Private Continual Release of Real-Valued Data Streams

Date: February 26, 2019

Victor Perrier,
Hassan Jameel Asghar,
Dali Kaafar
Macquarie University, Data61 (CSIRO), ISAE-SUPAERO

Streaming Data and Statistics

- Real-time monitoring of customer data can improve services
  - Real-time updates
  - Analysts/planners can optimize services

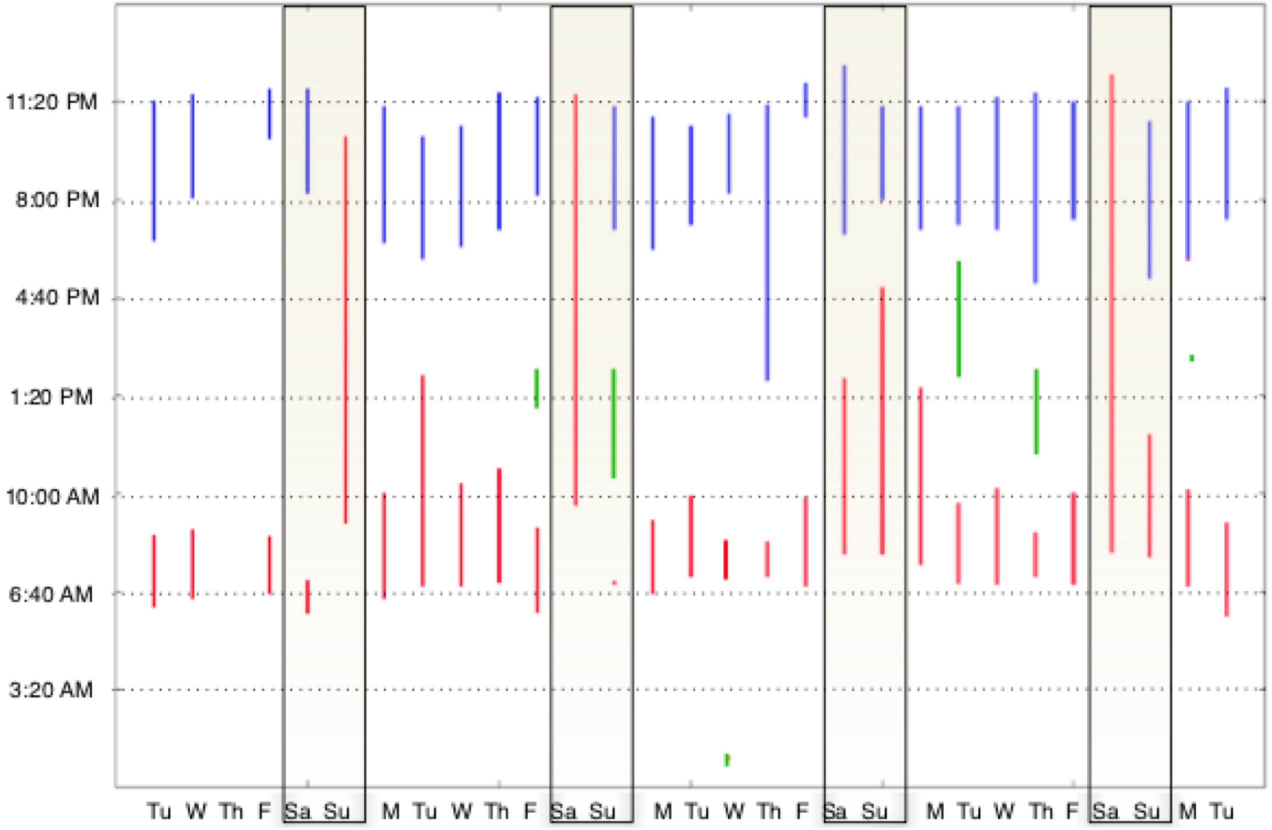
Service	Event	Real-time statistics
Energy	Smart-meter reading	Electricity usage in a neighborhood
Transport	Tap-on/off time	Peak hour commute times
Retail	Supermarket bill	Average expenditure in a supermarket
Location	Check-in/out time	Average time spent in a restaurant

Issue: Privacy

- Raw stats may reveal **sensitive** events
  - Unusual presence at home (smart meter)
  - Trip to beach instead of work (transport)

- Events (observations) can be linked to real-life activities [MSF+10]

Smart meter readings

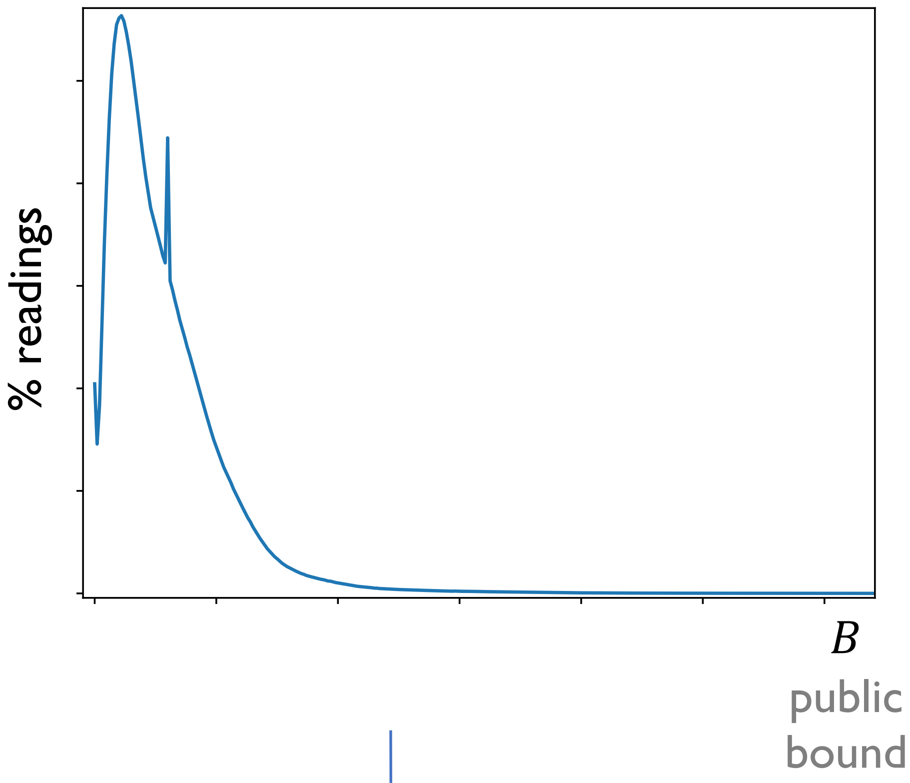


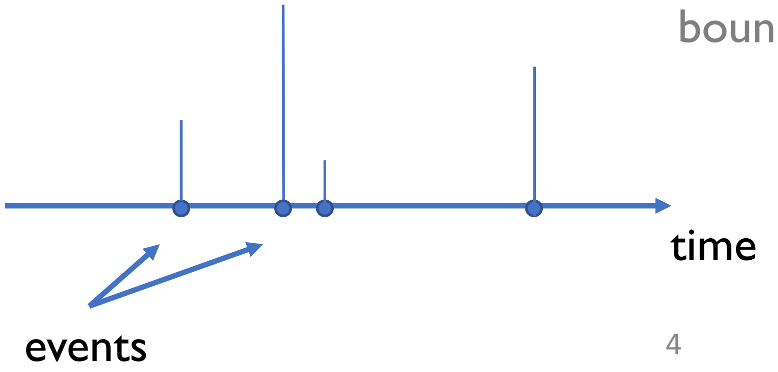
Unusual activity [MSF+10]

Privacy-Preserving Statistics

- Differential privacy a natural candidate
  - Most work on static databases
  - Some work on binary data streams [DNPR10, CSS11]
- Our problem
  - Data from an event is **real-valued** within a public upper bound B
  - Release updated sum/average at each event
  - ***Event-level*** privacy
    - Peculiar events protected

CDF of observations



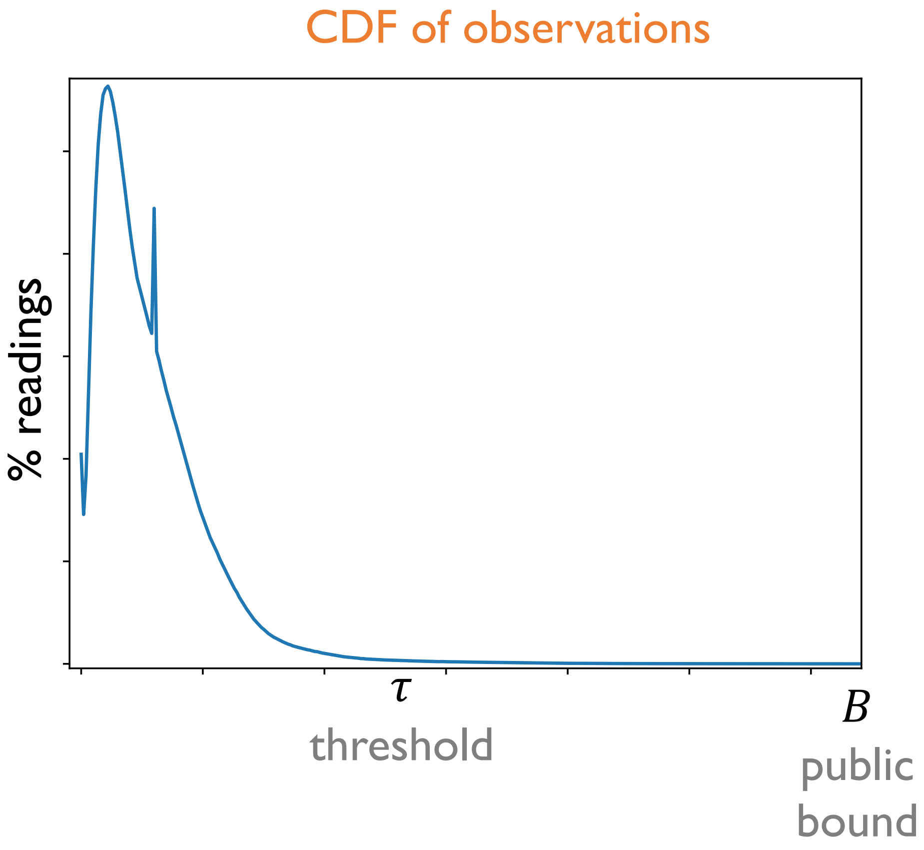


How to Release the Average?

- Basic: add Laplace noise of scale B to each observation
  - Error Bn after n events

- Generalized binary stream algorithm fairs better
  - Error $B \log_2 n$ [DNPR10, CSS11]

- Problem: error still proportional to B
  - In many situations B is too **loose** or **unknown**
    - E.g., Unlikely someone commuting for full 24 hours!
  - Most readings concentrated below a **threshold** τ
  - If τ known, error is only $\tau \log_2 n$
    - Significant if $B : \tau$ large

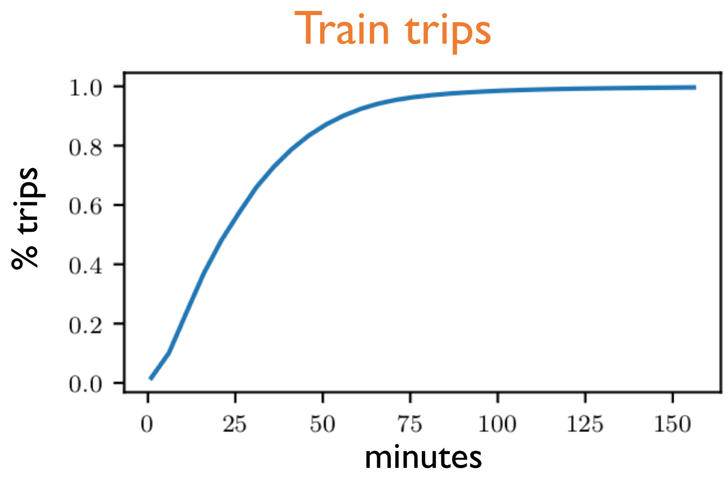


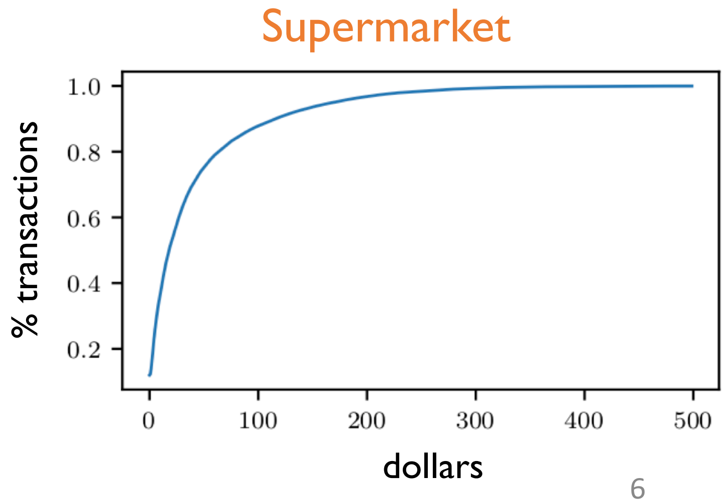
Validation of Data Concentration

- Is data really concentrated well below a conceivable B?

- Train trips dataset
  - 50 million trips over four weeks (Sydney, Australia)
  - Conceivable bound $B = 24$ hours

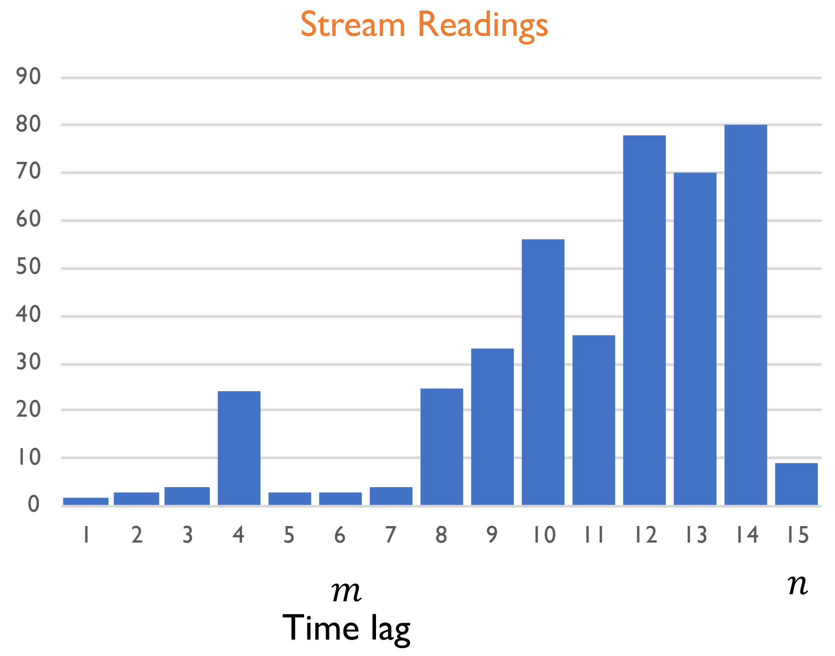
- Supermarket dataset
  - 140,000 transactions by 1,000 customers (Australia)
  - Conceivable bound $B = ?$





How to Estimate Threshold with Privacy?

- Need to observe a subset m of observations – **time lag**
- Time lag needs to be optimized for accuracy
  - Too early: high outlier error
  - Too late: marginal gain (may just use B as estimate)
- Naively estimating τ violates privacy
  - E.g., maximum of m observations is an exact event!



Our Work

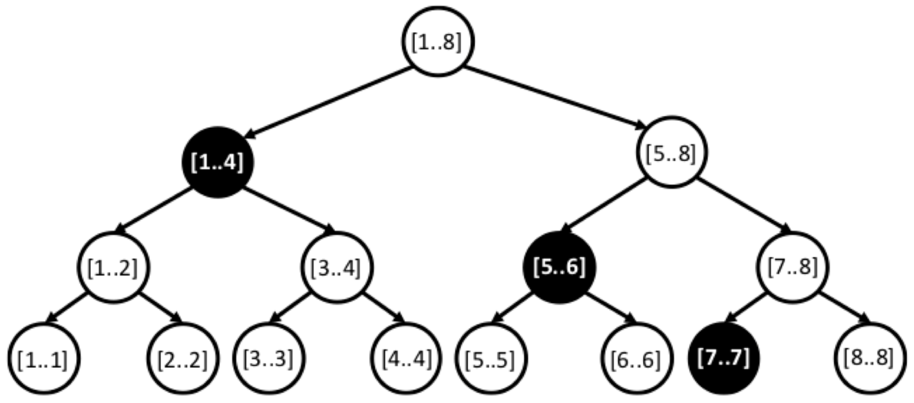
- A method to estimate threshold τ using a subset of observations
  - With differential privacy
  - and utility optimized for moving average

- Mechanism is **generic** – can also be used for
  - Average over a sliding window
  - Releasing histogram of streaming data
  - Estimating scale of distribution

Background: Binary Tree Algorithm

- Binary tree (BT) algorithm [DNPR10, CSS11]
  - Find at most $\log_2 n$ nodes in tree whose union equals sum up to i events
  - Add Laplace noise of scale $\frac{B\log_2 n}{\epsilon}$ instead of $\frac{Bn}{\epsilon}$

- **Goal:** Use BT as sub-module but noise scaled to τ instead of B



Computing private sum of first 7 observations

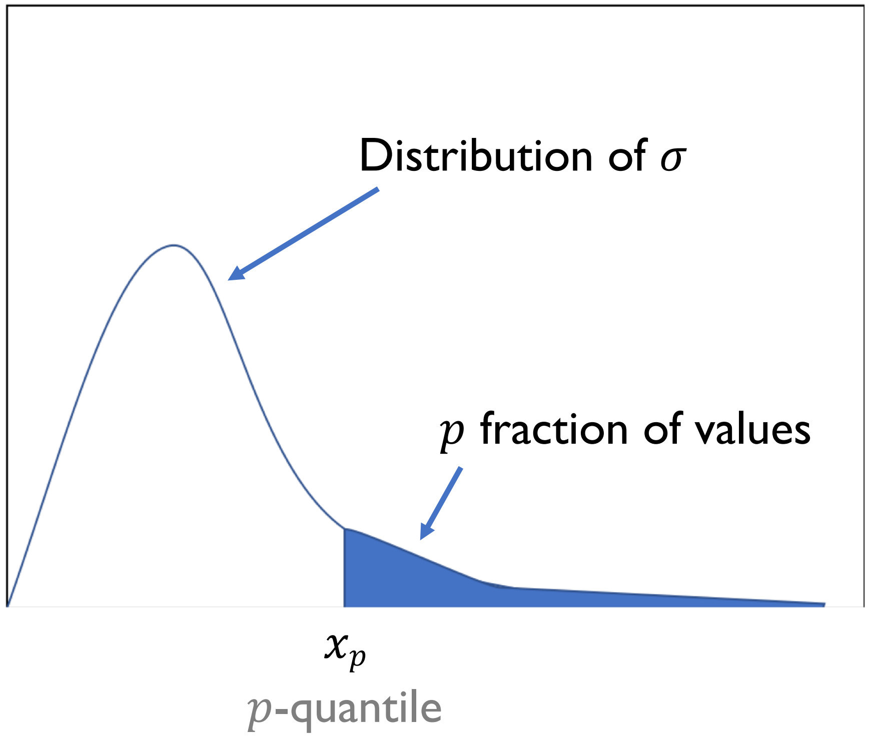
Global Mechanism

1. Estimate threshold τ using first m observations using budget ϵ_1
2. Use Laplace noise with scale $\frac{\tau}{\epsilon_2}$ to release sum of first m observations
3. Update & release sum for each event after m with Laplace noise of scale $\tau \log_2 n/\epsilon$ using BT algorithm

- Overall: (ϵ, δ)-differential privacy

What are the Choices for Threshold?

- False starts
  - Differentially private max of m values?
    - max function is highly sensitive
    - Adjacent streams can differ by any value in $[0, B]$
  - Standard deviation of distribution of σ?
    - Need to know distribution in advance
- Statistic of choice: p-quantile
  - E.g., $p = 0.005$ (0.5% of values)



Privately Estimating p-Quantile

- Need to estimate p-quantile through first m readings
    - Satisfying $n \gg m \gg 1/p$ ← required for stable estimate of p
- Roadmap
    - Obtain the empirical estimate $\hat{x}_p$ of x_p
    - Add differentially private noise to $\hat{x}_p$
    - Set the result as threshold τ
- **Complication:** cannot use Global Sensitivity (GS) for DP noise
    - Maximum change in function over **all** adjacent streams
    - GS of p-quantile is close to B

Using Smooth Sensitivity

- Local sensitivity (LS)
  - Maximum change in p-quantile over streams adjacent to input stream **only**
  - Unfortunately, LS itself can be sensitive
    - E.g., big differences in LS over nearby streams

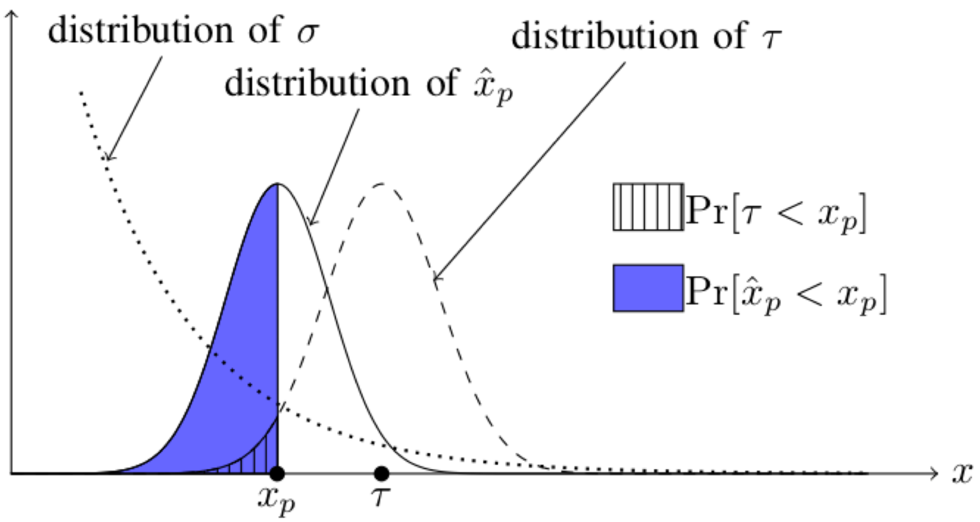
- Smooth sensitivity (SS) [NRS07]
  - $d(\sigma, \sigma')$: Hamming distance between streams σ and σ'
  - $\mathrm{SS}(\sigma, b) = \max_{\sigma'} \ \{e^{-bd(\sigma,\sigma')} \cdot \mathrm{LS}_{\sigma'}\}$
    - Smooths out change in LS as we move away from input stream

Privately Obtaining the Threshold

- Obtain threshold as

$$\tau = \hat{x}_p + \text{Laplace noise with SS}$$

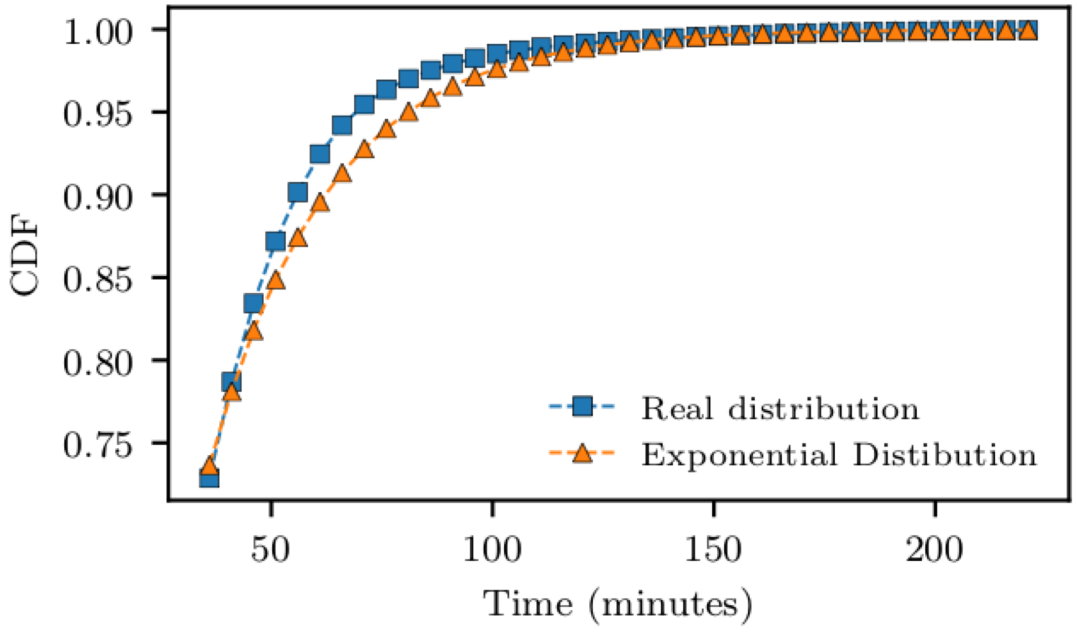
- We have swept some details under the rug
  - $\hat{x}_p$ and τ should be $\geq x_p$ to bound error
  - We assume $\hat{x}_p \geq x_p$



Utility Analysis

- Light-tailed distributions
  - Lighter than exponential distribution with the same p-quantile
- True for train trips and supermarket datasets for sufficiently small p
- If distribution is light-tailed
  - We show that error $\tau \log_2 n / \epsilon$ (**as required**)
  - **Note:** Privacy definition **not** dependent on distribution assumption

Train trips dataset



$p = 0.005$-quantile

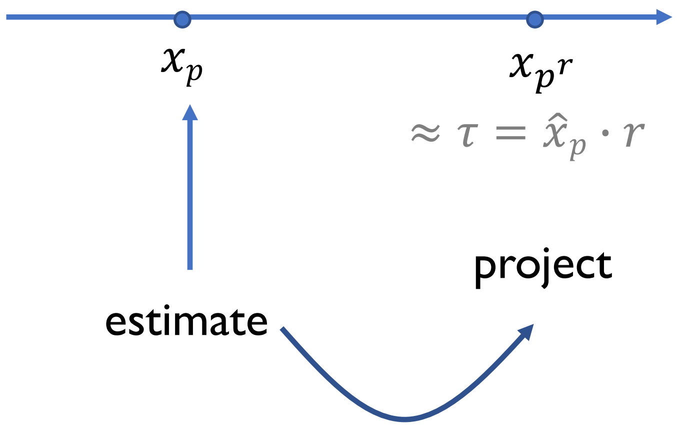
Utility Analysis for Light-tailed Distributions

- Exponential distribution has the property

$$x_p \cdot r \geq x_{p^r} \text{ for all } r \geq 1$$

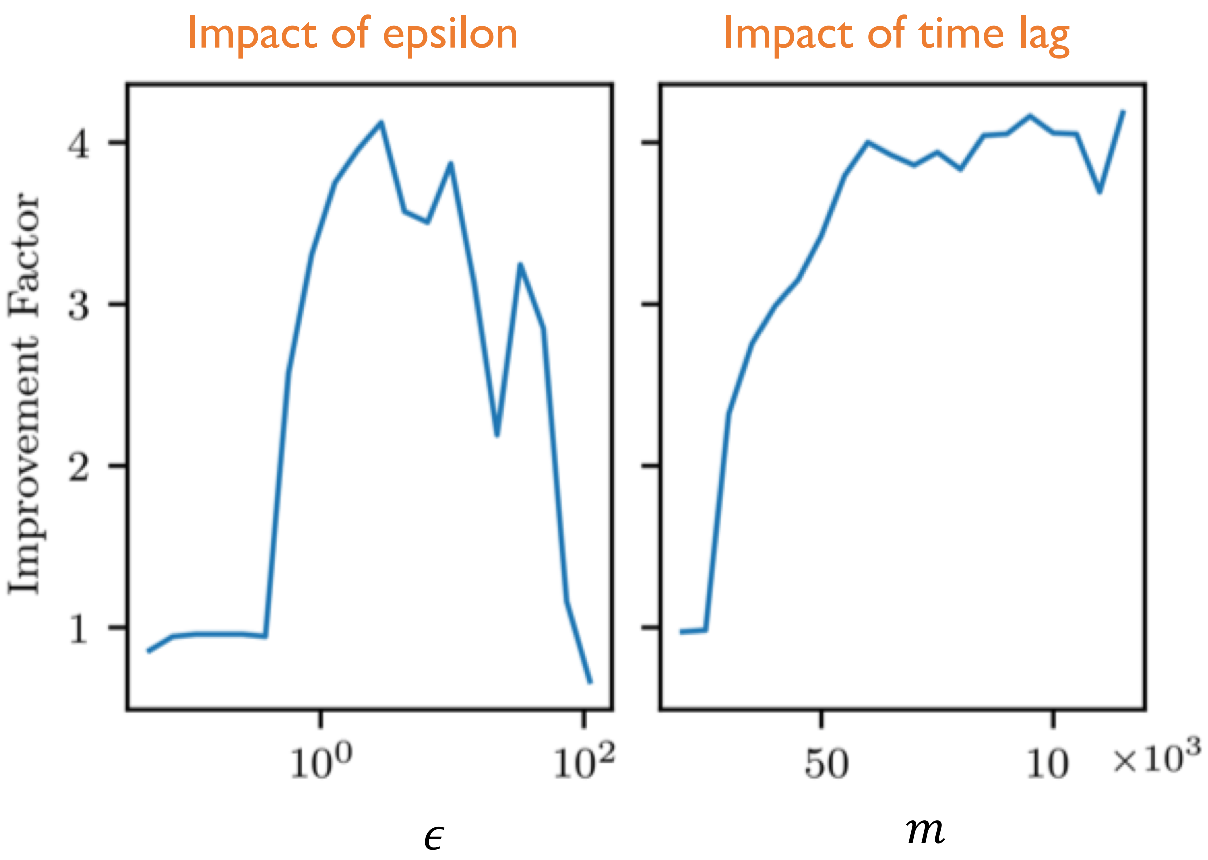
  - For **light-tailed** distributions: $\hat{x}_p \cdot r \geq x_{p^r}$

- Idea:
  - Estimate p-quantile using $1/p$ readings
  - Set threshold τ to $\hat{x}_p \cdot r$
  - Benefits:
    - Estimate threshold with a much smaller time lag m
    - Minimise outlier error
  - $O\left(\frac{\tau}{\epsilon}\log_2 n\right)$



What Values to Use in Practice?

- Improvement Factor (IF) metric
  - Ratio of error through BT versus our method
- **Epsilon:** IF increases with larger ϵ but then drops
  - Due to truncation: any value greater than threshold is fixed to threshold
- **Time lag:** Noticeable increase in impact factor with $m \approx 50{,}000$



Heuristics for Choosing Parameters

- Optimization suggests

Parameter	Interpretation	Value
p	p-quantile	0.005
r	Shifting p-quantile	Between 1 and 2
ϵ_1	Budget to estimate threshold	0.8 of overall privacy budget
ϵ_2	Budget to release sum of first m terms	Derive from ϵ_1
m	Time lag	50,000

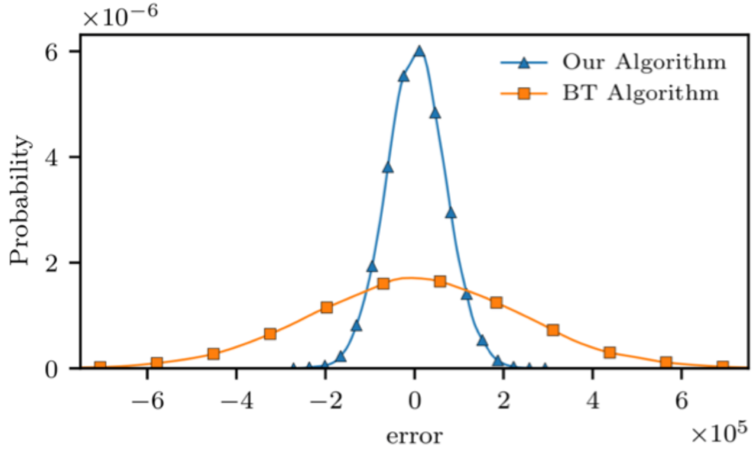
Experimental Evaluation

- Max error on the sum (at step n)
  - 20k repetitions

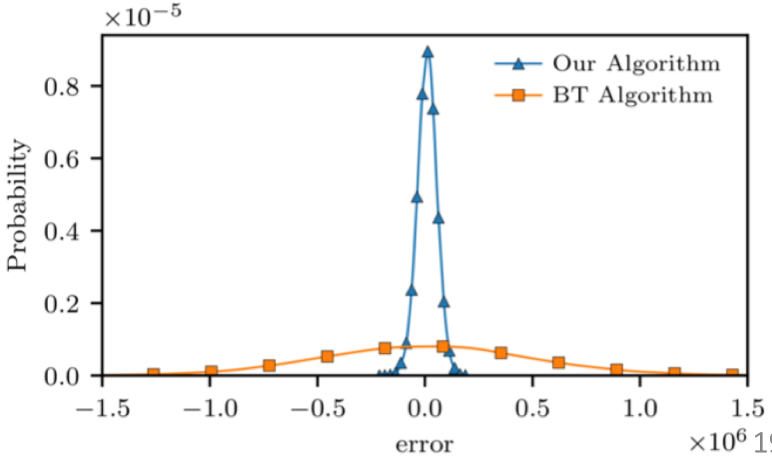
- Train trips
  - n = 250, 000, 000
  - m = 50,000
  - B = 1440 mins (24 hrs)
  - Improvement factor: 3.5

- Supermarkets
  - n = 150, 000
  - m = 50,000
  - B = 3,000 dollars
  - Improvement factor: 9

Train trips dataset



Supermarket dataset



Discussion

- Improved private release of moving average if distributions are light-tailed

- Question: which data have light-tailed distribution?
  - Any data coming from **short-lived, time constrained** events
    - Smart-meter data
    - Phone-call durations
    - Length of posts (on social media)
    - Daily average inter-arrivals of check-in times

- Heavy-tailed distributions are not “directly” time-constrained
  - Income distribution
  - File sizes in computer systems

Conclusion

- Shown a way to privately estimate the bulk of a distribution of streaming real-valued data
- Can be estimated by sacrificing a time lag
- Heuristics for choosing parameters in practice
- In worst-case, threshold is close to public bound B
    - We do not need to abort as in the propose-test-release approach [DL09]
- Moving average release is just one application – can be used in other applications

Questions

References

- [CSS11]Chan, T.H.H., Shi, E. and Song, D., 2011. Private and continual release of statistics. ACM Transactions on Information and System Security (TISSEC), 14(3), p.26.
- [DL09] Dwork, C. and Lei, J., 2009, May. Differential privacy and robust statistics. In STOC (Vol. 9, pp. 371-380).
- [DNPR10] Dwork, C., Naor, M., Pitassi, T. and Rothblum, G.N., 2010, June. Differential privacy under continual observation. In Proceedings of the forty-second ACM symposium on Theory of computing (pp. 715-724). ACM.
- [MSF+10] Molina-Markham, A., Shenoy, P., Fu, K., Cecchet, E. and Irwin, D., 2010, November. Private memoirs of a smart meter. In Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building (pp. 61-66). ACM.
- [NRS07] Nissim, K., Raskhodnikova, S. and Smith, A., 2007, June. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing* (pp. 75-84). ACM.