Differential Privacy on Fully Dynamic Streams

Yuan Qiu[1] and Ke Yi[2]

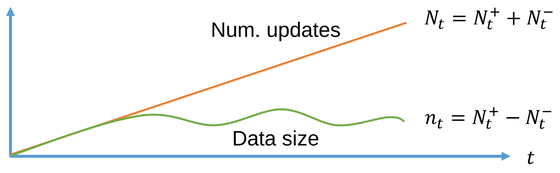
[1]Southeast University [2]Hong Kong University of Science and Technology

Background

■ **Stream Model**: A sequence of updates, (s_t, x_t) arrives at time t:

- No-op: $s_t = \perp \implies D_t = D_{t-1}$
- Insertion: $s_t = + \implies D_t = D_{t-1} \cup \{x_t\}$
- Deletion: $s_t = - \implies D_t = D_{t-1} - \{x_t\}$

➢ Insertion-only Stream: $s_t \in \{\perp, +\}$
➢ Fully Dynamic Stream: $s_t \in \{\perp, +, -\}$



In Fully Dynamic Streams, Number of updates $N_t \gg$ Data size n_t

■ **Differential Privacy (DP)**: Protect data against membership inference

- A randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if for any pair of neighboring update sequences $(s, x) \sim (s', x')$, and any subset of outputs $Y \subseteq \mathcal{Y}$,

$$\Pr[\mathcal{M}(s, x) \in Y] \le e^{\varepsilon} \cdot \Pr[\mathcal{M}(s', x') \in Y] + \delta\,.$$

- Two streams are neighbors if they differ by one timestamp.

$$(s, x): \quad (+, a), (+, b), (+, c), \perp, \qquad (-, b), \dots$$
$$(s', x'): \quad (+, a), (+, b), (+, c), (+, e), (-, b), \dots$$

➢ If $\mathcal{M}_i$ is $(\varepsilon_i, \delta_i)$-DP, then $(\mathcal{M}_1, \dots, \mathcal{M}_k)$ is $(\sum_i \varepsilon_i\,, \sum_i \delta_i)$-DP.
➢ If $\mathcal{M}_i$ is $(\varepsilon_i, \delta_i)$-DP, and any pair of neighbors only affects the output of exactly one $\mathcal{M}_i$, then $(\mathcal{M}_1, \dots, \mathcal{M}_k)$ is $(\max_i \varepsilon_i\,, \max_i \delta_i)$-DP.

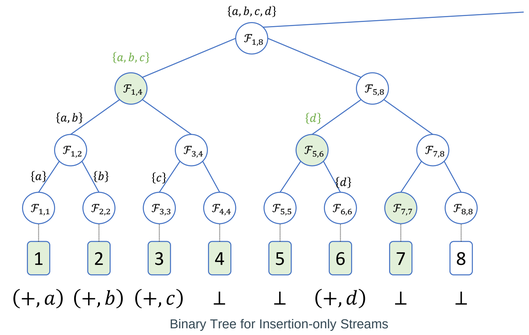
■ **Linear Query**: $f(D) = \sum_{x \in D} f(x)$, where $f(x) \in [0,1]$ is the weight of x

Query	SOTA Mechanism	Static Error
Basic Counting $\lvert D\rvert$	Laplace Mechanism	$\alpha = O\left(\frac{1}{\varepsilon}\right)$
Arbitrary Queries $\mathcal{F}$	Private Multiplicative Weights	$\alpha = \begin{cases} \tilde{O}\left(\lvert D\rvert^{\frac{1}{2}}\right), & \delta > 0 \\ \tilde{O}\left(\lvert D\rvert^{\frac{2}{3}}\right), & \delta = 0 \end{cases}$

$\tilde{O}(\cdot)$ suppresses dependencies on ε and polylogarithmic terms

Differential Privacy on Insertion-only Streams

■ **Binary Tree Mechanism**: STOA for Insertion-only Streams



Binary Tree for Insertion-only Streams

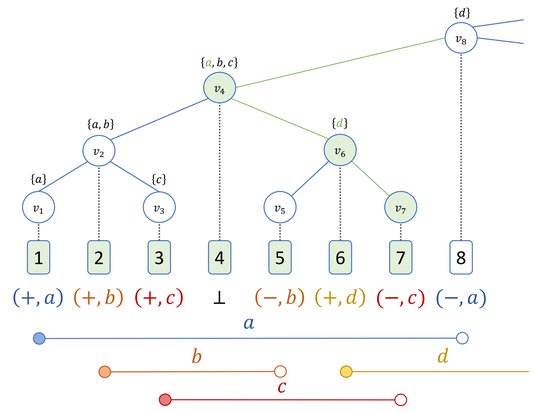
- Let $\alpha^{(k)}$ denote the error under a disjoint union of k mechanisms
- Clearly $\alpha^{(k)} \le k \cdot \alpha$, but the bound can be tightened, e.g. $\alpha^{(k)}_{\text{Lap}} = O\left(\frac{\sqrt{k}}{\varepsilon}\right)$

➢ If a static mechanism has error α, then binary tree mechanism has error

$$\alpha^{(\log t)}\left(\frac{\varepsilon}{\log t}, \frac{\delta}{\log t}, N_t, \dots\right)$$

➢ Extends to fully dynamic streams by separating D_t^+ and D_t^-
➢ **Problem: Error scales with N_t, but the data size is only n_t**

Differential Privacy on Fully Dynamic Streams



Online Interval Tree for Fully Dynamic Streams

Online Interval Tree

■ **Construction**: Each (Insertion, Deletion) pair is treated as an interval

- After observing the update (s_t, x_t), compute D_t
- $D(v_t) \subseteq D_t$ stores an item if its insert-node is in the subtree rooted at v_t
  - Example: At $t = 6$, $D_6 = \{a, c, d\}$, and $D(v_6) = \{d\}$
- When an item is deleted, we augment the item with its deletion time
  - Example: At $t = 2$, $D(v_2) = \{a, b\}$; At $t = 5$, $D(v_2)$ becomes $\{a, (b, 5)\}$
- **Difference in one update only affects $\log t$ nodes**

■ **Querying**: Report items in the dataset at timestamp q

- Given q, visit v_i on the root-to-v_q path where $i \le q$
- Report all "live" intervals $[l, r)$ stored in v_i where $r > q$
  - Example: Query $q = 7$, visit v_4, v_6, v_7, report $\{a, d\}$
- **Each queried item is reported exactly once**

Deletion-only Mechanism

- Each node v_i is initialized with $|D_i(v_i)| \le n_i$ items
- The items gets deleted as time goes by, $D_t(v_i) = D_i(v_i) - D_t^-(v_i)$
- We can track the deletions $D_t^-(v_i)$, and answer through

$$f(D_t(v_i)) = f(D_i(v_i)) - f(D_t^-(v_i))$$

- Our target error is n_t, so we want to bound both $|D_i(v_i)|$ and $|D_t^-(v_i)|$

➢ Solution: use separate privacy budget to track the number of deletions
➢ When more than half are deleted, restart mechanisms at the node
➢ There can be only $\log N_t$ restarts at time t

Takeaway

- If there is a static DP mechanism with error $\alpha(n)$, then there is a DP mechanism for fully dynamic streams with error $\tilde{O}(\alpha(n_t)) \ll \tilde{O}(\alpha(N_t))$.

References

- Binary Tree Mechanism: T. H. Chan, E. Shi, and D. Song. Private and continual release of statistics. In Proc. Automata, Languages and Programming, ICALP, volume 6199, pages 405–417. Springer, 2010.
- (Offline) Interval Tree: M. de Berg, O. Cheong, M. J. van Kreveld, and M. H. Overmars. Computational geometry: algorithms and applications, 3rd Edition. Springer, 2008. ISBN 9783540779735.