Differentially Private Histograms under Continual Observation: Streaming Selection into the Unknown

Adrian Rivera Cardoso and Ryan Rogers

Differentially Private Histograms: One-shot

- Our task is to release histograms $h \in \mathbb{N}^d$ subject to DP [DMNS, DKM$^+$].
- We assume that the ℓ_∞-sensitivity of the histograms is 1.
- Either the ℓ_0-sensitivity is bounded $\Delta_0 < d$ or not. If $\Delta_0 = d$, then we bound privacy loss by only releasing top-k results, not the full histogram.
- Either we have access to the full histogram and know its labels (known domain) or we do not (unknown domain).
- The following algorithms can be extended to use Gaussian noise instead of Laplace.

	ℓ_0 sensitivity Δ_0	unrestricted ℓ_0 sensitivity
Known domain	$\texttt{KnownLap}^{\Delta_0}$ [DMNS]	$\texttt{KnownGumb}^{k}$ [MT]
Unknown domain	$\texttt{UnkLap}^{\Delta_0,\bar{d}}$ [DR]	$\texttt{UnkGumb}^{k,\bar{d}}$ [DR]

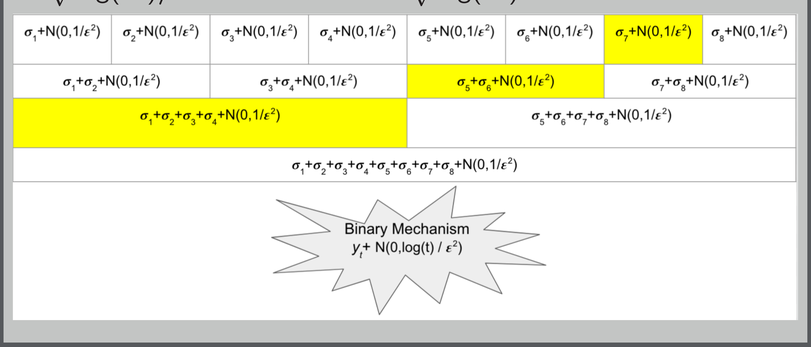
Table 1: DP algorithms for top-k distinct count queries

Continual Observation Setting

- Data is not stagnant, it continually changes and we want to provide a DP running count as data streams.
- Consider a bit stream $\sigma_1, \cdots, \sigma_T \in \{0,1\}$ and we want to release for each $t \in [T]$,

$$y_t = \sum_{\tau=1}^{t} \sigma_\tau.$$

- **Approach 1:** add noise to each σ_t, results in noise scale $\sqrt{T}/\varepsilon$ to the counts and ε-DP.
- **Approach 2:** add noise to each running count y_t, results in noise scale $1/\varepsilon$ to the counts and $\sqrt{T}\varepsilon$-DP.
- **Approach 3** [CSS, DNPR]: Binary Mechanism noise scales with $\sqrt{\log(T)}/\varepsilon$ to the counts and $\sqrt{\log(T)} \cdot \varepsilon$-DP.



Differentially Private Histograms: Continual Observation

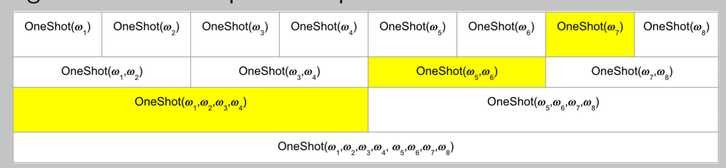
- Rather than a bit stream, assume we get a stream of items $\omega_1, \cdots, \omega_T \subseteq \mathcal{U}$ for some universe $\mathcal{U}$ with $|\mathcal{U}| = d$.
- We want to return a histogram $h_t = (h_t^{(u)} : u \in \mathcal{U})$ where $h_t^{(u)} = \sum_{\tau=1}^{t} \mathbb{1}\{u \in \omega_\tau\}$ for each $t \in [T]$.
- As in the one-shot setting, we consider the settings where $|\omega_t| \leq \Delta_0 < d$ or $\Delta_0 = d$ as well as the domain $\mathcal{U}$ being known or not.
- Largely unexplored problem:

	ℓ_0 sensitivity Δ_0	unrestricted ℓ_0 sensitivity
Known domain	`BinaryMech` [CSS, DNPR]	??
Unknown domain	??	??

Table 2: DP algorithms for top-k in continual observation setting

From One-shot to Continual Observation: `MetaAlgo`

- Using the idea of the binary mechanism, we can simply apply each one-shot DP algorithm in Table 3 on each subsequence of $\omega_1, \cdots, \omega_T$.
- Then any post-processing aggregation function could be used to get the private histogram at each round $t \in [T]$.
- The overall privacy loss would be the privacy loss of each one-shot DP algorithm and then composition could be applied at most $\log(T)$ many times.
- This is not very efficient, as it requires computing one-shot DP algorithms on multiple subsequences.

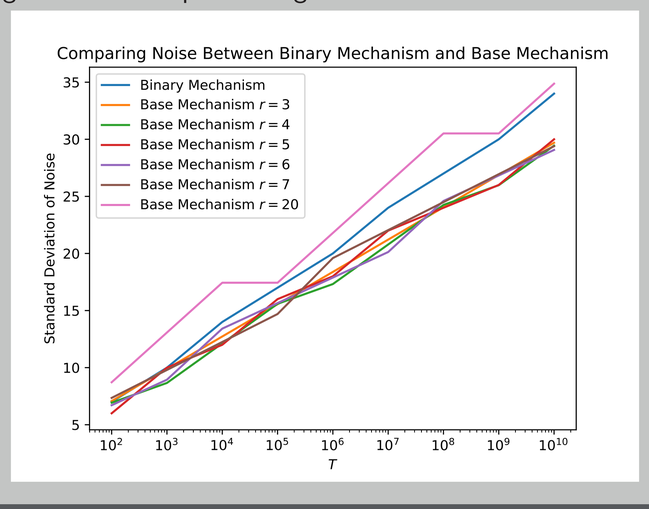


Scalable Setting

- The biggest limitation to using the meta algorithm approach is that it requires subsequences of the event stream.
- [RSP$^+$] describes an approach where existing infrastructure can quickly and efficiently compute aggregate histograms and apply DP algorithms only to the histograms.
- We want to provide similar capabilities in the continual observation setting, meaning that we want to design DP algorithms that only have access to the running histogram counts at each round, not the subsequence.

Known Domain, Restricted ℓ_0-sensitivity: `BinaryMech`

- The binary mechanism can be viewed as applying the $\texttt{KnownLap}^{\Delta_0}$ in the Meta Algorithm framework above.
- Furthermore, the binary mechanism can be implemented with access only to a running histogram counts of the stream, without storing the subsequences.
- We considered how the choice of base can change the overall scale of noise required for DP, as also considered in [QYL].
- Practically, a base of 3-5 should be used, not base 2, for a large range of stream sequence lengths T — we denote as `KnownBase`.



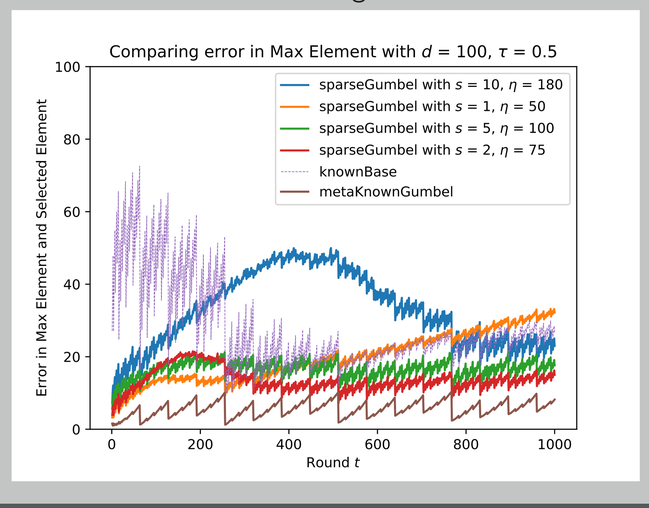
Results

	ℓ_0 sensitivity Δ_0	unrestricted ℓ_0 sensitivity
Known domain	`BinaryMech` [CSS, DNPR] $\texttt{KnownBase}^{\Delta_0}$	$\texttt{SparseGumb}^{s,k}$
Unknown domain	$\texttt{UnkBase}^{\Delta_0}$ [DR]	`MetaAlgo` with $\texttt{UnkGumb}^{k,\bar{d}}$

Table 3: DP algorithms for top-k in continual observation setting

Known Domain, Unrestricted ℓ_0-sensitivity: `SparseGumb`

- First to study the DP selection problem in the continual observation setting, despite the exponential mechanism [MT] being one of the fundamental algorithm of DP.
- Applying `BinaryMech` would have privacy loss scale with d.
- One would expect the top elements at a round to remain the top elements for several rounds in the stream.
- Only when the top elements need updating, i,e. a *switch*, will additional privacy loss be charged.
- We allow a certain number of switches s and once it is exhausted, we will show the same top elements with updated counts.
- Straightforward privacy analysis: combines sparse vector [DNR$^+$], `BinaryMech`, and the exponential mechanism.
- `SparseGumb` can empirically perform nearly as well as the less scalable version that uses `MetaAlgo` with $\texttt{KnownGumb}^{k}$.



Unknown Domain, Restricted ℓ_0-sensitivity: `UnkBase`

- Algorithm `UnkBase` simply uses the Binary Mechanism for the elements you have seen and imposes a threshold m_δ such that no element is released with a noisy count lower than m_δ where

$$m_\delta = 1/\varepsilon \cdot (\log_2(T) + 1)\, \Phi^{-1}(1 - \delta/T) + 1$$

- Analysis considers a binary mechanism that pads all cells with dummy variables $\{\top_j\}$, and when a new element appears in the sequence it replaces a dummy variable.
- When aggregating to get the histogram at round t, we replace any dummy variable with new elements.
- We know that dropping all dummy elements that appeared above the threshold does not impact the privacy loss (post processing).
- Further, adding independent noise to the dummy elements and dropping them is equivalent to never having considered them

References

[CSS] Chan, Shi, and Song. Private and continual release of statistics. In *ACM Trans. Inf. Syst. Secur.'11*.

[DKM$^+$] Dwork, Kenthapadi, McSherry, Mironov, and Naor. Our data, ourselves. In *EuroCrypt06*.

[DMNS] Dwork, McSherry, Nissim, and Smith. Calibrating noise to sensitivity in private data analysis. In *TCC '06*.

[DNPR] Dwork, Naor, Pitassi, and Rothblum. In *STOC '10*.

[DNR$^+$] Dwork, Naor, Reingold, Rothblum, and Vadhan. On the complexity of DP data release.

[DR] Durfee and Rogers. Practical DP top-k selection with pay-what-you-get composition. In *NeurIPS'19*.

[MT] McSherry and Talwar. Mechanism design via DP. In *FOCS'07*.

[QYL] Qardaji, Yang, and Li. Understanding hierarchical methods for dp histograms. In *VLDB Endowment'13*.

[RSP$^+$] Rogers, Subramaniam, Peng, Durfee, Lee, Kancha, Sahay, and Ahammad. Linkedin's audience engagements API.