Practical Differential Privacy

Georgios Kellaris

Ph.D. Dissertation at HKUST

Privacy Tools for Sharing Research Data

A National Science Foundation Secure and Trustworthy Cyberspace Project

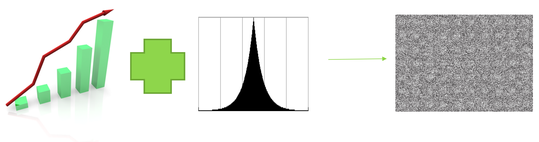
with additional support from the Sloan Foundation and Google, Inc.

ϵ-Differential Privacy [DMNS06]

- Ensure user privacy when publishing statistics, by hiding the presence of any user in the data
- How?
  - Add noise to the statistics before publishing
  - Noise is drawn form the Laplace Distribution
  - The noise scale must be proportional to the impact any user has on the statistics (sensitivity)

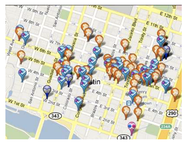
Utility Issues

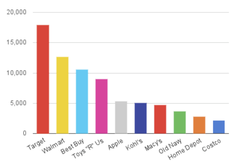
- The required noise may destroy the accuracy of the output
- For critical applications this might be unacceptable
  - E.g. for detecting disease outbreaks



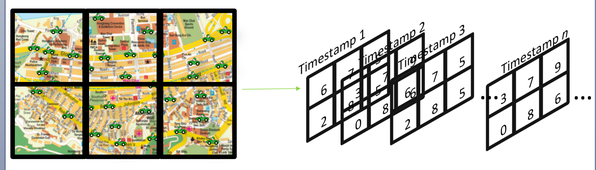
Targeted Settings and Challenges

- Historical Data
  - High Sensitivity



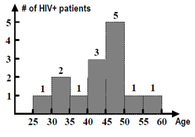


- Infinite Streams
  - Event-level privacy [DNPR10] not enough



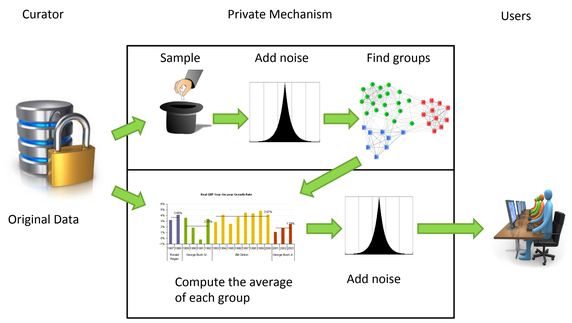
- Histograms for Range Queries
  - Noise error accumulates when computing ranges
  - Accurate methods are not efficient

Name	Age	HIV+
Alice	42	Yes
Bob	31	Yes
Carol	32	Yes
Dave	36	No
Ellen	43	Yes
Frank	41	Yes
Grace	26	Yes
...	...	...

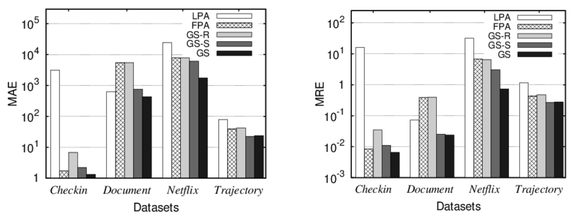


Solutions

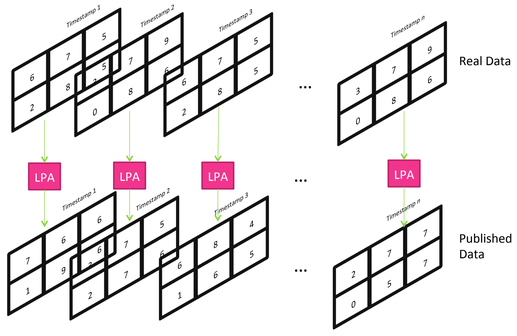
Historical Data [KP13]



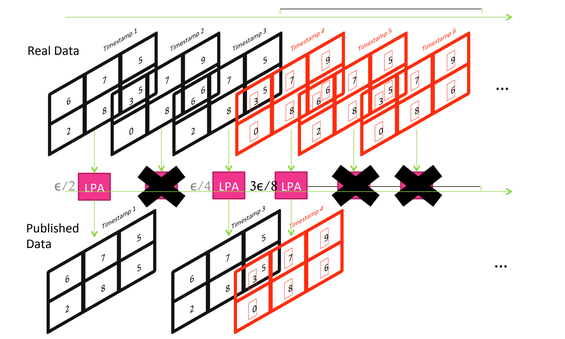
Results



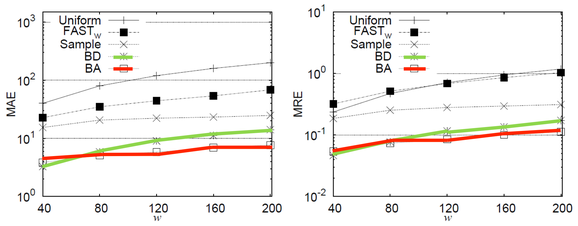
Infinite Streams [KPXP14]



- Event level
  - ϵ-differential privacy at any timestamp
  - it does not protect user movement
  - noise proportional to 1
- w-Event level
  - ϵ-differential privacy at any w consecutive timestamps
  - it protects user movement that lasts at most w timestamps
  - noise proportional to $w \ll T$
- User level
  - ϵ-differential privacy at all timestamps
  - it protects user movement
  - noise proportional to T



Results

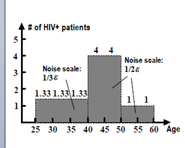


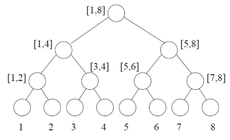
(a) MAE (b) MRE

Histograms for range queries

- Every existing method can be reproduced from three *modules* by parameterizing them

Smoothing	Hierarchical	Workload Aware
[ZCXMX14]	[QYL13]	[LHMW14]
Slow ($O(d^2)$); Accurate for small ranges	Fast ($O(d)$) Accurate for large ranges	Very slow ($O(d^3 \log d)$) Very accurate

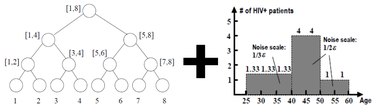




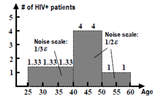
Given the range queries
- Add noise with different scale to each query answer
- Combine query answers that have overlaps to get more accurate results

- Combine modules in order to devise new efficient and accurate methods

SUB ($O(d)$)

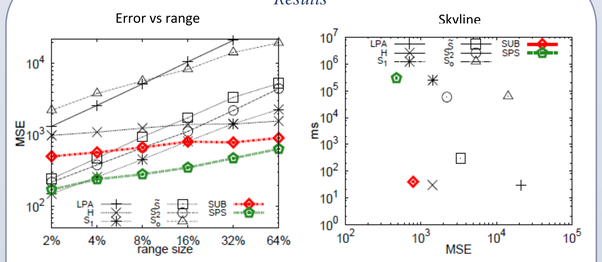


SPS ($O(d^2 \log d)$)



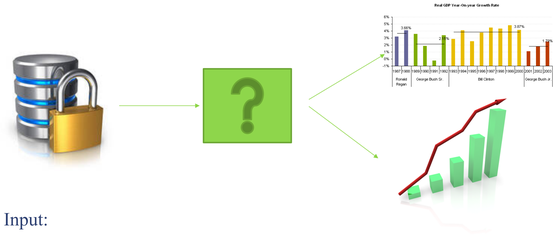
Workload Aware

Results



Goal

- Design a system that privately decides which method should be used in order to increase the accuracy of the result



- Input:
  - Original Data
  - Statistics
- Choose the best method/combination privately
- Compute the result
- Output noisy statistics

Challenges

- What are the data characteristics that render a method better than another in terms of accuracy?
- Can we determine the best method by utilizing non-private information on the input data?
- If not, can we utilize only statistics that have very low sensitivity?
- How can we choose the best method on-the-fly? (Streaming setting)

References

[DMNS06] Calibrating noise to sensitivity in private data analysis. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. TCC 2006.

[DNPR10] Differential privacy under continual observation. Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. STOC 2010.

[KP13] Practical Differential Privacy via Grouping and Smoothing. Georgios Kellaris, and Stavros Papadopoulos. VLDB 2013.

[KPXP14] Differentially private event sequences over infinite streams. Georgios Kellaris, Stavros Papadopoulos, Xiaokui Xiao, and Dimitris Papadias. VLDB 2015.

[LHMW14] A Data- and Workload-Aware Algorithm for Range Queries Under Differential Privacy. Chao Li, Michael Hay, Gerome Miklau, and Yue Wang. VLDB 2014.

[QYL13] Understanding Hierarchical Methods for Differentially Private Histograms. Wahbeh Qardaji, Weining Yang, and Ninghui Li. VLDB 2013.

[ZCXMX14] Towards accurate histogram publication under differential privacy. SDM 2014.

CRCS Center for Research on Computation and Society at Harvard School of Engineering and Applied Sciences

BERKMAN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The Institute for Quantitative Social Science at Harvard University

DATA PRIVACY LAB