$X_{ij} = 0, \quad b_{ij} = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2} \end{cases}$

$X_{ij} = 1, \quad b_{ij} = \begin{cases} 1 & \text{w.p. } \frac{1}{2} + \frac{\varepsilon}{2} \\ -1 & \text{w.p. } \frac{1}{2} - \frac{\varepsilon}{2} \end{cases}$

$\hat{f} = \left(\sum_i b_i\right) \cdot \frac{1}{\varepsilon}$

error $\approx \sqrt{n}$

Local Privacy has $\sqrt{n}$ error

+

$\left\{ \begin{array}{l} \text{Sublinear Algorithm} \\ \text{Sketching} \\ \text{"Big data Algorithm"} \end{array} \right.$ has lower communication computation

w/ $\lesssim \sqrt{n}$ error.

Count Sketch for dim reduction

Random Signed vector

$x_1 \longrightarrow [\,0 \cdots 0 \quad +1 \cdots 0\,] \quad H(x_1)$

$\vdots$

$x_n \longrightarrow \underbrace{[\,0 \;\; 0 \cdots 0 \;\; -1 \;\; 0 \cdots\cdots 0\,]}_{m} \quad H(x_n)$

hash function $\quad h : [d] \to [m] \quad (m \ll d.)$

random sign $\quad s : [d] \to \{\pm 1\}$

Pairwise Independence: $\quad \forall x \neq x', \quad \forall y, y'$

$$\underset{h}{P}[\, H(x) = y \text{ and } H(x') = y' \,] = \frac{1}{4m^2}$$

$\forall x \in [d].$

$$\hat{f}(x) = \sum_i \langle H(x), H(x_i) \rangle$$

$$= \sum_{i : x_i = x} \underbrace{\langle H(x), H(x_i) \rangle}_{1} + \underbrace{\sum_{i : x_i \neq x} \langle H(x), H(x_i) \rangle}_{\text{Collusion Noise.}}$$

$$= f(x) + \text{Collusion Noise.}$$

$$\mathbb{E}[\hat{f}(x)] \simeq f(x) + \sum_{i : x_i \neq x} \left(\underbrace{\frac{1}{m}\left(\frac{1}{2} \times 1 + \frac{1}{2} \times (-1)\right)}_{h(x) = h(x_i)} + \underbrace{\left(1 - \frac{1}{m}\right) \cdot 0}_{h(x) \neq h(x_i)} \right)$$

$$= f(x).$$

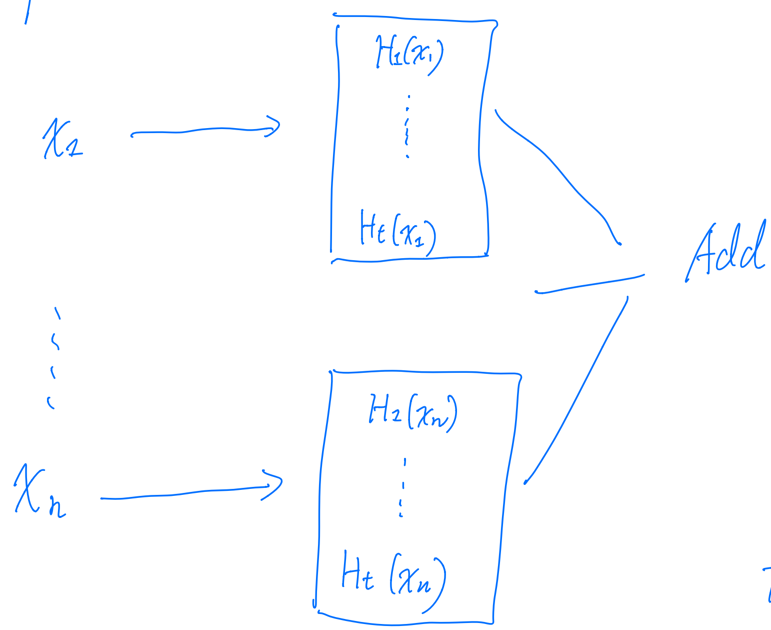
$\mathbb{P}[h(x) = h(x_i)] = \frac{1}{m}$

$\mathbb{P}[s(x) = s(x_i)] = \frac{1}{2}$

Set $m \geq O(\sqrt{n})$, w.p. $\frac{3}{4}$

$$|\hat{f}(x) - f(x)| \leq O\left(\frac{n^{3/4}}{\sqrt{m}}\right) \longrightarrow O(\sqrt{n})$$

Amplification



$$\begin{array}{ccc} H_1(x_1) & \cdots & H_t(x_1) \\ + & & + \\ \vdots & \vdots & \vdots \\ + & & + \\ H_1(x_n) & & H_t(x_n) \end{array}$$

$\forall x \in [d]$

Take median across the t inner products

Set $t = \log\left(\frac{1}{\beta}\right)$

$$\left|\hat{f}(x) - f(x)\right| \leq O\left(\frac{n^{\frac{3}{4}}}{\sqrt{m}}\right)$$

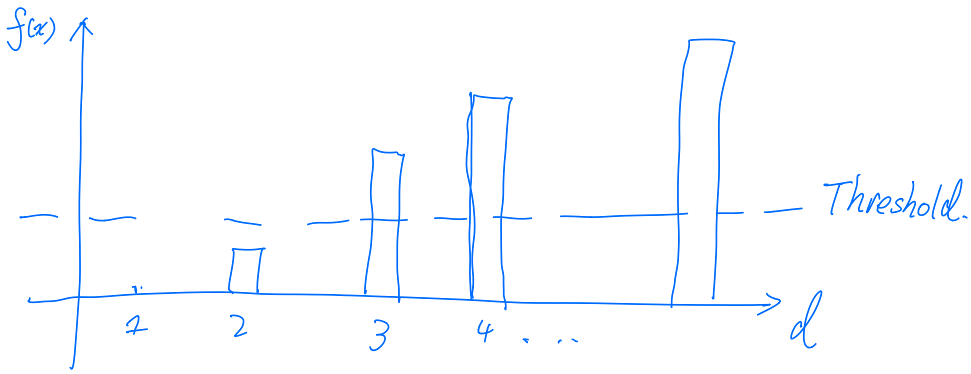
w.p. $1-\beta$

Communicate $m = O(\sqrt{n})$ bits

Count Sketch + Random Response on each bit.

Amplification

Heavy Hitters



$d \gg n$.

Heavy hitters $\approx$ Frequency Estimation ?

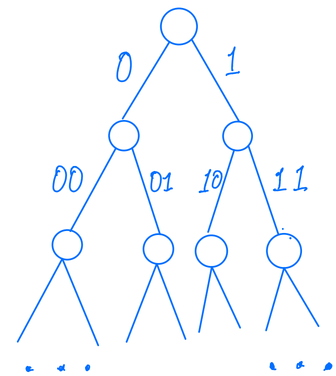
error $\approx$ Threshold.

Avoid linear Runtime in d ?

TreeHist

$x_i \in [d] \rightarrow \underbrace{(0\,1\,1\,0\,1 \cdots 0)}_{\log d}$ $\quad \mathcal{X} = \{0,1\}^{\log d}$.

Frequency oracle: give count for every sub-tree.

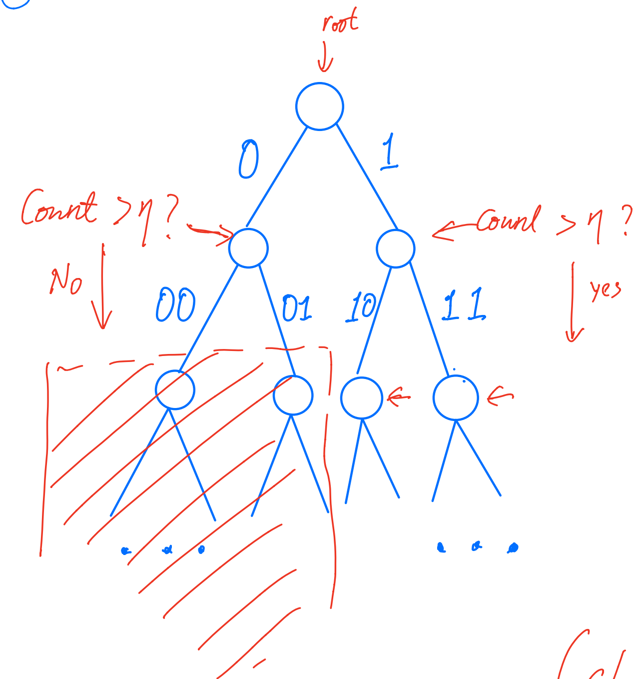


log d Depth.

Leaf $\underbrace{\qquad\qquad}_{\mathcal{X} = \{0,1\}^{\log d}}$

$\boxed{\text{Problem: find } x \text{ such that } f(x) \geq \eta \ (\text{e.g. } \sqrt{n})}$

Pruning



Top-Down.

At each level:

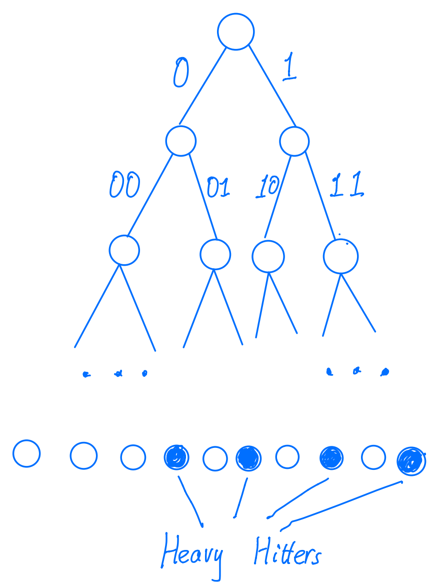
$\# \{\text{nodes} > \eta\} \le \frac{n}{\eta}$

$\#$ surviving nodes at each level $\le \frac{n}{\eta}$

$\#$ Total frequency queries $\le O\left(\frac{n}{\eta} \log d\right)$.

(Choose $\eta = \sqrt{n}$)

Heavy Hitters

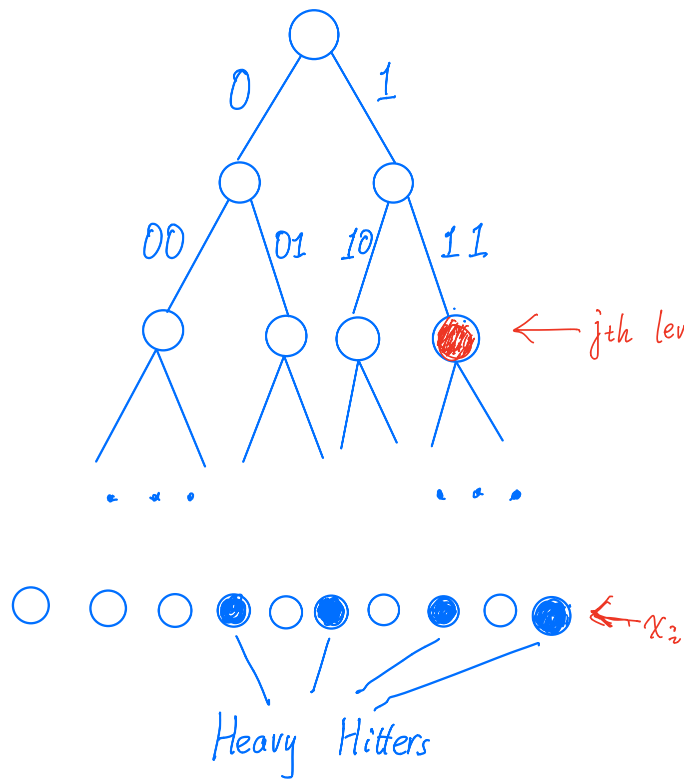


List = { (heavy hitter, its frequency) }

Frequency Estimation at each level.

$[n] \longrightarrow [\log d]$

Assign every x_i to a random level.



x_i's j-prefix $\overbrace{\boxed{0\ 1\ 1\ 0 \cdots 1}}^{j}$

$\Big\downarrow$ Count Sketch H

Random Signed Vector.

$\boxed{0\ 0\ 0\ \pm 1\ 0 \cdots 0}$

$\Big\downarrow$ ε-Randomization

$\underbrace{\boxed{1\ -1\ -1\ +1\ \cdots\ +1}}_{\sqrt{n}}$

If bit $= 0 \rightarrow \text{unif}\{\pm 1\}$

bit $= 1 \rightarrow \begin{cases} +1 & \text{w.p. } \frac{1}{2} + \frac{\varepsilon}{2} \\ -1 & \text{w.p. } \frac{1}{2} - \frac{\varepsilon}{2} \end{cases}$

Pruning

Heavy Hitters & Frequency.

Communication $\sqrt{n}$. bits

Computation $\sqrt{n} \log d$. for server.

· Improvement.

Hardamard Transform + Randomized Response.

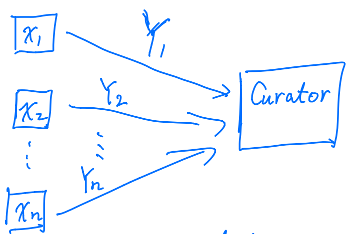
$O(\log d)$ bits

Final error : $O\left(\frac{\sqrt{n}\ \log d}{\varepsilon}\right)$.

Reducing Communication Further

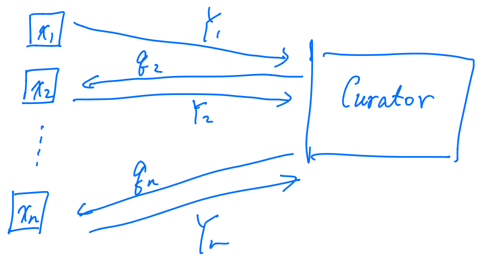
3 Sources of error

Role of Interactions



ε-Local randomizer.

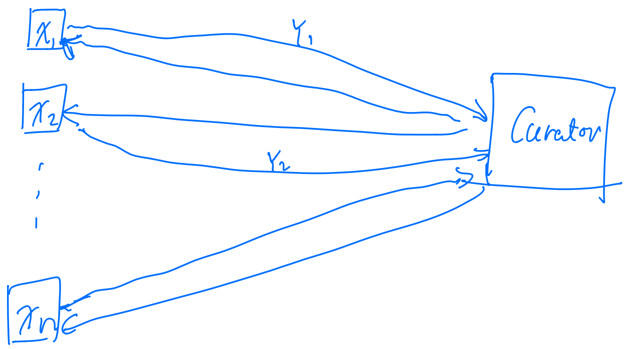
Non-interactive.



Some delay

Sequentially interactive

Every user reports once.



Less practical.

SGD

Fully interactive

Communication Complexity.