

On the Utility Gain of Iterative Bayesian Update for Locally Differentially Private Mechanisms

Héber H. Arcolezi, Selene Cerna, and Catuscia Palamidessi

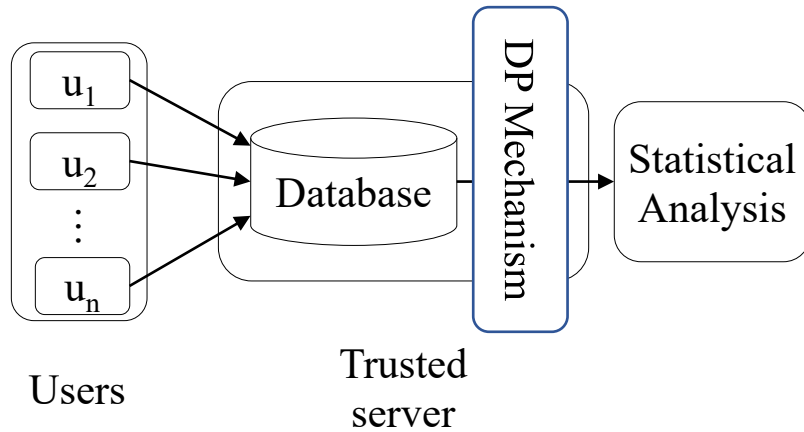
Inria and École Polytechnique (IPP), Palaiseau, France

`{heber.hwang-arcolezi,selene-leya.cerna-nahuis,catuscia.palamidessi}@inria.fr`

DBSec, July 20th, 2023

Context

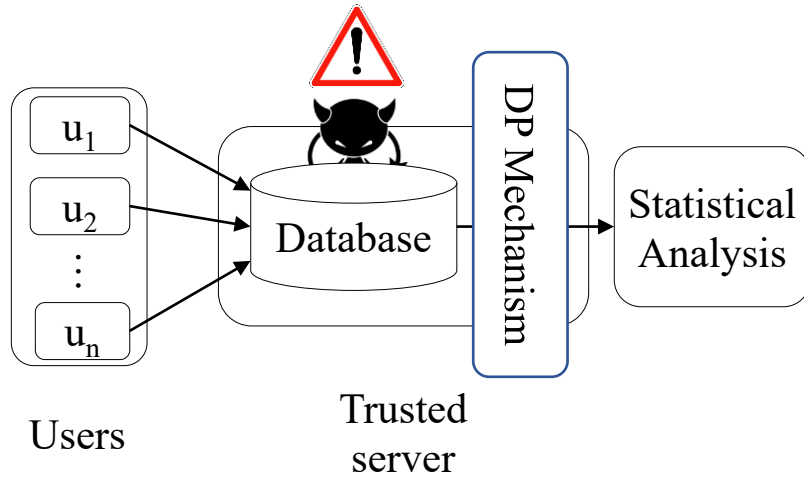
Differential Privacy (DP) [Dwork et al, 2006]



Centralized DP:

- ✓ High utility.
- ✗ Need to trust the server.

Differential Privacy (DP) [Dwork et al, 2006]



Centralized DP:



High utility.

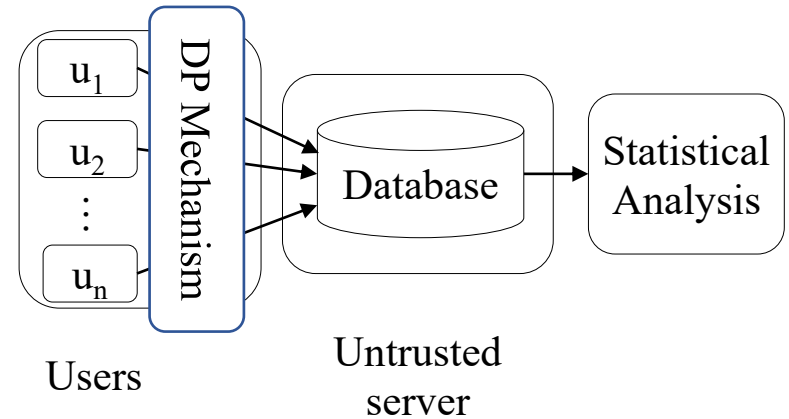
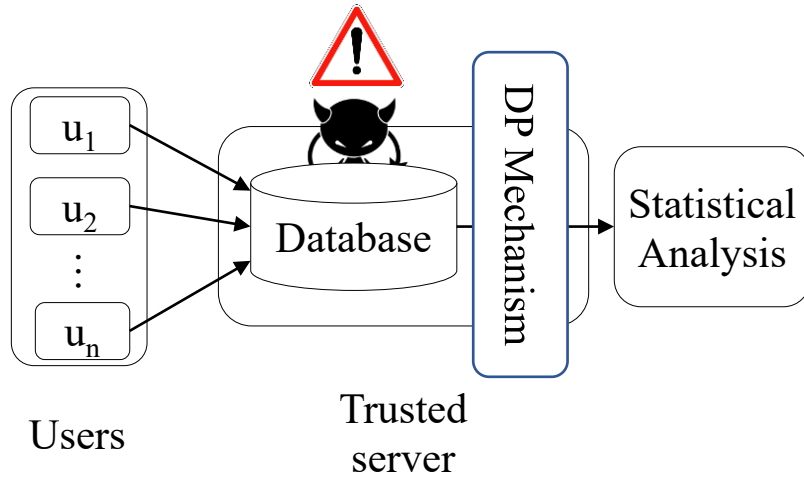


Need to trust the server.



Data breaches, data misuse, etc.

Differential Privacy (DP) [Dwork et al, 2006; Duchi et al, 2013]



Centralized DP:

- ✓ High utility.
- ✗ Need to trust the server.
- ✗✗ **Data breaches, data misuse, etc.**

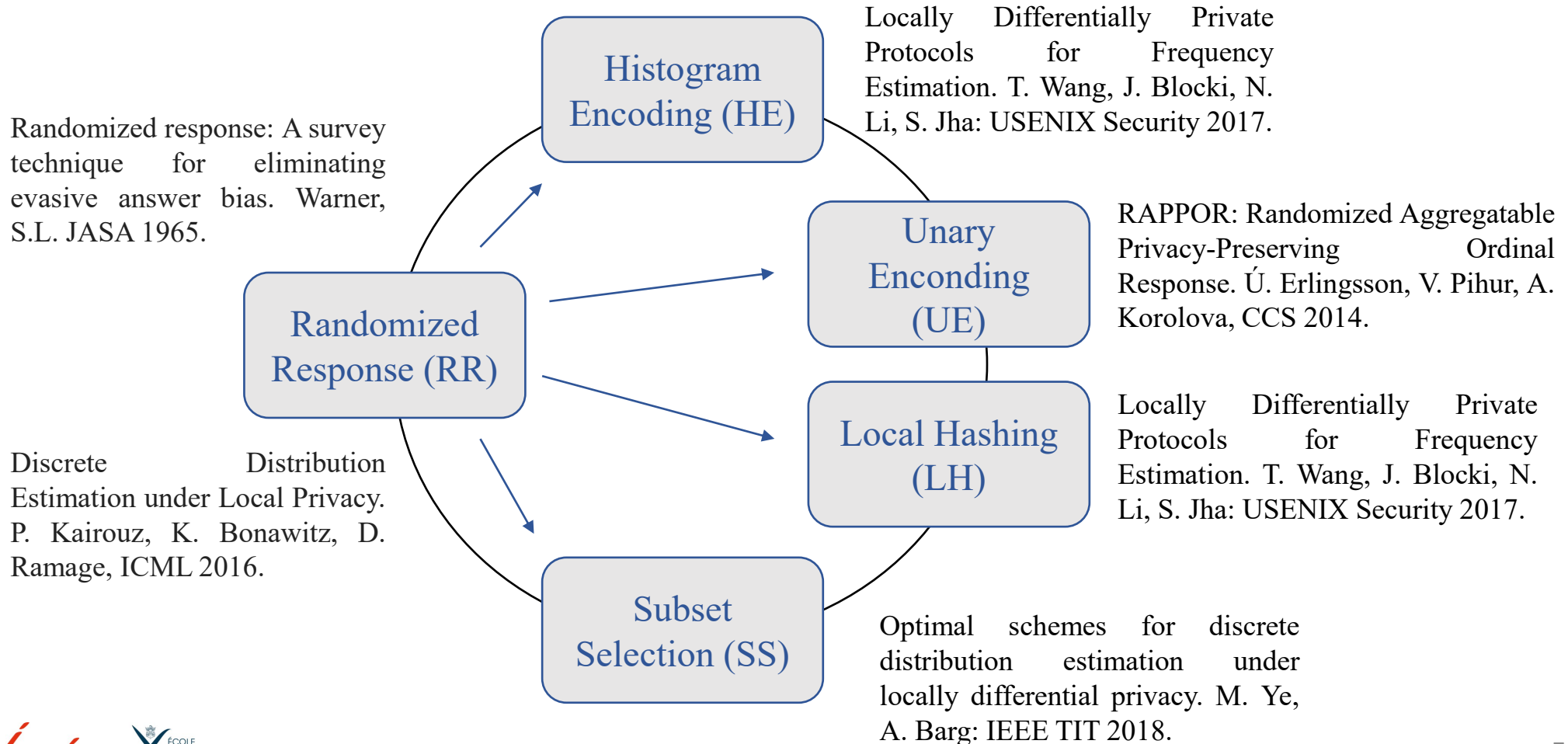
Local DP (LDP):

- ✓ No need to trust the server.
- ✗ Low utility.

Key Differences Between Central and Local DP

- Central DP concerns any two neighboring datasets;
 - Let f be the mean query on database D : $\tilde{\mu} = f(D) + \text{Lap}(s/\epsilon)$.
- Local DP concerns any two values;
 - Let the user's value x lies in range $[-1, 1]$: $y = x + \text{Lap}(2/\epsilon)$;
 - The server aggregates LDP data to estimate mean: $\tilde{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$.
- As a result, **the amount of noise is different** (each sample);
- Two lines of research to improve the privacy-utility trade-off:
 1. Design new LDP mechanisms;
 2. Improve the estimation at the server side.

State-of-the-Art LDP Distribution Estimation Mechanisms



Post-Processing Distribution Estimator for LDP Mechanisms

Paper	Estimator	Post-Processing	LDP Mechanisms Evaluated
Discrete Distribution Estimation under Local Privacy (ICML 2016)	Matrix Inversion (MI)	<ul style="list-style-type: none"> • Re-normalization • Projection onto the probability simplex 	<ul style="list-style-type: none"> • Generalized RR (GRR) • Symmetric UE (SUE)
Locally Differentially Private Frequency Estimation with Consistency (NDSS 2020)	MI	<ul style="list-style-type: none"> • 10 techniques (e.g., enforcing only non-negativity, re-normalization, ...) 	<ul style="list-style-type: none"> • Optimal LH (OLH)
Generalized iterative bayesian update and applications to mechanisms for privacy protection (Euro S&P 2020)	Iterative Bayesian Update (IBU)	<ul style="list-style-type: none"> • Generic IBU for personalized LDP 	<ul style="list-style-type: none"> • GRR • SUE
Reconstruction of the distribution of sensitive data under free-will privacy (arXiv 2022)			
Our (DBSec 2023)	MI vs IBU	<ul style="list-style-type: none"> • MI re-normalization 	<ul style="list-style-type: none"> • 7 one-time (e.g., GRR, SUE, ...) • 7 longitudinal (e.g., RAPPOR)

Outline

1. Context
- 2. Background & Problem Statement**
3. Experimental Results
4. Conclusion & Perspectives

LDP: Formal Definition & Properties [Duchi et al, 2013]

Def (ϵ -LDP). A randomized mechanism \mathcal{M} satisfies ϵ -LDP, where $\epsilon \geq 0$, if for **any two inputs** $v, v' \in \text{Domain}(\mathcal{M})$ and for **any output** $z \in \text{Range}(\mathcal{M})$:

$$\frac{\Pr[\mathcal{M}(v) = z]}{\Pr[\mathcal{M}(v') = z]} \leq e^\epsilon$$



LDP: Formal Definition & Properties [Duchi et al, 2013]

Def (ϵ -LDP). A randomized mechanism \mathcal{M} satisfies ϵ -LDP, where $\epsilon \geq 0$, if for **any two inputs** $v, v' \in \text{Domain}(\mathcal{M})$ and for **any output** $z \in \text{Range}(\mathcal{M})$:

$$\frac{\Pr[\mathcal{M}(v) = z]}{\Pr[\mathcal{M}(v') = z]} \leq e^\epsilon$$



Def (Pure ϵ -LDP) [Wang et al, 2017]. An ϵ -LDP mechanism \mathcal{M} is **pure** if there are two probability parameters $0 < q^* < p^* < 1$ such that for all $v \neq v' \in \text{Domain}(\mathcal{M})$:

$$\begin{aligned}\Pr[\mathcal{M}(v) \in \{z | v \in S(z)\}] &= p^*, \\ \Pr[\mathcal{M}(v') \in \{z | v \in S(z)\}] &= q^*,\end{aligned}$$

where $S(z)$ is the set of items that z ‘supports’.

LDP Distribution Estimation: MI and IBU

\mathbf{f} : Original distribution $\tilde{\mathbf{f}}$: Observed distribution

Matrix Inversion (MI)

$$\hat{\mathbf{f}} = \frac{\tilde{\mathbf{f}} - nq^*}{n(p^* - q^*)} = \tilde{\mathbf{f}}A_{vz}^{-1}$$

Iterative Bayesian Update (IBU)

$$\hat{\mathbf{f}}^{t+1} = \tilde{\mathbf{f}} \cdot \frac{\hat{\mathbf{f}}^t * A_{vz}}{\hat{\mathbf{f}}^t \cdot A_{vz}}$$

Channel matrix (probability of obtaining z given v):

$$A_{vz} = \begin{bmatrix} p^* & \cdots & q^* \\ \vdots & \ddots & \vdots \\ q^* & \cdots & p^* \end{bmatrix}$$

Outline

1. Context
- 2. Background & Problem Statement**
 - i. One-Time Distribution Estimation;**
 - ii. Longitudinal Distribution Estimation.
3. Experimental Results
4. Conclusion & Perspectives

Problem Statement #1: One-Time Distribution Estimation

Algorithm 1 General pure LDP procedure for distribution estimation.

Input : Original data of users, privacy parameter ϵ , mechanism $\mathcal{M}_{(\epsilon)}$.

Output : Estimated discrete distribution.

User-side

- 1: **for** each user $i \in [1..n]$ with input data $v^i \in V$ **do**
 - 2: **Encode**(v^i) into a specific format (**if needed**);
 - 3: **Obfuscate**(v^i) as $z^i = \mathcal{M}_{(\epsilon)}(v^i)$;
 - 4: Transmit z^i to the aggregator.
 - 5: **end for**
- # Server-side*
- 6: Obtain the support set $S(z)$ and probabilities p^* and q^* for $\mathcal{M}_{(\epsilon)}$.
 - 7: **Estimate** Aggregate the obfuscated data z^i ($i \in [1..n]$) to estimate $\{\hat{f}(v)\}_{v \in \mathcal{D}}$.
 - 8: **return** : Estimated discrete distribution $\hat{\mathbf{f}}$ (*i.e.*, a k -bins histogram).
-

f: Original distribution **$\hat{\mathbf{f}}$** : Estimated distribution

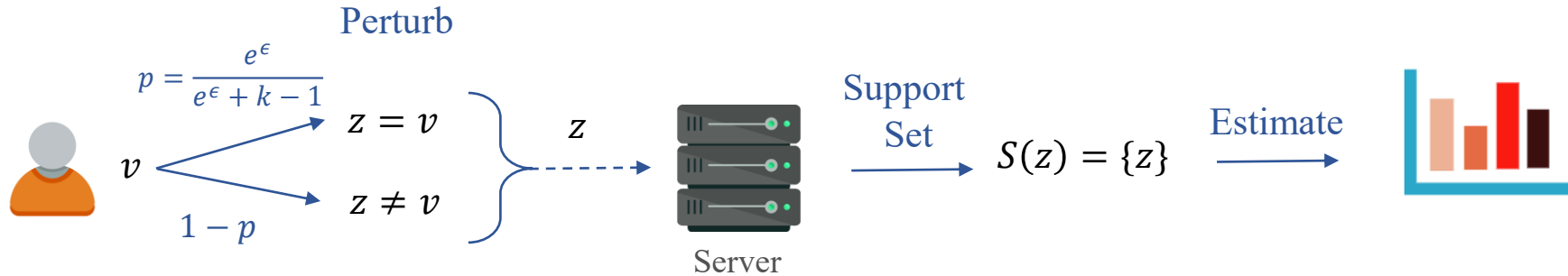
MSE(**f**, **$\hat{\mathbf{f}}$**)



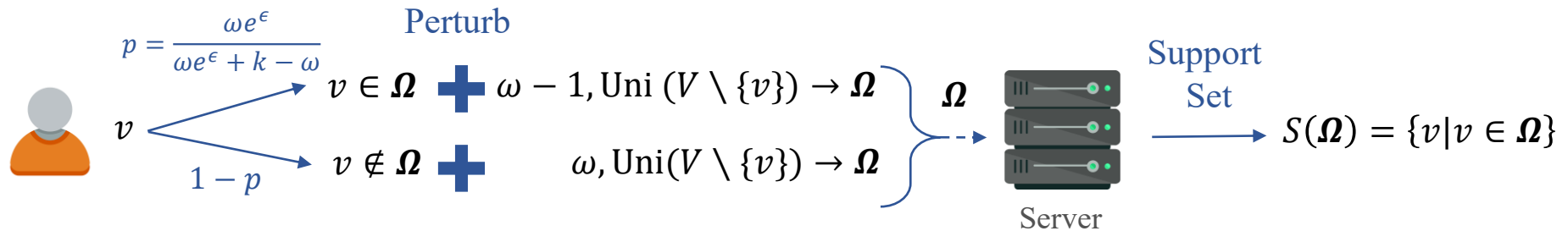
MAE(**f**, **$\hat{\mathbf{f}}$**)

One-Time LDP Distribution Estimation Mechanisms

Generalized Randomized Response (GRR)

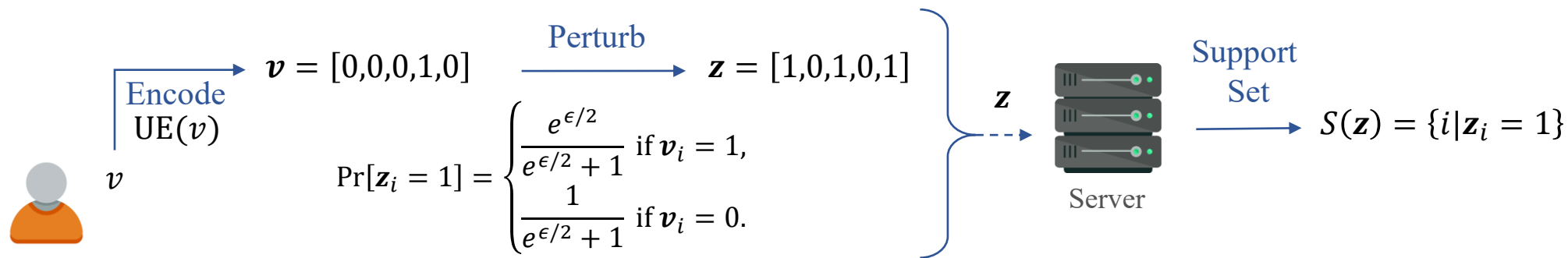


Subset Selection (SS)

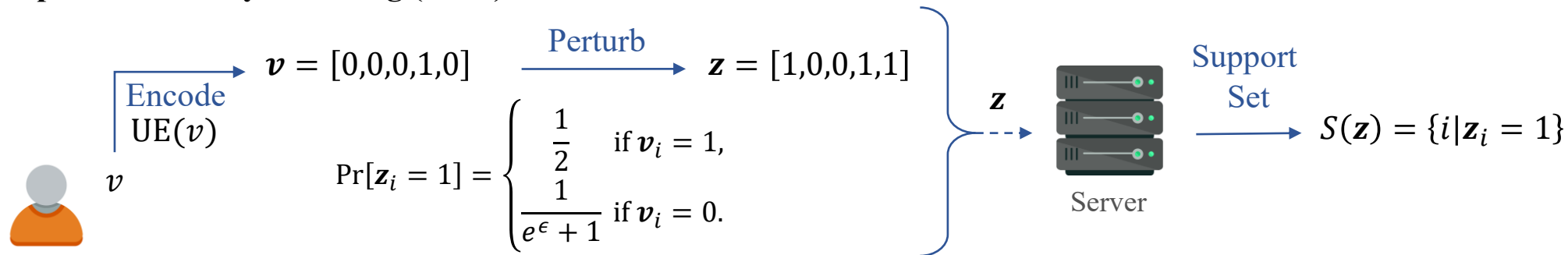


One-Time LDP Distribution Estimation Mechanisms

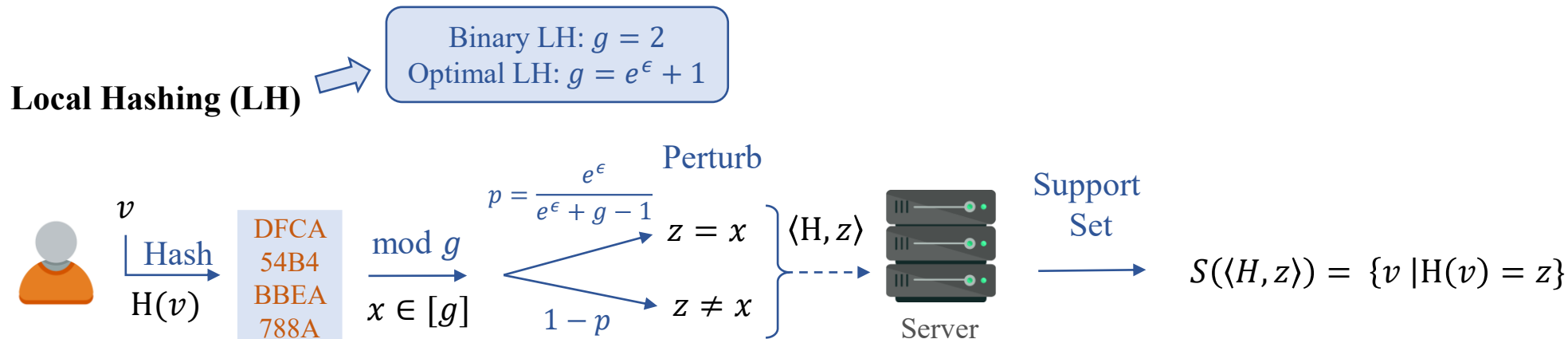
Symmetric Unary Encoding (SUE)



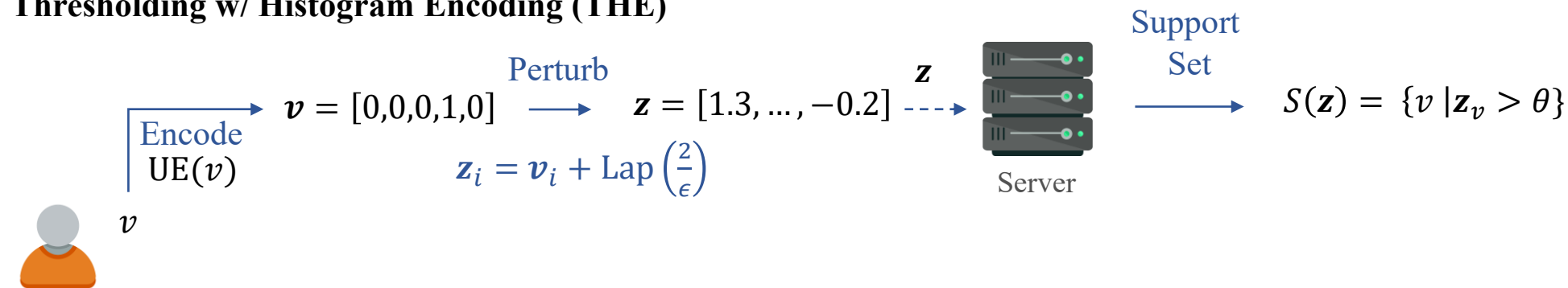
Optimized Unary Encoding (OUE)



One-Time LDP Distribution Estimation Mechanisms



Thresholding w/ Histogram Encoding (THE)



Outline

1. Context
2. **Background & Problem Statement**
 - i. One-Time Distribution Estimation;
 - ii. Longitudinal Distribution Estimation.**
3. Experimental Results
4. Conclusion & Perspectives

Problem Statement #2: Longitudinal Distribution Estimation

Algorithm 2 Memoization-based procedure for longitudinal distribution estimation under LDP guarantees.

Input : Original data of users, privacy parameters $\epsilon_\infty, \epsilon_1$, mechanisms $\mathcal{M}_1, \mathcal{M}_2$.

Output : Estimated discrete distribution $\hat{\mathbf{f}}$ at each $t \in [\tau]$.

User-side

- 1: **for** each user $i \in [1..n]$ with input data $v^i \in V$ **do**
- 2: **Encode**(v^i) into a specific format (**if needed**);
- 3: **Obfuscate**(v^i) as $z^i = \mathcal{M}_{1(\epsilon_\infty)}(v^i)$; \triangleright First obfuscation step: p_1^* and q_1^*
- 4: **Memoize**(z^i) for v^i .
- 5: **for** each time $t \in [\tau]$ **do**:
- 6: **Obfuscate**(z^i) as $z_t^i = \mathcal{M}_{2(\epsilon)}(z^i)$; \triangleright Second obfuscation step: p_2^* and q_2^*
- 7: Transmit z_t^i to the aggregator.
- 8: **end for**
- 9: **end for**

Server-side

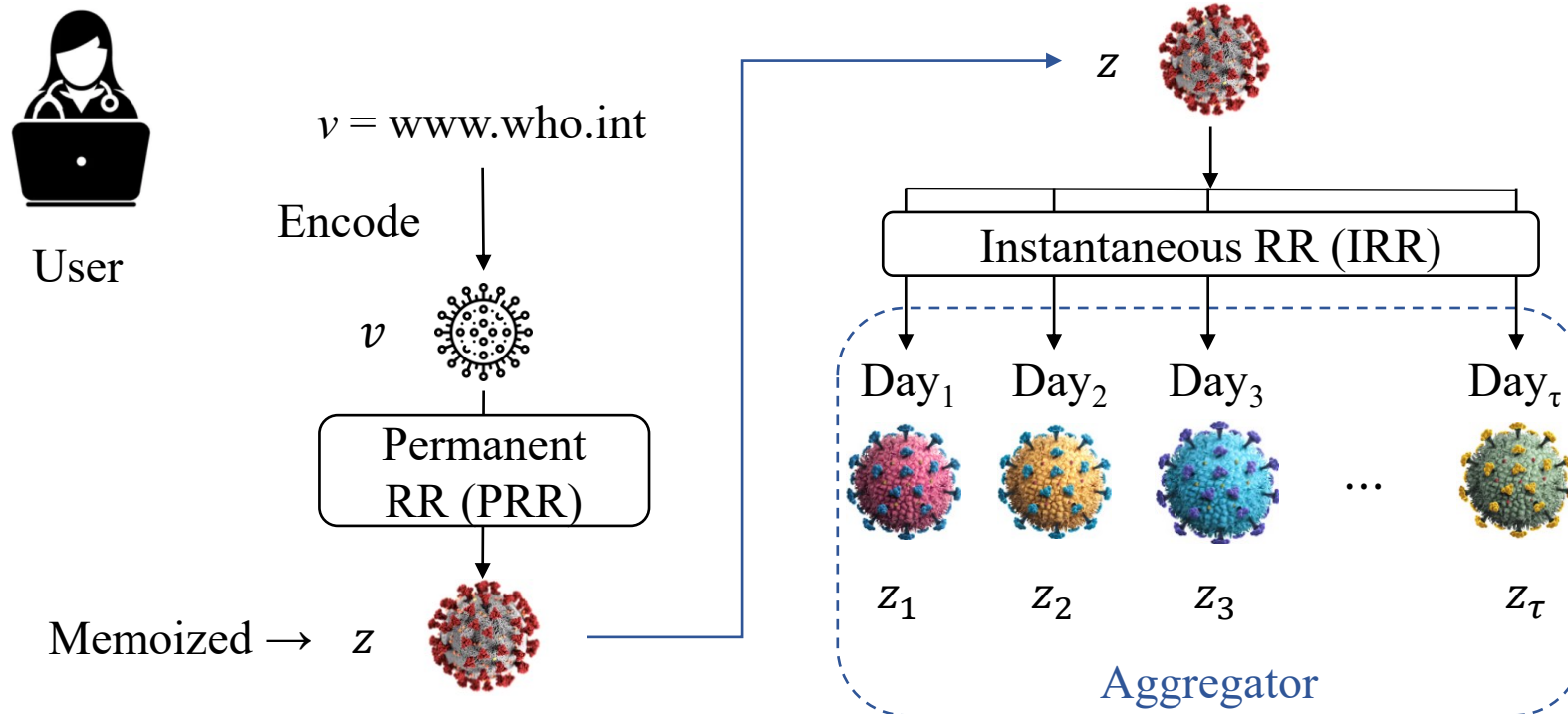
- 10: Obtain the support set $S(z)$ and probabilities p_1^*, q_1^*, p_2^* , and q_2^* for $\mathcal{M}_{1(\epsilon)}, \mathcal{M}_{2(\epsilon)}$.
 - 11: **for** each time $t \in [\tau]$ **do**:
 - 12: **Estimate** Aggregate the obfuscated data z_t^i ($i \in [1..n]$) to estimate $\{\hat{f}(v)\}_{v \in \mathcal{D}}$.
 - 13: **end for**
-

f: Original distribution **$\hat{\mathbf{f}}$** : Estimated distribution

$$\text{MSE}(\mathbf{f}, \hat{\mathbf{f}}) \quad \img alt="A blue double-headed arrow pointing left and right." data-bbox="468 854 532 932"/> \quad \text{MAE}(\mathbf{f}, \hat{\mathbf{f}})$$

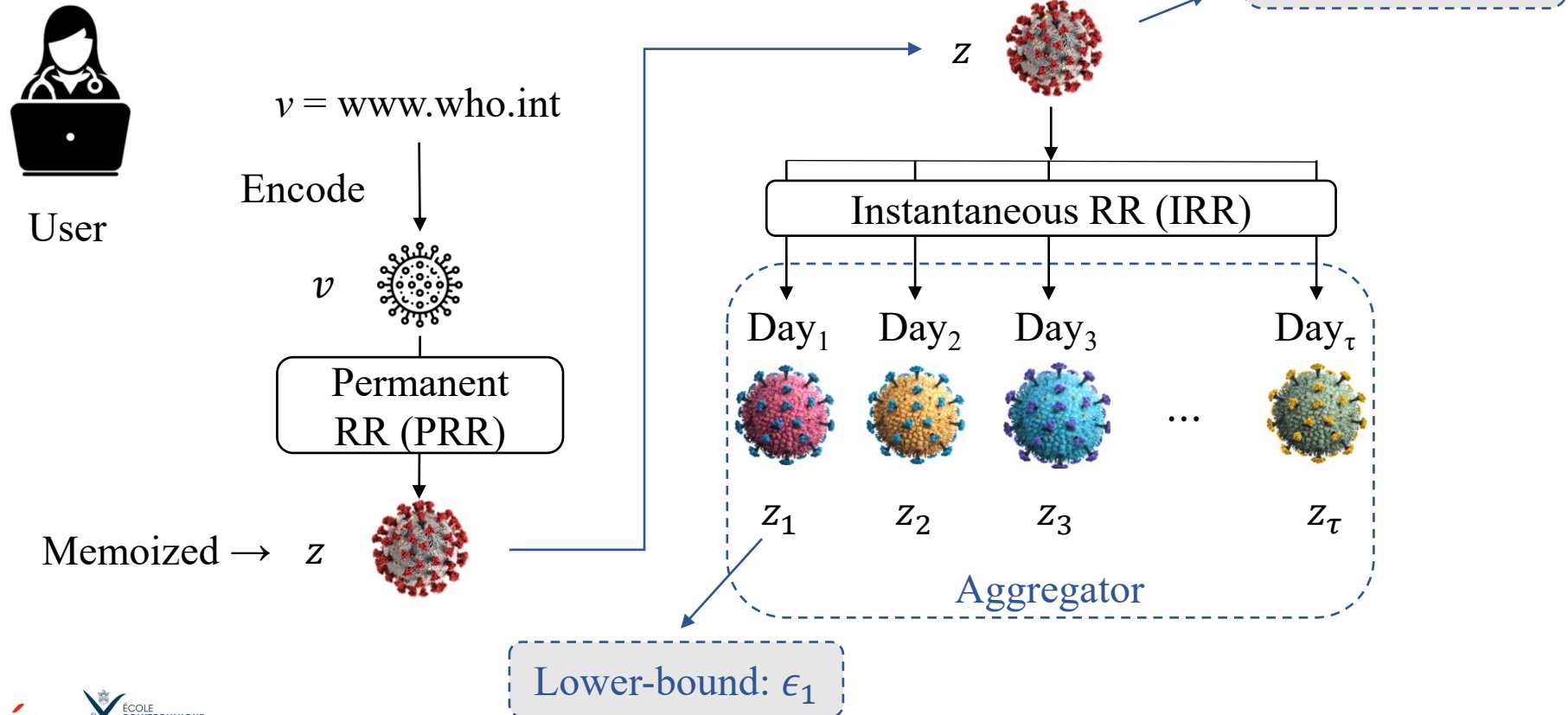
Longitudinal LDP Distribution Estimation Mechanisms

Memoization-based solution [Erlingsson, Pihur, Korolova, 2014]:



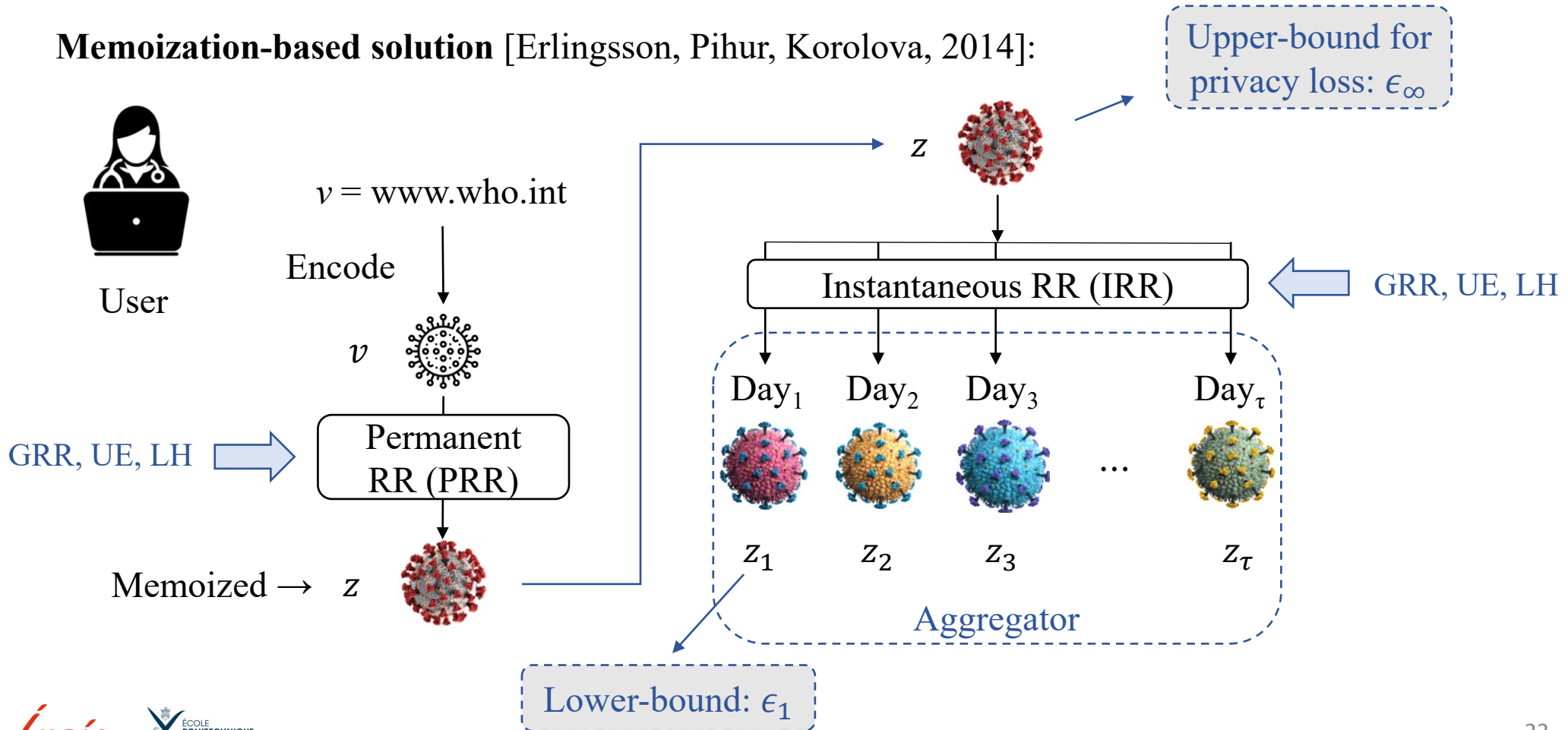
Longitudinal LDP Distribution Estimation Mechanisms

Memoization-based solution [Erlingsson, Pihur, Korolova, 2014]:



Longitudinal LDP Distribution Estimation Mechanisms

Memoization-based solution [Erlingsson, Pihur, Korolova, 2014]:



Outline

1. Context
2. Background & Problem Statement
- 3. Experimental Results**
4. Conclusion & Perspectives

Setting of Experiments

Six data distributions:

- Gaussian, Exponential, Uniform, Poisson, Triangular, Real.

Four domain size:

- $k \in \{2, 50, 100, 200\}$.

Two number of users:

- $n \in \{20000, 100000\}$.

Fourteen LDP mechanisms:

- **One-time:** GRR, SS, SUE, OUE, BLH, OLH, THE.
- **Longitudinal:** L-GRR, four L-UE, two L-LH.

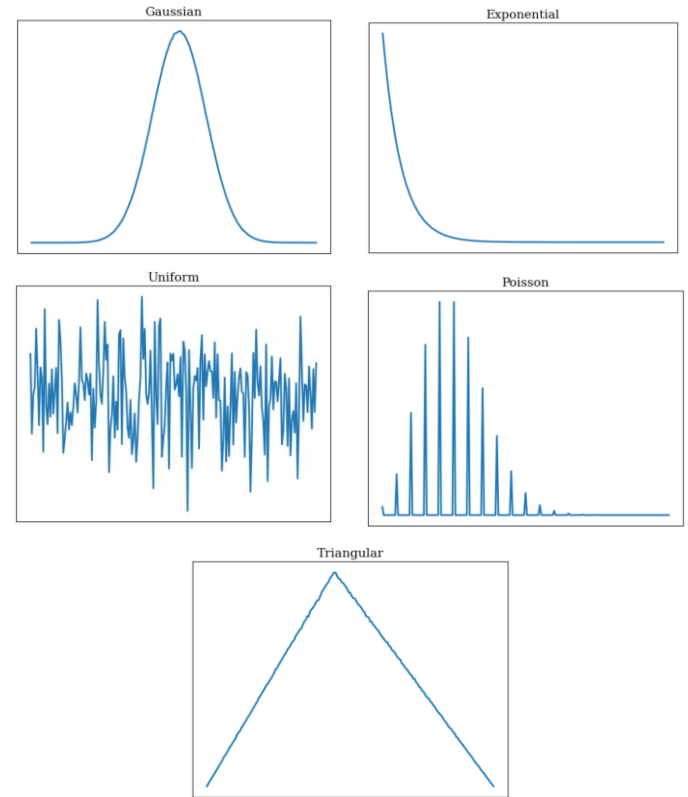
Two utility metrics:

- MSE and MAE.

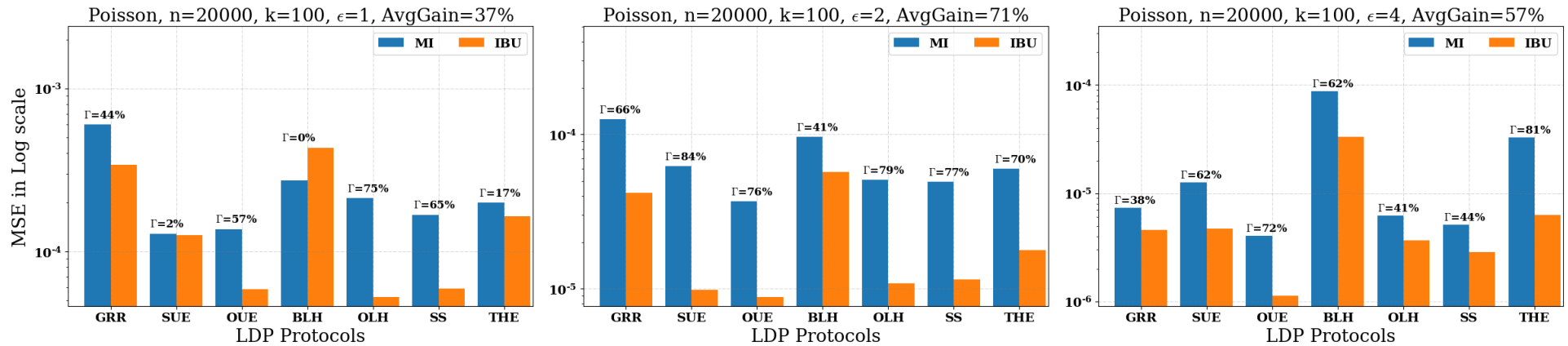


IBU utility gain

$$\Gamma(\%) = 100 \cdot \max\left(\frac{\text{Metric}_{\text{MI}} - \text{Metric}_{\text{IBU}}}{\text{Metric}_{\text{MI}}}, 0\right)$$



Instance of IBU Utility Gain: One-Time LDP Mechanisms



Summary of IBU Utility Gain: One-Time LDP Mechanisms

Averaged IBU gain in % considering all experimented k, n, ϵ .

Dist.	GRR		SUE		OUE		SS		THE		BLH		OLH		Avg.	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Gauss.	1	1	13	7	10	6	3	1	13	7	16	9	11	7	9	5
Exp.	16	11	26	15	27	16	19	11	26	15	16	10	27	16	22	13
Unif.	0	0	29	21	20	14	14	10	31	22	57	43	18	12	24	17
Poiss.	39	28	41	26	44	28	41	27	41	27	14	6	46	30	38	24
Triang.	0	0	21	13	15	9	10	6	23	14	36	21	15	9	17	10
Rea.l	31	21	40	23	42	25	34	19	42	25	21	11	44	27	36	21
Avg.	14	10	28	17	26	16	20	12	29	18	26	16	26	16	24	15

Mechanisms w/ highest IBU gain: SUE and THE

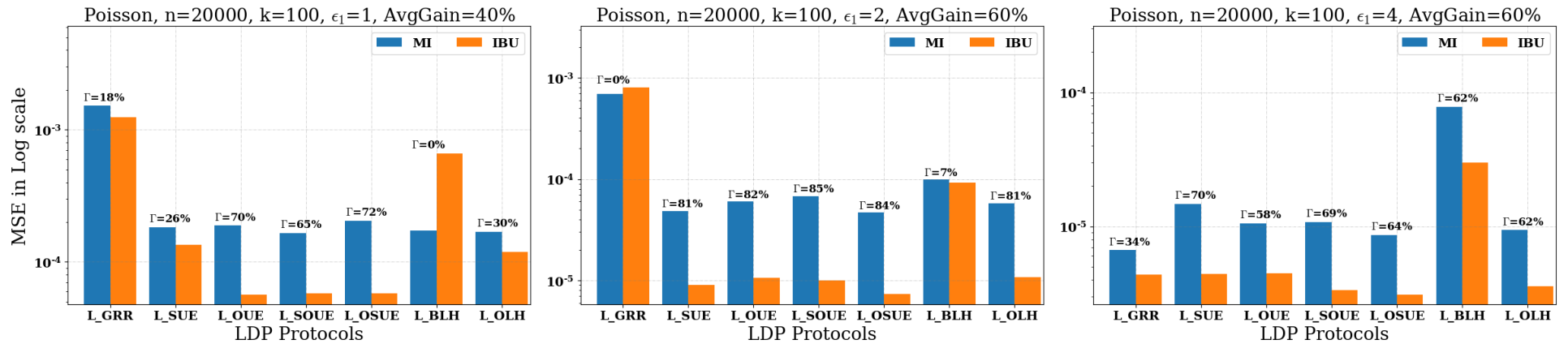
Summary of IBU Utility Gain: One-Time LDP Mechanisms

Averaged IBU gain in % considering all experimented k, n, ϵ .

Dist.	GRR		SUE		OUE		SS		THE		BLH		OLH		Avg.	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Gauss.	1	1	13	7	10	6	3	1	13	7	16	9	11	7	9	5
Exp.	16	11	26	15	27	16	19	11	26	15	16	10	27	16	22	13
Unif.	0	0	29	21	20	14	14	10	31	22	57	43	18	12	24	17
Poiss.	39	28	41	26	44	28	41	27	41	27	14	6	46	30	38	24
Triang.	0	0	21	13	15	9	10	6	23	14	36	21	15	9	17	10
Rea.l	31	21	40	23	42	25	34	19	42	25	21	11	44	27	36	21
Avg.	14	10	28	17	26	16	20	12	29	18	26	16	26	16	24	15

Distributions w/ highest IBU gain: Poisson and real

Instance of IBU Utility Gain for Longitudinal LDP Mechanisms



Summary of IBU Utility Gain: Longitudinal LDP Mechanisms

Averaged IBU gain in % considering all experimented k, n, ϵ .

Dist.	L-GRR		L-SUE		L-OUE		L-SOUE		L-OSUE		L-BLH		L-OLH		Avg.	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Gauss.	14	5	13	8	9	5	10	7	12	7	2	0	7	4	9	5
Exp.	4	1	27	16	26	15	27	16	27	16	4	2	20	12	19	11
Unif.	36	25	31	22	12	8	16	11	18	13	54	43	21	16	26	19
Poiss.	5	2	43	28	48	32	49	32	44	29	11	6	42	30	34	22
Triang.	28	17	24	15	11	7	13	9	16	10	26	14	14	9	18	11
Real.	4	1	43	25	43	27	44	27	43	25	9	4	34	22	31	18
Avg.	15	8	30	19	24	15	26	17	26	16	17	11	23	15	23	14

Mechanisms w/ highest IBU gain: L-SUE and L-SOUE

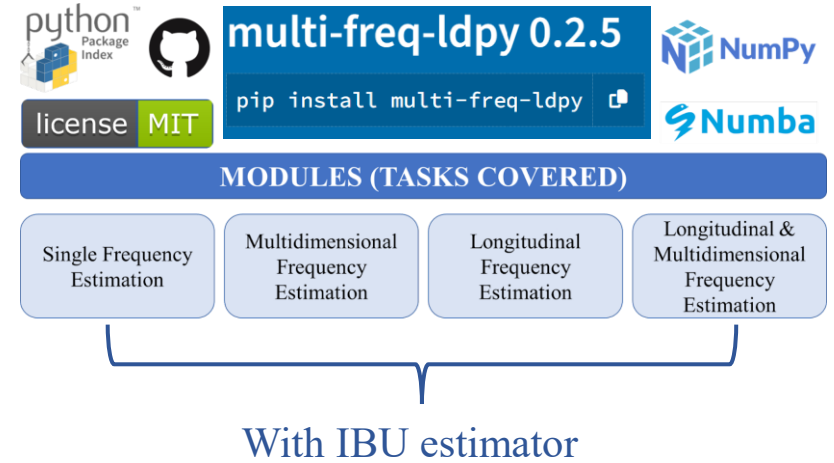
Summary of IBU Utility Gain: Longitudinal LDP Mechanisms

Averaged IBU gain in % considering all experimented k, n, ϵ .

Dist.	L-GRR		L-SUE		L-OUE		L-SOUE		L-OSUE		L-BLH		L-OLH		Avg.	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Gauss.	14	5	13	8	9	5	10	7	12	7	2	0	7	4	9	5
Exp.	4	1	27	16	26	15	27	16	27	16	4	2	20	12	19	11
Unif.	36	25	31	22	12	8	16	11	18	13	54	43	21	16	26	19
Poiss.	5	2	43	28	48	32	49	32	44	29	11	6	42	30	34	22
Triang.	28	17	24	15	11	7	13	9	16	10	26	14	14	9	18	11
Real.	4	1	43	25	43	27	44	27	43	25	9	4	34	22	31	18
Avg.	15	8	30	19	24	15	26	17	26	16	17	11	23	15	23	14

Distributions w/ highest IBU gain: Poisson and real

IBU Implementation into Multi-Freq-LDPy [Arcolezi et al, 2022]



IBU Implementation into Multi-Freq-LDPy [Arcolezi et al, 2022]

```
# Multi-Freq-LDPy functions for GRR protocol
from multi_freq_ldpy.pure_frequency_oracles.GRR import GRR_Client,
    GRR_Aggregator_IBU

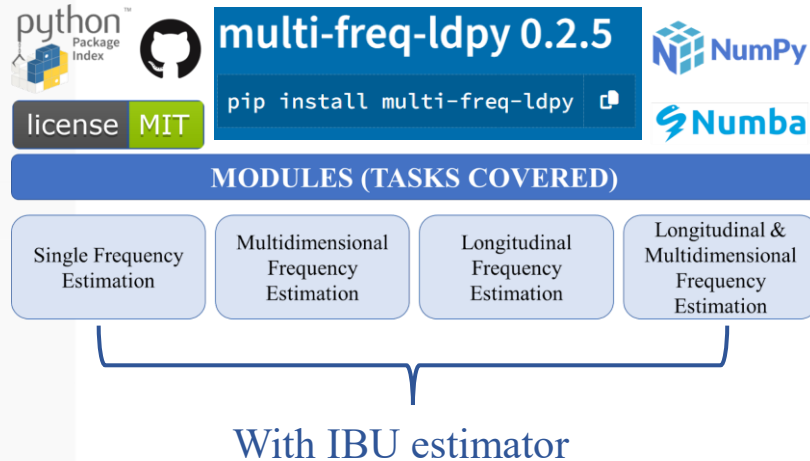
# NumPy library
import numpy as np

# Parameters for simulation
eps = 1 # privacy guarantee
n = int(1e6) # number of users
k = 5 # attribute's domain size

# Simulation dataset following Uniform distribution
dataset = np.random.randint(k, size=n)

# Simulation of client-side data obfuscation
rep = [GRR_Client(user_data, k, eps) for user_data in dataset]

# Simulation of server-side aggregation
GRR_Aggregator_IBU(rep, k, eps, nb_iter=10000, tol=1e-12, err_func="max_abs")
>>> array([0.199, 0.201, 0.199, 0.202, 0.199])
```



IBU Implementation into Multi-Freq-LDPy [Arcolezi et al, 2022]

```
# Multi-Freq-LDPy functions for GRR protocol
from multi_freq_ldpy.pure_frequency_oracles.GRR import GRR_Client,
GRR_Aggregator_IBU
```

```
# NumPy library
import numpy as np
```

```
# Parameters for simulation
eps = 1 # privacy guarantee
n = int(1e6) # number of users
k = 5 # attribute's domain size
```

```
# Simulation dataset following Uniform distribution
dataset = np.random.randint(k, size=n)
```

```
# Simulation of client-side data obfuscation
rep = [GRR_Client(user_data, k, eps) for user_data in dataset]
```

```
# Simulation of server-side aggregation
GRR_Aggregator_IBU(rep, k, eps, nb_iter=10000, tol=1e-12, err_func="max_abs")
>>> array([0.199, 0.201, 0.199, 0.202, 0.199])
```



Essentially just 2 lines of code to simulate the LDP data collection pipeline with IBU estimation



MODULES (TASKS COVERED)

Single Frequency Estimation

Multidimensional Frequency Estimation

Longitudinal Frequency Estimation

Longitudinal & Multidimensional Frequency Estimation

With IBU estimator

Outline

1. Context
2. Background & Problem Statement
3. Experimental Results
4. **Conclusion & Perspectives**

Takeaway Messages

Conclusions:

- We benchmarked IBU against MI in **several contexts** for 14 LDP mechanisms;
- IBU can **significantly improve** the utility of LDP distribution estimation;
- We implemented IBU into **multi-freq-ldpy**.

Takeaway Messages

Conclusions:

- We benchmarked IBU against MI in **several contexts** for 14 LDP mechanisms;
- IBU can **significantly improve** the utility of LDP distribution estimation;
- We implemented IBU into **multi-freq-ldpy**.

Perspectives:

- Investigate IBU for “**non-pure**” LDP mechanisms;
- Consider **different initialization and stopping criteria** for IBU;
- IBU for high-dimensional data (*i.e.*, $k \gg 200$);
- Implement Generalized IBU (GIBU) into **multi-freq-ldpy**.

On the Utility Gain of Iterative Bayesian Update for Locally Differentially Private Mechanisms

Héber H. Arcolezi, Selene Cerna, and Catuscia Palamidessi

Inria and École Polytechnique (IPP), Palaiseau, France

[PAPER](#)



[ARTIFACT](#)



[CONTACT](#)



hharcolezi.github.io



heber.hwang-arcolezi@inria.fr



[@hharcolezi](https://twitter.com/hharcolezi)