Incorporating Item Frequency for Differentially Private Set Union

Ricardo Silva Carvalho, Ke Wang, Lovedeep Gondara

Simon Fraser University, BC, Canada

SFU SIMON FRASER UNIVERSITY

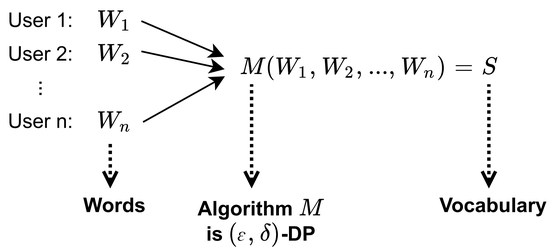
Summary

We propose greedy differentially private set union mechanisms that incorporate item frequency. Our contributions are the following:

- We incorporate the item frequency, which is typically available in set union problems, to boost the utility of private mechanisms.
- The sampling step to limit the number of items each user contributes to is eliminated: users keep all of their items for the set union.
- Novel DP greedy mechanisms with smaller output threshold compared to previous work.
- A mechanism variation with knowledge transfer, with improved utility on experiments.

Introduction

Motivating example: Set union mechanism to build a vocabulary while satisfying user privacy.



Differential Privacy (DP)

$\mathcal{M}$ is (ε, δ)-DP, if for all neighboring datasets D and D' that differ in the addition or removal of one user's data, and all sets S of possible outputs, we have:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta$$

Differentially Private Set Union

The problem of DPSU was formalized by [1].

Let U be some unbounded universe of items. Each user i has a set of items $W_i \subseteq U$, we want an (ε, δ)-DP mechanism $\mathcal{M}$ that outputs a subset $S \subseteq \cup_i W_i$ such that the size of S is as large as possible.

Previous DPSU Mechanisms

The previous mechanisms used for DPSU are the:

- COUNT mechanisms [2, 3]
- WEIGHTED mechanisms [1]
- POLICY mechanisms [1]

Overview: Iterations over users and items.

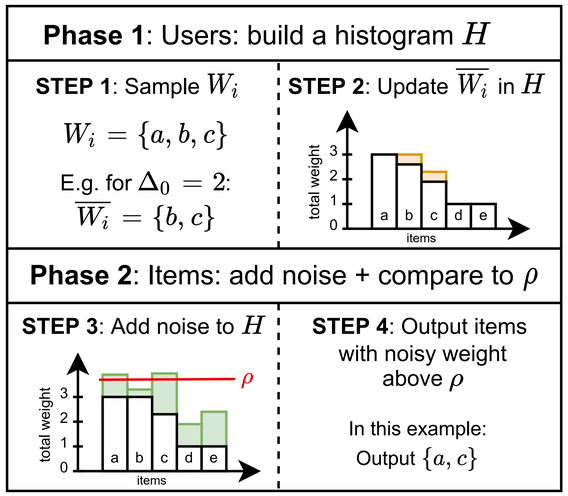


Table: Differences between our work and previous.

Mechanism	Sampling	User View	Updates	KT
COUNT	Yes	Independent	$1/\Delta_0$	No
WEIGHTED	Yes	Independent	$1/\lvert W_i \rvert$	No
POLICY	Yes	Dependent	Gap to Γ	No
This work	**No**	**Dependent**	**Greedy**	**Yes**

Our Main Mechanism: GW

GW: **G**reedy updates **W**ithout sampling: (ε, δ)-DP

Phase 1: Iterate over users to build histogram H:

- Each user keeps all items $W_i = \overline{W_i}$ along with item frequency C_i **(eliminated sampling step)**.
- The histogram is updated by each user using the greedy mechanism $\texttt{GU}(\overline{W_i}, C_i, \Gamma)$ showed below.

Phase 2: Iterate over the items, to output set union:

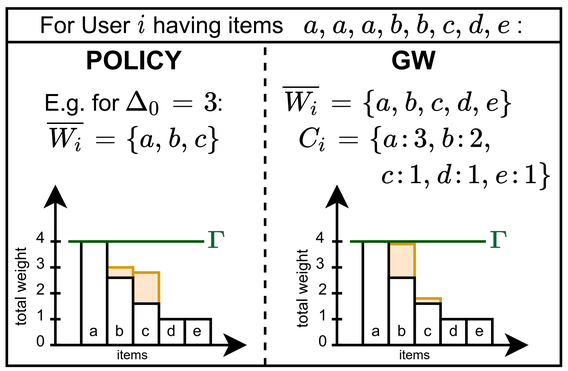
- Add noise $\xi = \texttt{Lap}(0, 1/\varepsilon)$ to each item in H.
- Output the items with noisy weight above the threshold $\rho = 1 - 1/\varepsilon \cdot \log 2\delta$.

GU: Greedy Update Algorithm

Input: H, $\overline{W_i}$, C_i, Γ
Output: Updated histogram H.

budget $= 1$, $S = \emptyset$, $O = \emptyset$
Filtering: Will only update items below Γ.
for $u \in \overline{W_i}$ **do**
  if $H[u] < \Gamma$ **then**
    $S \leftarrow S \cup u$
while budget > 0 **do**
  Item Selection: Get best item to update
  $u^\star = \operatorname{argmax}_{u \in S \setminus O} C_i[u]$
  Greedy update: Allocate budget to $u^\star$
  cost $= \Gamma - H[u^\star]$
  if cost $\leq$ budget **then** ▷ Use budget until Γ
    $H[u^\star] = H[u^\star] +$ cost
    budget $=$ budget $-$ cost
    $O = O \cup \{u^\star\}$
  else ▷ Cannot reach Γ with budget, use all
    $H[u^\star] = H[u^\star] +$ budget
    break
return H

Comparison to POLICY

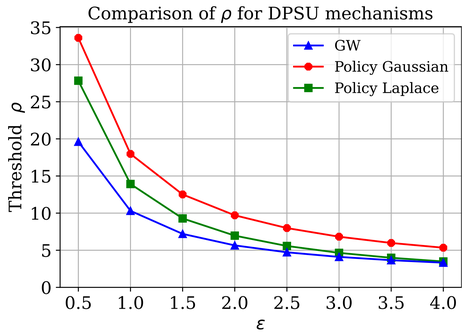


GW-KT: Knowledge Transfer

Replaces C_i in **GW** with C_i^{pub} from **public** data.

Idea: Gauge the common items in the private data and reach Γ quicker for them, leaving more budget to output the remaining items in the private data.

Threshold Comparison



Experiments

Dataset	Policy Lapl. (best Δ_0)	Policy Gaus. (best Δ_0)	GW
Reddit	15485 ± 63	16958 ± 37	**17051** ± 51
Twitter	33757 ± 15	34332 ± 49	**37697** ± 22
Finance	45323 ± 75	40724 ± 62	**49868** ± 23

Dataset	GW-KT (imdb)	GW-KT (covid)	GW-KT (songs)	GW-KT (wiki)	GW-KT (enron)
Reddit	18968	18811	18979	**19057**	19012
Twitter	40910	41332	41739	42060	**42808**
Finance	50451	**51059**	50409	50934	**51070**

- **GW** consistently outperforms POLICY mechanisms [1] by around 10% for Twitter and Finance datasets, whereas for Reddit it gets equivalent utility.
- **GW-KT** has best results as 12%, 25% and 13% better than [1] w.r.t. Reddit, Twitter and Finance.
- All public datasets tested were beneficial to DPSU with KT, even those from very diverse domains. Moreover, the larger the intersection between public and sensitive vocabularies, the better.

References

[1] Sivakanth Gopi, Pankaj Gulhane, Janardhan Kulkarni, Judy Hanwen Shen, Milad Shokouhi, and Sergey Yekhanin. Differentially private set union. *ICML*, 2020.

[2] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. *WWW*, 2009.

[3] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private sql with bounded user contribution. *PET*, 2020.