A Joint Exponential Mechanism for Differentially Private Top-*k*

Jennifer Gillenwater[1], Matthew Joseph[1],
Andrés Muñoz Medina[1], Mónica Ribero[1,2]

[1] Google New York
[2] UT Austin

Problem Formulation

- n users contribute a d-dimensional binary vector.

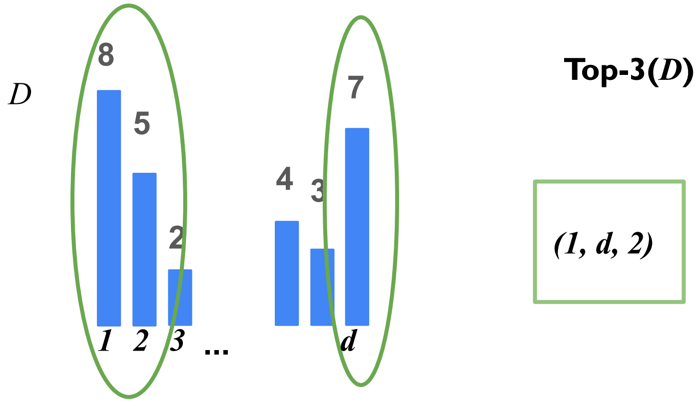
Database ***D***

$$u_1 = [0, 0, 1 \cdots, 0]$$
$$u_2 = [1, 0, 1 \cdots, 0]$$
$$u_3 = [1, 1, 1 \cdots, 0]$$

$$\text{item counts} = [2, 1, 3, \ldots, 0]$$

Problem Formulation

- n users contribute a d-dimensional binary vector.
- **Goal:** Identify the *sequence* with the highest counts from data domain with d elements.



Differential Privacy [DMNS06]

A randomized mechanism $\mathcal{M} : \mathcal{D} \to V$ is *(ε,δ)-**differentially private*** if for any pair of datasets D and D' differing in *only one record (add/remove)* and any output S

$$\mathbb{P}(\mathcal{M}(D) \in S) \leq e^{\epsilon} \mathbb{P}(\mathcal{M}(D') \in S) + \delta$$

Pure DP if $\delta = 0$

The exponential mechanism

Select item $x \in \mathcal{X}$ with highest utility given by function

$$u(D, x)$$

The *exponential mechanism* that selects item $x \in \mathcal{X}$ w.p.

$$\propto e^{\frac{\epsilon u(D,x)}{2\Delta u}}$$

$$\Delta u = \sup_{x \in \mathcal{X}, D \sim D'} |u(D, x) - u(D', x)|$$

is **ε- DP** [DR14].

State of the art

- Peeling mechanism [BST10,DR19]: applies a DP subroutine k times to repeatedly select and remove (or "peel" off) the highest-count item
  - Exponential mechanism: *Best approximate DP.*
  - Permute and flip: *Best pure DP*

Pure DP if $\delta = 0$

State of the art

- Peeling mechanism [BST10,DR19]: applies a DP subroutine k times to repeatedly select and remove (or "peel" off) the highest-count item

- Needs composition: (accessing raw data k times)
  - Using $\epsilon/\sqrt{k}$ each call and *concentrated differential privacy composition* provides *approximate* $(\epsilon, \delta) - DP$
  - Using ϵ/k each call and standard composition provides *pure* $\epsilon - DP$

Our contributions

JOINT: an exponential mechanism whose output space consists of all $O(d^k)$ length-k sequences.

- ε - differentially private top-k algorithm.
- Time $O(dk\log(k)+d\log(d))$ and space $O(dk)$.
- No composition needed!

JOINT

$$\mathbf{c} = [c_1, c_2, c_3, \cdots, c_d]$$
$$c_1 > c_2 > \ldots > c_d$$

Let c be the *ordered* vector of counts.

utility function over sequences:

$$u^*(D, (s_1, \ldots, s_k)) = -\max_{i \in [k]} c_i - c_{s_i}$$

Price of returning item s_i in position i

For a given sequence, pay the highest cost.

Negative because it has to represent utility.

Theorem 1

JOINT is ε-DP. Follows directly from the fact that JOINT is an instance of the exponential mechanism.

Theorem 2

JOINT samples a sequences from the exponential mechanism with utility u^* in time

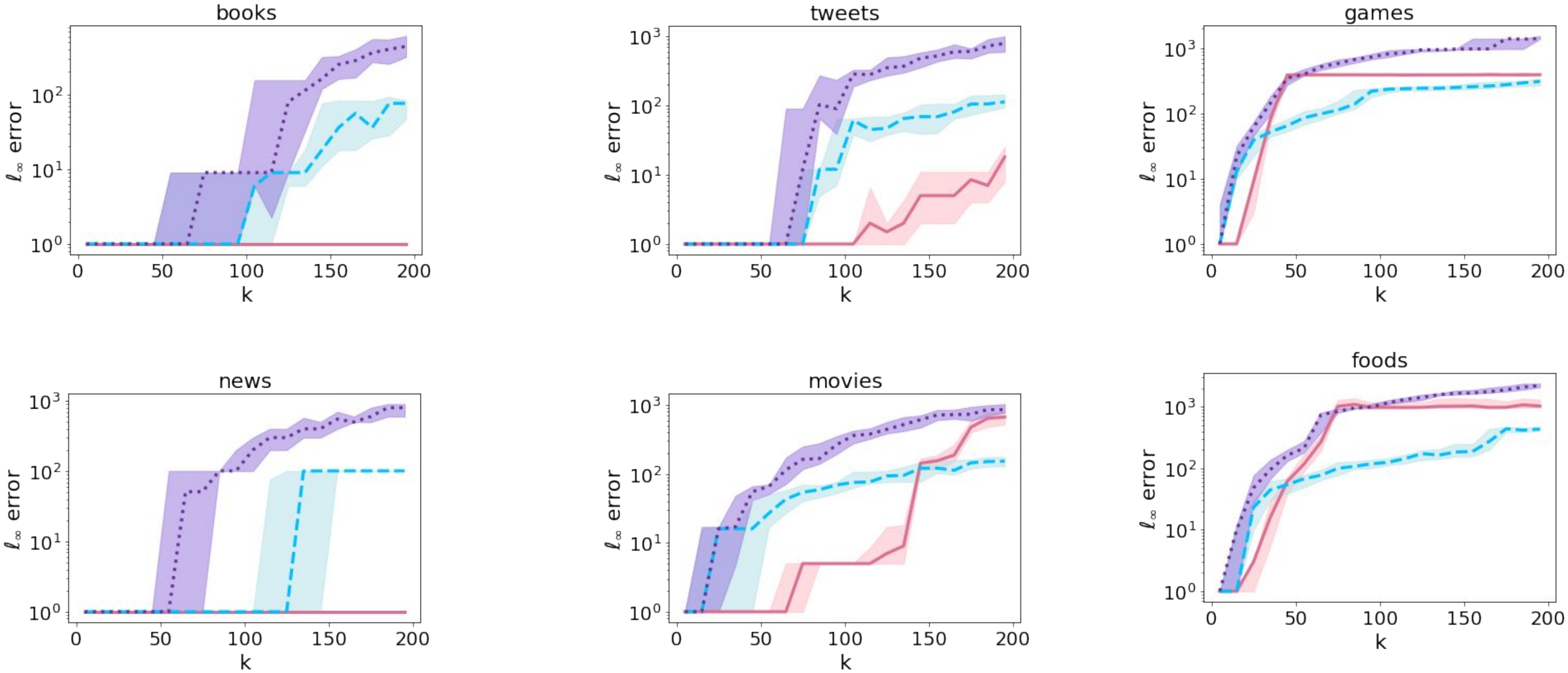
$$O(dk\log(k) + d\log(d)).$$

- $d \log(d)$ comes from sorting all d item counts values.
- $dk \log(k)$ comes from sorting all the dk utility values
  - Use k-way merging to sort dk utility values, that are already organized in k length d arrays.

Experiments

- 6 datasets
- 2 baselines:
  - PNF-PEEL: Peeling with permute-and-flip [MS20] (Best pure DP)
  - CDP-PEEL: Peeling with exponential mechanism [DR19] (Best approx DP)
- Metrics:
  - ℓ_∞- norm
  - See paper for more metrics.

ℓ_∞ − norm



Conclusion

- **JOINT** improves on existing pure DP methods and often improves on existing approximate DP methods when k is not large.

- The best approach for the case where users can contribute to some number of items larger than 1 but less than d is an interesting topic for future work.

Thank you!