Orthogonal Mechanism for Answering Batch Queries with Differential Privacy

Presented by

Huang Dong

01 July, 2015

A*STAR I²R

Outline

- Background
- Motivation
- Proposed Work
- Performance Evaluation
- Conclusions

A*STAR I²R

Background

- The use of personal data has grown vastly in the past few years and privacy protection is a major issue
- Differential privacy is a promising technique in achieving data privacy guarantee
- Noise magnitude affects the accuracy significantly, leading to unmeaningful results
- Recent works attempt to reduce noise magnitude but cause high computational complexity, inapplicable to large-scale datasets

A*STAR I²R

Motivation

- Correlation among multiple queries causes high noise magnitude
- Decompose the original queries into new queries can reduce noise magnitude
- Existing works rarely focus on the correlation analysis
- The decomposition based on query orthogonality have two distinct advantages:
  - Smaller required noise magnitude
  - Lower computational complexity

A*STAR I²R

Proposed Work

- Scenario: Data analysts want to make queries on the count of individuals in a dataset under differential privacy framework
- Suppose a query set consists of m queries expressed as: $Q(D) = WD$
- The Laplace mechanism (LM) is:

$$\mathcal{K}(Q, \mathcal{D}) = W\mathcal{D} + Lap(S(Q)/\epsilon)^m$$

- The noisy results obtained by LM may be unmeaningful due to high noise magnitude

A*STAR I²R

Proposed Work (cont'd)

- We decompose the query matrix W by $W = B\tilde{W}$, where $\tilde{W}$ is the new query matrix.
- $\tilde{W}$ is constructed first, then derive B. The construction of $\tilde{W}$ is based on orthogonality
- The proposed orthogonal mechanism (OM) is

$$\mathcal{F}(Q, D) = B(\tilde{W}D + Lap(S(\tilde{Q})/\epsilon)^s)$$

- Construction procedure of $\tilde{W}$:
  - Suppose $rank(W) = r$, then randomly select r independent queries

Proposed Work (cont'd)

- **Construction Procedure of $\tilde{W}$:**
  - **Given r independent queries, count the number of occurrence of each domain x_i and find the index with the largest count**
  - **Find the query set containing domain x_i from the r independent queries**
  - **Find the intersect of the above query set**
  - **Construct a new query set $\tilde{Q}$, consisting of s queries, from the intersect to represent the original query set.**
- **When the new query matrix, $\tilde{W}$, is derived, the matrix B is easy to be resolved**

A Practical Example

- **Consider a query set Q with workload matrix**

$$W = \begin{bmatrix} 0.3657 & 0 & 0.9812 & 0 \\ 0 & 0.0645 & 0 & 0 \\ 0 & 0.5879 & 0.7602 & 0 \\ 0 & 0 & 0 & 0.7310 \\ 0 & 0.7313 & 0 & 0 \\ 0 & 0 & 0.7122 & 0.9053 \end{bmatrix}$$

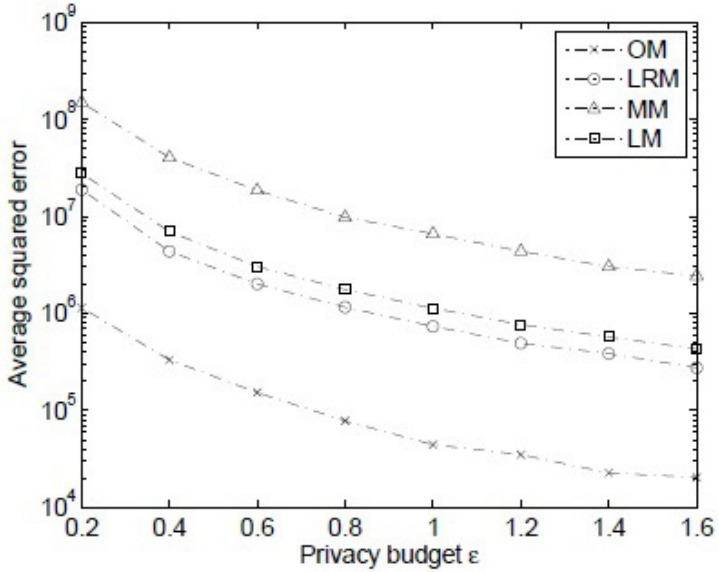
- **Decomposition results:**
- $$B = \begin{bmatrix} 0 & 0.9812 & 0.3657 & 0 \\ 0.0645 & 0 & 0 & 0 \\ 0.5879 & 0.7602 & 0 & 0 \\ 0 & 0 & 0 & 0.7310 \\ 0.7313 & 0 & 0 & 0 \\ 0 & 0.7122 & 0 & 0.9053 \end{bmatrix} \quad \tilde{W} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- **Noise variance comparison:**
  - **Before decomposition:** $\approx 75/\epsilon^2$ **since** $S(Q) \approx 2.5$
  - **After decomposition:** $< 18/\epsilon^2$ **due to** $S(\tilde{Q}) = 1$

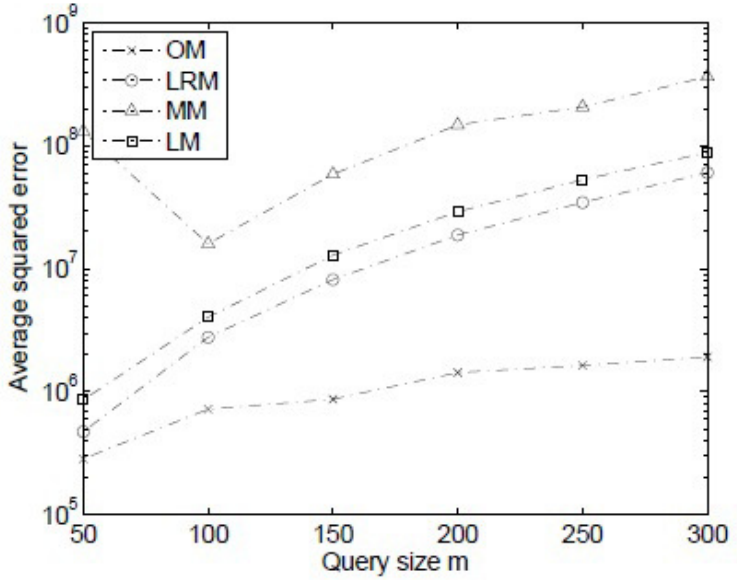
A*STAR I²R

Performance Evaluation

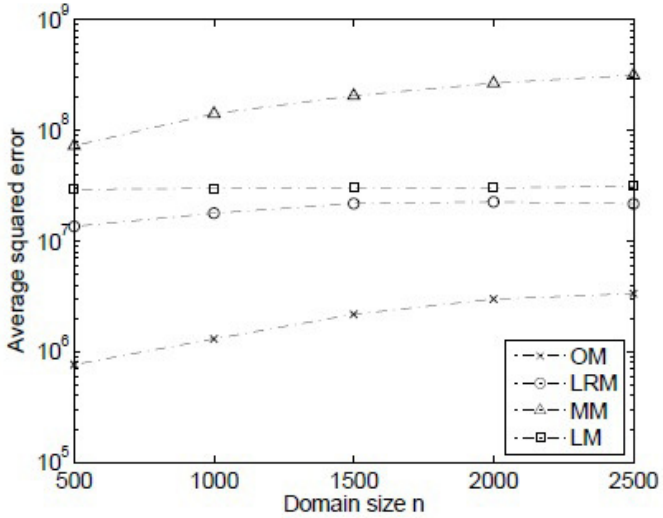
- **Accuracy comparison**



(b) W with $\tau = 0.4$



(b) W with $\tau = 0.4$

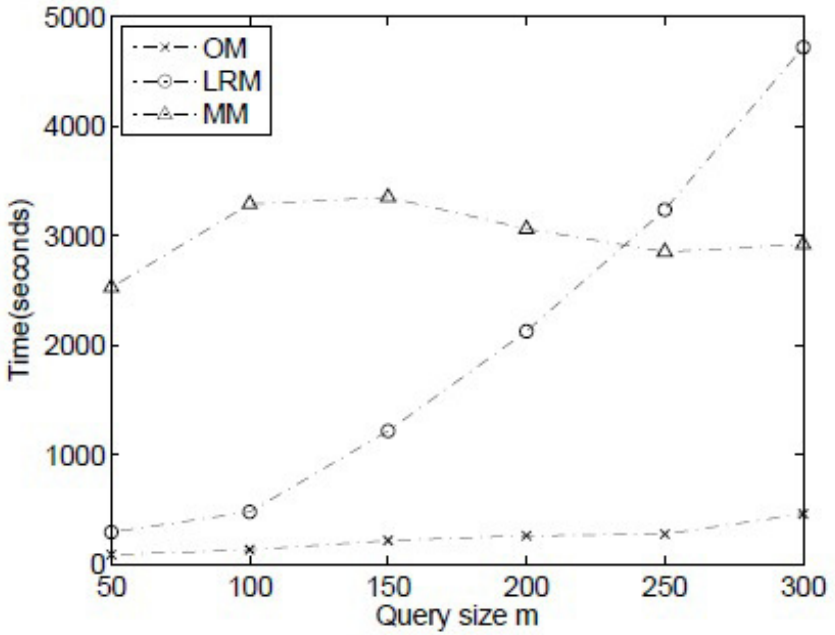


(b) W with $\tau = 0.4$

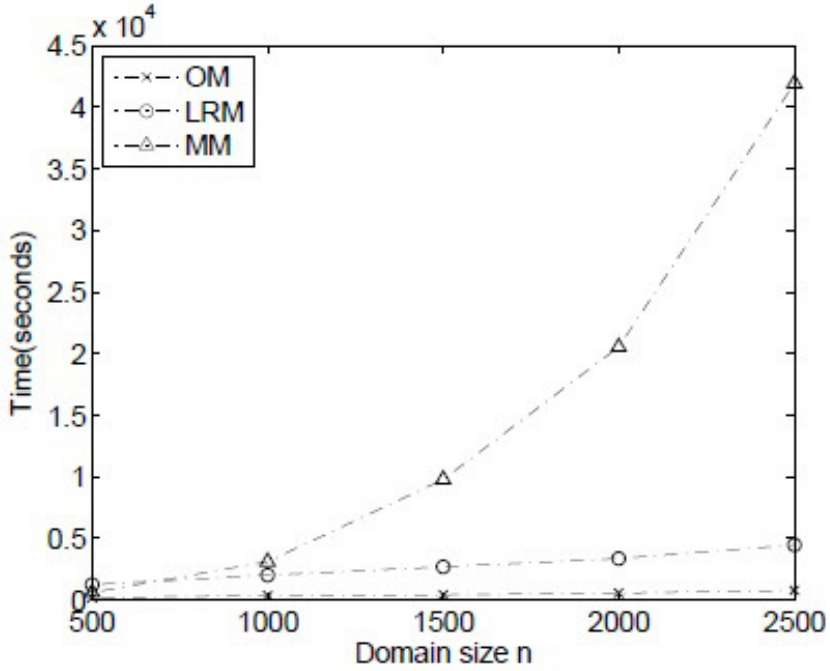
A*STAR I²R

Performance Evaluation (cont'd)

- **Execution time comparison**



(b) W with $\tau = 0.4$



(b) W with $\tau = 0.4$

A*STAR I²R

Conclusions

- We propose a novel mechanism, orthogonal mechanism (OM), for answering a batch of queries with differential privacy.
- The proposed OM significantly reduces the noise magnitude by removing the correlation between queries.
- The computational complexity of the proposed OM is much lower than that of existing work.

Thank You! Q&A

A*STAR I²R