On the Risks of Collecting Multidimensional Data Under Local Differential Privacy

Héber H. Arcolezi
Inria and École Polytechnique (IPP)
heber.hwang-arcolezi@inria.fr

Sébastien Gambs
Université du Québec à Montréal, UQAM
gambs.sebastien@uqam.ca

Jean-François Couchot
Femto-ST Institute, Univ. Bourg. Franche-Comté, CNRS
jean-francois.couchot@univ-fcomte.fr

Catuscia Palamidessi
Inria and École Polytechnique (IPP)
catuscia@lix.polytechnique.fr

Introduction

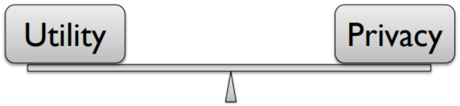
ÉCOLE POLYTECHNIQUE Inria

Local Differential Privacy (LDP): Definition & Properties

Def (ϵ-LDP) [1]. 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** $z \in \text{Range}(\mathcal{M})$:

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

Privacy Loss



Fundamental (L)DP properties [2]:

- **Post-processing** → if $\mathcal{M}$ is ϵ-LDP, then the composition $f(\mathcal{M})$ is ϵ-LDP for any f.
- **Composition** → Let $\mathcal{M}_1$ be a ϵ_1-LDP mechanism and $\mathcal{M}_2$ a ϵ_2-LDP mechanism. Then, the composed mechanism $\mathcal{M} = \big(\mathcal{M}_1(v), \mathcal{M}_2(v)\big)$ is $(\epsilon_1 + \epsilon_2)$-LDP.

[1] Duchi et al. *Local privacy and statistical minimax rates*. FOCS 2013.
[2] Dwork et al, 2006. *Calibrating noise to sensitivity in private data analysis*. TCC 2006.

Motivation for Attack-Based Approaches

Why? → Challenging, under-explored, and crucial problem.

Impact:

- Attacks allow interpreting privacy claims;
- Enable vulnerability discovery;
- Help practitioners to adequately select the privacy mechanism.

Motivation for Attack-Based Approaches

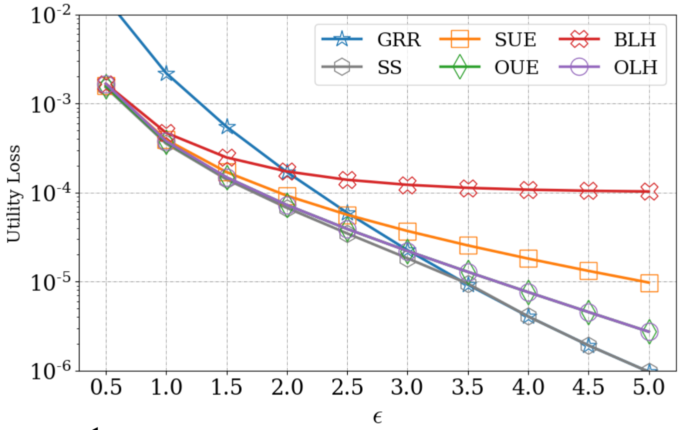
Why? → Challenging, under-explored, and crucial problem.

Impact:

- Attacks allow interpreting privacy claims;
- Enable vulnerability discovery;
- Help practitioners to adequately select the privacy mechanism.

Usual approach: Privacy-Utility Trade-off

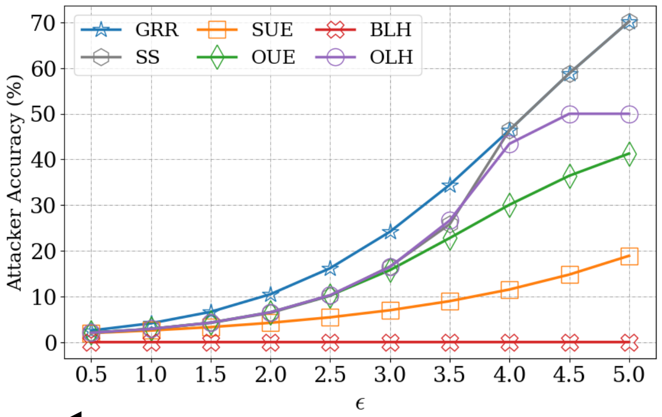
More utility



More privacy

Our approach: Privacy-Robustness Trade-off

More robust to attack

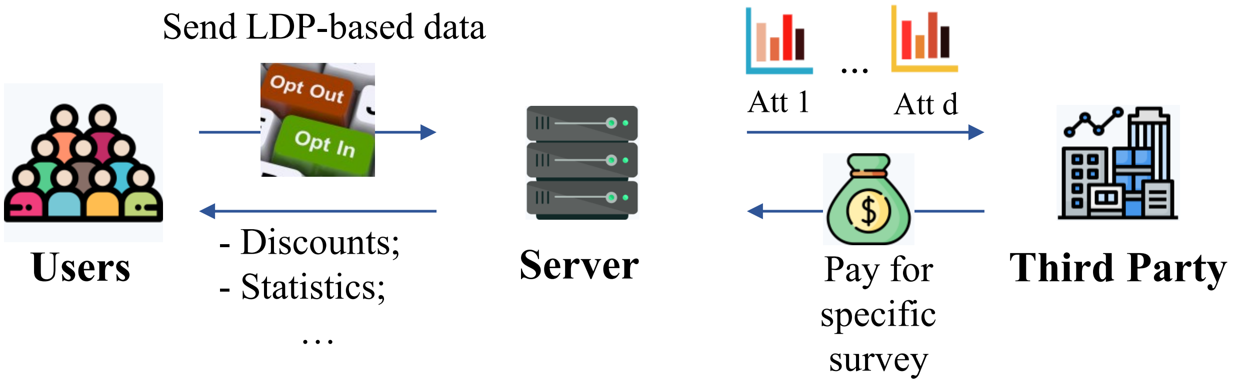


More privacy

Problem Statement & Assumptions

Motivating example:

- Server collects multidimensional data ($d \geq 2$) under LDP;
- Server surveys the population multiple times (*e.g.*, different attributes);
- Server's utility goal → independent histogram estimation (no correlation).



Problem Statement & Assumptions

Motivating example:

- Server collects multidimensional data ($d \geq 2$) under LDP;
- Server surveys the population multiple times (*e.g.*, different attributes);
- Server's utility goal → independent histogram estimation (no correlation).

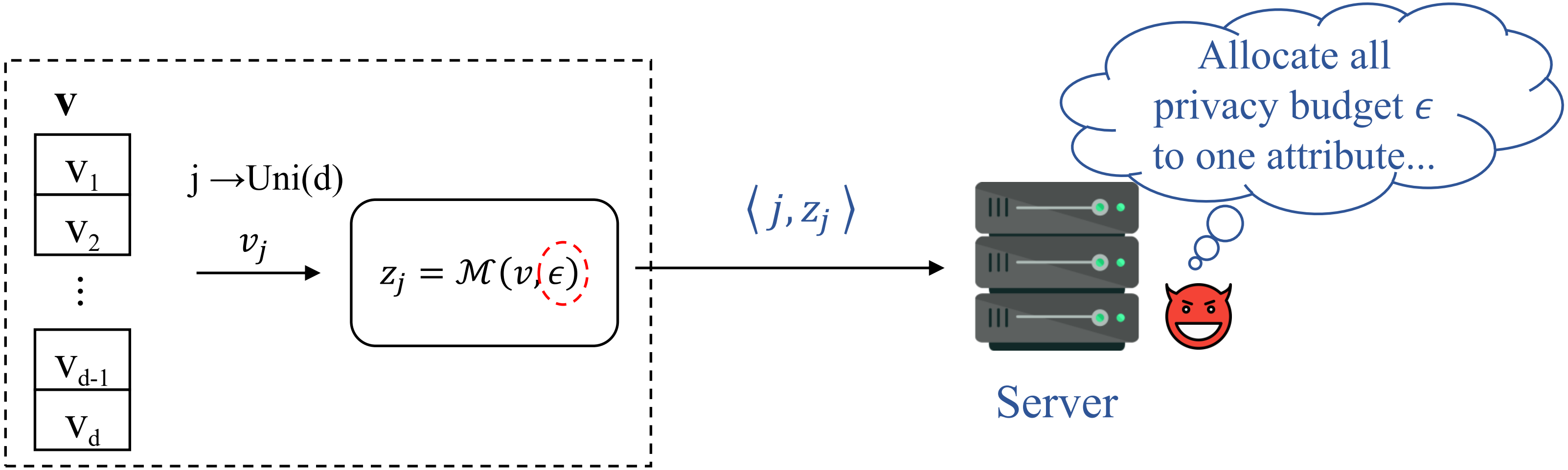
Server assumptions:

- Knows the users' pseudonymized IDs;
- Has **no knowledge** about the real data distributions;
- Has access to background knowledge (*e.g.*, Census data);
- Uses state-of-the-art solutions: SMP [3] or RS+FD [4].

[3] Wang *et al*. *Collecting and analyzing multidimensional data with local differential privacy*. ICDE 2019.
[4] Arcolezi *et al. RS+FD: Multidimensional frequency estimates with local differential privacy*. CIKM 2021.

State-of-the-Art Solutions for Multidimensional Data

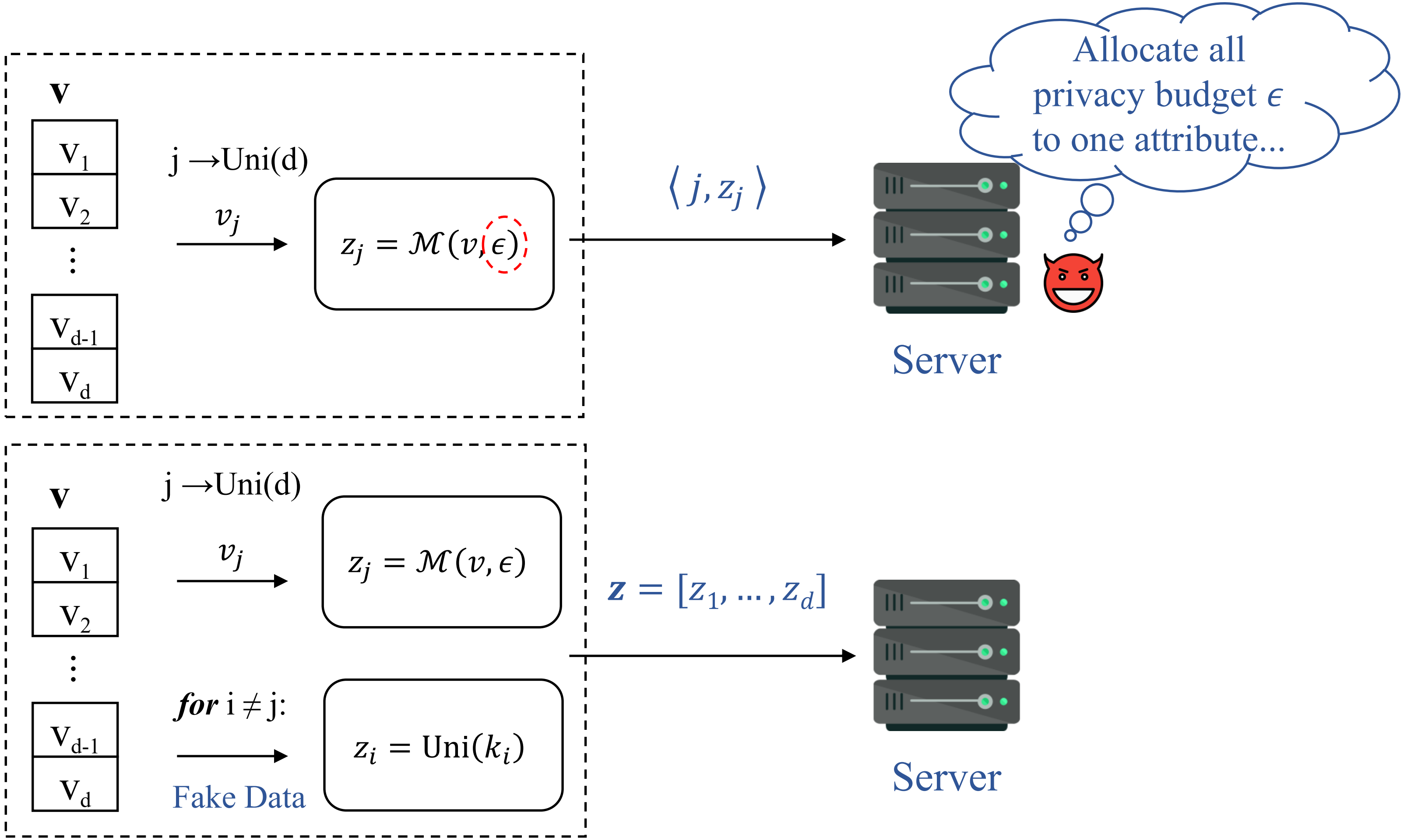
Random Sampling (SMP)



State-of-the-Art Solutions for Multidimensional Data

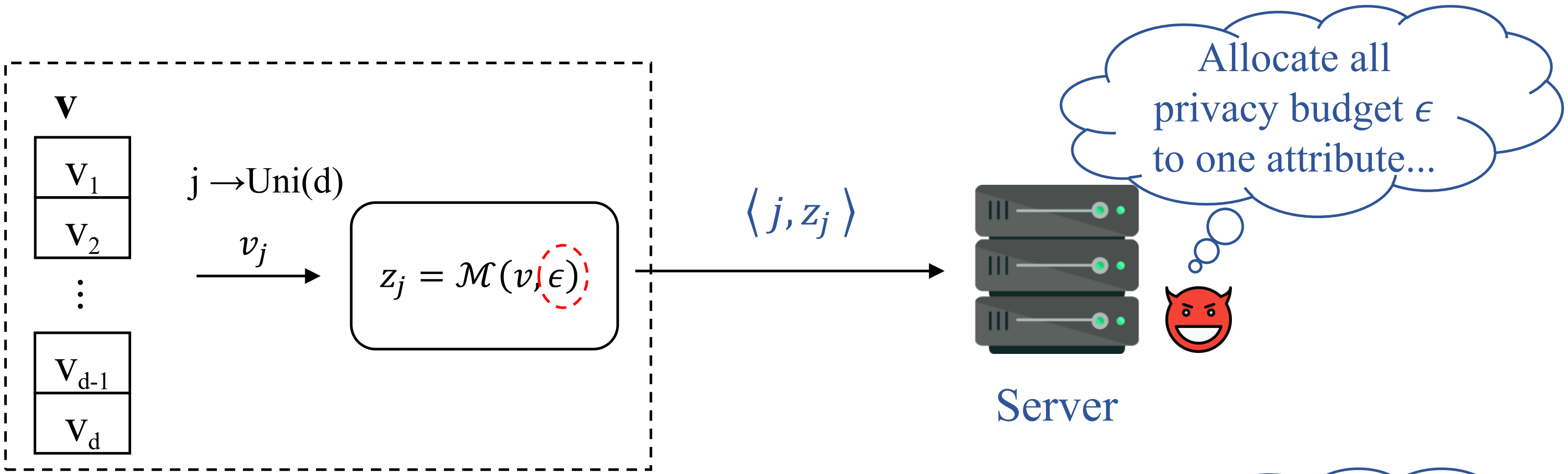
Random Sampling (SMP)

Random Sampling Plus Fake Data (RS+FD)

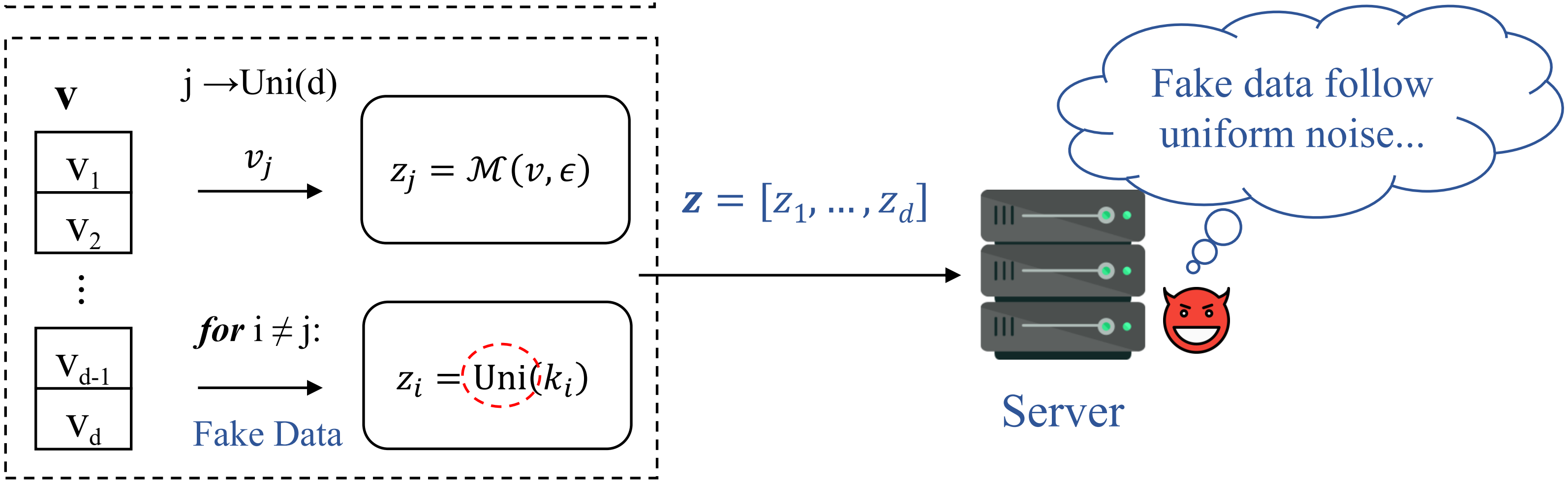


State-of-the-Art Solutions for Multidimensional Data

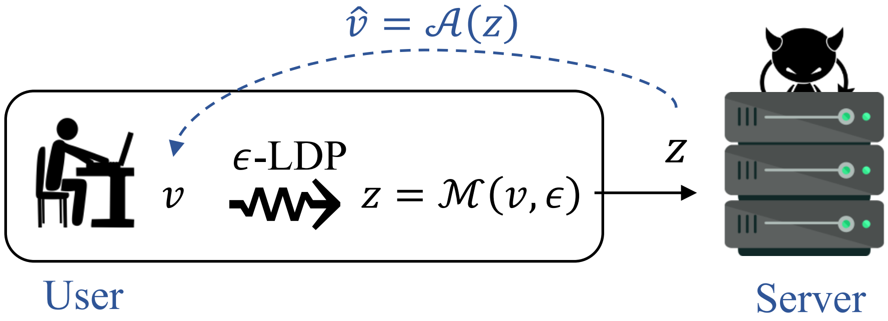
Random Sampling (SMP)



Random Sampling Plus Fake Data (RS+FD)



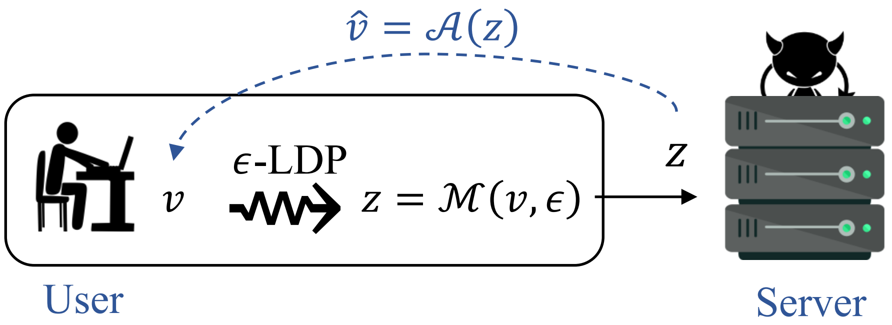
Summary of Our Contributions



Distinguishability attack:

- Value distinguishability;

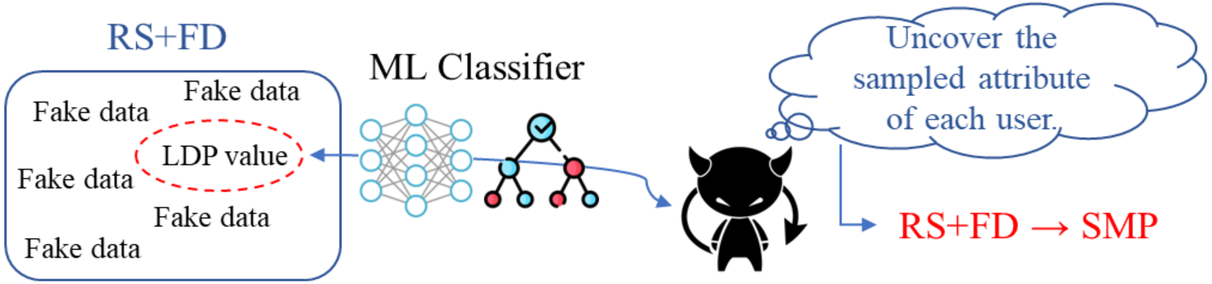
Summary of Our Contributions



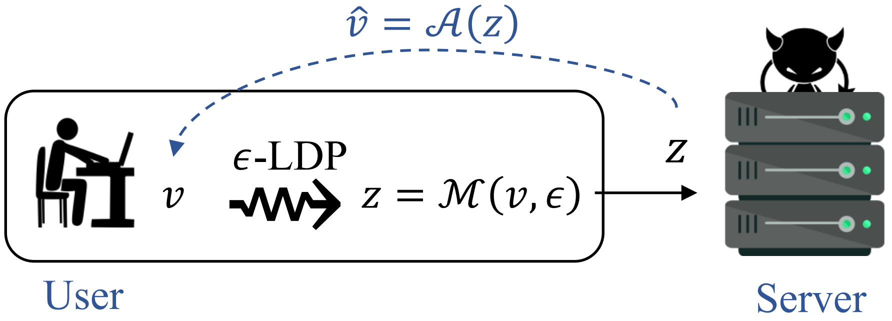
Distinguishability attack:

- Value distinguishability;

- Fake data distinguishability.

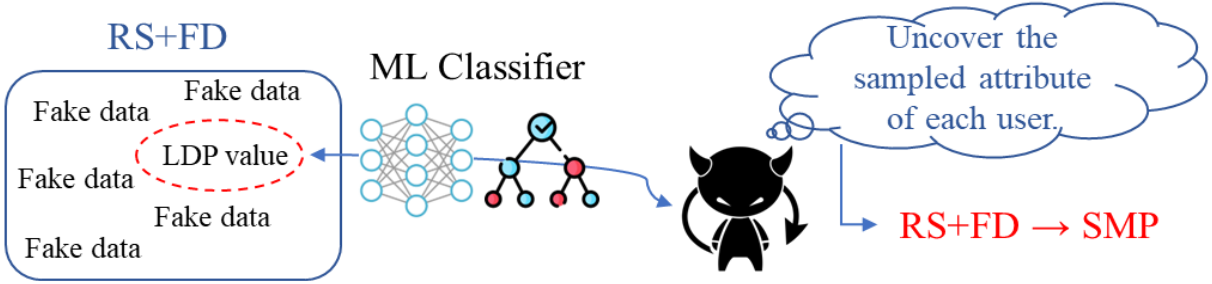


Summary of Our Contributions



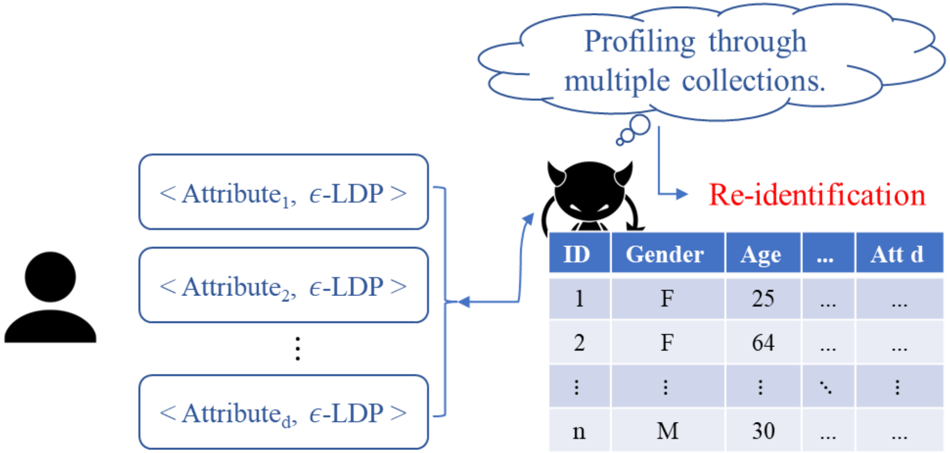
Distinguishability attack:

- Value distinguishability;
- Fake data distinguishability.



Re-identification attack:

- Profiling users + background knowledge.



ID	Gender	Age	...	Att d
1	F	25	...	...
2	F	64	...	...
⋮	⋮	⋮	⋱	⋮
n	M	30	...	...

École Polytechnique Inria

Outline

1. Introduction
2. **Attack-Based Approaches to LDP**
3. Conclusion & Perspectives

Outline

Value Distinguishability Attack

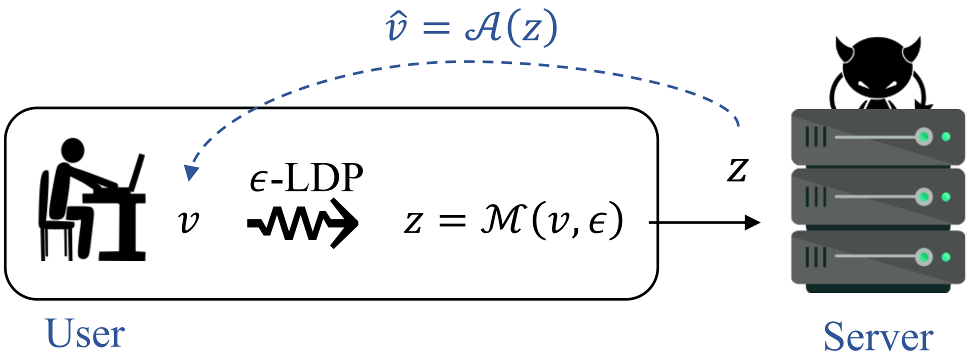
Assumption: Each user has a value $v \in V$, where $k = |V|$.

LDP mechanism: SMP solution.

Adversary's goal: Predict v given $z = \mathcal{M}(v, \epsilon)$, *i.e.*, $\hat{v} = \mathcal{A}(z)$.

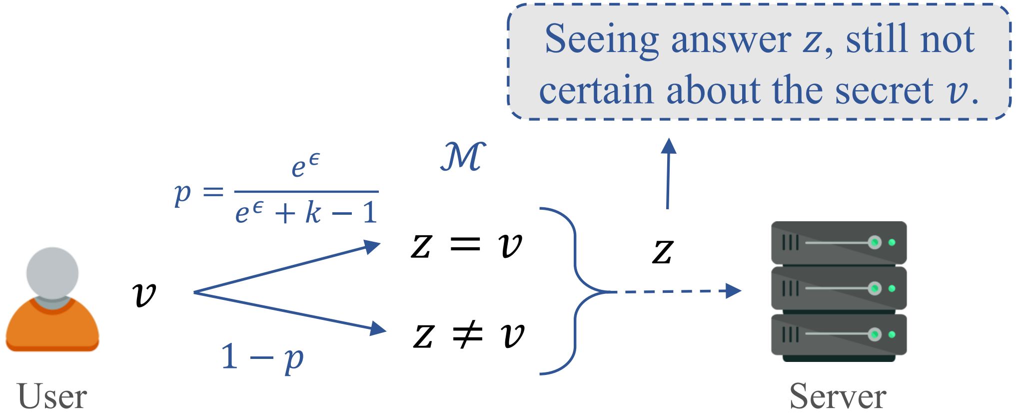
Metric: Accuracy (ACC).

Baseline: Uniform random guess $\text{ACC} = {}^{1}/_{k}$.



Generalized Randomized Response (GRR)

- No encoding required;
- Report $z = v$ with prob. $p = \frac{e^{\epsilon}}{e^{\epsilon}+k-1}$;
- Otherwise, report any other value $z = \text{Uni}\ (V \setminus \{v\})$ with prob. $q = \frac{1-p}{k-1}$ [5, 6].

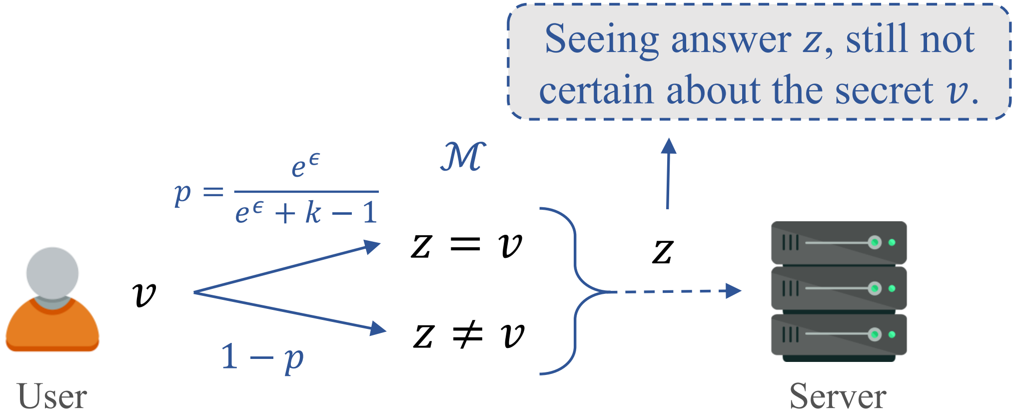


Plausible deniability: Let v be an embarrassing value of V. As long as $\Pr[\mathcal{M}(v) = v] < 1$, the user can deny to have v.

[5] Warner. *Randomized response: A survey technique for eliminating evasive answer bias*. JASA 1965.
[6] Kairouz *et al*. *Discrete distribution estimation under local privacy*. ICML 2016.

Generalized Randomized Response (GRR)

- No encoding required;
- Report $z = v$ with prob. $p = \frac{e^{\epsilon}}{e^{\epsilon}+k-1}$;
- Otherwise, report any other value $z = \text{Uni}\,(V \setminus \{v\})$ with prob. $q = \frac{1-p}{k-1}$ [5, 6].



Attacker $\mathcal{A}$: Since $p > q$, predict reported value as the true one:

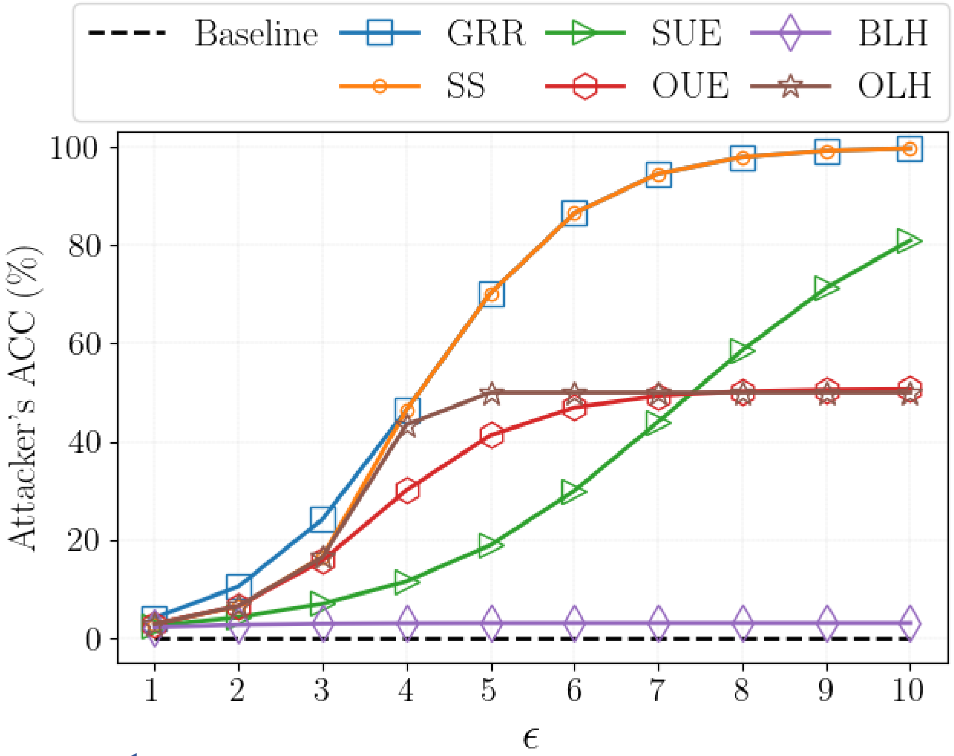
- $\hat{v} = \mathcal{A}(z) = z$.

[5] Warner. *Randomized response: A survey technique for eliminating evasive answer bias*. JASA 1965.
[6] Kairouz *et al*. *Discrete distribution estimation under local privacy*. ICML 2016.

Instance of Value Distinguishability Attack Results

Attacker's ACC w/ domain size $k = 64$ **and** $\epsilon \in \{1,2,\ldots,9,10\}$**.**

More robust to attack



More privacy

Outline

Fake Data Distinguishability Attack

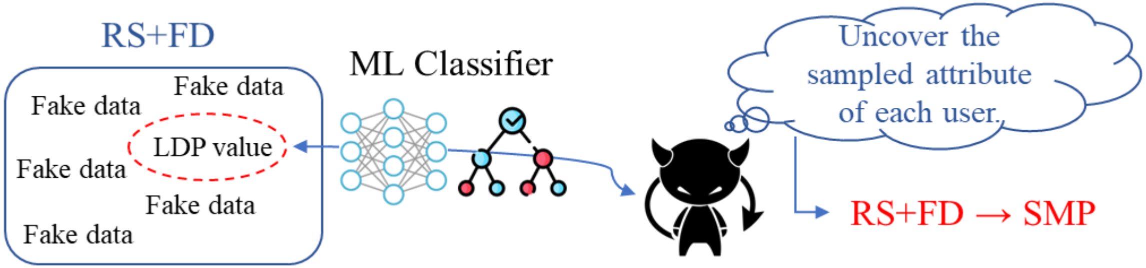
Assumption: Each user has a tuple $\boldsymbol{v} = [v_1, \cdots, v_d]$ of $d \geq 2$ attributes.

LDP mechanism: RS+FD solution.

Adversary's goal: Predict sampled attribute given $\boldsymbol{z} = [z_1, \cdots, z_d]$.

Metric: Attribute Inference Accuracy (AIF-ACC).

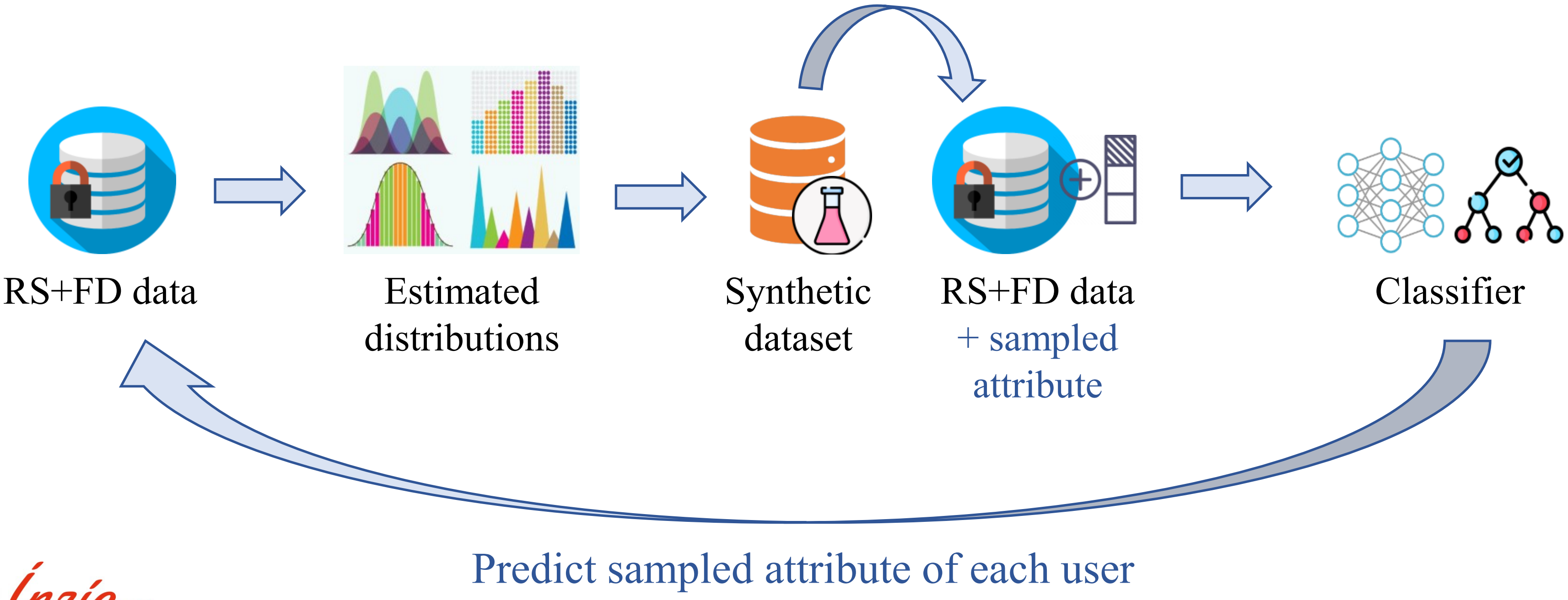
Baseline: Uniform random guess AIF-ACC $= {}^{1}/_{d}$.



Attack Model

No Knowledge (NK) model:

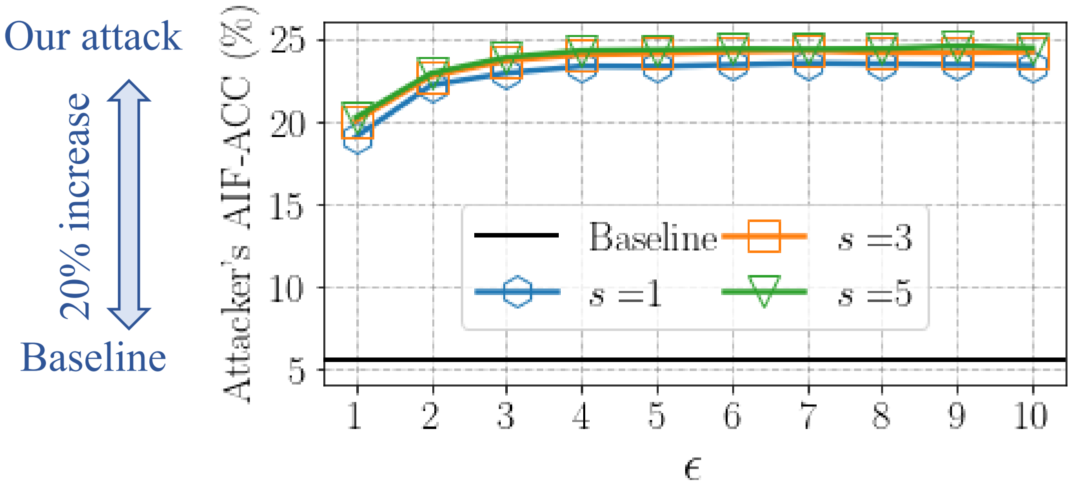
- Training a classifier over *s* synthetic profiles;
- Has knowledge about the RS+FD mechanism and ϵ used by users.



Instance of Fake Data Dinstinguishability Results: RS+FD

Setting:

- Average over 20 runs for stability;
- RS+FD solution with **GRR**;
- Number of synthetic profiles $s \in \{1n, 3n, 5n\}$.



Outline

1. Introduction
2. **Attack-Based Approaches to LDP**
    I. Value Distinguishability;
    II. Fake Data Distinguishability;
    III. **Re-Identification;**
    IV. Countermeasure Solution.
3. Conclusion & Perspectives

Re-Identification Attack

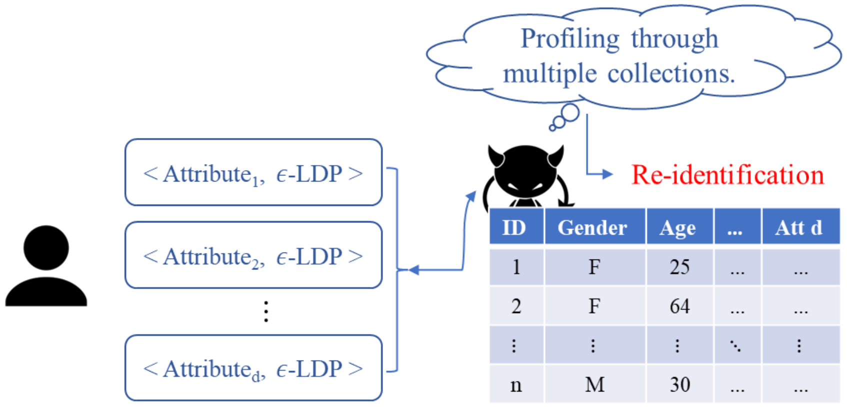
Aassumptions: Collect multidimensional data multiple times (sample different attributes).

LDP mechanism: SMP and RS+FD solutions.

Adversary's goal: Profile and re-identify user in top-$k \in \{1, 10\}$ guesses.

Metric: Re-Identification Accuracy (RID-ACC).

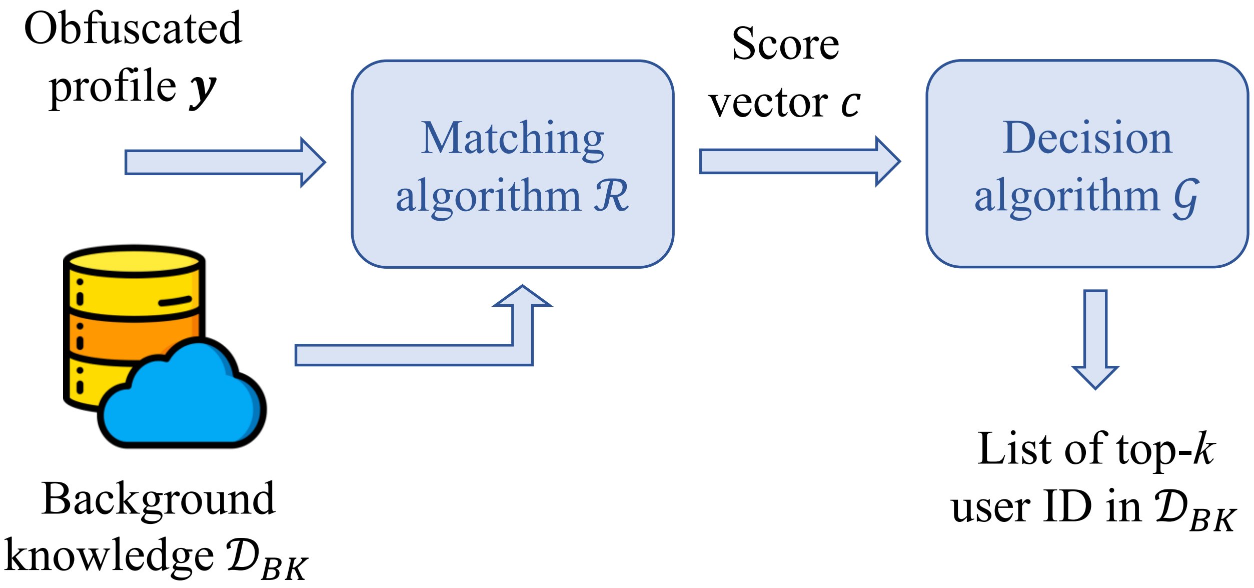
Baseline: Uniform random guess RID-ACC $= {}^{\text{top}-k}/_{n}$.



Attack Model

Adversary has access to side information $\mathcal{D}_{BK}$:

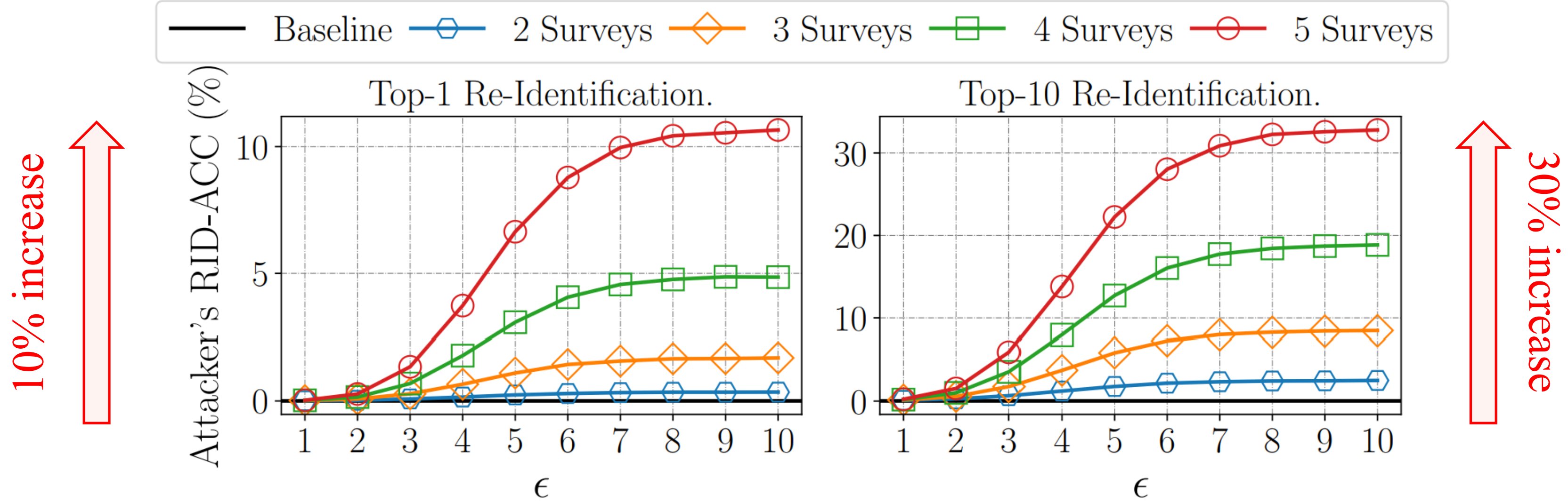
- $\mathcal{R}$: compute distance between inferred profile $\boldsymbol{y}$ and all users in $\mathcal{D}_{BK}$.
- $\mathcal{G}$: takes score vector c and outputs list of top-k guesses.



Instance of Re-Identification Results: SMP

Setting:

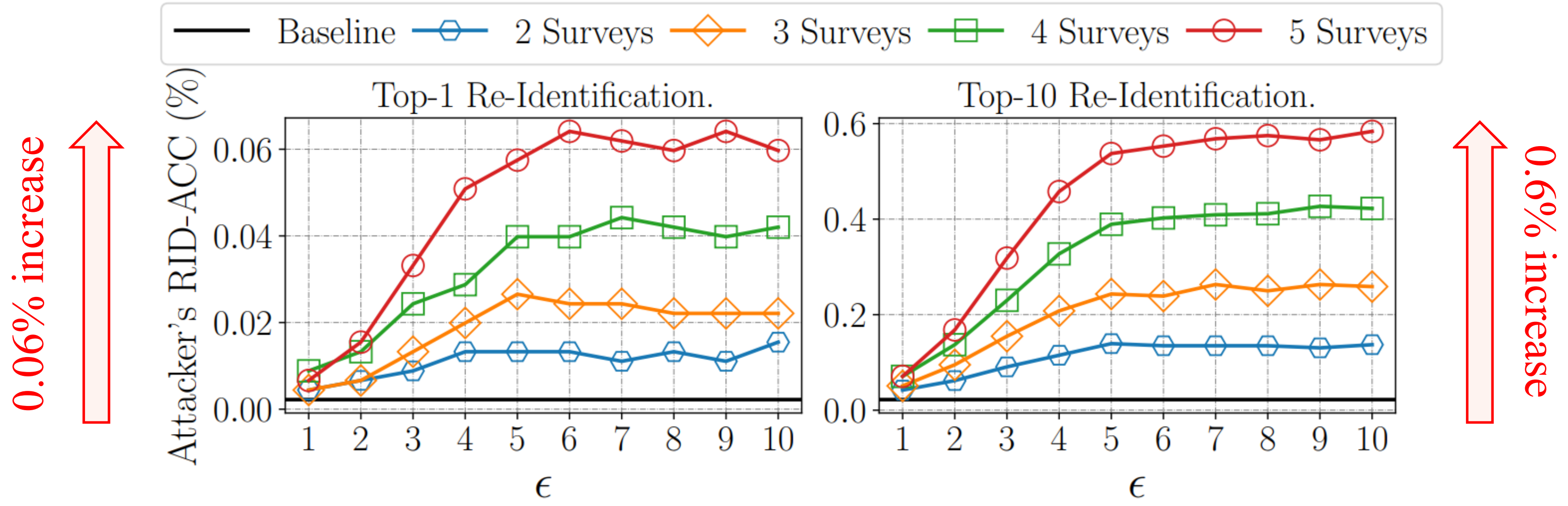
- Average over 20 runs for stability;
- **SMP** solution with **GRR**;
- Number of data collections #Surveys ∈ {1,2, … , 5}.



Instance of Re-Identification Results: RS+FD

Setting:

- Average over 20 runs for stability;
- **RS+FD** solution with **GRR**;
- Number of data collections #Surveys ∈ {1,2, … , 5}.



Outline

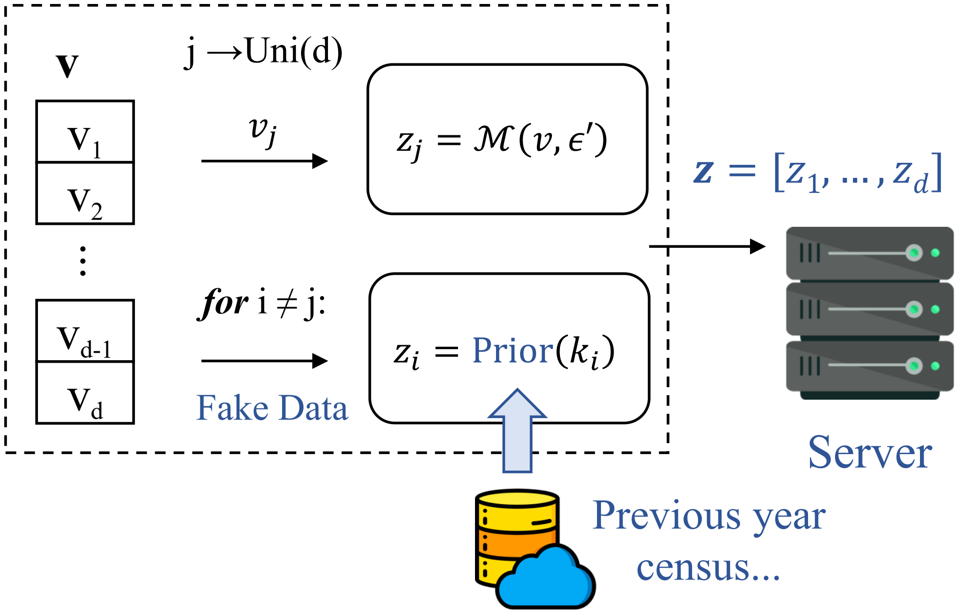
1. Introduction
2. **Attack-Based Approaches to LDP**
    I. Value Distinguishability;
    II. Re-Identification;
    III. Fake Data Distinguishability.
    IV. **Countermeasure Solution.**
3. Conclusion & Perspectives

Countermeasure Solution for Fake Data Distinguishability

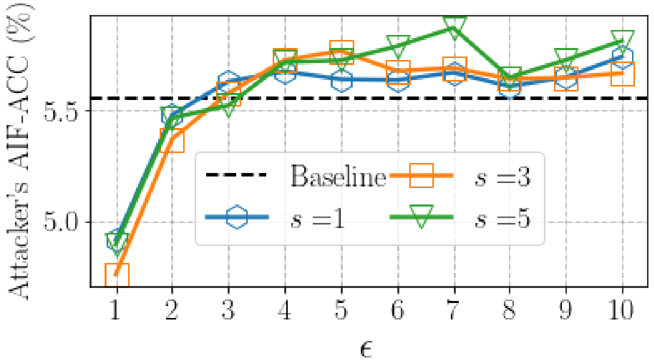
Insights:

- RS+FD is a natural countermeasure to re-identification attacks;
- Chained errors on data distinguishability attacks.
- Uniform fake data of RS+FD is distinguishable.

Random Sampling Plus Realistic Fake Data (RS+RFD)



Attack ~ Baseline



Outline

1. Introduction
2. Attack-Based Approaches to LDP
3. **Conclusion & Perspectives**

Takeaway Messages

Conclusion:

- Identified new privacy threats for LDP mechanisms (*i.e.*, SMP and RS+FD);
- Distinguishability & re-identification attacks;
- RS+FD → Natural countermeasure against re-identification attacks;
- RS+RFD → Countermeasure solution against fake data distinguishability;

Takeaway Messages

Conclusion:

- Identified new privacy threats for LDP mechanisms (*i.e.*, SMP and RS+FD);
- Distinguishability & re-identification attacks;
- RS+FD → Natural countermeasure against re-identification attacks;
- RS+RFD → Countermeasure solution against fake data distinguishability;

Perspectives:

- Use privacy attacks for DP auditing [7];
- Privacy risks of local *d*-privacy mechanisms [8];
- Design of new countermeaure solutions.

[7] Jagielski, Ullman, Oprea. *Auditing differentially private machine learning*. NeurIPS 2020.
[8] Chatzikokolakis *et al. Broadening the scope of differential privacy using metrics*. PETS 2013.

On the Risks of Collecting Multidimensional Data Under Local Differential Privacy

Héber H. Arcolezi
Inria and École Polytechnique (IPP)
heber.hwang-arcolezi@inria.fr

Sébastien Gambs
Université du Québec à Montréal, UQAM
gambs.sebastien@uqam.ca

Jean-François Couchot
Femto-ST Institute, Univ. Bourg. Franche-Comté, CNRS
jean-francois.couchot@univ-fcomte.fr

Catuscia Palamidessi
Inria and École Polytechnique (IPP)
catuscia@lix.polytechnique.fr

PAPER

ARTIFACT

CONTACT

hharcolezi.github.io heber.hwang-arcolezi@inria.fr @hharcolezi