Providing Input-Discriminative Protection for Local Differential Privacy

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

*University of Arizona #Emory University †Kyoto University

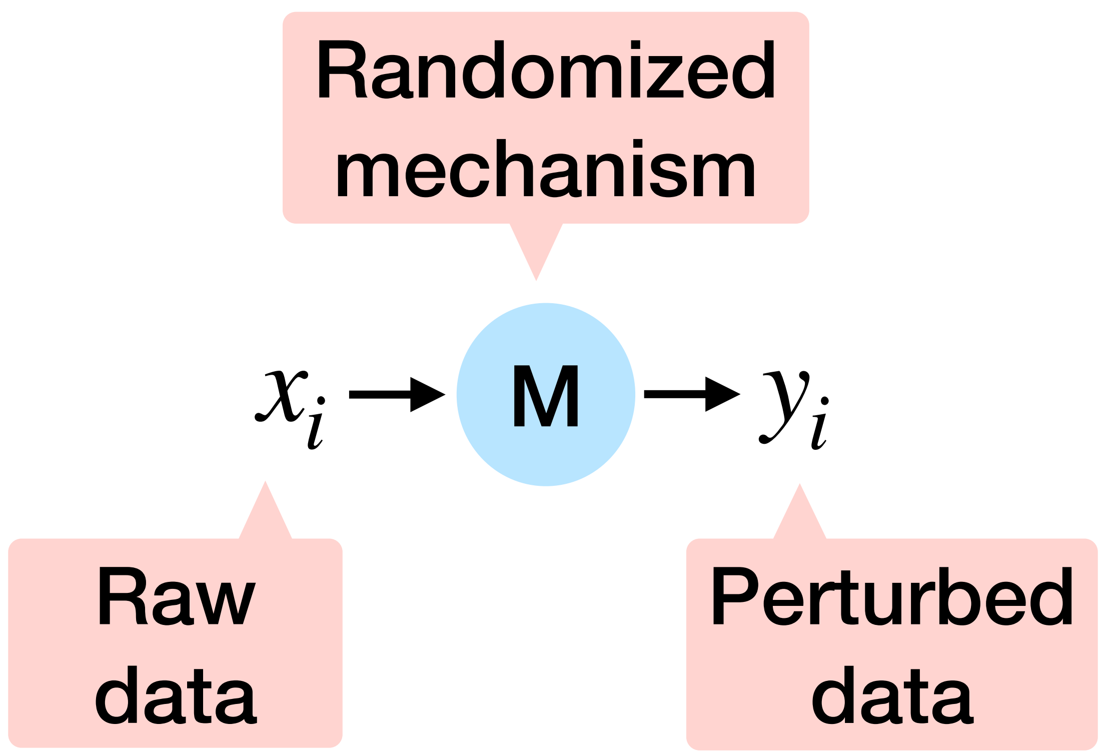
IEEE International Conference on Data Engineering (ICDE), April 2020

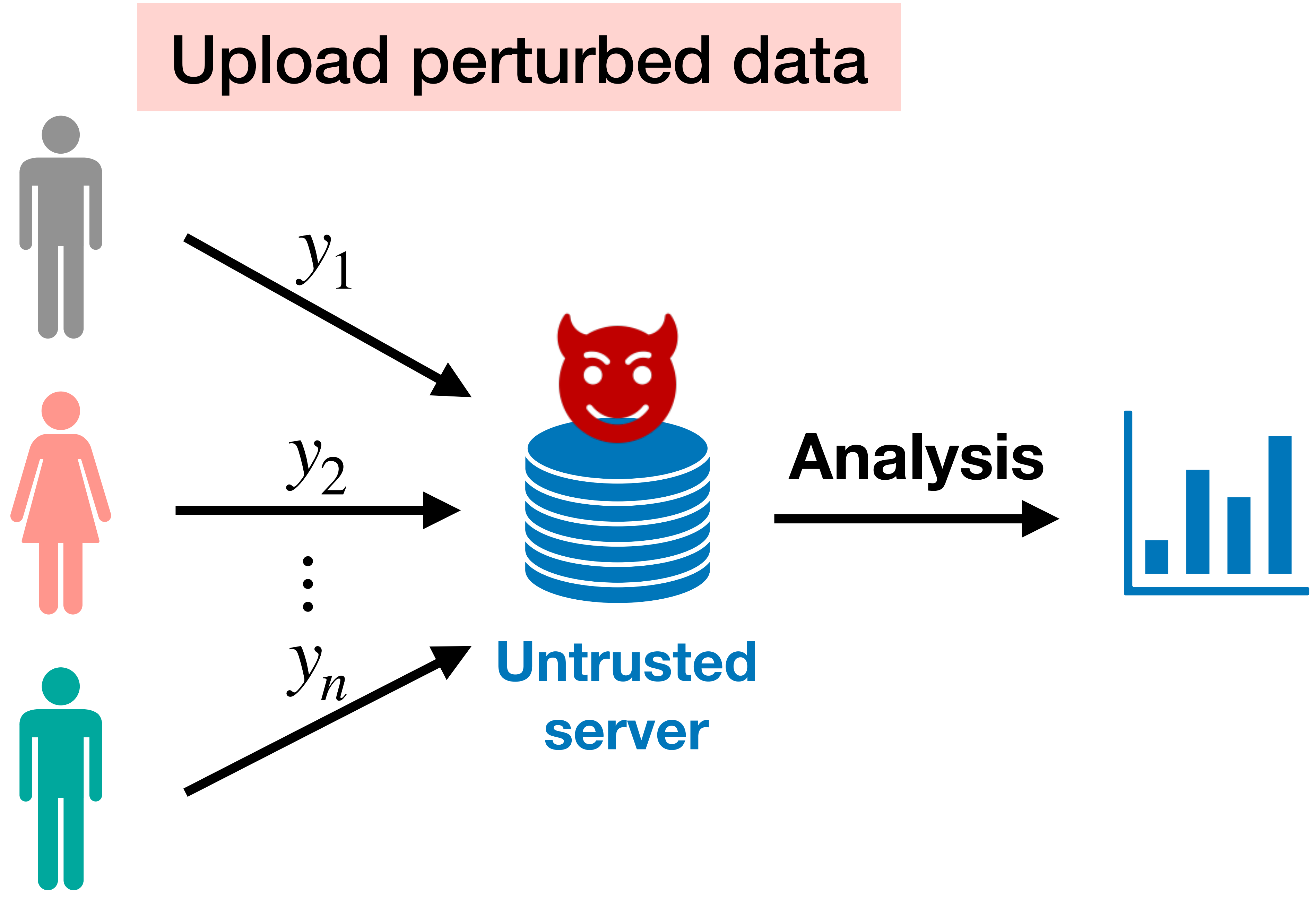
Overview

- Background on LDP
- Our Privacy Notion: ID-LDP
- Our Privacy Mechanism on ID-LDP
- Evaluation
- Conclusion

Background

- Companies are collecting our private data to provide better services (Google, Facebook, Apple, Yahoo, Uber, …)
- However, privacy concerns arise
  - Yahoo: massive data breaches impacted 3 billion user account, 2013
  - Facebook: 267 million users' data has reportedly been leaked, 2019
  - …
- Possible solution: locally private data collection model





Local Differential Privacy (LDP) [Duchi et al, FOCS' 13]

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

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

- x, x' : the possible input (raw) data (generated by the user)
- y : the output (perturbed) data (public and known by adversary)
- ϵ : privacy budget (a smaller ϵ indicates stronger privacy)

An adversary cannot infer whether the input is x or x' with high confidence (controlled by ϵ)

Applications of LDP

Google Developers

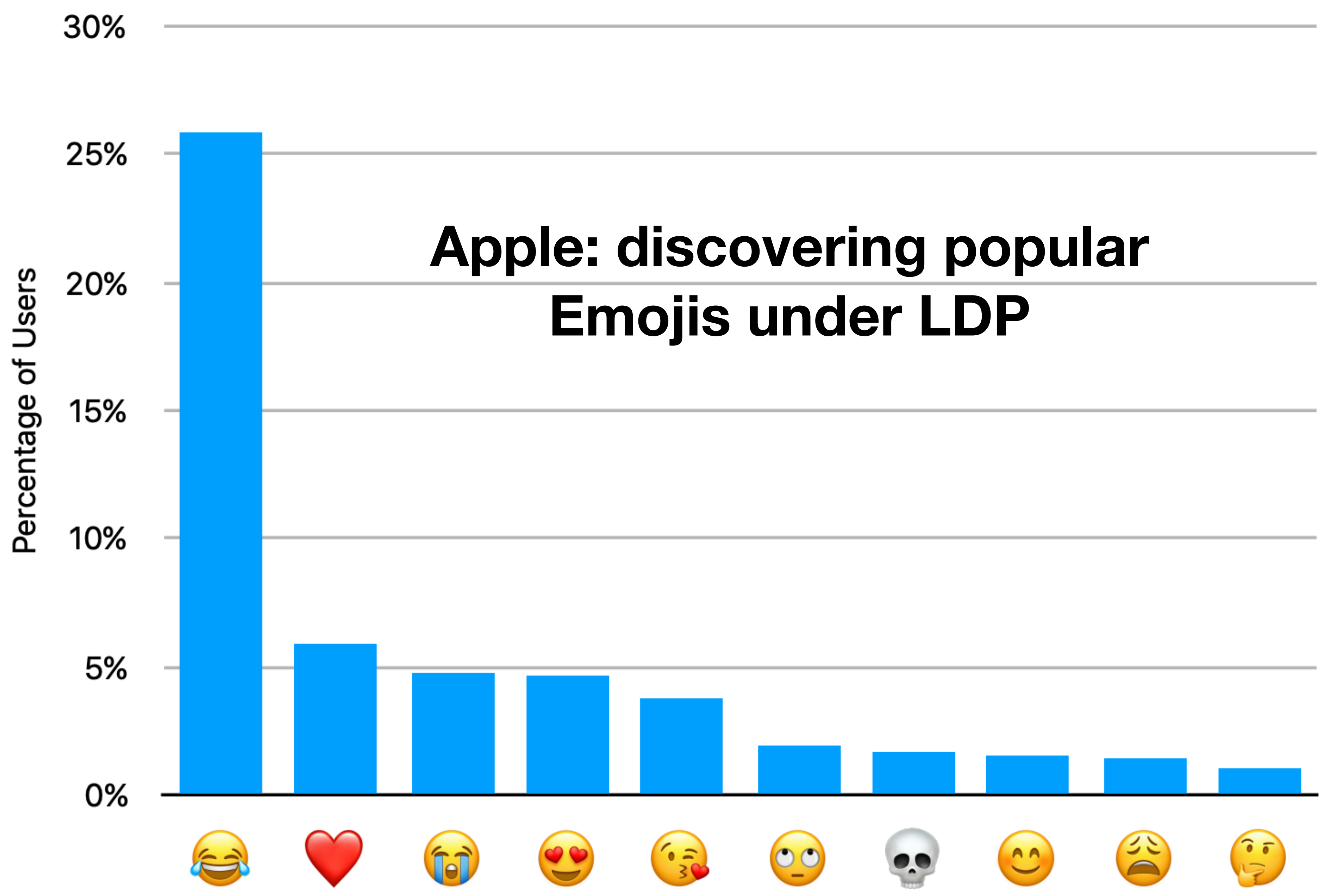
Blog of our latest news, updates, and stories for developers

Enabling developers and organizations to use differential privacy

Thursday, September 5, 2019

Posted by Miguel Guevara, Product Manager, Privacy and Data Protection Office

Source:
https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html



Source:
https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html

Limitations of LDP

- LDP notion requires the same privacy budget for all pairs of possible inputs
- Existing LDP protocols perturb the data in the same way for all inputs
- However, in many practical scenarios, different inputs have different degrees of sensitiveness, thus require distinct levels of privacy protection.

Scenarios	High sensitiveness	Low sensitiveness
Website-click records	Politics-related	Facebook and Amazon
Medical records	HIV and cancer	Anemia and headache

- LDP protocols can provide excessive protection for some inputs that do not need such strong privacy (leading to an inferior privacy-utility tradeoff)

Our Privacy Notion: Input-Discriminative LDP (ID-LDP)

ϵ_x is the privacy budget of an input x

- Given a privacy budget set $\mathscr{E} = \{\epsilon_x\}_{x \in \mathscr{D}}$, a randomized mechanism M satisfies $\mathscr{E}$-ID-LDP if and only if for any pair of inputs $x, x' \in \mathscr{D}$ and output $y \in Range(M)$

$$\frac{\Pr(M(x) = y)}{\Pr(M(x') = y)} \leqslant e^{r(\epsilon_x, \epsilon_{x'})}$$

$r(\cdot, \cdot)$ is a function of two privacy budgets

- In this paper, we focus on an instantiation called MinID-LDP with $r(\epsilon_x, \epsilon_{x'}) = \min\{\epsilon_x, \epsilon_{x'}\}$

Intuition: for any pair of inputs x, x', MinID-LDP guarantees the adversary's capability of distinguishing them would not exceed the bound controlled by both ϵ_x and $\epsilon_{x'}$ (thus achieving differentiated privacy protection for each pair)

MinID-LDP has Sequential Composition like LDP, which guarantees the overall privacy for a sequence of mechanisms.

Relationships with LDP

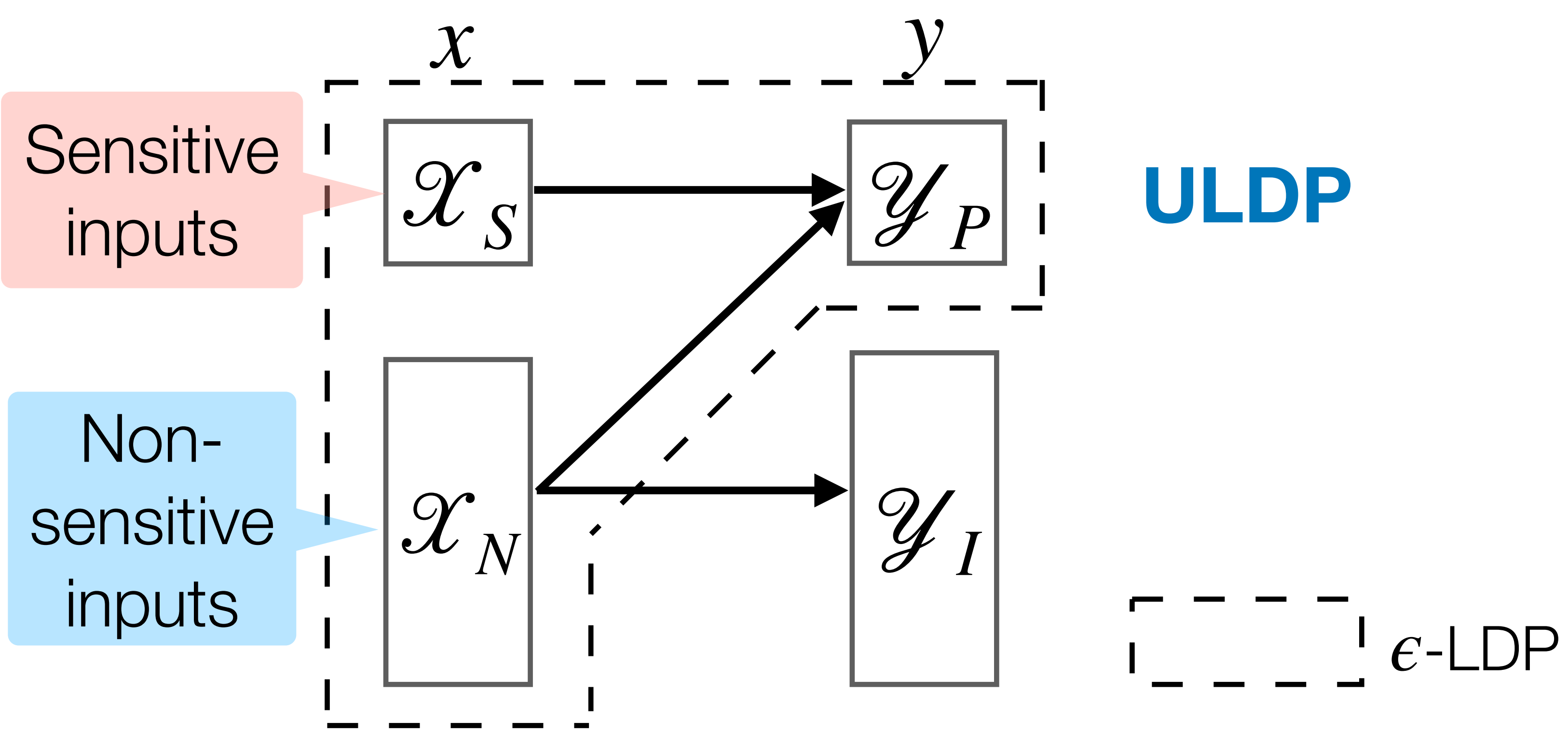
1. If $\epsilon_x = \epsilon$ for all $x \in \mathscr{D}$, then $\mathscr{E}$-MinID-LDP $\Leftrightarrow$ ϵ-LDP
2. If $\min\{\mathscr{E}\} \geqslant \epsilon$, then ϵ-LDP $\Rightarrow$ $\mathscr{E}$-MinID-LDP
3. If $\epsilon \geqslant \min\{\max\{\mathscr{E}\}, 2\min\{\mathscr{E}\}\}$, then $\mathscr{E}$-MinID-LDP $\Rightarrow$ ϵ-LDP

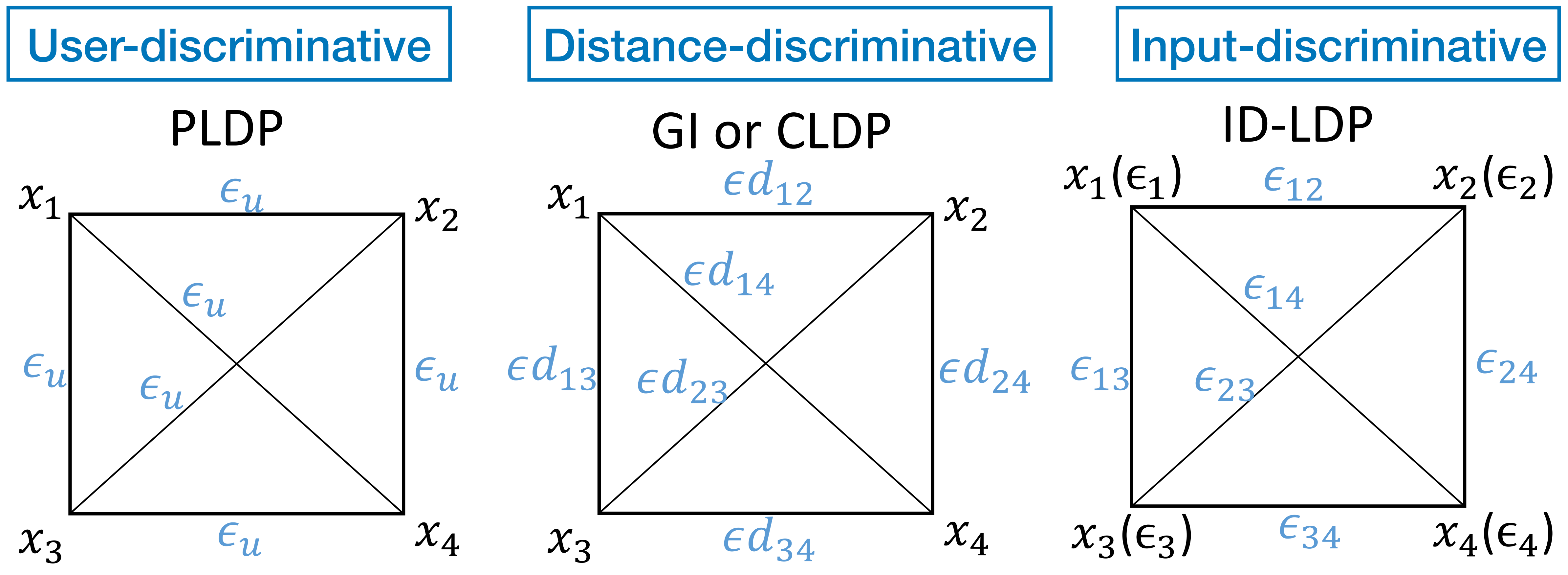
Factor 2 is due to the symmetric property of the indistinguishability definition

MinID-LDP can be regarded as a relaxation compared with LDP. It captures user's **fine-grained privacy requirement**, when LDP is too strong (i.e., provides overprotection).

Related Privacy Notions

- Personalized LDP (PLDP) [Chen et al, ICDE' 16]
- Geo-indistinguishability (GI) [Andres et al, CCS' 13]
- Condensed LDP (CLDP) [Gursoy et al, TDSC' 19]
- Utility-optimized LDP (ULDP)
  [Murakami and Kawamoto, USENIX Security' 19]





ϵ_u: the privacy budget of a user u for all pairs of inputs (different user may have different ϵ_u)

ϵd_{ij}: the privacy budget for a pair of inputs x_i, x_j
d_{ij}: distance between x_i, x_j

ϵ_i: the privacy budget of x_i
ϵ_{ij}: the privacy budget of a pair of inputs x_i, x_j for all users
MinID-LDP: $\epsilon_{ij} = \min\{\epsilon_i, \epsilon_j\}$

Privacy budget of a pair of inputs in several related notions

ULDP does not guarantee the indistinguishability between the sensitive and non-sensitive inputs when observing some outputs, thus ULDP does not guarantee LDP.

Privacy Mechanism Design under ID-LDP

Problem Statement

- Data types: categorical (two cases: each user has only one item or an item-set)
- Analysis Task/Application: frequency estimation (which is the building block for many applications)
- Objectives: minimize MSE of frequency estimation while satisfying ID-LDP

Challenges

ID-LDP protocols perturb inputs with different probabilities

- The number of variables (perturbation parameters) and privacy constraints (to be satisfied for any x, x', y) can be very large (especially for a large domain or item-set data).
- Objective function (MSE) is dependent on the unknown true frequencies;

Example: assume domain size m, then m^2 variables and m^3 constraints

Preliminaries: LDP protocols

- Randomized Response
- Unary Encoding

Our protocol satisfying ID-LDP is based on this

LDP Protocol: Randomized Response

- Randomized Response (RR) [Warner, 1965]: reports the truth with some probability (for binary answer: yes-or-no)

Advanced versions: Unary Encoding, Generalized RR, …

- Example: Is your annual income more than 100k?

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

Frequency of response y

Frequency estimation: $\hat{f} = \dfrac{f-(1-p)}{2p-1}$

Unbiasedness: $\mathbb{E}[\hat{f}] = f^*$

True frequency

To satisfy ϵ-LDP: $p = \dfrac{e^\epsilon}{e^\epsilon + 1}$ **(since** $\dfrac{p}{1-p} = e^\epsilon$ **)**

$\mathbb{E}[f] = f^*p + (1-f^*)(1-p) = (2p-1)f^* + (1-p)$

LDP Protocol: Unary Encoding (UE)

- To handle more general case (domain size is d), UE represents the input/output by multiple bits.
- Step 1. encode the input $x = i$ into vector $\mathbf{x} = [0,\cdots,0,1,0,\cdots,0]$ with length d
- Step 2. perturb each bit independently

By minimizing the approximate MSE of frequency estimation

$\mathbf{x}[k]$	$\mathbf{y}[k]$	RAPPOR [Erlingsson et al, CCS' 14]	OUE [Wang et al, USENIX Security' 17]
1	1	w.p. p	w.p. 0.5
1	0	w.p. $1-p$	w.p. 0.5
0	1	w.p. $1-p$	w.p. q
0	0	w.p. p	w.p. $1-q$

To satisfy ϵ-LDP:

$$p = \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}, \quad q = \frac{1}{e^{\epsilon}+1}$$

Overview of Our Protocol for ID-LDP

Recall the two challenges:

1) High complexity of the optimization problem.
2) MSE depends on unknown true frequencies.

For single-item data: IDUE (Input-Discriminative Unary Encoding)

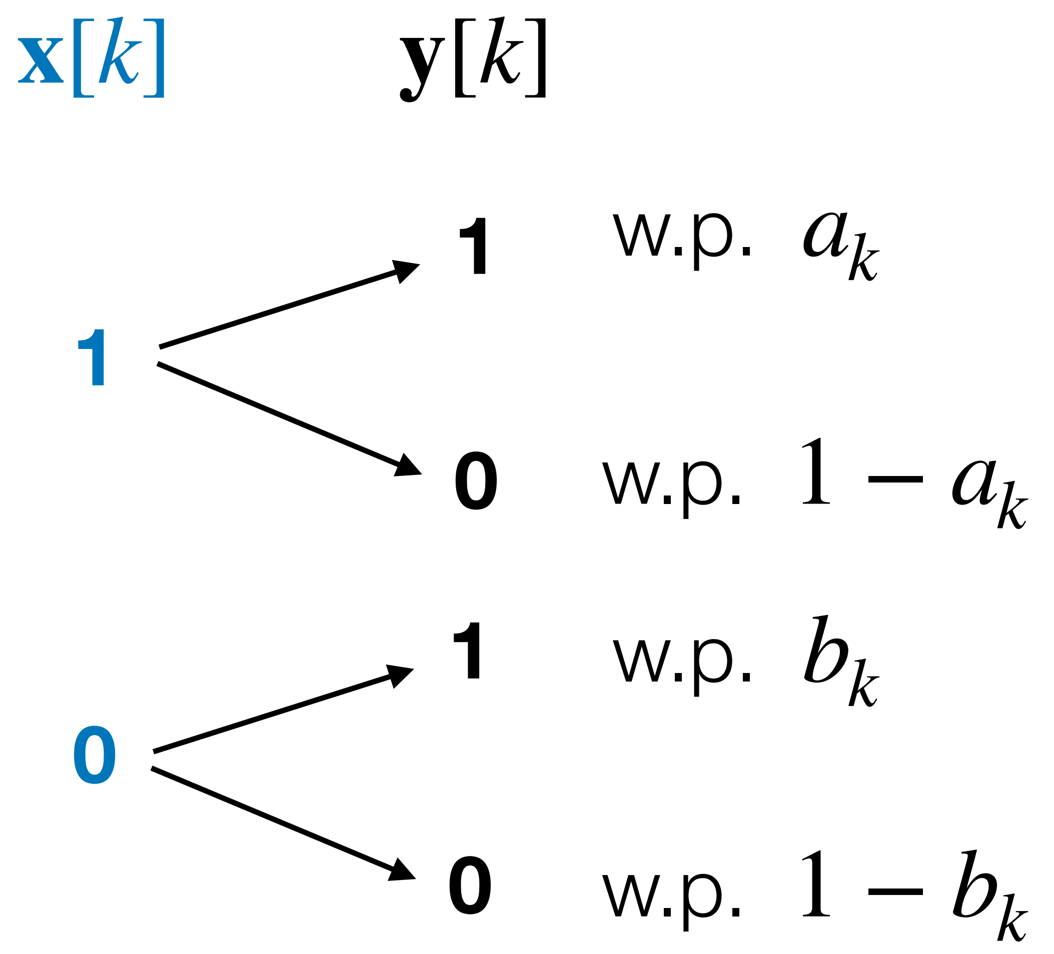
1. We propose Unary Encoding based protocol with only $2m$ variables and m^2 constraints
2. We address the second challenge by developing three variants of optimization models (some models can further reduce the problem complexity)

For item-set data: IDUE-PS (with Padding-and-Sampling protocol)

1. We extend IDUE for item-set data (by combining with a sampling protocol) to solve the scalability issue
2. We show IDUE-PS also satisfies MinID-LDP (if the base protocol IDUE satisfies MinID-LDP)

Privacy Mechanism for Single-Item Data

- Step 1, encode the input $x = i$ into $\mathbf{x} = [0,\cdots,0,1,0,\cdots,0]$
- Step 2, perturb each bit independently (with different probabilities)
- Step 3, estimate frequency/counting by $\hat{c}_i = \dfrac{\sum_u \mathbf{y}_u[i] - nb_i}{a_i - b_i}$



n — number of users
a_i, b_i — perturbation probabilities
c_i^* — true frequency
$\hat{c}_i$ — estimated frequency

$$\text{MSE}_{\hat{c}_i} = \text{Var}[\hat{c}_i] = \frac{nb_i(1-b_j)}{(a_i - b_i)^2} + \frac{c_i^*(1 - a_i - b_i)}{a_i - b_i}$$

$$\frac{a_i(1-b_j)}{b_i(1-a_j)} \leqslant e^{r(\epsilon_i, \epsilon_j)} \ (\forall i, j)$$

Benefits

1. The optimization problem only has $2m$ variables and m^2 constraints
2. The frequency estimator is unbiased, and its MSE can be composed by two terms, where only the second term is dependent on the true frequencies c_i^*

Comparison with LDP Protocols

Example: a health organization is taking a survey which asks n participants to return a response perturbed from categories {HIV, anemia, headache, stomachache, toothache}, where HIV ($i = 1$) is more sensitive, thus we set different privacy budgets, such as $\epsilon_1 = \ln 4$ and $\epsilon_i = \ln 6$ $(i = 2,\cdots,5)$.

TABLE I: Utility comparison in the toy example, where $\epsilon_1 = \ln 4$ and $\epsilon_i = \ln 6$ $(i \neq 1)$.

Mechanisms	Privacy Notions	Probability of flipping the i-th bit				Variance of frequency estimation		**Total variance**
		$1 - a_i$ (if $\mathbf{x}[i] = 1$)		b_i (if $\mathbf{x}[i] = 0$)		$\mathrm{Var}[\hat{c}_i]$		$\sum_i \mathrm{Var}[\hat{c}_i]$
		$i = 1$	$i = 2 \sim 5$	$i = 1$	$i = 2 \sim 5$	$i = 1$	$i = 2 \sim 5$	
RAPPOR [4]	LDP	0.33	0.33	0.33	0.33	$2n$	$2n$	$10n$
OUE [6]	LDP	0.5	0.5	0.2	0.2	$1.78n + c_i$	$1.78n + c_i$	$9.9n$
IDUE	MinID-LDP	0.41	0.33	0.33	0.28	$3.27n + 0.31c_i$	$1.32n + 0.13c_i$	$8.68n \sim 8.86n$

More perturbation noise for $i = 1$

Less perturbation noise for $i \neq 1$

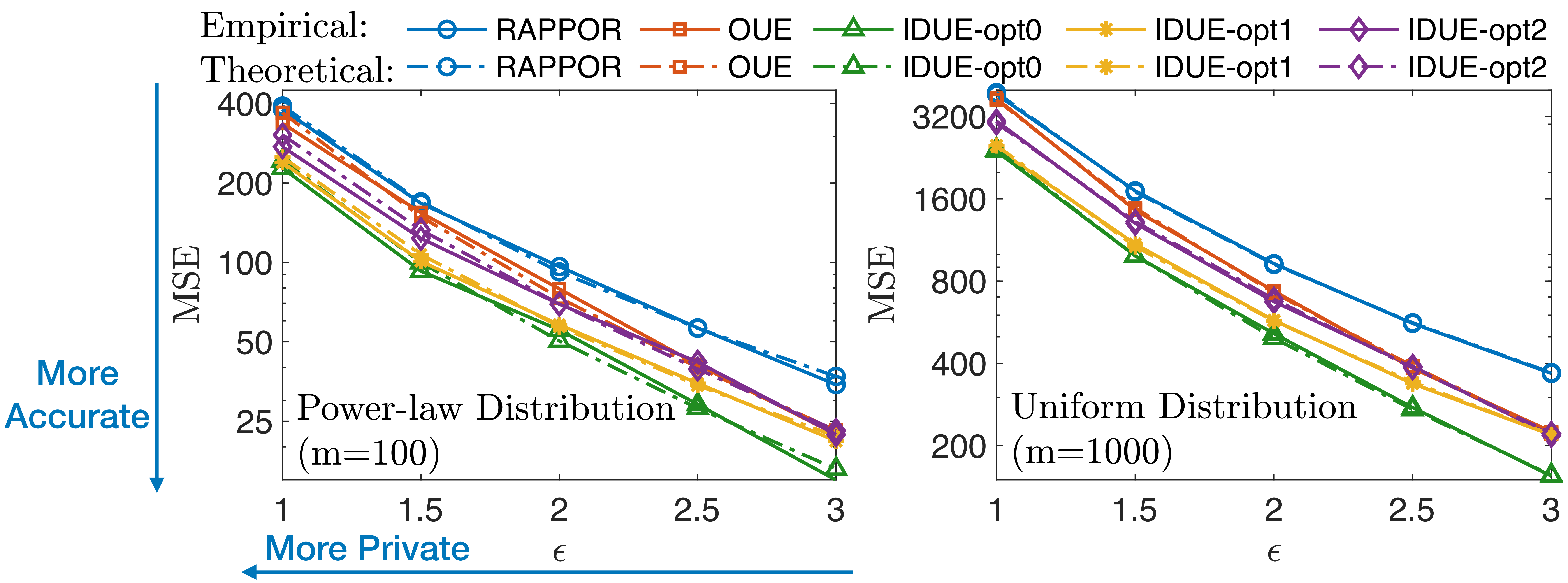
The total variance of IDUE is in a range because it depends on the distribution of true input data, and the upper bound is still less than that of RAPPOR and OUE.

Evaluation

TABLE II: Synthetic and Real-world Datasets

Datasets	# Records	# Users (n)	# Items (m)
Power-law	100,000	100,000	100
Uniform	100,000	100,000	1,000
Retail [27]	908,576	88,162	16,470
Kosarak [27]	8,019,015	990,002	41,270
Clothing [28]	192,544	105,508	5,850

We compare the frequency estimation results of our mechanisms (IDUE and IDUE-PS) with RAPPOR and OUE using two synthetic datasets and three real-world datasets.



Comparison of Empirical (dashed lines) and Theoretical(solid lines) results of synthetic data (single-item input).

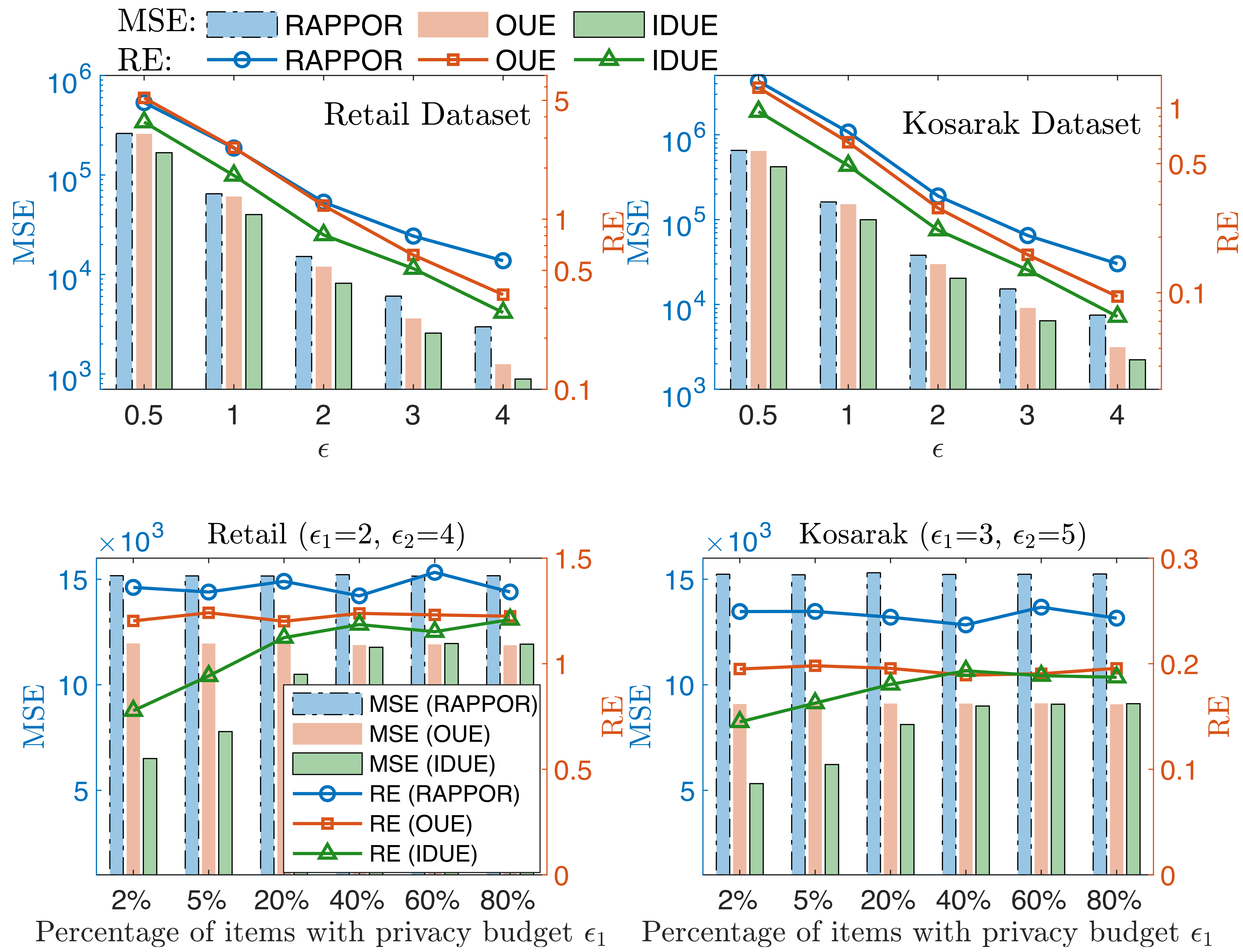
Empirical results are very close to theoretical results

IDUE has smaller MSE than RAPPOR and OUE

opt0: has the smallest MSE

opt1 and opt2: not good as opt0, but better than RAPPOR and OUE

Real-World Data (Single-Item)



IDUE has smallest MSE and RE (relative error)

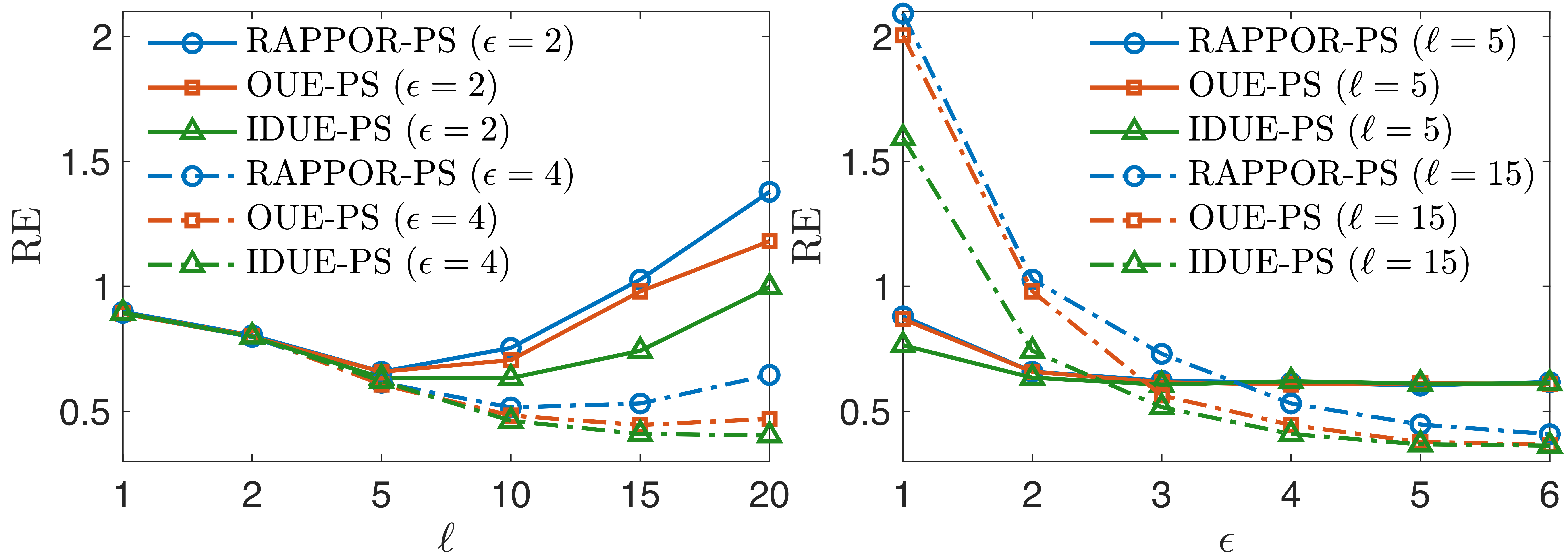
$$\textbf{RE} = \frac{1}{|S|}\sum_{i\in S}\frac{|\hat{c}_i - c_i^*|}{c_i^*}$$

If only small portion of inputs are more sensitive (i.e., have the smallest privacy budget), then IDUE has smaller estimation error.

Otherwise, IDUE has similar performance compared with OUE

Item-Set Data

Varying ℓ and ϵ in Kosarak item-set data



The optimal ℓ (parameter of Padding-and-Sampling protocol) depends on both data distribution and privacy budget (the original paper only mentioned data-dependent). We leave this as our further work.

Conclusion

1. Privacy notion ID-LDP provides input-discriminative protection in the local setting
2. Its instantiation MinID-LDP is a fine-grained version of LDP
3. The proposed protocol IDUE outperforms LDP protocols
4. The advanced version IDUE-PS solves the scalability problem for item-set data

Future work:

- Extend our work to handle more complex data types and analysis tasks;
- Study the strategy of finding the optimal ℓ based on the data distribution and privacy budget.

Thanks for your attention !

Q&A