Supporting both Range Queries and Frequency Estimation with Local Differential Privacy

IEEE Conference on Communications and Network Security(CNS), June 2019

Xiaolan Gu*, Ming Li*, Yang Cao† and Li Xiong#

* University of Arizona

† Kyoto University

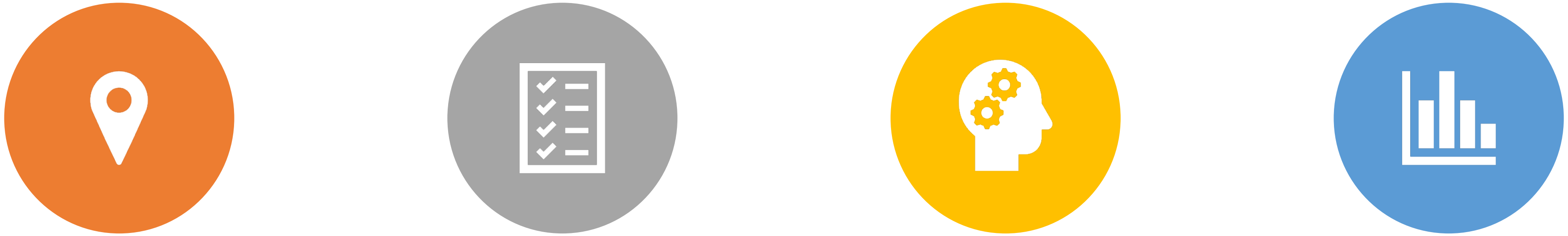
Emory University

Overview

- Background
- Privacy Notions and Mechanisms
- Problem Formulation
- The Proposed Mechanism
- Frequency Estimation Protocol
- Evaluation
- Conclusion

Background

- Companies are collecting our private data to provide better services.



Provide location-based services

Provide recommendations

Learn user preferences patterns

Provide statistical information

- However, private data collection raises privacy concerns.

Privacy Concerns

Moreover, privacy leakage might be occurred even erasing users' identifiers before releasing the data.

Companies need to carefully release users' data for analysis.

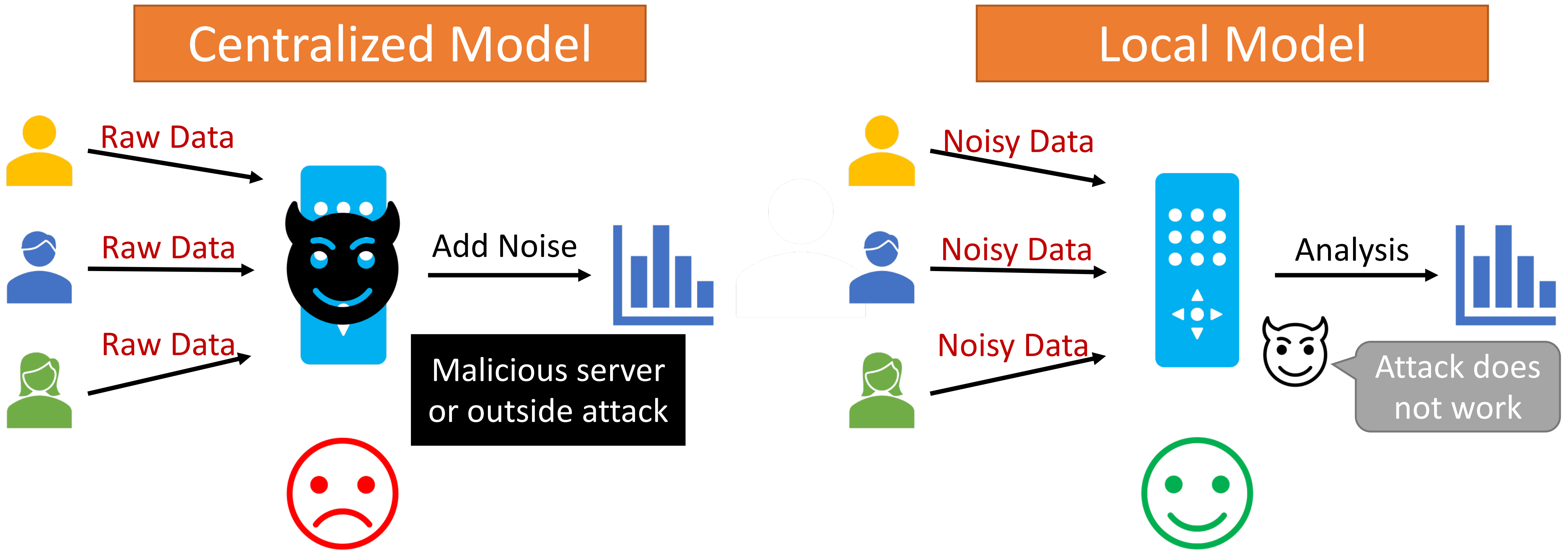
- Privacy lawsuit of Netflix Prize

Source: Wikipedia

Note: These percentages reflect all respondents who, on a scale of 1-5 rated their concern as a 5 (Extremely concerned) or 4 (Very Concerned) with
Source: Consumer Perceptions of Privacy in the Internet of Things, Altimeter Group, 2015 Base: n=2062 respondents

Source: https://jessgroopman.wordpress.com/2015/07/30/how-does-your-business-perform-against-consumers-biggest-privacy-concerns/

Centralized and Local Privacy Models



- The local model is more secure than the centralized one.

Local Differential Privacy (LDP)

A mechanism M satisfies ϵ-LDP iff for any pair of inputs x, x' and any output y

$$\boxed{\frac{\Pr(M(x) = y)}{\Pr(M(x') = y)} \le e^{\epsilon}}$$

- x, x': the raw data (only in user-side);
- y: the perturbed data (can be published and known by adversary).
- ϵ: privacy budget (the smaller ϵ indicates the stronger privacy)

Intuitively, given any output y of a mechanism M, an adversary cannot infer whether the input is x or x' with high confidence (controlled by ϵ).

Mechanisms under LDP [for frequency estimation]

- Randomized Response (RR) [Warner, 1965]: reports the truth with specified probability; (binary answer)

- Example: are you wearing glasses?

Truth	Response	
1	1	w.p. p
	0	w.p. $1-p$
0	1	w.p. $1-p$
	0	w.p. p

Expectation of observed frequency:

$$E[f] = f^*p + (1-f^*)(1-p)$$
$$= (2p-1)f^* + (1-p)$$

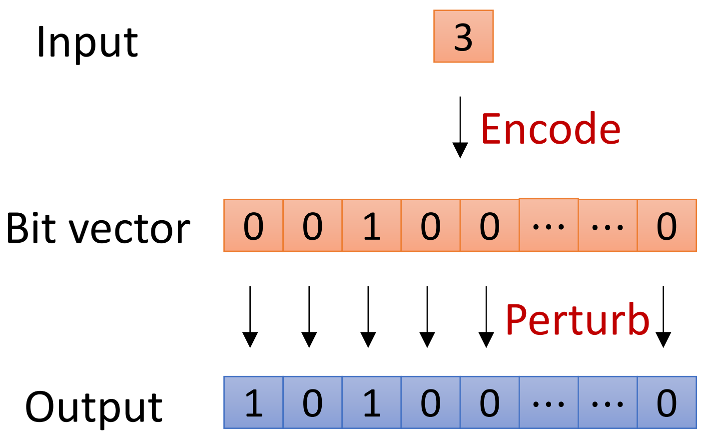
Calibration: $\hat{f} = \frac{f-(1-p)}{2p-1}$

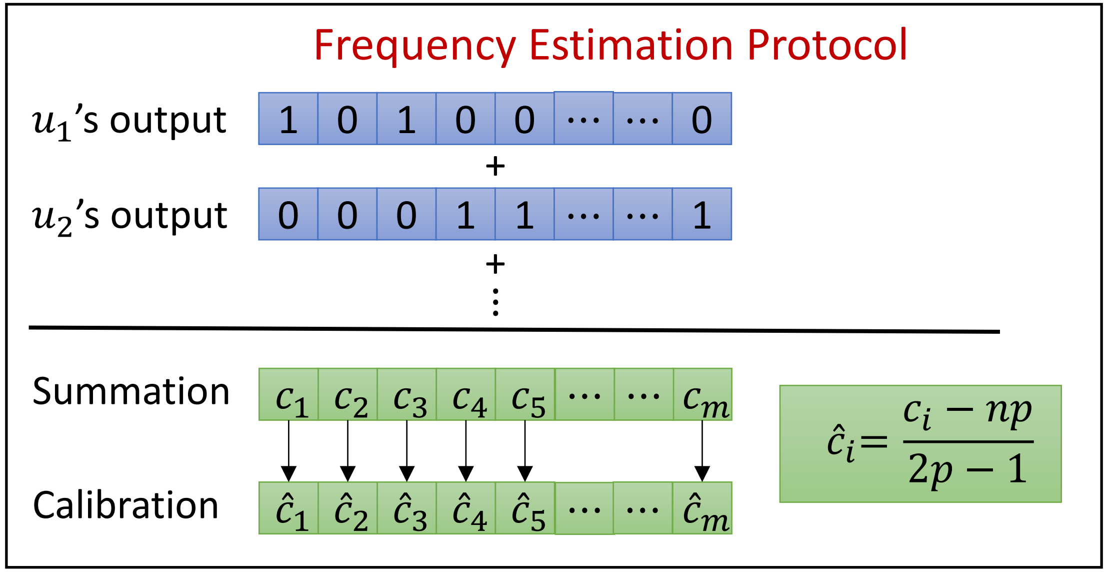
Unbiased estimator: $E[\hat{f}] = f^*$

Extend RR for General Cases

- RAPPOR [CCS'14]: encodes the input into a bit vector and flips each bit with specified probability;

- Example: what's your favorite color ? (among 10 options)





Motivation

Answer one type of queries	Answer the both

- Individual queries
  - Typing-words recommendation
  - Nearest restaurant
- Aggregate queries
  - The most popular emojis
  - Population distribution

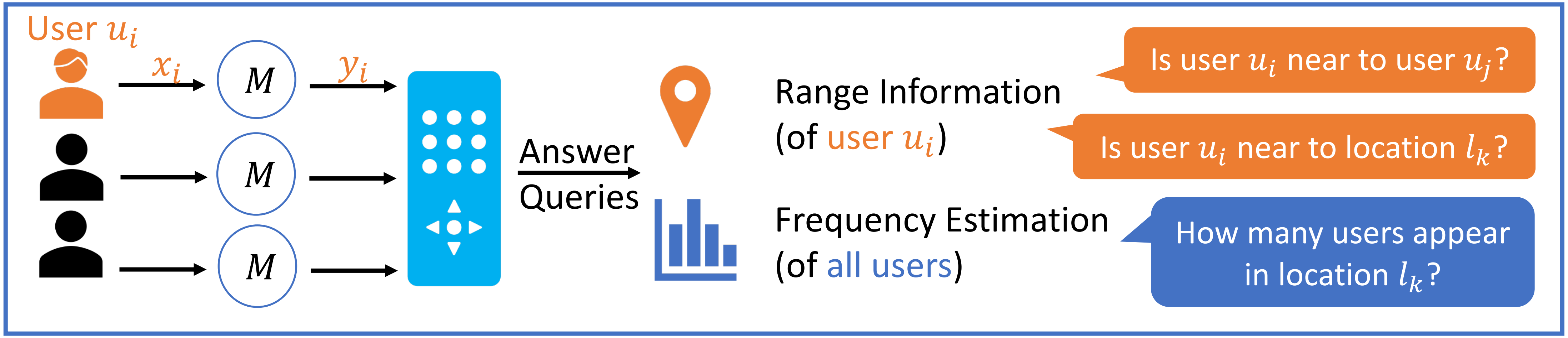
- Social relationship study
  - Co-location of two users
  - Location frequency
- Movie recommendation
  - User's history record
  - Popular movies

Existing mechanisms only support one type of queries with high utility

- Planar Laplace: range queries (individual);
- RR, RAPPOR, and OUE: frequency estimation (aggregate).

Problem Formulation

- The input/output domain contains m items, indexed by $I = \{1,2,\cdots,m\}$;
- There are n users, where each user u_i independently perturbs her raw data x_i into y_i and uploads y_i to the server;
- The server answers two types of queries: 1. range query such as POI (points of interest) search; 2. frequency estimation (aggregate information).



Privacy Notions: LDP and local d-privacy

- Problem of LDP notion: it does not consider the distance between items (thus it will lead to bad utility of range query).
- Solution: local d-privacy (a variant of LDP with distance metric).

$$\boxed{\frac{\Pr(M(x) = y)}{\Pr(M(x') = y)} \leq e^{\epsilon \cdot d(x, x')}}$$

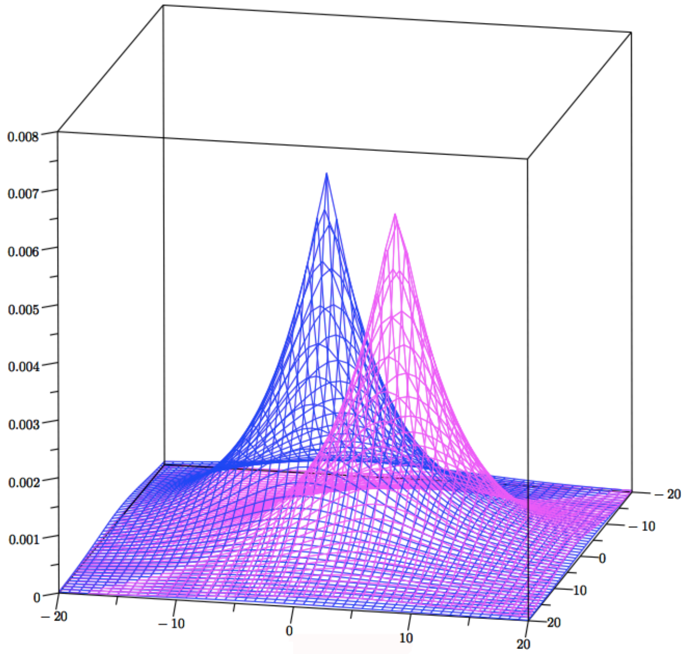
where $d(\cdot,\cdot)$ can be any distance metric with triangle inequality, i.e., $d(x, x') + d(x, x'') \geq d(x', x'')$.

The distance metric in local d-privacy relaxes the strong privacy constraint of LDP, thus can provide better utility

Existing Mechanisms under Local d-privacy

- Planar Laplace [CCS' 13]

  - Only under Euclidean distance (i.e., the notion of geo-indistinguishability)
  - Not designed for frequency estimation



- Optimization-based mechanism

[ICDCIT 2015]

Idea: obtain the optimal perturbation probability matrix by minimizing the linear objective function with required privacy constraints.

Problem: 1. the dimension issue of solving the optimization problem; 2. for frequency estimation, the theoretical utility depends on true frequencies with non-linear correlation.

Problem of Existing Mechanisms

- Randomized Response (RR) based mechanisms (under LDP) do not consider distance; bad utility on range query
- Planar Laplace Mechanism (under local d-privacy) does not consider aggregate information; bad utility on frequency estimation
- In optimization-based mechanism, the objective function (frequency estimation) is hard to evaluate and solving the optimization problem takes high computation cost. impractical

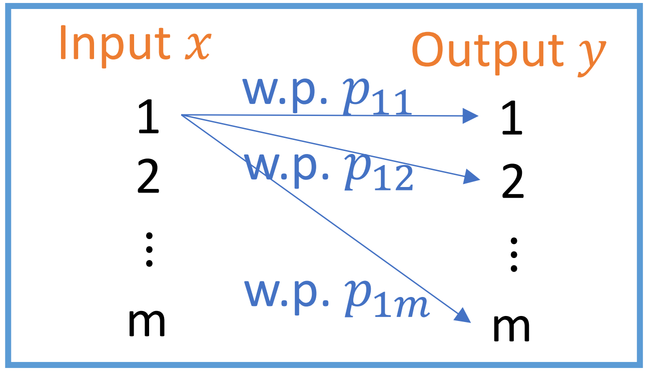
Our solution: combine the idea of RR and Planar Laplace.

Our Mechanism

Privacy constraint of local d-privacy:

$$p_{ik}/p_{jk} \leq e^{\epsilon d_{ij}} \ (\forall i, j, k)$$

where $p_{ik} = \Pr(y = k | x = i)$



- Let $p_{jk} = e^{-\epsilon d_{jk}} \cdot p_{kk}$ for $j \neq k$. (intuition: make p_{jk} as small as possible)
- Let $\sum_{k=1}^{m} p_{jk} = 1$ for all j. (the summation of probabilities from the same input should be 1)

Combining the above two equations, we can first compute p_{kk} by solving m-dimensional linear equations, then compute $p_{jk} = e^{-\epsilon d_{jk}} \cdot p_{kk}$.

Properties of Our Mechanism

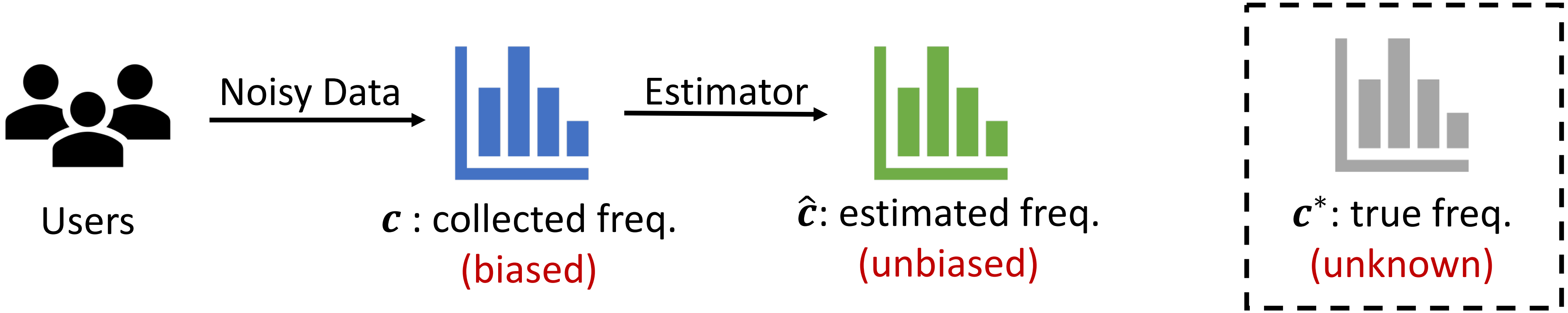
It satisfies local d-privacy (by triangle inequality of distance metric).

$$\frac{p_{ik}}{p_{jk}} = \frac{p_{kk}}{p_{jk}} \cdot \frac{p_{ik}}{p_{kk}} = e^{\epsilon(d_{jk}-d_{ik})} \leq e^{\epsilon d_{ij}}$$

It is the optimal solution when the objective function is $f = \sum_{k=1}^{m} p_{kk}$.

- This objective function corresponds to co-location queries;
- This property is proved by Karush-Kuhn-Tucker (KKT) Conditions.

Frequency Estimation



- Result needs calibration (by estimator) since perturbation is biased;
- The frequency estimation protocol in LDP does not work in our case.

Frequency estimator for our mechanism: $\hat{\boldsymbol{c}} = (\mathrm{P}^T)^{-1}\boldsymbol{c}$, where $\mathrm{E}[\hat{\boldsymbol{c}}] = \boldsymbol{c}^*$.

Theoretical mean square error: MSE=Var[$\hat{\boldsymbol{c}}$] (related to true freq. $\boldsymbol{c}^*$).

Evaluation

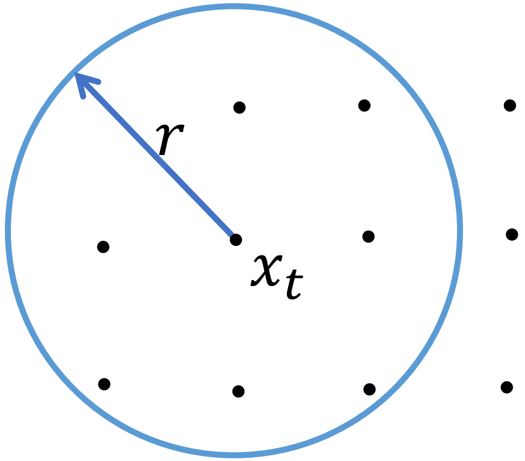
Theoretical

- $\text{Error}_{\text{range}} = \sum_{t=1}^{n}\left(1 - \sum_{y_t \in R(x_t,r)} p_{x_t y_t}\right)$
- $\text{MSE}_{\text{freq}} = \sum_{k=1}^{m}\sum_{j=1}^{m}\left(c_j^* \sum_{i=1}^{m} q_{ki}^2 p_{ji}\right) - n$

Simulation

- $\text{Error}_{\text{range}} = \sum_{t=1}^{n}\left(1 - \mathbf{1}_{R(x_t,r)}(y_t)\right)$
- $\text{MSE}_{\text{freq}} = \sum_{k=1}^{m}(\hat{c}_k - c_k^*)^2$

Range query evaluation:
how accurate of output



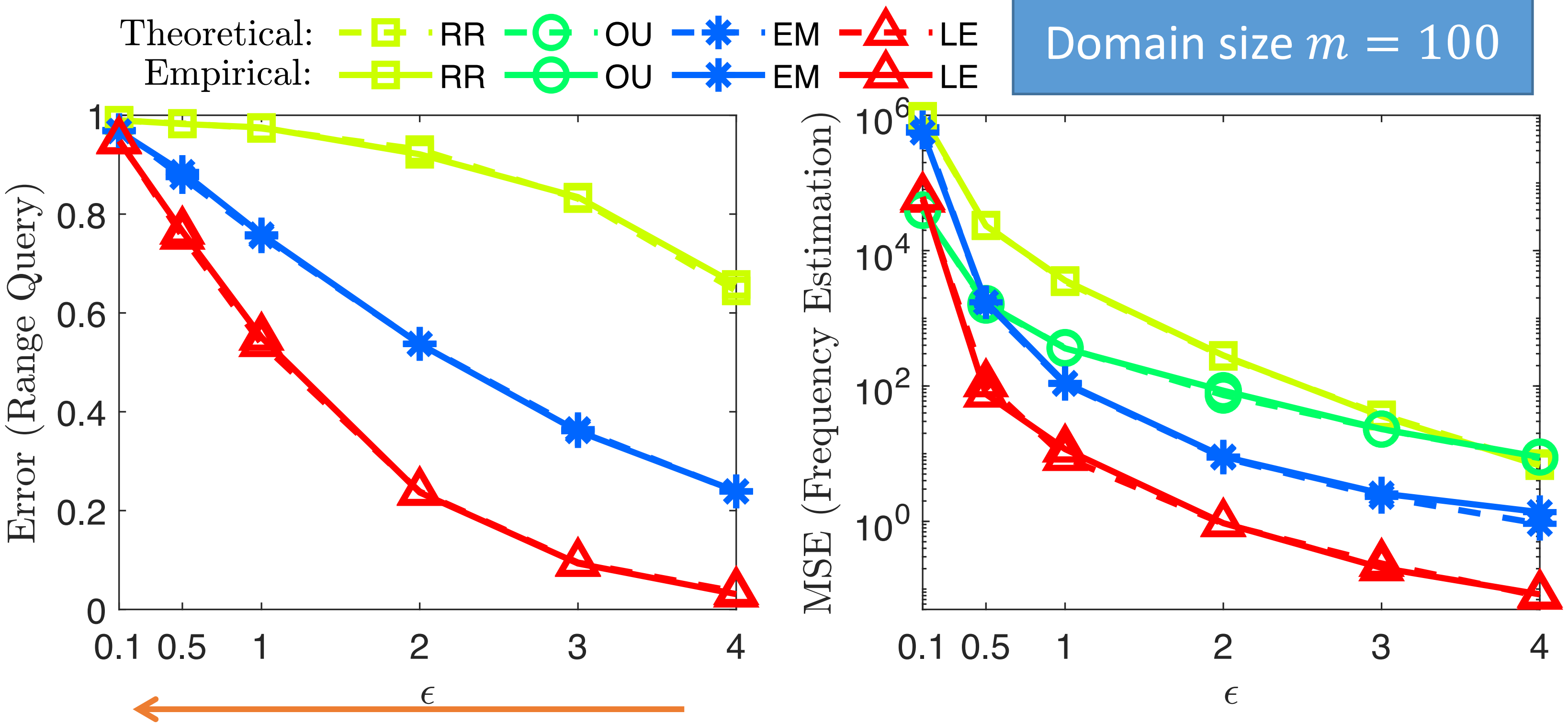
$$R(x_t, r) = \{k \mid k \in I, d(k, x_t) \le r\}$$

$$\mathbf{1}_{R(x_t,r)}(y_t) = \begin{cases} 1, y_t \in R(x_t, r) \\ 0, y_t \notin R(x_t, r) \end{cases}$$

Synthetic Data

Domain size $m = 100$

More accuracy

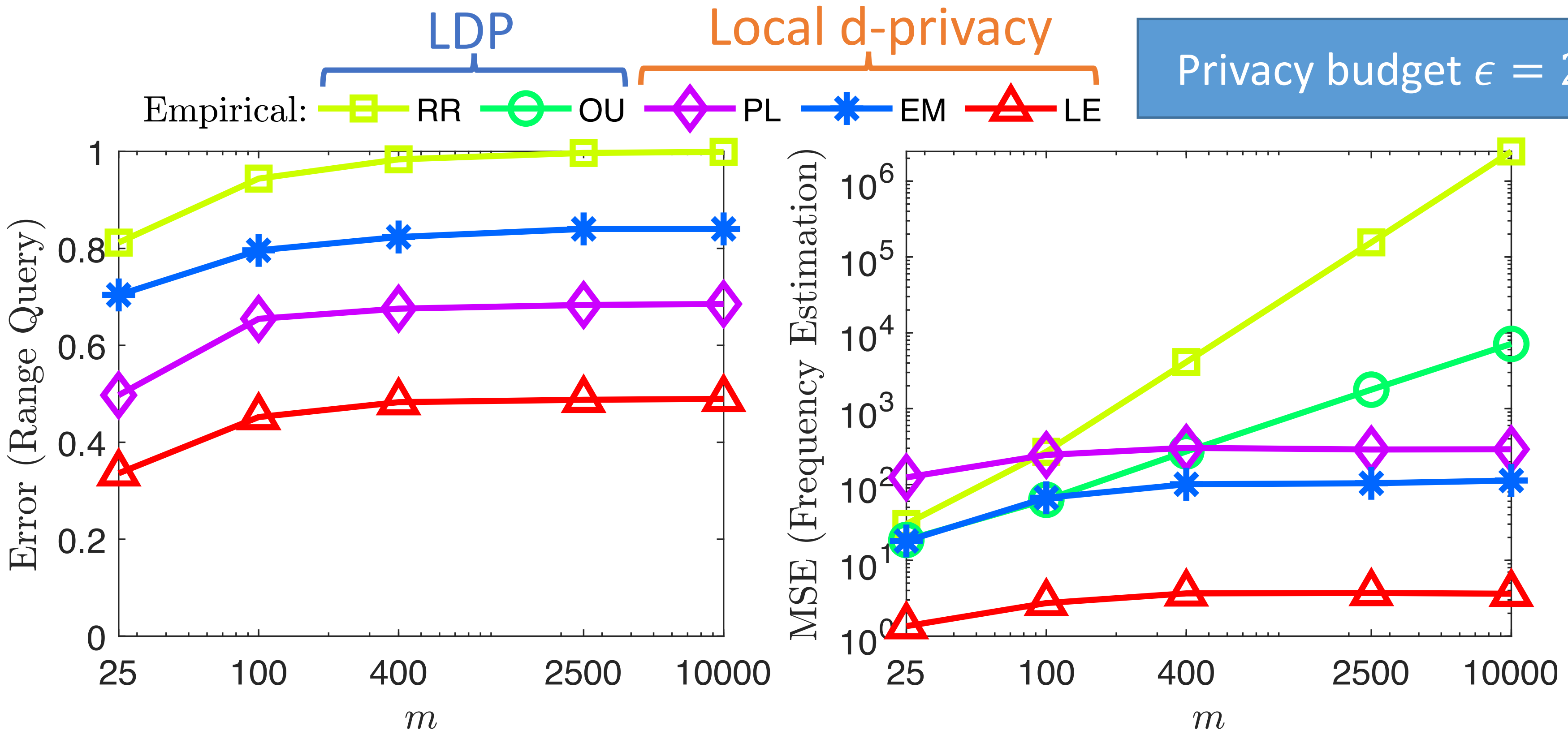


More privacy

- LE: our mechanism
- OU only supports freq. estimation

Observations: 1. The theoretical analysis is effective;
2. Our mechanism (LE) outperforms other ones.

Influence of Domain Size



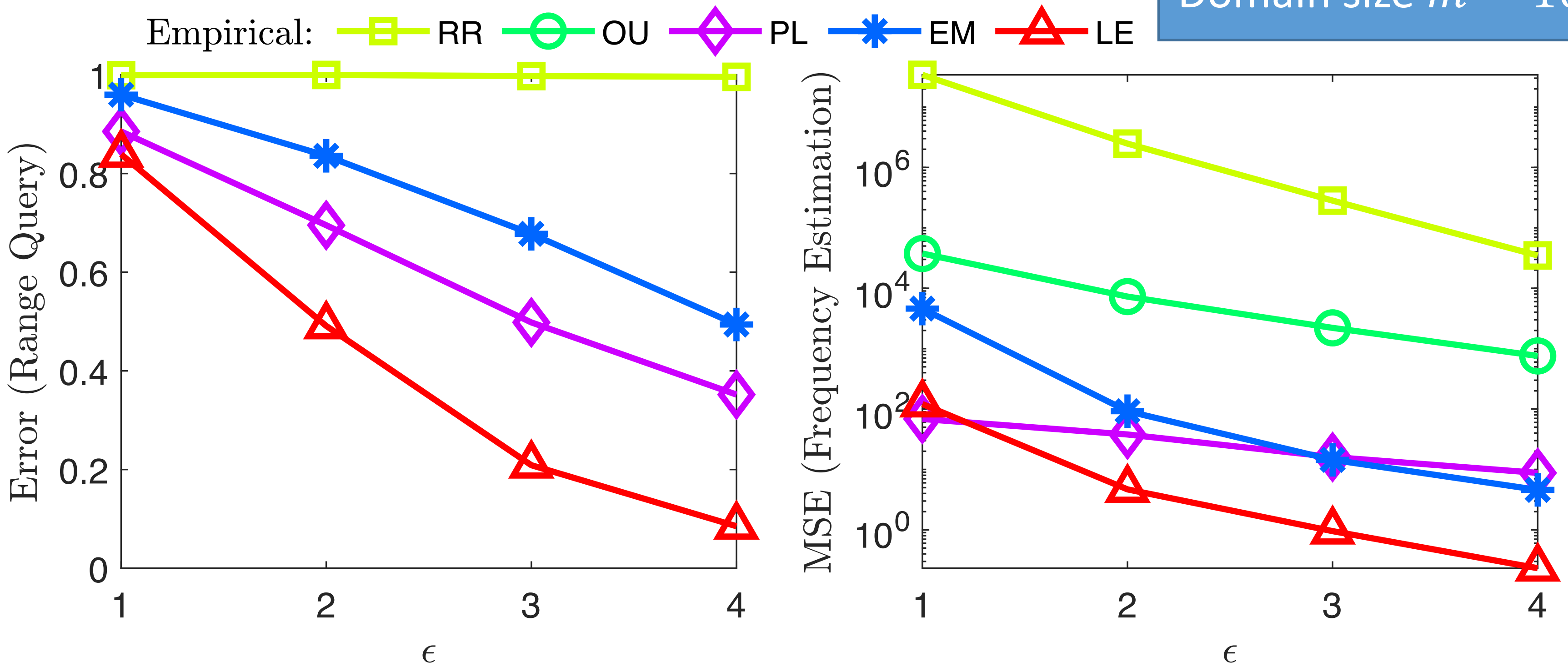
Privacy budget $\epsilon = 2$

More experiments (on estimator and range size) can be found in our paper.

Observations: 1. For LDP mechanisms, $\mathrm{MSE}_{\mathrm{freq}}$ is proportional to domain size;
2. For local d-privacy mechanisms, domain size has little influence.

Real-world Location Data [Gowalla]

Domain size $m = 10{,}000$



Similar observations!

Conclusion

- We tackle the problem of supporting both range query and frequency estimation with high utility at the the same time;
- We adopt the notion of local d-privacy instead of LDP to obtain better utility and design our mechanism by combining the idea of randomized response mechanism and Planar Laplace;
- Our mechanism only needs to solve a linear equation which has less computation complexity than optimization-based mechanism.
- Simulation results show the effectiveness of our mechanism and the advantage of local d-privacy (the notion with distance metric)

Future work: support complex data types (e.g., set-valued data) and complex analysis tasks (e.g., frequent items mining).

Thank You!

Questions & Answers