Cache Me If You Can: Accuracy-Aware Inference Engine for DP Data Exploration

Miti Mazmudar, Thomas Humphries, Matthew Rafuse, Xi He

SETUP

- Interactive DP.
- Data owner specifies privacy budget ε.
- Data analysts care about accuracy; they do not know DP techniques.
- Accuracy: A set of queries (workload) has two accuracy parameters: α, β.

MOTIVATION

Status quo: Ge et al.'s APEx [1] provides accurate responses to workloads, while minimizing ε.
Drawbacks:
- Does not consider prior noisy responses ⇒ more ε spent than required.
- Analysts may not know how to plan workloads for minimial ε usage.

CONTRIBUTIONS

An inference engine, **CacheDP** for interactive DP queries that offers the following:
- Accurate responses to all analysts' workloads using lower privacy budget.
- Removes the need for analysts to plan subsequent workloads by maintaining a cache.
- Uses the cache to save privacy budget for related workloads.

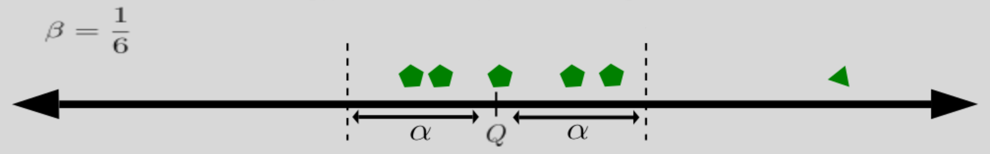
WORK-IN-PROGRESS

- Choosing a cache-aware optimal strategy matrix.
- Implementing our algorithm and cache structures.
- Evaluating our implementation for the following use-cases:
  - Entity resolution tasks.
  - Private spatial data exploration.
- Extending to multiple dimensions [4].

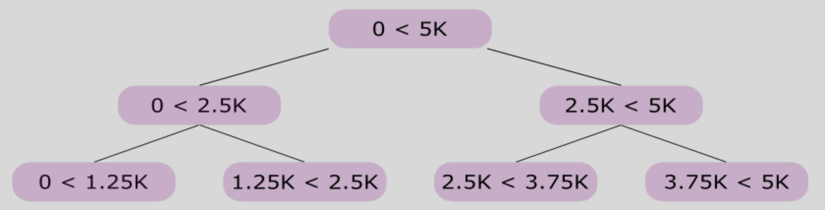
BACKGROUND

Accuracy Params: error bound α, failure rate β.

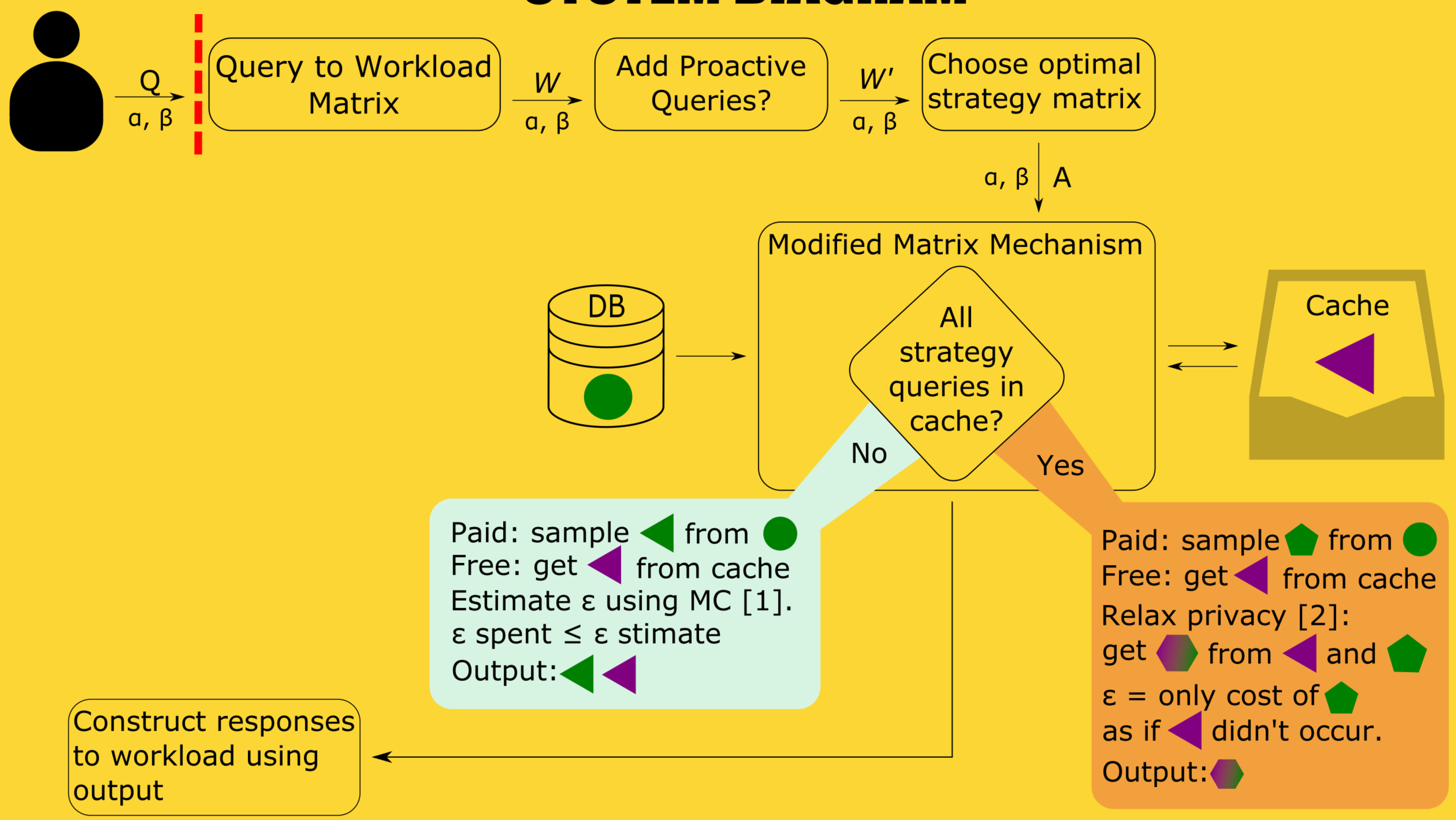
$$Pr[\|Q - \tilde{Q}\|_\infty \geq \alpha] \leq \beta$$



Matrix Mechanism [3]: Example Strategy



SYSTEM DIAGRAM



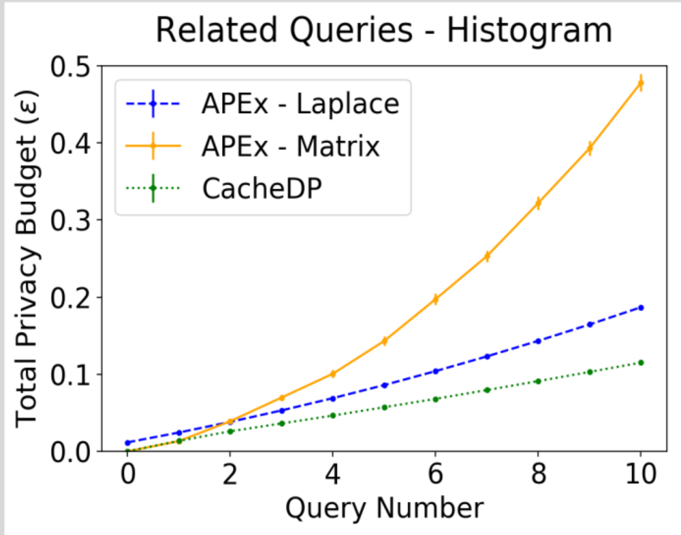
PROACTIVE APPROACH

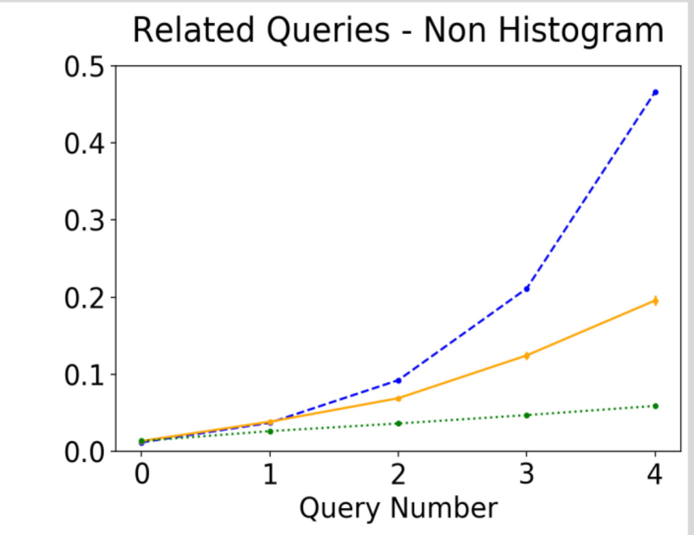
- Includes disjoint queries within the workload.
- Exploits the parallel composition theorem.
- Threshold to bound a minor increase in the privacy budget (for meeting the accuracy guarantee of overall workload)

RELATED QUERIES EXAMPLES

$$\text{Histogram: } Q_1 = [0, 5000), Q_2 = \{[0, 2500), [2500, 5000)\},$$
$$Q_3 = \{[0, 1250), [1250, 2500), [2500, 3750), [3750, 5000)\} \dots$$

$$\text{Non-Histogram: } Q_1 = [0, 5000), Q_2 = \{[0, 2500), [0, 5000)\},$$
$$Q_3 = \{[0, 1250), [0, 2500), [0, 3750), [0, 5000)\} \dots$$

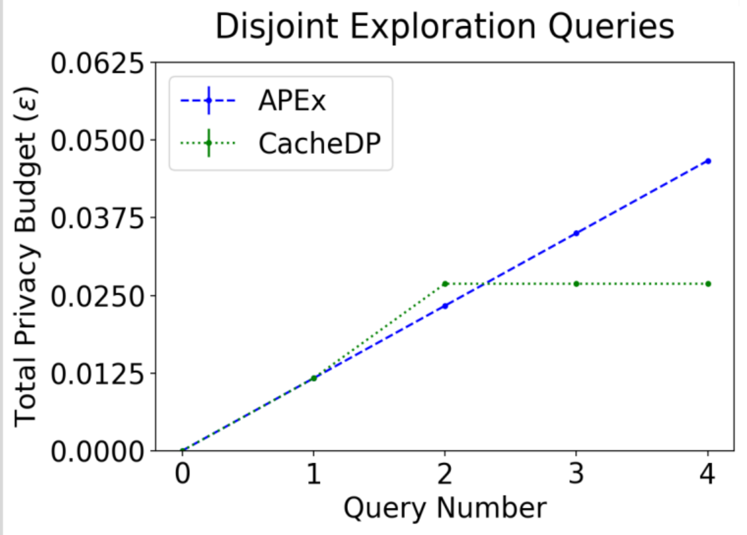


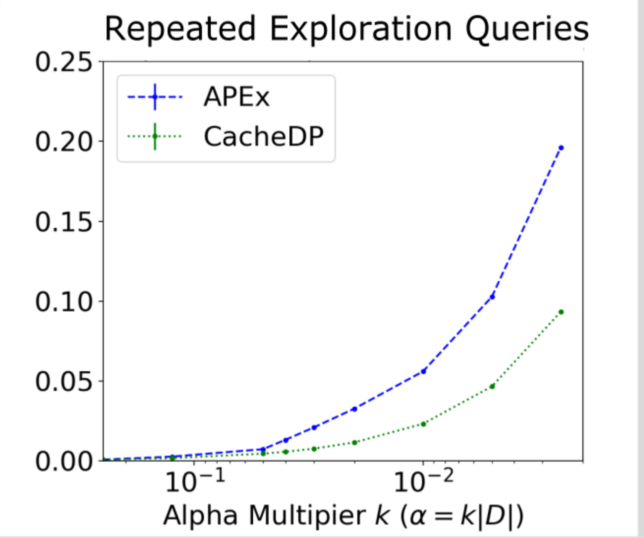


DISJOINT AND REPEATED EXAMPLES

$$\text{Disjoint: } Q_1 = [0, 5000), Q_2 = [0, 500), Q_3 = [500, 1000),$$
$$Q_4 = [500, 1000) \dots Q_{11} = [4500, 5000)$$

$$\text{Repeated: } Q_1 = [0, 5000), \alpha = 0.25|D|, \beta = 0.0005.$$
$$Q_2 = [0, 5000), \alpha = 0.125|D|, \beta = 0.0005.$$





REFERENCES

[1] Chang Ge, Xi He, Ihab F. Ilyas, and Ashwin Machanavajjhala. 2019. *APEx: Accuracy-Aware Differentially Private Data Exploration.* SIGMOD '19.
[2] Fragkiskos Koufogiannis, Shuo Han, and George J. Pappas. 2016. *Gradual Release of Sensitive Data under Differential Privacy.* Journal of Privacy & Confidentiality.
[3] Chao Li, Gerome Miklau, Michael Hay, Andrew Mcgregor, and Vibhor Rastogi. 2015. *The Matrix Mechanism: Optimizing Linear Counting Queries under Differential Privacy.* The VLDB Journal.
[4] Ryan McKenna, Gerome Miklau, Michael Hay, and Ashwin Machanavajjhala. 2018. *Optimizing Error of High-Dimensional Statistical Queries under Differential Privacy.* VLDB '18.

CrySP
Cryptography, Security, and Privacy Research Group
DSg Data Systems Group
UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE