Binary Tree Mechanism:

Specific case of counting query: interval queries

Say $X = [M]$, i.e., each entry in database in $[M]$.

Interval query: $f_{s,t}(D) =$ # users w/ entry in $[s,t]$

$f_{s,t}: \mathcal{D} \to \mathbb{R}$

$F(D) = (f_{s,t}(D))_{1 \le s \le t \le M}$

$F: \mathcal{D} \to \mathbb{R}^{\binom{M}{2}}$

(captures things like CDF: $\phi = (f_{1,t})_{t \in [M]}$

$\Rightarrow$ median: $t: f_{1,t}(D) = n/2$

How can we return $F(D)$ privately?

Obvious option 1: Laplace mechanism

$\Delta F = \Theta(M^2)$: consider adding one row at median

$\Rightarrow$ Laplace mechanism adds $\mathrm{Lap}\left(\frac{M^2}{\varepsilon}\right)$-noise to every interval query

to get $a_{s,t} = f_{s,t}(D) + Z_{s,t}$

$\Rightarrow \mathbb{E}\left[\max_{1 \le s \le t \le M} |a_{s,t} - f_{s,t}(D)|\right] = \Theta\left(\frac{M^2 \log M}{\varepsilon}\right)$

Improvement: get better bounds by answering fewer queries, but enough to allow us to reconstruct all others!

Idea 2: threshold queries

Let $\hat{F}(D) = (f_{1,t})_{t \in [M]}$

$\Rightarrow \Delta \hat{F} = \Theta(M)$

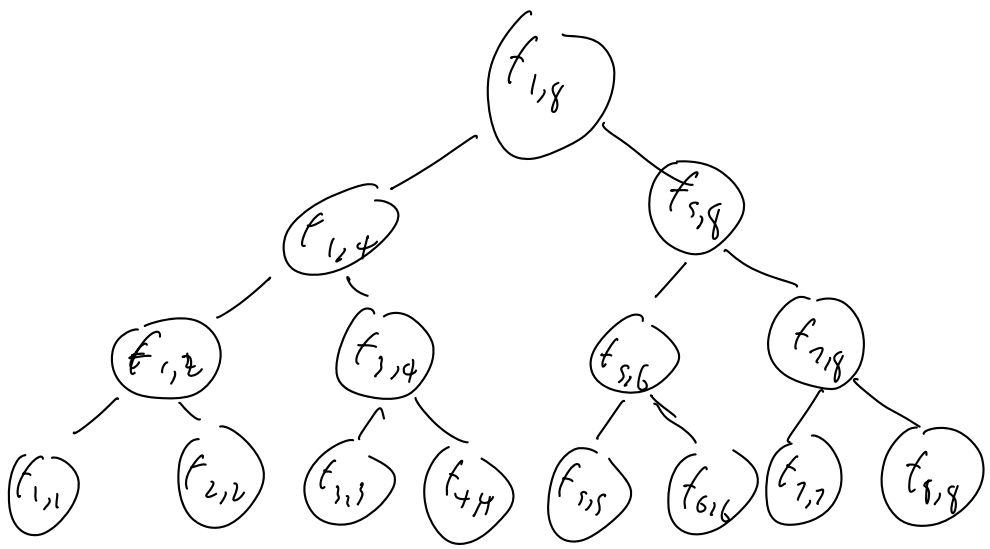
$\Rightarrow O\left(\frac{M \log M}{\varepsilon}\right)$-error on each

$f_{s,t}(D) = f_{1,t}(D) - f_{1,s}(D)$

$\Rightarrow$ error $\leq O\left(\frac{M \log M}{\varepsilon}\right)$.

Even better: binary tree!

Work M is power of 2.



$T = \{(u,v) : u = j \cdot 2^{i-1}$ (+1 except $j=1$), $v = (j+1) \cdot 2^{i-1}$ for $1 \le i \le \log M$, $1 \le j \le M/2^{i-1}\}$

$$|T| = \sum_{i=1}^{\log M} \frac{M}{2^{i-1}} = 2M-1$$

Call this query $F^{BT}(D) = (f_{s,t}(D))_{(s,t)\in T}$

Lemma: $\Delta F^{BT} \le \log M$

Pf: Adding/removing one entry changes entries on one root-leaf path

$\Rightarrow$ Can add $\mathrm{Lap}(\frac{\log M}{\varepsilon})$ noise to each coordinate of F^{BT}

$\forall (u,v) \in T$: $Z_{u,v} \sim \mathrm{Lap}(\frac{\log M}{\varepsilon})$ $a_{u,v} = f_{u,v}(D) + Z_{u,v}$

$\Rightarrow \mathbb{E}[\max_{(u,v)\in T} |Z_{u,v}|] = O(\frac{\log^2 M}{\varepsilon})$

Claim: $\forall t \in M$, $\exists S_t \subseteq T$, $|S_t| \le \log M$, s.t.

$$f_{1,t}(D) = \sum_{(u,v)\in S_t} f_{u,v}(D)$$

Pf: Binary representation of t

Answer $f_{l,t}(D)$ is $\sum_{(r,v)\in S_t} q_{r,v} = b_{l,t}$

$$\Rightarrow |b_{l,t} - f_{l,t}(D)| = \left| \sum_{(r,v)\in S_t} q_{r,v} - f_{l,t}(D) \right|$$

$$= \left| \sum_{(r,v)\in S_t} Z_{r,v} \right| \le \sum_{(r,v)\in S_t} |Z_{r,v}|$$

$$\le |S_t| \cdot \max_{(r,v)\in T} |Z_{r,v}|$$

$$\le \log M \cdot \max_{(r,v)\in T} |Z_{r,v}|$$

$$\Rightarrow \mathbb{E}\left[\max_{1\le s\le t\le M} |b_{s,t} - f_{s,t}(D)| \right] \le \log M \cdot \frac{\log^2 M}{\varepsilon} = \frac{\log^3 M}{\varepsilon}$$

$\Rightarrow$ set $b_{s,t} = b_{1,t} - b_{1,s}$, max error $= O\left(\frac{\log^3 M}{\varepsilon}\right)$

Thm: $\exists$ ε-DP alg. that answers all $\binom{M}{2}$ interval queries s.t.

$$\mathbb{E}\left[\max_{1\le s\le t\le M} |f_{s,t}(D) - b_{s,t}| \right] = O\left(\frac{1}{\varepsilon}\log^3 M\right)$$

Notes: - can do more careful analysis to get $O(\frac{1}{\varepsilon}\log^{2.5} M)$.

$|\sum Z_{r,v}| \le \sum |Z_{r,v}|$ a bit loose: more like $\left(\sum Z_{r,v}^2\right)^{1/2}$

- Multiple ways of reconstructions, e.g.

$$f_{1,6} = f_{1,4} + f_{5,6} = f_{1,8} - f_{7,8}.$$

can help in practice!

- True answers to threshold queries $f_{1,t}$ monotonic, but our answers might not be due to noise. Post-processing to make monotonic, can help in practice, and sometimes in theory!