
PRIVACY

Privacy Issues in Data Publishing

- Governments and organizations publish anonymous personal information for research, analytics and services

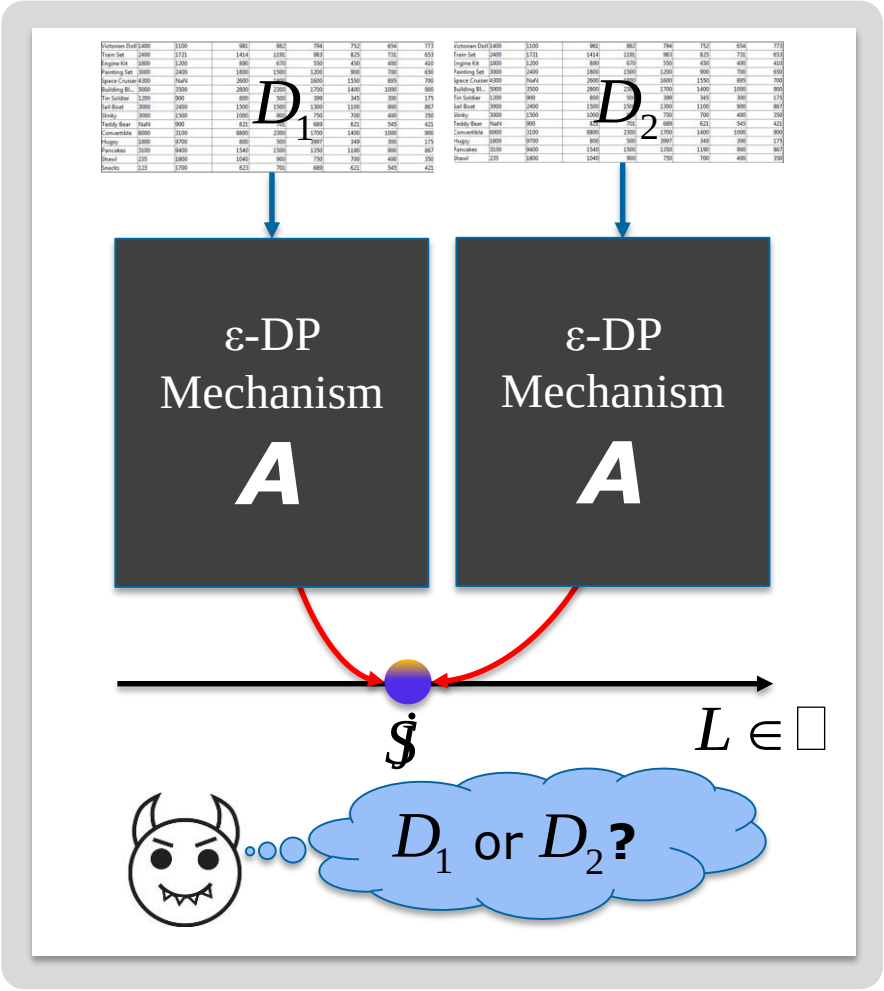
	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

www.netflix.com

- Privacy leak
  - identify a person from internet databases
    - de-anonymize Netflix Price dataset [A. Narayanan '08]
  - discover an individual's record by comparing databases
    - your record was not in the database last month, but now it is…

USC Viterbi School of Engineering

Differential Privacy (DP)



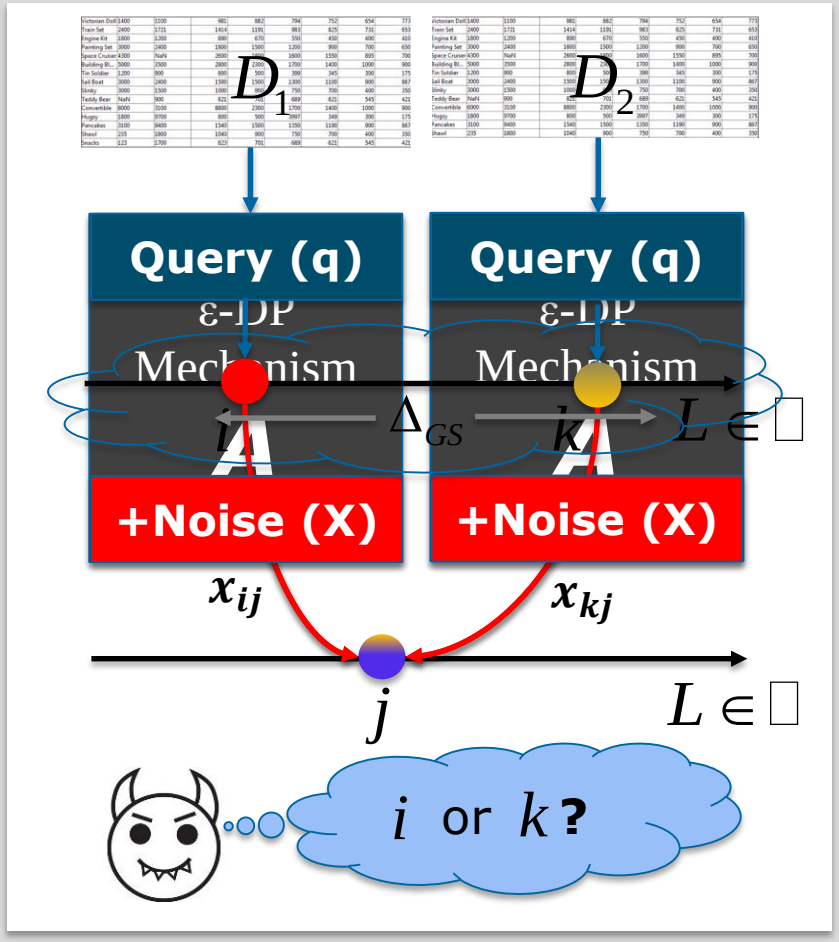
- ε-Differential Privacy [C. Dwork '06]

$$e^{-\varepsilon} \leq \frac{P_r[A(D_1) \in S]}{P_r[A(D_2) \in S]} \leq e^{\varepsilon}$$

  - privacy ➔ information loss
- ε-DP Mechanisms
  - DP Noise-adding mechanisms
    - *Laplacian, Geometric*
  - other DP mechanisms
    - *Matrix* [C. Li '10], *K-norm* [M. Hardt '09]
  - non-numeric DP mechanism
    - *Exponential* [F. McSherry '07]

USC Viterbi School of Engineering

DP Noise-Adding Mechanism



- DP Noise-Adding Mechanism

$$A(D) = q(D) + X(D)$$

- Oblivious Mechanism

$$e^{-\varepsilon} \le \frac{x_{ij}}{x_{kj}} \le e^{\varepsilon}, \quad \forall i,k \in L$$

$$|i-k| \le \Delta_{GS}$$

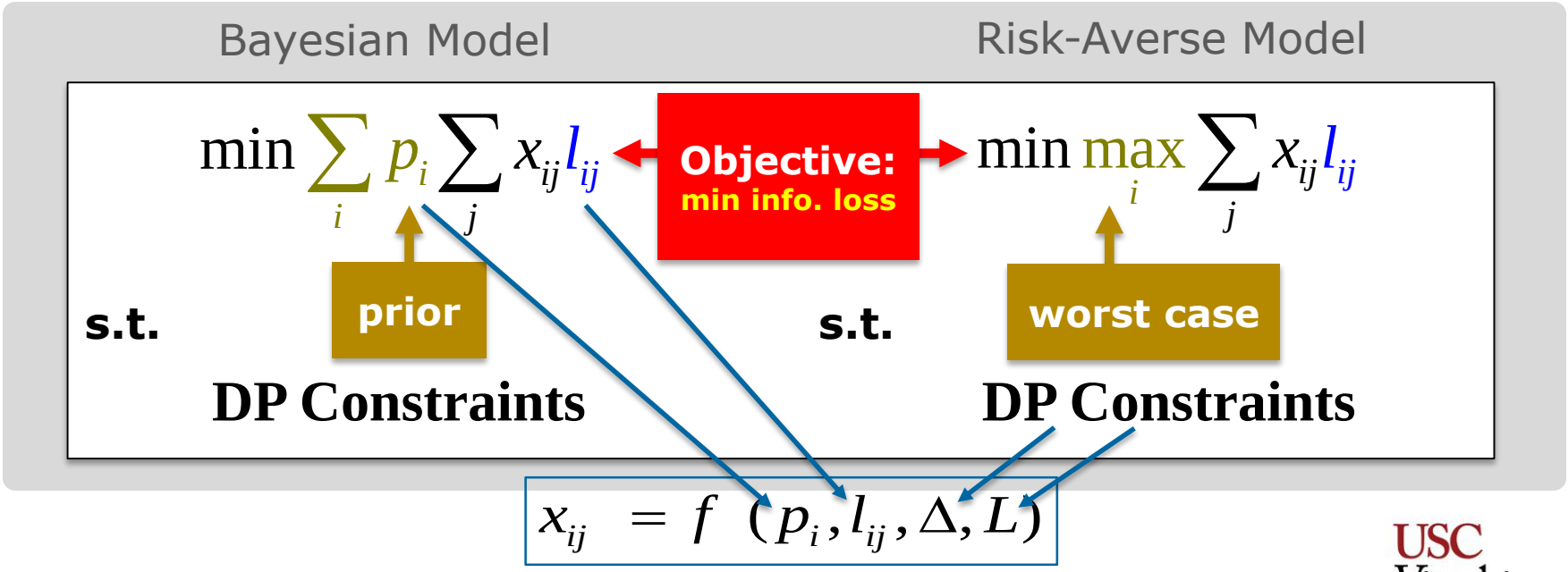
- Global Sensitivity

$$\Delta_{GS} = \max_{\substack{\forall D_1, D_2 \in D^n: \\ d_H(D_1, D_2)=1}} |q(D_1) - q(D_2)|$$

USC Viterbi School of Engineering

Optimal DP Mechanism

- Widely-used information loss function: $l_{ij} = l_{|j-i|}$
- A DP mechanism is called *optimal* if it **minimizes information loss** and preserves DP.
- Data managers solve the optimization problem for mechanism x_{ij}

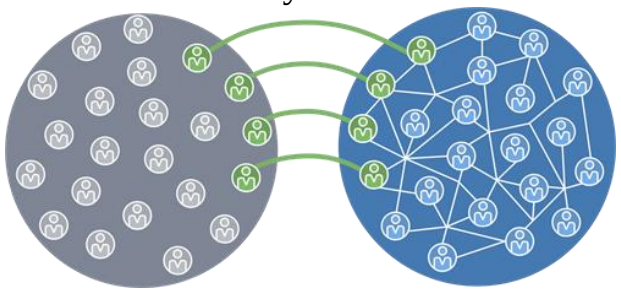


USC Viterbi School of Engineering

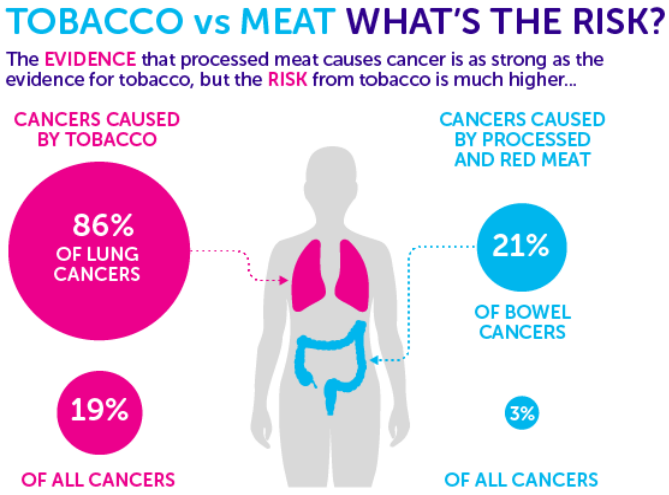
Presence of Side-Information

- Side-information exists everywhere…
  - auxiliary databases
  - research studies, common knowledge
  - mathematical theories
    - central limit theorem
    - transformations of random variables
- The presence of side-information is important and cannot be neglected.
- Side-information ➔ Prior probability

Auxiliary Databases



Research Studies

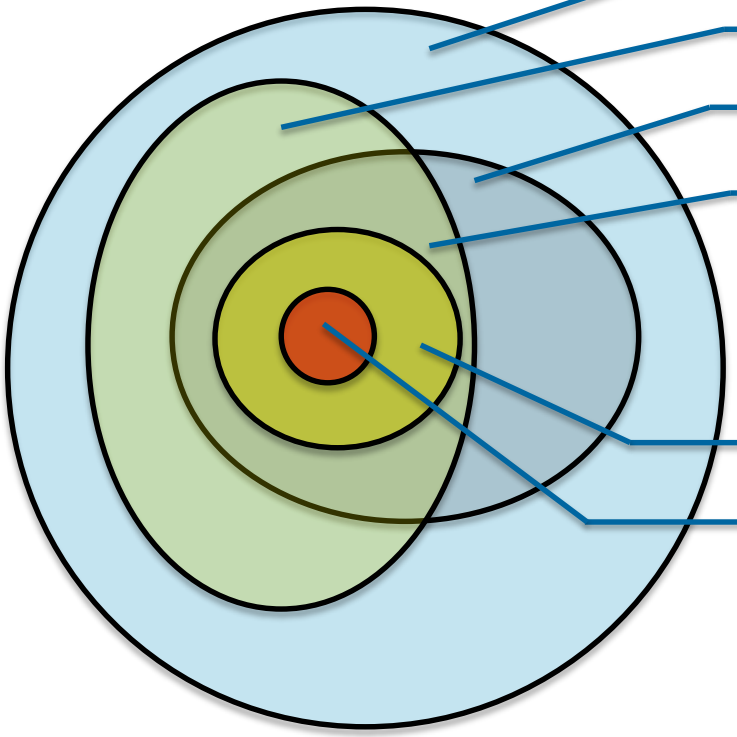


scienceblog.cancerresearchuk.org

USC Viterbi School of Engineering

State-of-the-Art and Open Questions

solution space = $(p_i, L, \Delta, l_{ij} = l_{|j-i|})$



- Optimal DP Mechanism (Bayesian)
- Optimal DP Mechanism (Risk-Averse)
- Optimal DP Mechanism (Bayesian, $\Delta = \Delta_{GS}$)
- Optimal DP Mechanism (Risk-Averse, $\Delta = \Delta_{GS}$)

:Staircase Mechanism [Q. Geng '14]

Optimal in **Risk-Averse** model

Optimal for **unbounded domain *L***

- Universally Optimal DP Mechanism (unknown)
- Universally Optimal DP Mechanism ($L \in Z$, Δ=1)

:Geometric Mechanism [M. Gupte '10] [A. Ghosh '12]

Universally optimal in both **Risk-Averse** and **Bayesian** model

- A *universally optimal* mechanism is optimal for all priors p_i and all loss functions l_{ij}.

USC Viterbi School of Engineering

Main Contributions

Propose **open questions** in DP mechanism design

For (Bayesian, $\Delta=\Delta_{GS}$), we propose a heuristic design

- optimal design for general priors is difficult
- we start with heuristic design, and it surprisingly leads to significant improvement in utility-privacy tradeoffs

Show via experiments, the importance of the optimal Bayesian mechanism design

- optimal Bayesian design is **non-trivial** when side-information substantially narrows down the outputs of the query

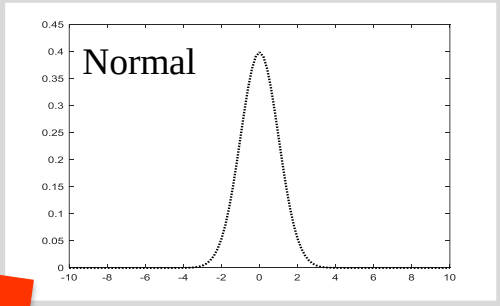
USC Viterbi School of Engineering

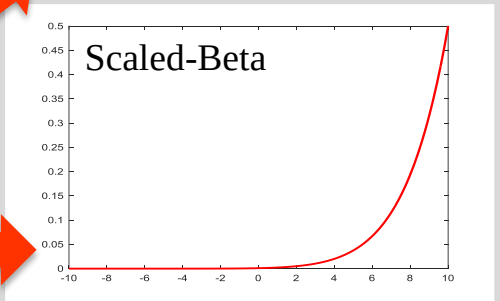
Experimental Context and Settings

- Queries - **Mean** and **Max**
- Oblivious mechanism ➔ database independent ➔ synthetic data

– public information
  - known: domain $L \in [-10,10]$
  - each entity is independent and uniformly distributed

– mathematical theories
  - central limit theorem
    ➔ **mean** value is approximately *normal* distributed
  - transformations of random variables
    ➔ the **max** value is scaled-*beta* distributed over L

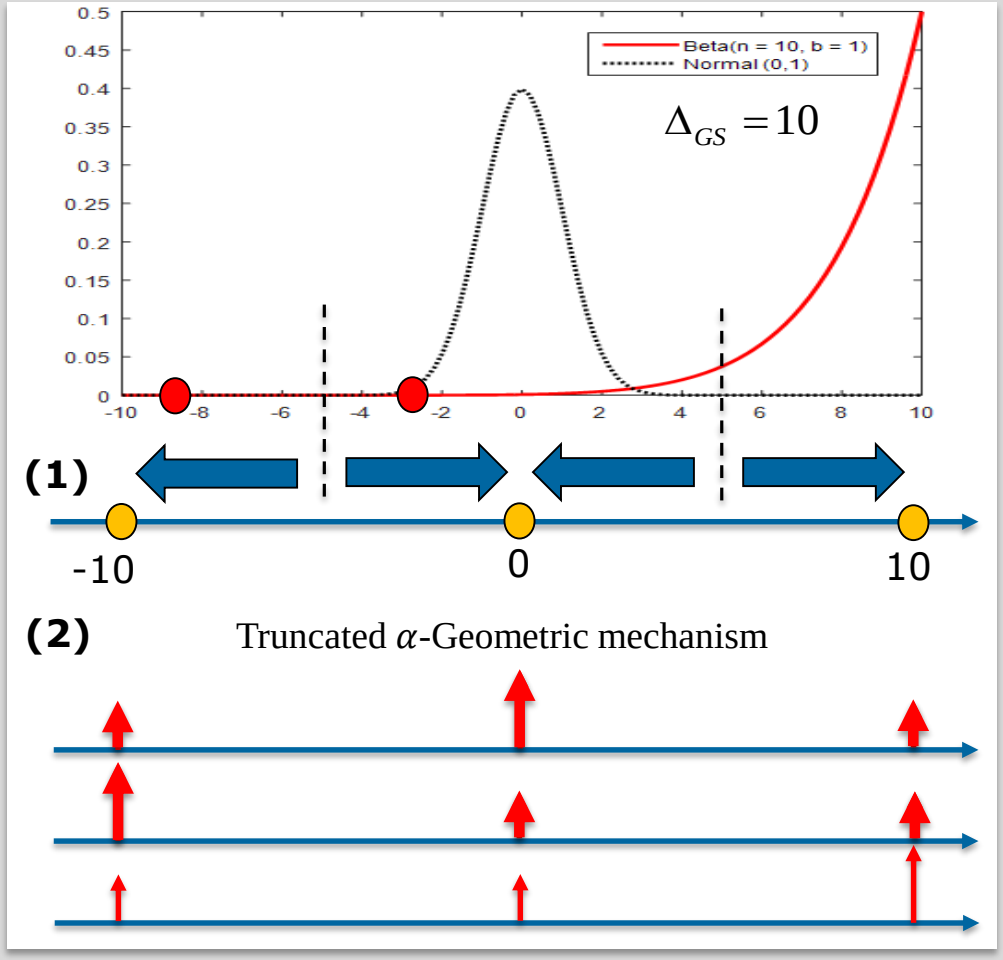




- Global sensitivity = 10
- Gaussian is truncated and normalized in L

USC Viterbi School of Engineering

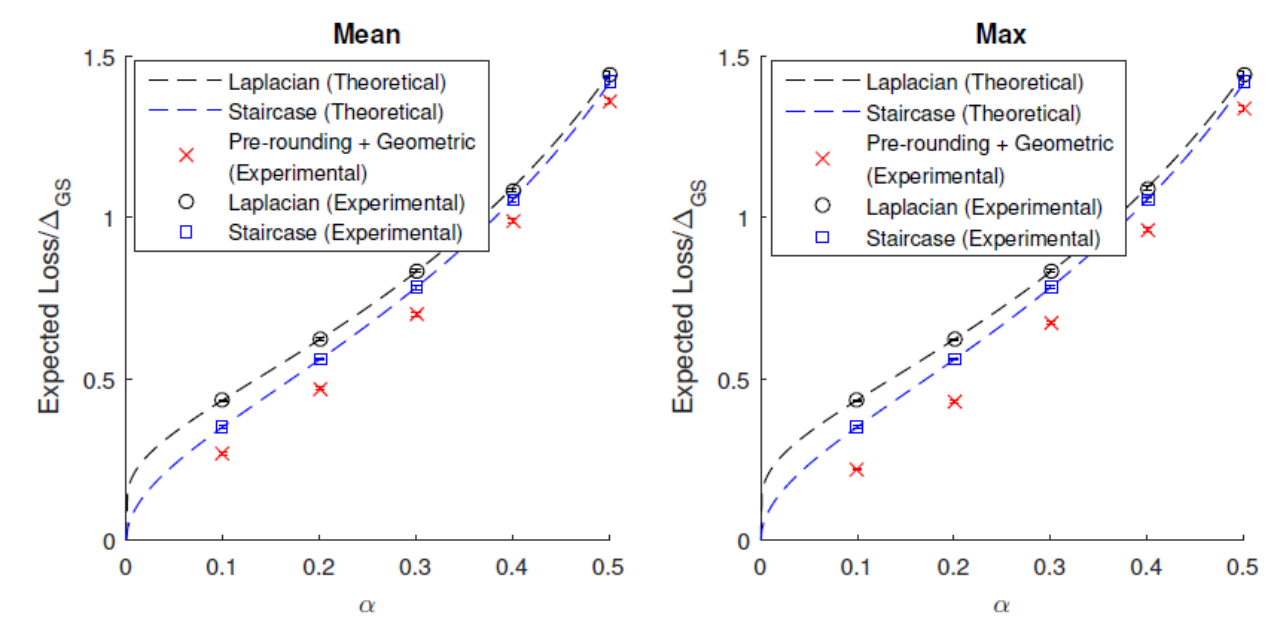
Our Heuristic DP Mechanism



- (1): Pre-rounding
  - only outputs {-10,0,10}
- (2): Add truncated α-Geometric Noise ($\alpha = e^{-\varepsilon}$)
  - $P_r[X>10]$ goes to $P_r[X=10]$
  - $P_r[X<-10]$ goes to $P_r[X=-10]$
- The heuristic mechanism satisfies ε-DP ($\alpha = e^{-\varepsilon}$)
- Mechanism designed for **low-variance** priors

USC Viterbi School of Engineering

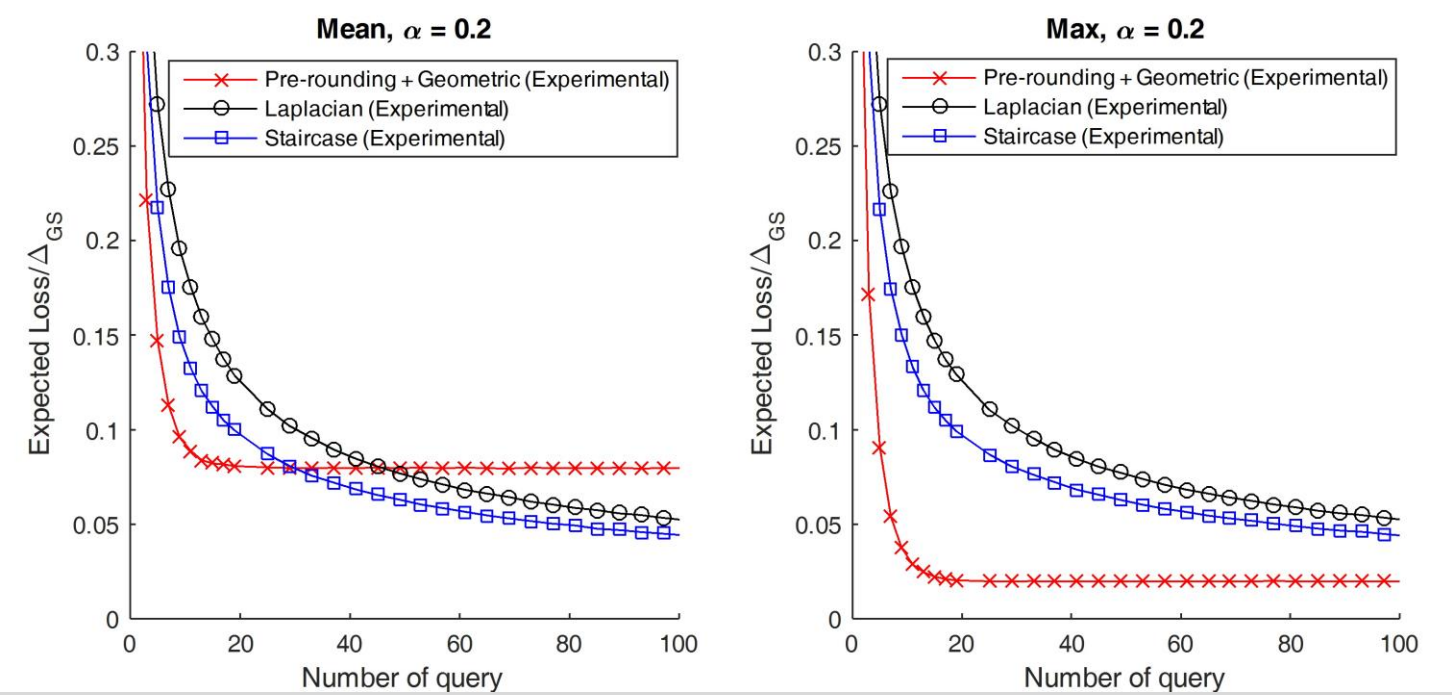
Utility-Privacy Tradeoff Performance



- Significant improvement in low & intermediate privacy regime *(the red 'x')*.
- In the high privacy regime tend towards convergence
  - DP mechanism adds extremely large noise to maintain privacy
  - noise dominates the performance

USC Viterbi School of Engineering

Our Mechanism is Collusion-Proof !



- Users collude in perturbed results (based on MLE)
- The heuristic design is collusion-proof *(the red curve)*

USC Viterbi School of Engineering

Design Insights

When query outputs are substantially narrowed down by side-information, discretizing the domain and adding truncated Geometric noise is a good idea

A robust, simple, and efficient Bayesian design is possible!

A collusion-proof Bayesian design is also feasible

USC Viterbi School of Engineering

Future Directions

Optimal Bayesian design mechanism

- so that we know how good our design is
- new heuristic methods and design insights
- studies of implementation complexity

Applications of the optimal Bayesian design

- applying Bayesian design to practical problems with side-information
- many practical issues will be involved

Optimal Bayesian design in approximate DP

- more efficient, but less robust

Q&A

Thank you!

Email: **chienlun@usc.edu**