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

ABSTRACT

The private collection of multiple statistics from a population is a fundamental statistical problem. One possible approach to realize this is to rely on the local model of differential privacy (LDP). Numerous LDP protocols have been developed for the task of frequency estimation of single and multiple attributes. These studies mainly focused on improving the utility of the algorithms to ensure the server performs the estimations accurately. In this paper, we investigate privacy threats (re-identification and attribute inference attacks) against LDP protocols for multidimensional data following two state-of-the-art solutions for frequency estimation of multiple attributes. To broaden the scope of our study, we have also experimentally assessed five widely used LDP protocols, namely, generalized randomized response, optimal local hashing, subset selection, RAPPOR and optimal unary encoding. Finally, we also proposed a countermeasure that improves both utility and robustness against the identified threats. Our contributions can help practitioners aiming to collect users' statistics privately to decide which LDP mechanism best fits their needs.

PVLDB Reference Format:
Héber H. Arcolezi, Sébastien Gambs, Jean-François Couchot, and Catuscia Palamidessi. On the Risks of Collecting Multidimensional Data Under Local Differential Privacy. PVLDB, 16(1): XXX-XXX, 2023.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/hharcolezi/risks-ldp.

1 INTRODUCTION

Private and public organizations regularly collect and analyze digital data about their collaborators, volunteers, clients, etc. However, due to the sensitive nature of this personal data, the collection of users' raw data on a centralized server should be avoided. The distributed version of Differential Privacy (DP) [19–21], known as Local DP (LDP) [18, 27], aims to address such a challenge. Indeed, using an LDP mechanism, a user can sanitize her profile locally before transmitting it to the server, which leads to strong privacy protection even if the server used for the aggregation is malicious. The LDP model has a close connection with the concept of randomized response [50], which provides "plausible deniability" to users' reports. For this reason, LDP has been already implemented in large-scale systems by Google [23], Microsoft [14] and Apple [42].

A fundamental task under LDP guarantees is frequency estimation [12, 14, 23, 25, 26, 42, 46, 47], in which the data collector estimates the number of users for each possible value of one attribute based on the sanitized data of the users. More recently, a new line of research started investigating security and privacy threats to LDP protocols [7, 10, 28, 31, 51]. For instance, from a security point of view, some of these works aim at maliciously changing the estimated statistic on the *server-side* with untargeted manipulation attacks [10] or targeted data poisoning attacks [7, 31, 51]. Taking a privacy perspective, Murakami and Takashi [34] were the first ones to investigate privacy risks for the *user* in the LDP model, introducing a relaxed version of LDP that accounts for the risk of re-identification [24, 33–35]. More recently, the work of Gursoy and co-authors [22] introduced a Bayesian adversary to analyze the privacy relationships of LDP protocols for single frequency estimation, in which an adversary can infer the user' true value.

In this paper, we also investigate the **privacy threats for the users** when the server aims to perform frequency estimation of multiple attributes under LDP guarantees [5, 36, 43, 44, 47]. Formally, the profile of each user is characterized by d attributes $\mathcal{A} = \{A_1, A_2, \ldots, A_d\}$, in which each attribute A_j has a discrete domain of size $k_j = |A_j|$, for $j \in [1, d]$. There are n users $\mathcal{U} = \{u_1, \ldots, u_n\}$, and each user u_i, for $i \in [1, n]$, holds a private tuple $\mathbf{v}^{(i)} = \left[v_1^{(i)}, v_2^{(i)}, \ldots, v_d^{(i)}\right]$, in which $v_j^{(i)}$ represents the value of attribute A_j in record $\mathbf{v}^{(i)}$. Thus, for each attribute $A_j \in \mathcal{A}$, for $j \in [1, d]$, the aggregator's goal is to estimate a k_j-bins histogram, including the frequency of all values in A_j.

To the best of our knowledge, for the task considered[1], there are mainly three solutions for satisfying LDP by randomizing the user's tuple $\mathbf{v} = [v_1, v_2, \ldots, v_d]$[2], which are described in the following:

- **Splitting (SPL).** This naïve solution directly splits the privacy budget ϵ over d attributes and report all attributes with $\frac{\epsilon}{d}$-LDP [6, 36, 44, 47].

Proceedings of the VLDB Endowment, Vol. 16, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

[1] This is a different task of joint distribution estimation under LDP guarantees [37, 53].
[2] For the sake of simplicity, we will omit the index notation $\mathbf{v}^{(i)}$ in the analysis as we focus on one arbitrary user u_i here.

- **Sampling (SMP).** Instead of splitting the privacy budget, one state-of-the-art solution allows users to randomly sample a single attribute and report it with ϵ-LDP [6, 36, 44, 47].
- **Random Sampling Plus Fake Data (RS+FD) [5].** One of the weakness of the SMP solution is that it discloses the sampled attribute, which might not be fair to all users (*e.g.*, some users will sample age but others will sample sensitive attribute such as disease). The objective of the RS+FD solution is precisely to enable users to "hide" the sampled attribute (*i.e.*, ϵ-LDP value) by also generating one uniformly random fake data for each non-sampled attribute. Thus, RS+FD creates *uncertainty* on the server-side.

Focusing first on these solutions, we empirically demonstrate through extensive experiments that **the SMP solution is vulnerable to re-identification attacks** by collecting users' multidimensional data several times with ϵ values commonly used by industry nowadays [13, 38]. For instance, assume that a user has multiple mobile applications each surveying the user with the SMP solution on different attributes. Another possible scenario is the situation in which the same mobile application is used on a regular basis but surveys users with different attributes. This enables the user to sample a (possibly different) attribute each time, thus resulting in sending their sampled attribute along with their ϵ-LDP report. Nevertheless, we show that an adversary who can see every tuple containing $\langle$sampled attribute and ϵ-LDP report$\rangle$ can construct a partial or complete profile of the user, which can possibly be unique (or in a small anonymity set of k individuals) in the population considered. Therefore, once the set of k individuals (referred to as top-k in this paper) is re-identified, one can leverage this through well-known re-identification attacks (*e.g.*, homogeneity) [11, 29, 32, 39–41].

More specifically, to attack the SMP solution, our adversarial analysis focuses on the reduced **"plausible deniability"** [16, 50] of using the whole privacy budget ϵ to report a single attribute. In this situation, the adversary has a higher chance to retrieve the users' true value for each data collection performed, leading to partial and/or complete profiles. However, this depends on the LDP protocol being used as the encoding and randomization vary across them [22, 47]. In our experiments, we have assessed five widely used LDP protocols for frequency estimation (*a.k.a.* frequency oracle protocols [48, 49]), namely Generalized Randomized Response (GRR) [25, 26], Optimal Local Hashing (OLH) [47], Subset Selection (SS) [45, 52] and two Unary Encoding (UE) protocols (Basic One-time RAPPOR [23] and Optimal UE [47]). To assess the risks of re-identification we have also considered two privacy models, standard LDP and the relaxed version of LDP developed in [34] for measuring re-identification risks.

Moreover, we observe that **since the RS+FD solution generates fake data uniformly at random in [5, 43], it is possible to uncover the sampled attribute of users** in certain conditions. In this context, we evaluated the effectiveness of the RS+FD solution on hiding the sampled attribute to the aggregator by varying the privacy budget ϵ, the LDP protocol and the fake data generation procedure. In particular if the aggregator is able to break RS+FD into the SMP solution, the RS+FD solution might also be subject to the same vulnerability to re-identification attacks. Thus, we have proposed three attack models to uncover the sampled attribute of users using the RS+FD solution and evaluated its risks to re-identification attacks. Lastly, as shown in our results, RS+FD is, to some extent, a natural countermeasure to re-identification attacks. Building on this, we have designed a stronger countermeasure that adapts RS+FD to generate fake data following non-uniform distributions, **almost fully preventing the inference of the sampled attribute while preserving utility**.

To summarize, this paper makes the following contributions:

- We investigate privacy threats against LDP protocols for multidimensional data following two state-of-the-art solutions for frequency estimation of multiple attributes, SMP [6, 36, 44, 47] and RS+FD [5], providing insightful adversarial analysis that can help in LDP protocol selection.
- We demonstrate through extensive experiments that the SMP solution is vulnerable to re-identification attacks due to the disclosure of the sampled attribute and lower "plausible deniability" when using the whole privacy budget to report a single attribute.
- We propose three attack models to predict the sampled attribute of users that collect multidimensional data with the RS+FD solution with about a 2-20 fold increment over a random baseline model.
- We show through empirical results that the RS+FD solution can prevent (to some extent) re-identification attacks
- Finally, we present an improvement of the RS+FD as a countermeasure solution, which improves both privacy and utility. More precisely, our solution achieves less estimation error on multidimensional frequency estimation and almost fully prevents the inference of the sampled attribute.

Outline. The remainder of this paper is organized as follows. In Section 2, we review the privacy models, the LDP protocols and solutions for collecting multidimensional data investigated in this paper. Afterwards in Section 3, we present the system overview and adversarial setting for both SMP and RS+FD solutions. In Section 4, we present our experimental evaluation and discuss our results before in Section 5 presenting an improvement of the RS+FD as a countermeasure. Finally in Section 6, we review related work before concluding with future perspectives of this work in Section 7.

2 PRELIMINARIES

In this section, we first introduce the privacy models we consider, *i.e.*, LDP and a relaxed version of LDP developed in [34]. Then, we briefly review state-of-the-art LDP protocols for frequency estimation before describing three solutions for frequency estimation of multiple attributes.

2.1 Privacy Models

In this paper, we use LDP (Local Differential Privacy) [18, 27] as the privacy model considered, which is formalized as:

Definition 1 (ϵ-Local Differential Privacy). *A randomized algorithm $\mathcal{M}$ satisfies ϵ-local-differential-privacy (ϵ-LDP), where $\epsilon > 0$, if for any pair of input values $v_1, v_2 \in Domain(\mathcal{M})$ and any possible output y of $\mathcal{M}$:*

$$\Pr[\mathcal{M}(v_1) = y] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(v_2) = y]. \quad (1)$$

In essence, LDP guarantees that it is unlikely for the data aggregator to reconstruct the data source regardless of the prior knowledge. The privacy budget ϵ controls the privacy-utility trade-off for which lower values of ϵ result in tighter privacy protection. Similar to the centralized model of DP, LDP also enjoys several important properties, such as immunity to post-processing ($F(\mathcal{M})$ is ϵ-LDP for any function F) and composability [21].

In addition, we also use a recent relaxation of LDP known as PIE (Personal Information Entropy) [34], which aims at quantifying the re-identification risks in the local model. PIE is defined as the mutual information between user U (random variable representing a user in $\mathcal{U}$) and perturbed data Y (random variable representing perturbed data) as PIE $= I(U;Y)$ $(bits)$. As $I(U;Y)$ approaches 0, almost no information about user U can be obtained through the perturbed data Y. Based on this observation, the authors in [34] defined the privacy metric $(\mathcal{U},\alpha)$-PIE privacy that guarantees that the PIE is upper bounded by a parameter α for a set of users $\mathcal{U}$. More formally, let $p_{U,V}$ be the joint distribution of U and V (random variable representing personal data), Ψ be a finite set of all humans and $\mathcal{U} \subseteq \Psi$ be a finite set of users reporting attribute A_j of size k_j, the definition of $(\mathcal{U},\alpha)$-PIE privacy is:

Definition 2 (($\mathcal{U},\alpha$)-PIE privacy [34]). *Let $\mathcal{U} \subseteq \Psi$ and $\alpha \in \mathbb{R}_{\geq 0}$. An obfuscation mechanism $\mathcal{M}$ provides $(\mathcal{U},\alpha)$-PIE privacy if*

$$\underset{p_{U,V}}{sup}\ I(U;Y) \leq \alpha\ (bits). \tag{2}$$

Since the inequality in Eq. (2) holds for any distribution $p_{U,V}$, the PIE is upper bounded by α irrespective of the adversary's background knowledge [34]. The parameter α plays a role similar to the privacy budget ϵ in LDP and can be selected by fixing the lowest possible Bayes error probability $\beta_{U|S}$ (given a score vector S) as:

Corollary 1. *(Bayes error and PIE privacy [34]). Let $\mathcal{U} \subseteq \Psi$ and $\alpha \in \mathbb{R}_{\geq 0}$. If an obfuscation mechanism $\mathcal{M}$ provides $(\mathcal{U},\alpha)$-PIE privacy and if U is uniformly distributed (*i.e.*, one data per user), then*

$$\beta_{U|S} \geq 1 - \frac{\alpha + 1}{\log_2(n)}. \tag{3}$$

Lastly, the relationship between LDP and PIE is:

Proposition 1. *(LDP and PIE [34]). If an obfuscation mechanism $\mathcal{M}$ provides ϵ-LDP, then it provides $((\mathcal{U},\alpha)$-PIE privacy) for any $\mathcal{U} \subseteq \Psi$ such that $|\mathcal{U}| = n$, where*

$$\alpha = min\left\{\epsilon \log_2(e), \epsilon^2 \log_2(e), \log_2(n), \log_2(k_j)\right\}. \tag{4}$$

As stated in [34], Proposition 1 holds for any LDP protocol.

2.2 LDP protocols

In this subsection, we review five state-of-the-art LDP protocols, which enables the aggregator to estimate the frequency of any value $v_i \in A_j$, for $i \in [1, k_j]$, under LDP guarantees.

2.2.1 Generalized Randomized Response.

Randomized response (RR) is the classical technique for achieving LDP, which is a surveying technique proposed by Warner [50] to provide "plausible deniability" for individuals responding to embarrassing (binary) questions in a survey. The Generalized RR (GRR) [25, 26] protocol extends RR to the case of $k_j \geq 2$ while satisfying ϵ-LDP. Given a value $v_i \in A_j$, for $i \in [1, k_j]$, *GRR(v_i)* outputs the true value with probability p, and any other value $v \in A_j \setminus \{v_i\}$ with probability $1 - p$. More formally, the perturbation function is:

$$\forall y \in A_j: \quad \Pr[y = a] = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon + k_j - 1}, & \text{if } a = v \\ q = \frac{1}{e^\epsilon + k_j - 1}, & \text{otherwise,} \end{cases}$$

in which y is the perturbed value sent to the aggregator from a user with input value v. The GRR protocol has been proved to satisfy ϵ-LDP since $\frac{p}{q} = e^\epsilon$. To estimate the normalized frequency of $v_i \in A_j$, for $i \in [1, k_j]$, one counts how many times v_i is reported, expressed as $C(v_i)$, and then computes [47]:

$$\hat{f}(v_i) = \frac{C(v_i) - nq}{n(p - q)}, \tag{5}$$

in which n is the total number of users. In [47], it was proven that Eq. (5) is actually an unbiased estimator (*i.e.*, $\mathbb{E}(\hat{f}(v_i)) = f(v_i)$).

2.2.2 Optimal Local Hashing.

Local hashing (LH) protocols can handle a large domain size k_j by first using hash functions to map an input value to a smaller domain of size g_j (typically $g_j \ll k_j$), and then applying GRR to the hashed value in the smaller domain.

The authors in [47] have proposed Optimal LH (OLH), which selects $g_j = e^\epsilon + 1$. Given a value $v_i \in A_j$, for $i \in [1, k_j]$, in OLH, one reports $\langle H, GRR(H(v_i))\rangle$ in which H is randomly chosen from a family of universal hash functions that hash each value in A_j to $[g_j] = \{1, \ldots, g_j\}$, which is the domain that GRR($\cdot$) will operate on. The hash values will remain unchanged with probability p' and switch to a different value in $[g_j]$ with probability q', as:

$$\forall y \in [g_j]: \quad \Pr[y = (H, a)] = \begin{cases} p' = \frac{e^\epsilon}{e^\epsilon + g_j - 1}, & \text{if } a = H(v) \\ q' = \frac{1}{e^\epsilon + g_j - 1}, & \text{otherwise,} \end{cases}$$

in which y is the hash function and perturbed value sent to the aggregator from a user with input value v. From this, the aggregator can obtain the unbiased estimation of $v_i \in A_j$, for $i \in [1, k_j]$, with Eq. (5) by setting $p = p'$ and $q = \frac{1}{g_j} \cdot p' + \left(1 - \frac{1}{g_j}\right) \cdot q' = \frac{1}{g_j}$ [47].

2.2.3 Subset Selection.

The main idea of ω-Subset Selection (ω-SS) [45, 52] is to randomly select ω items within the input domain to report a subset of values (*i.e.*, $\Omega \subseteq A_j$). The user's true value $v_i \in A_j$, for $i \in [1, k_j]$, has higher probability of being included in the subset Ω, compared to other values in $A_j \setminus \{v_i\}$ that are sampled uniformly at random (without replacement). The optimal subset size $\omega = |\Omega|$ that minimizes the variance is $\omega = \frac{k_j}{e^\epsilon + 1}$ [45, 52].

Given a value $v_i \in A_j$, for $i \in [1, k_j]$, the ω-SS protocol starts by initializing an empty subset Ω. Afterwards, the true value v_i is added to Ω with probability $p = \frac{\omega e^\epsilon}{\omega e^\epsilon + k_j - \omega}$. Finally, it adds values to Ω as follows [22, 45, 52]:

- If v_i has been added to Ω in the previous step, then $\omega - 1$ values are sampled from $A_j \setminus \{v_i\}$ uniformly at random (without replacement) and are added to Ω;
- If v_i has not been added to Ω in the previous step, then ω values are sampled from $A_j \setminus \{v_i\}$ uniformly at random (without replacement) and are added to Ω.

From this, the aggregator can obtain the unbiased estimation of $v_i \in A_j$, for $i \in [1, k_j]$, with Eq. (5) by setting $p = \frac{\omega e^{\epsilon}}{\omega e^{\epsilon} + k_j - \omega}$ and $q = \frac{\omega e^{\epsilon}(\omega - 1) + (k_j - \omega)\omega}{(k_j - 1)(\omega e^{\epsilon} + k_j - \omega)}$ [45, 52].

2.2.4 *Unary Encoding Protocols.*

Unary encoding (UE) protocols interpret the user's input $v_i \in A_j$, for $i \in [1, k_j]$ as a one-hot k_j-dimensional vector. More precisely, $B = UE(v_i)$ is a binary vector with only the bit at the position v_i sets to 1 and the other bits set to 0. One well-known UE-based protocol is the Basic One-time RAPPOR [23], hereafter referred to as symmetric UE (SUE) [47], which randomizes the bits from B independently with probabilities:

$$\forall i \in [1, k_j]: \quad \Pr[B'_i = 1] = \begin{cases} p = \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}, & \text{if } B_i = 1 \\ q = \frac{1}{e^{\epsilon/2}+1}, & \text{if } B_i = 0. \end{cases} \tag{6}$$

Afterwards, the client sends B' to the aggregator. More recently, to minimize the variance of the SUE protocol, the authors in [47] proposed Optimal UE (OUE), which selects probabilities $p = \frac{1}{2}$ and $q = \frac{1}{e^{\epsilon}+1}$ in Eq. (6) asymmetrically (*i.e.*, $p + q \neq 1$). The estimation method used in Eq. (5) applies equally to both SUE and OUE protocols, in which both satisfy ϵ-LDP for $\epsilon = ln\left(\frac{p(1-q)}{(1-p)q}\right)$ [23, 47].

2.3 Multidimensional Frequency Estimation

Let n be the total number of users, $d \geq 2$ be the total number of attributes, $\mathbf{k} = [k_1, k_2, \ldots, k_d]$ be the domain size of each attribute, $\mathcal{M}$ be a local randomizer and ϵ be the privacy budget. Each user holds a tuple $\mathbf{v} = [v_1, v_2, \ldots, v_d]$, (*i.e.*, a private discrete value per attribute). The two next subsections describes the SPL, SMP and RS+FD solutions for frequency estimation of multiple attributes.

2.3.1 *Standard Solutions.*

In the centralized setting of DP, the privacy budget ϵ is split when answering multiple queries (*i.e.*, SPL solution). Previous works in the local setting either followed this traditional SPL approach or have proposed to split the users into d disjoint groups of n/d users to answer a single query (*i.e.*, SMP solution) [6, 36, 44, 47]. More specifically:

- **SPL.** On the one hand, due to the sequential composition theorem [21], users can split the privacy budget ϵ over the number of attributes d and send all randomized values y_j, for $j \in [1, d]$, with $\frac{\epsilon}{d}$-LDP to the aggregator (*i.e.*, a tuple $\mathbf{y} = [y_1, y_2, \ldots, y_d]$). However, this naïve SPL solution leads to high estimation error [6, 36, 44, 47].
- **SMP.** Instead of splitting the privacy budget ϵ, one state-of-the-art solution is to divide the users into d groups to answer a single attribute [6, 36, 44, 47]. More precisely, each user samples a single attribute $j \in [1, d]$ (we slightly abuse the notation and use j for A_j) at random and uses all the privacy budget to send it with ϵ-LDP. In this case, each user tells the aggregator which attribute is sampled, and what is the perturbed value for it ensuring ϵ-LDP (*i.e.*, $\langle j, y_j \rangle$).

2.3.2 *Random Sampling Plus Fake Data (RS+FD).*

Because the SMP solution discloses the sampled attribute, one can say that it is not fair to all users (*e.g.*, some users will sample age while others will sample disease). To address this issue, the recently proposed RS+FD [5] solution is composed of two steps, namely local randomization and fake data generation. More precisely, each user samples a unique attribute uniformly at random $j = Uniform([1, d])$ and uses an ϵ-LDP protocol to sanitize its value v_j. Next, for each non-sampled attribute $i \in [1, d] \setminus \{j\}$, the user generates random fake data following A_i. Finally, each user sends the (LDP or fake) value of each attribute to the aggregator (*i.e.*, a tuple $\mathbf{y} = [y_1, y_2, \ldots, y_d]$). In this manner, the sampling result is not disclosed to the aggregator, thus increasing the *uncertainty*. For this reason, to satisfy ϵ-LDP, following the parallel composition theorem [21] and the amplification by sampling result [30], RS+FD utilizes an amplified privacy budget $\epsilon' = \ln(d \cdot (e^{\epsilon} - 1) + 1)$ for the sampled attribute [5].

With the RS+FD solution, the estimator should remove the bias introduced by the local randomizer $\mathcal{M}$ and uniform fake data. In [5], the authors used GRR and OUE as LDP protocols within the RS+FD solution, which results in RS+FD[GRR], RS+FD[OUE-z] and RS+FD[OUE-r]. We briefly recall how these three protocols, generalizing OUE to UE as one can select either SUE or OUE (*cf.* Section 2.2.4) as local randomizers [5, 43].

For all three protocols, on the *client-side*, each user randomly samples an attribute j and uses $\mathcal{M}$ to sanitize the value v_j with an amplified privacy parameter $\epsilon' = \ln(d \cdot (e^{\epsilon} - 1) + 1)$. Next, the fake data generation procedure and the unbiased estimator for the frequency of each value $v_i \in A_j$, for $i \in [1, k_j]$, are as follows:

- **RS+FD[GRR] [5].** For each non-sampled attribute $i \in [1, d] \setminus \{j\}$, the user generates fake data uniformly at random according to the domain size k_i. On the *server-side*, the unbiased estimator for this protocol is: $\hat{f}(v_i) = \frac{C(v_i)dk_j - n(d-1+qk_j)}{nk_j(p-q)}$, in which $C(v_i)$ is the number of times v_i has been reported, $p = \frac{e^{\epsilon'}}{e^{\epsilon'}+k_j-1}$ and $q = \frac{1-p}{k_j-1}$.
- **RS+FD[UE-z] [5].** For each non-sampled attribute $i \in [1, d] \setminus \{j\}$, the user generates fake data by applying an UE protocol to zero-vectors (*i.e.*, $[0, 0, \ldots, 0]$) of size k_i. On the *server-side*, the unbiased estimator for this protocol is: $\hat{f}(v_i) = \frac{d(C(v_i) - nq)}{n(p-q)}$, in which $C(v_i)$ is the number of times v_i has been reported and parameters p and q can be selected following the SUE [23] or OUE [47] protocols.
- **RS+FD[UE-r] [5].** For each non-sampled attribute $i \in [1, d] \setminus \{j\}$, the user generates fake data by applying an UE protocol to one-hot-encoded fake data (uniform at random) of size k_i. On the *server-side*, the unbiased estimator for this protocol is: $\hat{f}(v_i) = \frac{C(v_i)dk_j - n[qk_j + (p-q)(d-1) + qk_j(d-1))]}{nk_j(p-q)}$, in which $C(v_i)$ is the number of times v_i has been reported and parameters p and q can be selected following the SUE [23] or OUE [47] protocols.

3 SYSTEM OVERVIEW & PRIVACY THREATS

In this section, we first describe the system overview and adversary model. Afterwards, we present our adversarial analyses of the solutions SMP and RS+FD.

3.1 System Overview

We consider the situation in which a (possibly untrusted) server collects users' multidimensional data for frequency estimation under ϵ-LDP guarantees multiple times. More precisely, in each data collection (*i.e.*, survey), the server can select a different number of attributes d. For instance, through a mobile app the server may collect private frequency estimation for different users' demographic data and different application usage (*e.g.*, how much time spent on the application, preferred widget, etc). Users could be encouraged to share their private data through the exchange of discount coupons, statistics to compare usage with other users, etc. For the sake of simplicity, we assume that the set of users $\mathcal{U}$ is unique across all surveys, although this can be relaxed in real-life allowing users to opt-in or opt-out of a given survey.

Adversary model. Following the LDP assumptions [18, 27], we assume that the server knows the users' pseudonymized IDs, *but not their private data or their real identity.* This also implies that the server has no knowledge about the real data distributions. However, we assume that the server might have some background knowledge $\mathcal{D}_{BK}$ coming from public available source, such as Census data [2]. This background knowledge could for instance contain partial or complete profiles of users along with their true identities. Thus, the adversary could be for example the server itself, an attacker who intercepts the communication between the client and the server (*e.g.*, through a man-in-the-middle attack) or a third-party analyst with whom the server may have shared the collected data.

3.2 Attacking SMP: Plausible Deniability and Risks of Re-Identification

Plausible deniability. Let v_y be an embarrassing value of $A_j = \{v_y, v_n\}$ (*e.g.*, a value "Yes" for an attribute A_j denoting whether someone cheated on their partner). As long as $\Pr\left[\mathcal{M}(v_y) = v_y\right] < 1$, the user can deny to have $A_j = v_y$.

The LDP protocols of Section 2.2 are based on RR [50], which provides "plausible deniability" for users' reports. However, increasing ϵ to improve utility of LDP protocols compromises the "plausible deniability" of the users' reports. Indeed, common ϵ values used daily by users in high-scale industrial systems nowadays range from small $\epsilon < 1$ to high values $\epsilon \geq 8$ [13, 38]. Thus, we conduct an adversarial analysis to the SMP solution (*cf.* Section 2.3.1) in which the user randomly samples a single attribute among $d \geq 2$ ones and uses the whole privacy budget ϵ to report it. Consequently, since the whole privacy budget will be allocated to a single attribute, the "plausible deniability" for this attribute will be lower, which can lead an attacker to predict the users' true value as the most likely value after randomization (see details in Section 3.2.1). In this case, if many surveys are proposed by the server to the same set of users with possibly different number of attributes d (*e.g.*, demographic, users' preference, application usage, etc), an attacker knowing the tuple ⟨sampled attribute and ϵ-LDP report⟩ will be able to profile each user throughout time. Therefore, once a profile of the target user is deduced, the attacker could use her background knowledge $\mathcal{D}_{BK}$ to possibly re-identify the user in a unique population [29, 32, 39–41] and infer all other available demographic attributes. The next four subsections analyze the "plausible deniability" of LDP protocols in a single collection and in multiple collections, and describes the proposed re-identification attack models.

3.2.1 *Plausible Deniability of LDP protocols.*

Given a user's true value $v_i \in A_j$, for $i \in [1, k_j]$, different LDP protocols $\mathcal{M}$ have different type of output $y_i = \mathcal{M}(v_i, \epsilon)$ [47]. For instance, UE protocols output unary encoded vectors, ω-SS outputs a subset Ω of ω non-encoded values and so on (*cf.* Section 2.2). Thus, for each user $u_i \in \mathcal{U}$, for $i \in [1, n]$, given y_i, the adversary's goal is to predict v_i, which is denoted as $\hat{v}_i$. The attacker's accuracy (ACC) for LDP protocols is measured by the number of correct predictions $v = \hat{v}$ over the number of users n: $ACC_{FO}(\%) = 100 \cdot \frac{\sum_{i=1}^{n} f(v_i, \hat{v}_i)}{|\mathcal{U}|}$, in which $f(v, \hat{v}) = 1$ if $v = \hat{v}$ and 0 otherwise. Following the "plausible deniability" intuition and the fact that for all LDP protocols the probability p of reporting the true value v_i (or bit i) is higher than any other value $v \in A_j \setminus \{v_i\}$, we have the following:

Plausible Deniability of GRR. Since no specific encoding is used with GRR, the most likely value after randomization is the user's u_i own true value v. Thus, an attacker can assume that the reported value y is the true one (*i.e.*, $\hat{v} = y$), which gives on expectation an $ACC_{GRR}(\%) = 100 \cdot \frac{e^{\epsilon}}{e^{\epsilon} + k_j - 1}$. This result has also been observed with the Bayesian adversary of [22].

Plausible Deniability of OLH. Since the output of OLH for user u_i is the hash function H_i used to hash the user's value v and the hashed value $h_i = H_i(v)$, the most likely value after randomization is one within the subset of all values $v \in A_j$ that hash to h_i (*i.e.*, $A_{j_H} = \{v | v \in A_j, H_i(v) = h_i\}$). Thus, the attacker's best guess is a random choice $\hat{v} = Uniform\left(A_{j_H}\right)$. This attack strategy has also been observed and then formalized in [22]. Using the standard parameter $g_j = e^{\epsilon} + 1$ in OLH, on expectation, one achieves: $ACC_{OLH}(\%) = 100 \cdot \frac{1}{2 \cdot \max\left(\frac{k_j}{e^{\epsilon}+1}, 1\right)}$ [22].

Plausible Deniability of ω-SS. Since the output of ω-SS for user u_i is a set $\Omega \subseteq A_j$, the most likely value after randomization is one within the subset Ω. Thus, the attacker's best guess is a random choice $\hat{v} = Uniform(\Omega)$. This attack strategy has also been observed and then formalized in [22]. Using the standard parameter $\omega = \frac{k_j}{e^{\epsilon}+1}$ [45, 52] in ω-SS, on expectation, one achieves: $ACC_{\omega\text{-}SS}(\%) = 100 \cdot \frac{e^{\epsilon}+1}{2k_j}$ [22].

Plausible Deniability of UE protocols. Since the output of UE protocols for user u_i is a sanitized unary encoded vector B_i of size k_j, there are three possibilities: 1) a single bit b in B is set to 1, in which the attacker's best guess is to predict the bit as the true value as $\hat{v} = B_b$; 2) more than one bit in B is set to 1, in which the attacker's best guess is a random choice of the bits set to 1 as $\hat{v} = Uniform\left(\{b | b \in [1, k_j] \text{ if } B_b = 1\}\right)$; and 3) no bit in B is set to 1, in which the attacker's best guess is a random choice of the domain $\hat{v} = Uniform\left(A_j\right)$. This attack strategy has also been observed and then formalized in [22]. Therefore, on expectation, the attacker's accuracy for SUE is [22]: $ACC_{SUE}(\%) = 100 \cdot \frac{1}{k_j(e^{\epsilon/2}+1)} \cdot \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}^{k_j-1} + \sum_{i=1}^{k_j} \frac{e^{\epsilon/2}}{(e^{\epsilon/2}+1)i} \cdot \text{Bin}\left(i-1; k_j-1, \frac{1}{e^{\epsilon/2}+1}\right)$, in which Bin(.) denotes a Binomial distribution with $k_j - 1$ trials, success probability $\frac{1}{e^{\epsilon/2}+1}$ and exactly $i-1$ successes. On the other hand, on expectation, the attacker's accuracy for OUE is [22]: $ACC_{OUE}(\%) = 100 \cdot \frac{1}{2k_j} \cdot \frac{e^{\epsilon}}{e^{\epsilon}+1}^{k_j-1} + \sum_{i=1}^{k_j} \frac{1}{2i} \cdot \text{Bin}\left(i-1; k_j-1, \frac{1}{e^{\epsilon}+1}\right)$.

3.2.2 *Plausible Deniability on Multiple Data Collections: Uniform Privacy Metric.*

When collecting multidimensional data with the SMP solution multiple times, the server could implement that all users sample attributes without replacement. This way, each user will randomly select a new attribute in each data collection (*i.e.*, survey), ensuring a uniform privacy metric across all users. Since for all LDP protocols the expected ACC_{FO} depends on ϵ and k_j, our analysis focuses on a generic LDP protocol here. Therefore, depending on the LDP protocol, the expected ACC with uniform privacy metric after d surveys (*i.e.*, collecting d attributes per user), denoted as ACC_{FO}^{U}, now follows:

$$ACC_{FO}^{U}(\%) = 100 \cdot \prod_{j=1}^{d} ACC_{FO}\left(\epsilon, k_j\right). \qquad (7)$$

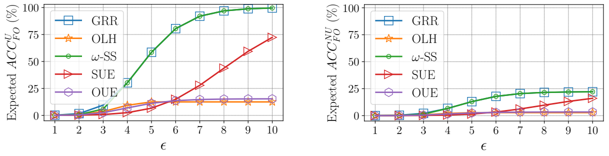
3.2.3 *Plausible Deniability on Multiple Data Collections: Non-Uniform Privacy Metric.*

On the other hand, when collecting multi-dimensional data with the SMP solution multiple times, the server can allow users to sample attributes with replacements in each data collection (*i.e.*, survey). In case of a repeated attribute (*e.g.*, demographic attribute), the user simply reports the previous randomized value (*a.k.a.* memoization [6, 14, 23]). This way, users will have a non-uniform privacy metric. Therefore, depending on the LDP protocol, the expected ACC with non-uniform privacy metric after d surveys, denoted as ACC_{FO}^{NU}, now follows:

$$ACC_{FO}^{NU}(\%) = 100 \cdot \prod_{j=1}^{d} \frac{d+1-j}{d} ACC_{FO}\left(\epsilon, k_j\right), \qquad (8)$$

in which Eq. (8) denotes the overall attacker's accuracy only considering users that reports a different attribute in each survey (*i.e.*, all d attributes). This is due to the fact that we are interested in analyzing the risks of re-identification by using the unique combination of the users' collected attributes in the background knowledge $\mathcal{D}_{BK}$.

Analytical analysis of expected ACC. In Fig. 1, we illustrate the expected ACC_{FO}^{U} (left-side plot) following Eq. (7) and the ACC_{FO}^{NU} (right-side plot) following Eq. (8) of each LDP protocol with the following parameters (taken from Section 4): $\epsilon = [1, 2, 3, \ldots, 10]$, $d = 3$, $\mathbf{k} = [74, 7, 16]$, $n = 45,222$, $\#surveys = d$. From Fig. 1, one can notice that GRR, ω-SS and SUE have the highest attacker's accuracy, which would enable an adversary to infer a complete profile of d attributes after #surveys. Moreover, allowing users to have non-uniform privacy metrics in the plot (b), minimizes the attacker's accuracy to infer complete profiles as the probability of selecting different attributes in all d surveys is $\frac{d!}{d^d}$.



(a) Uniform privacy metric. **(b) Non-uniform privacy metric.**

Figure 1: Analytical attacker's accuracy when collecting multidimensional data (d = 3) with the SMP solution multiple times ($\#surveys$ = 3) with attributes' domain size $\mathbf{k}$ = [74, 7, 16]: (a) uniform privacy metric across users with Eq. (7) and (b) non-uniform privacy metric across users with Eq. (8).

3.2.4 *Re-Identification Attack Models.*

Following the system overview of Section 3.1, we consider two re-identification attack models: **full-knowledge re-identification (FK-RI)** and **partial knowledge re-identification (PK-RI)**, that we detail in the following. The first FK-RI model considers that the attacker has access to the complete background knowledge $\mathcal{D}_{BK}$ to re-identify the target user. The latter PK-RI model considers that the attacker only has access to a subset $\mathcal{D}_{PK} \subseteq \mathcal{D}_{BK}$ for her re-identification attack. The re-identification success of both FK-RI and PK-RI models will depend on the results of Sections 3.2.2 and 3.2.3 to accurately profile the target user, which is impacted by the LDP protocols considered.

More precisely, after #surveys, the attacker will have a profile $\mathbf{y}_i$ of at most #surveys sanitized values for the target user $u_i \in \mathcal{U}$. The number of attributes inferred per target user depends on the setting used (*i.e.*, uniform or non-uniform privacy metrics). Therefore, the re-identification attack starts with a matching algorithm $\mathcal{R}$, which takes as input the sanitized profile $\mathbf{y}_i$ and the background knowledge $\mathcal{D}_{BK}$ (or $\mathcal{D}_{PK}$ for PK-RI), and outputs a score $c_i \in \mathbb{R}$. More precisely, the score c_i measures the distance between the target $\mathbf{y}_i$ and all samples $\mathbf{r} \in \mathcal{D}_{BK}$. Since the LDP protocols from Section 2.2 do not have a notion of "distance" when randomizing a value, when an attribute in $\mathbf{y}_i \neq \mathbf{r}$ the distance is 1 and 0 otherwise. A smaller distance between $\mathbf{y}_i$ and a profile in $\mathcal{D}_{BK}$ indicates that is highly likely that $\mathbf{y}_i$ has been re-identified through the uniqueness combination of #surveys attributes [29, 32, 39–41]. Finally, a decision algorithm $\mathcal{G}$ based on the distances computed outputs a list of top-k possible profiles (or IDs) in $\mathcal{D}_{BK}$ that corresponds to the target user $u_i \in \mathcal{U}$. The attacker's re-identification accuracy (RID-ACC) is measured by the number of correct re-identification $u_{id} = \hat{u}_{id}$ over the number of users n: $RID\text{-}ACC(\%) = 100 \cdot \frac{\sum_{i=1}^{n} f(u_{id_i}, \hat{u}_{id_i})}{|\mathcal{U}|}$, in which $f\left(u_{id}, \hat{u}_{id}\right) = 1$ if $u_{id} = \hat{u}_{id}$ and 0 otherwise.

3.3 Attacking RS+FD: Uncovering the Sampled Attribute (→ SMP)

Since the objective of the RS+FD solution is to hide the LDP value among fake data, discovering the sampled attribute of each user would convert RS+FD into the SMP solution again. Even more, RS+FD utilizes an amplified $\epsilon' > \epsilon$, which would decrease the "plausible deniability" of user's report (*cf.* Section 3.2.1) and could also be leveraged for re-identification attacks (*cf.* Section 3.2.4) under multiple data collections.

For instance, consider the scenario in which a given user $u \in \mathcal{U}$ whose sampled attribute is $t \in [1, d]$ produces an RS+FD's output tuple as $\mathbf{y} = [y_1, y_2, \ldots, y_d]$. In this situation, the **baseline classification model** is just a random guess $\hat{t} = Uniform(\{1, 2, \ldots, d\})$. In addition, we consider a **classifier learning setting** in which an attacker aims to train a classifier over a learning dataset $\mathbf{D}_l = \{(\mathbf{y}_i, t_i) | \ i \in [1, r]\}$ of r rows and $c = d + 1$ columns. More precisely, for each user u_i, $\mathbf{y}_i$ is the output tuple of the RS+FD solution (LDP/fake values, *i.e.*, a full profile of d attributes) and t_i is the sampled attribute (target is a class within $[1, d]$). **Because the sampled attribute t_i of users should be unknown to the attacker**, in this work, we propose three settings to build a learning dataset $\mathbf{D}_l$, which depends on the attack model. In all these settings, we assume that the attacker has the knowledge of the privacy budget ϵ and the LDP protocol used by users with the RS+FD solution. Finally, the attacker's attribute inference accuracy (AIF-ACC) is measured by the number of correct predictions $t = \hat{t}$ over the number of users in the testing dataset n_t: $AIF\text{-}ACC(\%) = 100 \cdot \frac{\sum_{i=1}^{n_t} f(t_i, \hat{t}_i)}{n_t}$, in which $f\left(t, \hat{t}\right) = 1$ if $t = \hat{t}$ and 0 otherwise.

3.3.1 *No Knowledge: Training a Classifier Over Synthetic Profiles.*

With no knowledge of the real sampled attribute of the n users $u \in \mathcal{U}$ and after aggregating users' LDP data, an attacker could use the estimated frequencies $\hat{\mathbf{f}} = [\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_d]$ to generate s **synthetic profiles** $\mathbf{s}_i = [s_1, s_2, \ldots, s_d]$, for $i \in [1, s]$, *i.e.*, mimic the real profiles with one value per attribute. Afterwards, for all s synthetic profiles, the attacker could follow the same protocol used by the real users (*i.e*, RS+FD with an LDP protocol) to generate the learning set $\mathbf{D}_l$. Notice that the attacker has full control over the training set size s, which can be seen as a trade-off between computational costs (*i.e.*, generating s synthetic profiles and use as training set) and the attacker's AIF-ACC. In this **no knowledge (NK)** model, the testing set $\mathbf{D}_t$ is composed of all the real RS+FD's sanitized tuples $\mathbf{y}$ of users $u \in \mathcal{U}$, and the objective is to accurately classify their sampled attribute $t \in [1, d]$.

3.3.2 *Partial-Knowledge: Training a Classifier Over Real (Known) Profiles.*

This second setting considers the scenario in which the attacker has knowledge about the sampled attribute of $n_{pk} < n$ real users, *i.e.*, the subset $\mathcal{U}_{pk} \subset \mathcal{U}$[3]. This setting corresponds in situations in which some users disclose the sampled attribute by preference (*e.g.*, less "sensitive" attributes) or due to security breaches. In this **partial-knowledge (PK)** model, the learning set $\mathbf{D}_l$ depends on the number of (compromised) profiles n_{pk} the attacker has access to and the testing set $\mathbf{D}_t$ has $n - n_{pk}$ sanitized tuples $\mathbf{y}$ of users $u \in \mathcal{U} \backslash \mathcal{U}_{pk}$, in which the objective is to accurately classify their sampled attribute $t \in [1, d]$.

3.3.3 *Partial-Knowledge Plus Synthetic Profiles.*

This last setting combines both NK and PK models, in which the attacker has knowledge about the sampled attribute of $n_{pk} < n$ real users and augments the subset $\mathcal{U}_{pk} \subset \mathcal{U}$ with s synthetic profiles. In this **hybrid model (HM)**, the learning set $\mathbf{D}_l$ is dependent on both the number of synthetic profiles s the attacker generates and the number of (compromised) profiles n_{pk} the attacker has access to. Similarly to the PK model, the testing set $\mathbf{D}_t$ has $n - n_{pk}$ sanitized tuples $\mathbf{y}$ of users $u \in \mathcal{U} \setminus \mathcal{U}_{pk}$, and the goal is to accurately classify their sampled attribute $t \in [1, d]$.

4 EXPERIMENTAL EVALUATION

In this section, we introduce the general setup of our experiments. Next, we present the experimental setting and results on the risks of re-identification of the SMP solution. Afterwards, we describe the setup of experiments carried out to uncover the sampled attribute of the RS+FD solution. Finally, we detail the experimental setting and results on the risks of re-identification of the RS+FD solution.

4.1 Experimental Setup

Environment. All algorithms were implemented in Python 3 with Numpy, Numba and Ray libraries. In all experiments, we report the results averaged over 20 runs.

Datasets. For ease of reproducibility, we conduct our experiments on two census-based multidimensional and open datasets.

- **ACSEmployement.** This dataset is generated from the Folktables Python package [15] that provides access to datasets derived from the US Census. We have selected the "Montana" state only, which results in $n = 10,336$ samples with $d = 18$ discrete attributes (target included) and domain size $\mathbf{k} = [92, 25, 5, 2, 2, 9, 4, 5, 5, 4, 2, 18, 2, 2, 3, 9, 3, 6]$.
- **Adult.** This is a classical dataset from the UCI ML repository [17] with $n = 45,222$ samples after cleaning. We selected $d = 10$ attributes ("age", "workclass", "education", "marital-status", "occupation", "relationship", "race", "sex", "native-country" and "salary") with domain size $\mathbf{k} = [74, 7, 16, 7, 14, 6, 5, 2, 41, 2]$, respectively.

4.2 Re-identification Risk of the SMP Solution

Methods evaluated. We consider for evaluation all five LDP protocols described in Section 2.2: GRR, OLH, ω-SS, SUE and OUE.

Privacy protection. On the one hand, we vary the privacy budget in the interval $\epsilon = [1, 2, \ldots, 9, 10]$, which corresponds to range of values used by industry nowadays [13, 38]. Besides, we also vary the Bayes error probability in Eq. (3) in the interval $\beta_{U|S} = [0.95, 0.9, 0.85, \ldots, 0.55, 0.5]$ (*i.e.*, from tighter privacy regimes to lower ones) to bound the PIE with α. In this case, following Eq. (4) and [34, Proposition 9], when k_j (attribute's domain size) is small, we will not use an LDP protocol (*i.e.*, $y = v$).

Attack performance metric. We rely on the attacker's re-identification accuracy (RID-ACC) metric to measure the quality of the re-identification attack, which corresponds to how many times the user is correctly re-identified in the top-k groups, for top-$k \in \{1, 5, 10\}$.

Baseline. For each top-k, the baseline re-identification model follows top-k random guesses (*i.e.*, $\hat{u}_{id} = Uniform(\{1, 2, \ldots, n\})$) without replacement with accuracy: top-k/n.

Experimental evaluation. We set the number of surveys #survey = 5, in which each survey sv, for $sv \in [1, \#\text{survey}]$, has a different number of attributes $d_{sv} = Uniform\left(\frac{d}{2}, \ldots, d\right)$ (*i.e.*, with at least $\frac{d}{2}$ attributes). The attributes are also selected at random per survey. We run the experiments with all LDP protocols with both FK-RI and PK-RI models (*cf.* Section 3.2.4) in which the background knowledge $\mathcal{D}_{BK}$ and $\mathcal{D}_{PK} \subseteq \mathcal{D}_{BK}$ (a random subset with at least $\frac{d}{2}$ attributes) is the own Adult and ACSEmployement datasets. In addition, we measure the attacker's RID-ACC after #survey ≥ 2. Finally, we also considered both uniform and non-uniform privacy metric settings from Sections 3.2.2 and 3.2.3.

Results. Fig. 2 illustrates the attacker's RID-ACC metric on the Adult dataset for top-k re-identification using the SMP solution with the GRR protocol by varying the uniform privacy metric (*i.e.*, ϵ for LDP and $\beta_{U|S}$ for α-PIE), the attack model (*i.e.*, FK-RI and PK-RI) and the number of surveys (*i.e.*, collections). Additional results with all LDP protocols, both Adult and ACSEmployement datasets, both

[3] If $\mathcal{U}_{pk} \subseteq \mathcal{U}$, this will correspond to a full-knowledge model in which the adversary has knowledge of all users' sampled attribute (*i.e.*, SMP solution).

FK-RI and PK-RI models as well as both uniform and non-uniform privacy metric settings are presented in Appendix C of [1].

Analysis. From Fig. 2, we can observe that our re-identification attacks present significant improvement over a random baseline model that has $RID\text{-}ACC \ll 1\%$ (*i.e.*, top-k/n). For instance, with a single shot (*i.e.*, top-1), the attacker's RID-ACC is already significant (*i.e.*, $RID\text{-}ACC \geq 5\%$) after about #survey ≥ 4 when considering the FK-RI model (plots (a) and (b)). On the other hand, when there is a set of top-10 possibilities that an attacker could re-identify the target user, the attacker achieves $RID\text{-}ACC \geq 2.5\%$ after only 2 surveys with an upper bound of about $RID\text{-}ACC \sim 33\%$ after 5 surveys. Although the user is not uniquely re-identified, this still represents a threat due to the possibility of performing homogeneity attacks [11, 29, 32, 39–41]. In addition, when comparing both LDP and PIE [34] privacy models, as shown in plots (b) and (d) of Fig. 2, even using a high Bayes error probability as $\beta_{U|S} = 0.95$ for each attribute and collection, the $(\mathcal{U}, \alpha)$-PIE privacy metric leads to higher attacker's RID-ACC in comparison with plots (a) and (c) in which $\epsilon = 1$. Indeed, these higher *RID-ACC* when using $\beta_{U|S}$ can be explained due to not using a local randomizer when k_j is small [34, Proposition 9] (the case for several attributes in the Adult dataset), with more difference for other protocols such as SUE, OUE and OLH (see Appendix C of [1]).

Overall, these high re-identification rates may be explained by many factors. First, the combination of multiple attributes within the Adult dataset leads to several unique people or small groups of people (this is also the case for the ACSEmployement dataset in Fig. 9). The GRR protocol is the more easily attacked LDP protocol as shown in Fig. 1. With other protocols, such as OLH and OUE, the attacker cannot infer a partial or complete profile of users when using ϵ-LDP as privacy model, which leads to lower risks of re-identification (*cf.* Appendix C of [1]). Also, in the FK-RI model with uniform privacy metrics across users, the attacker can use the whole background knowledge $\mathcal{D}_{BK}$ when she has collected all attributes from the users (*i.e.*, complete profiles). For instance, the attacker's *RID-ACC* metric decreased by almost half when considering the PK-RI model (plots (c) and (d) in Fig. 2) since there is less attributes as background information to use in the input of the matching algorithm $\mathcal{R}$ (see Section 3.2.4). On the other hand, although the PK-RI model is more realistic, it will allow a small portion of users to sample a different attribute in each survey, leading to higher information leakage than other users. Lastly, we used the same dataset for private data collection and background knowledge. A different set of experiments could mix demographic attributes and (synthetic) application usage in each survey, limiting the number of demographic attributes per user to constitute a profile.

4.3 Uncovering the Sampled Attribute of the RS+FD Solution (→ SMP)

Classifier. We use the state-of-the-art XGBoost [9] algorithm to predict the sampled attribute of users in a multiclass classification framework (*i.e.*, d attributes) with default parameters.

Methods evaluated. We consider for evaluation five protocols within the RS+FD solution from Section 2.3.2, namely RS+FD[GRR], RS+FD[SUE-z], RS+FD[SUE-r], RS+FD[OUE-z] and RS+FD[OUE-r].

Metrics. Similar to Section 4.2, we vary the privacy budget in the interval $\epsilon = [1, 2, \ldots, 9, 10]$. Besides, we use the attacker's attribute inference accuracy (AIF-ACC) metric to measure the quality of the attack, which corresponds to how many times the attacker can correctly predict the users' sampled attribute.

Baseline. The baseline classification model is a random guess $\hat{t} = Uniform(\{1, 2, \ldots, d\})$ with an attacker's AIF-ACC of $1/d$.

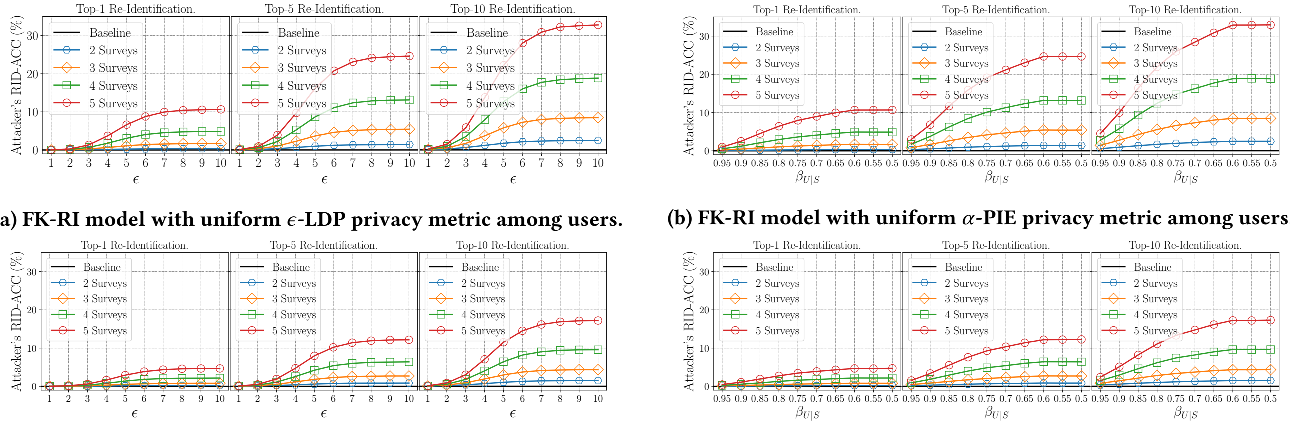
Experimental evaluation. All five protocols are evaluated with the three settings of Section 3.3, namely No Knowledge (NK), Partial-Knowledge (PK) and Hybrid Model (HM). For the NK model, we vary the number of synthetic profiles s the attacker generates in the interval $s = [1n, 3n, 5n]$. For the PK model, we vary the number of compromised profiles n_{pk} the attacker has access to in the interval $n_{pk} = [0.1n, 0.3n, 0.5n]$. Finally, for the HM setting, we combined both interval, *i.e.*, $(s, n_{pk}) = [(1n, 0.1n), (3n, 0.3n), (5n, 0.5n)]$.

Results. Fig. 3 illustrates the attacker's AIF-ACC metric on the ACSEmployement dataset with the three attack models (*i.e.*, NK, PK and HM) and all five protocols (*i.e.*, RS+FD[GRR], RS+FD[SUE-z], RS+FD[OUE-z], RS+FD[SUE-r] and RS+FD[OUE-r]), varying ϵ, the number of synthetic profiles s and the number of compromised profiles n_{pk}. Additional results are presented in Appendix D of [1].

Analysis. From Fig. 3, one can notice that the proposed attack models, namely, NK, PK and HM present significant 2-20 fold increments in the attacker's AIF-ACC over the Baseline model. Surprisingly, even under an NK model in which the attacker has access only to the estimated frequencies satisfying ϵ-LDP, generating $s = [1n, 3n, 5n]$ synthetic profiles to train a classifier provides higher attacker's AIF-ACC than having compromised $n_{pk} = 0.5n$ profiles in the PK model. On the other hand, increasing the number of synthetic profiles s that the attacker generates in the NK model has less impact than increasing the number of compromised profiles n_{pk} that the attacker has access to in the PK model. Due to this, results for both NK and HM models are quite similar.

In this adversarial analysis, the attacker's AIF-ACC now depends on both the LDP protocol and how fake data are generated. For the former (*i.e.*, different LDP protocols), the difference between RS+FD[GRR] and RS+FD[UE-r] protocols lies in the encoding and randomization steps, which directly affects the attacker's AIF-ACC with a difference of about 5% favoring the RS+FD[GRR] protocol. Since GRR requires no particular encoding, there is less noise compared to a randomized unary encoded vector. Furthermore, with respect to different fake data generation procedures, when fake data are generated with a uniformly random (encoded) value (*i.e.*, RS+FD[GRR] and RS+FD[UE-r]), the attacker's AIF-ACC is upper-bounded by about 25%. On the other hand, generating fake data through applying a UE protocol on zero-vectors led to an attacker's AIF-ACC of about 50% with RS+FD[OUE-z] and almost 100% with RS+FD[SUE-z] when $\epsilon = 10$. This high accuracy with RS+FD[UE-z] protocols is because there is only one parameter to perturb each bit when generating fake data, *i.e.*, $\Pr[0 \rightarrow 1] = q$ (*cf.* Section 2.2.4). When using different UE protocols, the randomization parameters p and q (*cf.* Section 2.2.4) also influence the attacker's AIF-ACC, which led RS+FD[SUE] protocols to have lower attacker's AIF-ACC when ϵ is small, but higher attacker's AIF-ACC in low privacy regimes.

Lastly, we remark that due to the original formulation of RS+FD in [5] to generate fake data uniformly at random, a classifier was



(a) FK-RI model with uniform ϵ-LDP privacy metric among users.

(b) FK-RI model with uniform α-PIE privacy metric among users.

(c) PK-RI model with uniform ϵ-LDP privacy metric among users.

(d) PK-RI model with uniform α-PIE privacy metric among users.

Figure 2: Attacker's re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification using the SMP solution with the GRR protocol by varying the uniform privacy metric (*i.e.*, ϵ for LDP and $\beta_{U|S}$ for α-PIE), the attack model (*i.e.*, full knowledge – FK-RI, partial knowledge – PK-RI) and the number of surveys (*i.e.*, data collections).

able to learn the sampled attribute from the users, as the distribution of the attributes was not always uniform with the ACSEmployement dataset. Nevertheless, when the attributes follow uniform-like distribution, none of the three attack models NK, PK or HM achieves a meaningful increment over the Baseline model (*cf.* results with the Nursery dataset [17] in Appendix D of [1]).

4.4 Re-identification Risk of the RS+FD Solution

In this section, we experiment with multiple data collections following the RS+FD solution to measure the attacker's RID-ACC. We follow a similar experimental evaluation of Section 4.2 with the addition of the attribute's inference attack (*cf.* Section 4.3) in each data collection (*i.e.*, survey). To this end, we use the NK model by generating $s = 1n$ profiles as accuracy did not substantially increased with higher s (*cf.* Fig. 3). We selected the RS+FD[GRR] [5] protocol as it provides an intermediate guarantee between RS+FD[UE-r] (lower bound) and RS+FD[UE-z] (upper bound) protocols. We only evaluated the FK-RI model with $\mathcal{D}_{BK}$ and uniform ϵ-LDP privacy metric across users (*i.e.*, users select a new attribute for each survey) as they led to higher re-identification rates using the SMP solution.

Results. Fig. 4 illustrates the attacker's RID-ACC metric on the Adult dataset for top-k re-identification using the FK-RI model and the RS+FD[GRR] protocol and by varying the uniform ϵ-LDP privacy metric and the number of surveys.

Analysis. From Fig. 4, one can note that the re-identification rates with RS+FD has drastically decreased in comparison with the results of the SMP solution in Fig. 2. Re-identification attacks on the RS+FD solution are not trivial, as the attacker has no guarantee that the predicted attribute is correct. Indeed, from Fig. 15 in Appendix D of [1], the attacker's AIF-ACC on the Adult dataset with the RS+FD[GRR] protocol is upper bounded in 40%, which leads to chained errors when profiling a target user in multiple collections. For instance, the attacker's RID-ACC for the top-1 group is nearly equal the random Baseline model. Even for the top-5 and top-10 groups the attacker's RID-ACC has meaningful improvement over the Baseline model. These results with the RS+FD[GRR] protocol indicates that RS+FD is already a countermeasure to re-identification attacks, except for the RS+FD[SUE-z] protocol in which the attacker can predict the attribute with high confidence when ϵ is high.

5 COUNTERMEASURE

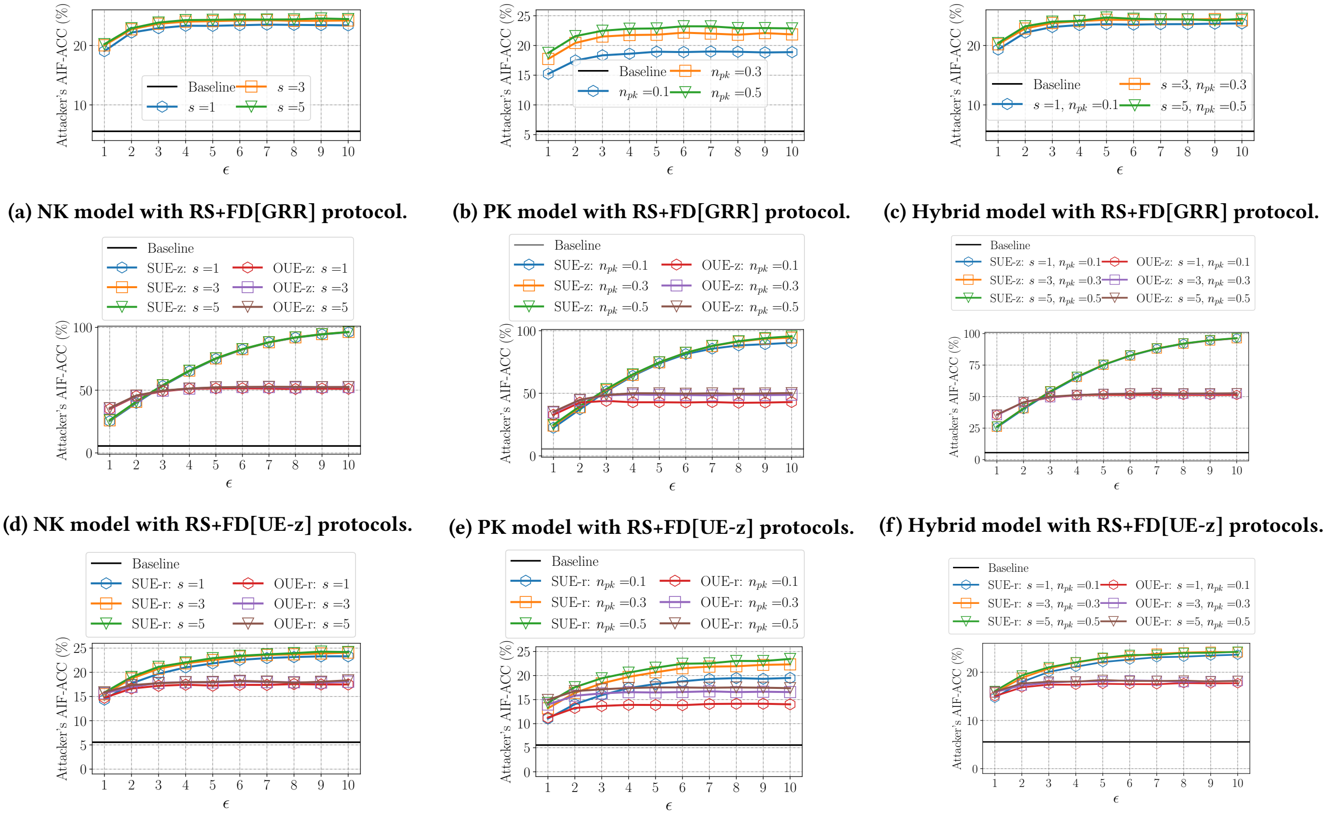
As shown in Section 4.4, the RS+FD solution already provides some resistance to re-identification attacks. Thus, we now present an improvement of the RS+FD solution and the experimental results.

5.1 Random Sampling Plus Realistic Fake Data

As briefly described in Section 2.3, the client-side of RS+FD [5] is split into two steps (*i.e.*, local randomization and *uniform* fake data generation). We now present an improvement of RS+FD, which we call Random Sampling Plus Realistic Fake Data (RS+RFD) as fake data will follow (potentially prior) *non-uniform* distributions. For instance, several demographic attributes have national statistics released by the Census [2] the previous year. Therefore, more "realistic" profiles can be generated by users to counter the inference of the sampled attribute and consequently the risk of re-identification.

Client-Side. Alg. 1 displays the pseudocode of our RS+RFD solution at the client-side. The input of RS+RFD is the user's true tuple of values $\mathbf{v} = [v_1, v_2, \ldots, v_d]$, the domain size of attributes $\mathbf{k} = [k_1, k_2, \ldots, k_d]$, the attributes' prior distributions $\tilde{\mathbf{f}} = [\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_d]$ (transmitted by the server in advance), the privacy parameter ϵ and a local randomizer $\mathcal{M}$. The output is a tuple $\mathbf{y} = [y_1, y_2, \ldots, y_d]$ of values (LDP and fake). In Alg. 1, line 6, *Sample* means a random sample is generated following prior $\tilde{f}_i$ of the attribute $i \in [1, d] \setminus \{j\}$.

Server-Side. The aggregator performs multiple frequency (or histogram) estimation on the collected data by removing bias introduced by the local randomizer $\mathcal{M}$ and fake data. The new estimators of using RS+RFD with GRR or UE-based protocols (*e.g.*, SUE [23] or OUE [47]) as local randomizer $\mathcal{M}$ in Alg. 1 is presented in the



(a) NK model with RS+FD[GRR] protocol.

(b) PK model with RS+FD[GRR] protocol.

(c) Hybrid model with RS+FD[GRR] protocol.

(d) NK model with RS+FD[UE-z] protocols.

(e) PK model with RS+FD[UE-z] protocols.

(f) Hybrid model with RS+FD[UE-z] protocols.

(g) NK model with RS+FD[UE-r] protocols.

(h) PK model with RS+FD[UE-r] protocols.

(i) Hybrid model with RS+FD[UE-r] protocols.

Figure 3: Attacker's AIF-ACC on the ACSEmployement dataset with three attack models (*i.e.*, NK, PK and hybrid) and five protocols (*i.e.*, RS+FD[GRR], RS+FD[SUE-z], RS+FD[OUE-z], RS+FD[SUE-r] and RS+FD[OUE-r]), varying ϵ, the number of synthetic profiles s the attacker generates and the number of compromised profiles n_{pk} the attacker has access to.

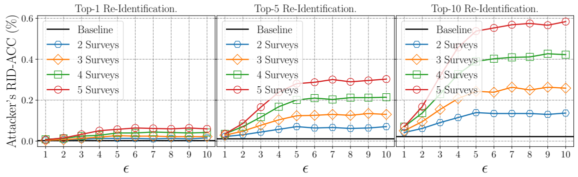


Figure 4: Attacker's re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification using the FK-RI model and the RS+FD[GRR] protocol and by varying the uniform ϵ-LDP privacy metric and the number of surveys.

following. For each attribute $j \in [1, d]$, the aggregator estimates $\hat{f}(v_i)$ for the frequency of each value $i \in [1, k_j]$ as:

- **RS+RFD[GRR].** The RS+RFD[GRR] estimator is:

$$\hat{f}_{\text{GRR}}(v_i) = \frac{dC(v_i) - n\left(q + (d-1)\tilde{f}_j(v_i)\right)}{n(p-q)}, \tag{9}$$

in which $C(v_i)$ is the number of times v_i has been reported, $\tilde{f}_j(v_i)$ is the prior distribution of value $v_i \in A_j$, $\epsilon' = \ln\left(d \cdot (e^{\epsilon} - 1) + 1\right)$, $p = \frac{e^{\epsilon'}}{e^{\epsilon'} + k_j - 1}$ and $q = \frac{1-p}{k_j - 1}$. The probability tree of the RS+RFD[GRR] protocol, the proof that the estimator in Eq. (9) is unbiased and its variance computation are provided in Appendix A of [1].

Algorithm 1 Random Sampling plus Realistic Fake Data (RS+RFD)

```
Input : tuple $\mathbf{v} = [v_1, v_2, \ldots, v_d]$, domain size of attributes $\mathbf{k} = [k_1, k_2, \ldots, k_d]$, prior distribution of attributes $\tilde{\mathbf{f}} = [\tilde{f_1}, \tilde{f_2}, \ldots, \tilde{f_d}]$, privacy parameter $\epsilon$ and local randomizer $\mathcal{M}$.
Output : sanitized tuple $\mathbf{y} = [y_1, y_2, \ldots, y_d]$.
1: $\epsilon' = \ln\left(d \cdot (e^{\epsilon} - 1) + 1\right)$          ▷ Amplification by sampling [30]
2: $j \leftarrow Uniform(\{1, 2, \ldots, d\})$          ▷ Selection of attribute to sanitize
3: $B_j \leftarrow \mathsf{Encode}(v_j)$          ▷ Encode (if needed)
4: $y_j \leftarrow \mathcal{M}(B_j, k_j, \epsilon')$          ▷ Sanitize data of the sampled attribute
5: for $i \in [1, d] \setminus \{j\}$ do          ▷ For each non-sampled attributes
6:     $y_i \leftarrow \mathsf{Sample}(\{1, \ldots, k_i\}, \tilde{f_i})$          ▷ Generate one fake data
7: end for
return : $\mathbf{y} = [y_1, y_2, \ldots, y_d]$          ▷ Sanitized tuple
```

- **RS+RFD[UE-r].** Similar to the RS+FD[UE-r] protocol in Section 2.3.2, in Line 6 of Alg. 1, for each non-sampled attribute i, for $i \in [1, d] \setminus \{j\}$, the user generates fake data by applying an UE protocol to one-hot-encoded random data following prior distribution $\tilde{f}_i$. The RS+RFD[UE-r] estimator is:

$$\hat{f}_{\text{UE-r}}(v_i) = \frac{dC(v_i) - n\left(q + (p-q)(d-1)\tilde{f}_j(v_i) + q(d-1)\right)}{n(p-q)}, \tag{10}$$

in which $C(v_i)$ is the number of times v_i has been reported, $\epsilon' = \ln\left(d \cdot (e^{\epsilon} - 1) + 1\right)$ and $\tilde{f}_j(v_i)$ is the prior distribution of value $v_i \in A_j$. Parameters p and q can be selected following the SUE [23] protocol ($p = \frac{e^{\epsilon'/2}}{e^{\epsilon'/2}+1}$ and $q = \frac{1}{e^{\epsilon'/2}+1}$) or OUE [47] protocol ($p = \frac{1}{2}$ and $q = \frac{1}{e^{\epsilon'}+1}$). The probability tree of the RS+RFD[UE-r] protocol, the proof that the estimator in Eq. (10) is unbiased and its variance calculation is provided in Appendix B of [1].

Privacy analysis. Similar to the RS+FD solution [5], let $\mathcal{M}$ be any existing LDP mechanism, Alg. 1 satisfies ϵ-LDP, in a way that $\epsilon' = \ln\left(d \cdot (e^{\epsilon} - 1) + 1\right)$, in which d is the number of attributes.

Limitations. In addition to known limits of the RS+FD solution [5, 43], RS+RFD adds a limitation on being dependent on the underlying prior distributions $\tilde{\mathbf{f}}$ to generate realistic fake data. Yet, many demographic attributes have Census-based data released annually [2] and other attributes' priors can be defined following domain expert knowledge.

5.2 Experimental Results

In this section, we present the general setup of experiments with the RS+RFD solution, which includes: the frequency estimation of multiple attributes and the inference attack of the sampled attribute.

5.2.1 General Experimental Setup. We use the same Adult and ACSEmployement datasets described in Section 4.1.

Prior distribution. To simulate prior distributions $\tilde{\mathbf{f}} = [\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_d]$ to be used to generate realistic fake data with RS+RFD, we perturb the real frequency of each attribute $j \in [1, d]$ with the standard Laplace mechanism [19–21] in centralized DP satisfying $\epsilon = 0.1/d$ (*i.e.*, split $\epsilon = 0.1$ by d attributes).

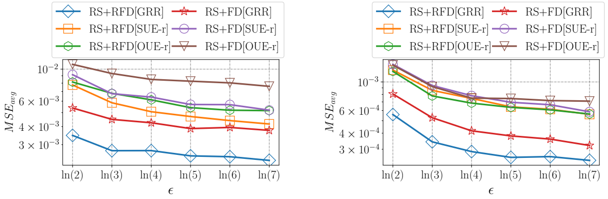
Methods evaluated. We consider for evaluation three protocols within the RS+RFD solution from Section 5.1, namely, RS+RFD[GRR], RS+RFD[SUE-r] and RS+RFD[OUE-r].

5.2.2 Frequency Estimation of Multiple Attributes. We compare the results of our RS+RFD protocols with their respective version within the RS+FD [5] solution, *i.e.*, RS+FD[GRR], RS+FD[SUE-r] and RS+FD[OUE-r] (*cf.* Section 2.3.2).

Evaluation metrics. To compare with [5], we vary ϵ in the interval $\epsilon = [\ln(2), \ln(3), \ldots, \ln(7)]$ and we measure the quality of the estimated frequencies with the averaged mean squared error metric: $MSE_{avg} = \frac{1}{d}\sum_{j\in[1,d]} \frac{1}{|A_j|}\sum_{v\in A_j}(f(v) - \hat{f}(v))^2$.

Results. Fig. 5 illustrates for all methods and both Adult and ACSEmployement datasets the MSE_{avg} metric (y-axis) according to the privacy parameter ϵ (x-axis).

Analysis. For both datasets, one can observe that the MSE_{avg} metric of our proposed RS+RFD protocols outperforms the utility of their respective version within the RS+FD solution. The intuition is that since random noise is drawn from realistic prior distributions, the fake data also contributes to the estimation of the attribute. On the other hand, when random noise follows uniform distributions as with RS+FD, fake data can only increase the estimation of non-correct items. Finally, we have observed in our experiments that SUE [23] outperforms its optimized version OUE [47] in some conditions, with both RS+RFD and RS+FD solutions.



(a) ACSEmployement dataset.

(b) Adult dataset.

Figure 5: Averaged MSE metric varying ϵ on the ACSEmployement (a) and Adult (b) datasets for multidimensional frequency estimation with the RS+RFD and RS+FD solutions.

5.2.3 Uncovering the Sampled Attribute of the RS+RFD Solution (→ SMP). This section follows similar parameters (dataset, ϵ range and attacker's AIF-ACC metric) used in the experiments of Section 4.3.

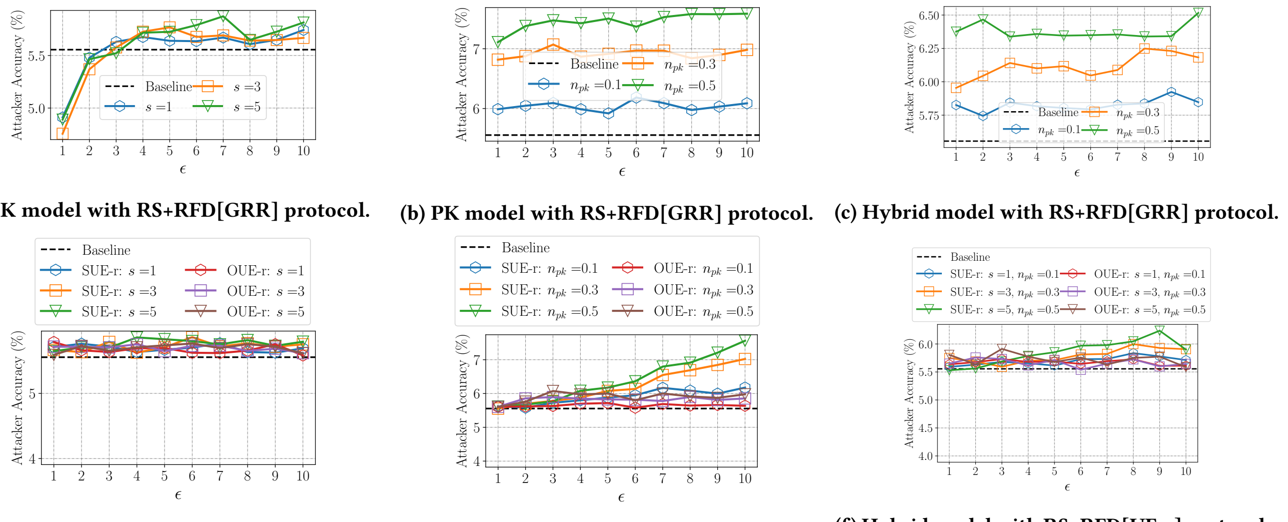
Results. Fig. 6 illustrates the attacker's AIF-ACC metric on the ACSEmployement dataset with three attack models (*i.e.*, NK, PK and hybrid) and our three protocols (*i.e.*, RS+RFD[GRR], RS+RFD[SUE-r] and RS+RFD[OUE-r]), varying ϵ, the number of synthetic profiles s as well as the number of compromised profiles n_{pk}.

Analysis. We highlight that the non-stability in the plots of Fig. 6 is due to different sources of randomness: $\epsilon = 0.1$-DP prior distributions $\tilde{\mathbf{f}}$, ϵ-LDP randomization, fake data generation and the XGBoost algorithm. From Fig. 6, one can remark that our RS+RFD protocols considerably decrease the attacker's AIF-ACC when comparing with their respective RS+FD version in Fig. 3. In contrast with the results of Section 4.3, the results with the PK model has higher attacker's AIF-ACCs than the NK model. This is intuitive since the attacker gained "real" information of the sampled attribute, increasing the attacker's AIF-ACC as the number of compromised profiles n_{pk} gets higher. Nevertheless, for all three NK, PK and HM models, the accuracy gain over a random Baseline model is still minor, highlighting the benefits of our RS+RFD proposal.

6 RELATED WORK

The literature in the local DP model has largely explored the issue of improving the utility of LDP protocols [5, 6, 14, 18, 23, 25, 26, 36, 42, 44, 46–49]. Recently, a few works have started to design attacks to LDP protocols. Some authors focused on maliciously modifying the estimated statistic on the server through targeted or untargeted attacks [7, 10, 31, 51]. To counter such kinds of attacks, some works [4, 28] investigated cryptography-based approaches.

These targeted or untargeted attacks raise awareness of potential security vulnerabilities of LDP protocols. However, these attacks do not aim to attack users' privacy as initially investigated in [22, 34] and in this work. By the time of completing this paper, we learned



(a) NK model with RS+RFD[GRR] protocol. **(b) PK model with RS+RFD[GRR] protocol.** **(c) Hybrid model with RS+RFD[GRR] protocol.**

(d) NK model with RS+RFD[UE-r] protocols. **(e) PK model with RS+RFD[UE-r] protocols.** **(f) Hybrid model with RS+RFD[UE-r] protocols.**

Figure 6: Attacker's AIF-ACC on the ACSEmployement dataset with three attack models (*i.e.*, NK, PK and hybrid) and our three protocols (*i.e.*, RS+RFD[GRR], RS+RFD[SUE-r] and RS+RFD[OUE-r]), varying ϵ, the number of synthetic profiles s the attacker generates and the number of compromised profiles n_{pk} the attacker has access to.

that an adversarial analysis of state-of-the-art LDP protocols for single-frequency estimation had been recently published in the concurrent and independent work by Gursoy and co-authors [22]. In this paper, we have used the formalized Bayesian adversary from [22] only as an analytical measure of our "plausible deniability" interpretation in Section 3.2.1. Thus, we have extended this analysis to multiple collections in Sections 3.2.2 and 3.2.3, which were intended to analytically evaluate the consequent risks of re-identification [24, 33–35]. In this context, Murakami and Takahashi [34] first investigated the re-identification risks in the LDP model for a single attribute and proposed the $(\mathcal{U}, \alpha)$-PIE privacy metric described in Section 2.1. While the setting considered in [34] focused on a single attribute, such as location traces, our work considers multiple attributes being collected multiple times.

On the other hand, Arcolezi and co-authors [5] introduced the RS+FD solution focusing only on the utility of the protocols, which was also later studied in [43]. In this work, we have proposed three attack models to the RS+FD solution, showing it is possible to distinguish the ϵ-LDP report from fake data. This has critical implications, since the sampled attribute utilizes an amplified $\epsilon' > \epsilon$ due to amplification by sampling [30], thus minimizing the "plausible deniability" of the report also leading to (reduced) re-identification risks (*cf.* Section 4.4). For this reason, we proposed a countermeasure in this paper, which relies on generating non-uniform fake data using RS+FD (*i.e.*, RS+RFD of Section 5.1). As shown in the results of Section 5.2, **our RS+RFD solution improved both utility and privacy**. More specifically, our RS+RFD protocols minimize the estimation error compared to their respective RS+FD [5] versions (see Fig. 5). Besides, our RS+RFD protocols almost fully prevent the inference of the sampled attribute (see Fig. 6), which also naturally prevent the risk of re-identification.

7 CONCLUSION AND PERSPECTIVES

In this paper, we have studied privacy threats against LDP protocols for multidimensional data following two state-of-the-art solutions for frequency estimation of multiple attributes, *i.e.*, SMP and RS+FD [5]. On the one hand, we presented inference attacks based on "plausible deniability" [50] of five widely used LDP protocols (*i.e.*, GRR [25, 26], OLH [47], ω-SS [45, 52], RAPPOR [23] and OUE [47]) under multiple collections following the SMP solution. This analysis also empirically clarifies the risks of re-identification when an attacker is able to build complete and/or partial profiles of users and can correlate them with prior knowledge.

In addition, we proposed three attack models to infer the sampled attribute of the RS+FD solution, which allowed us to still reconstruct complete and/or partial profiles of users and lead to re-identification (although to a much lesser extent than the SMP solution). Finally, we have proposed a countermeasure solution named RS+RFD (*i.e.*, RS+FD with realistic fake data) that improves both utility and privacy. That is, in our experiments, RS+RFD minimized the estimation error in comparison with the RS+FD solution, as well as almost fully mitigated the inference of the sampled attribute.

For future work, we will evaluate the SMP's risks of re-identification on d-privacy [3, 8] and LDP [46] protocols designed for ordered data (*e.g.*, age and location) for which there exists a non-binary notion of distance between profiles.

ACKNOWLEDGMENTS

This work was partially supported by the ERC project HYPATIA with grant agreement Nº 835294 and by the EIPHI-BFC Graduate School (contract "ANR-17-EURE-0002"). Sébastien Gambs is supported by the Canada Research Chair program as well as a Discovery Grant from NSERC. All computations have been performed on the "Mésocentre de Calcul de Franche-Comté".

REFERENCES

[1] Full Version. On the Risks of Collecting Multidimensional Data Under Local Differential Privacy. Available online: https://github.com/hharcolezi/risks-ldp.

[2] John M. Abowd. 2018. The U.S. Census Bureau Adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM. https://doi.org/10.1145/3219819.3226070

[3] Mario Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. 2018. Invited Paper: Local Differential Privacy on Metric Spaces: Optimizing the Trade-Off with Utility. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE. https://doi.org/10.1109/csf.2018.00026

[4] Andris Ambainis, Markus Jakobsson, and Helger Lipmaa. 2004. Cryptographic Randomized Response Techniques. In *Public Key Cryptography – PKC 2004*. Springer Berlin Heidelberg, 425–438. https://doi.org/10.1007/978-3-540-24632-9_31

[5] Héber H. Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. 2021. Random Sampling Plus Fake Data: Multidimensional Frequency Estimates With Local Differential Privacy. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 47–57. https://doi.org/10.1145/3459637.3482467

[6] Héber H. Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. 2022. Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates. *Digital Communications and Networks* (2022). https://doi.org/10.1016/j.dcan.2022.07.003

[7] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2021. Data Poisoning Attacks to Local Differential Privacy Protocols. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 947–964.

[8] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the Scope of Differential Privacy Using Metrics. In *Privacy Enhancing Technologies*, Emiliano De Cristofaro and Matthew Wright (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 82–102. https://doi.org/10.1007/978-3-642-39077-7_5

[9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. https://doi.org/10.1145/2939672.2939785

[10] Albert Cheu, Adam Smith, and Jonathan Ullman. 2021. Manipulation Attacks in Local Differential Privacy. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE. https://doi.org/10.1109/sp40001.2021.00001

[11] Aloni Cohen. 2022. Attacks on Deidentification's Defenses. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 1469–1486. https://www.usenix.org/conference/usenixsecurity22/presentation/cohen

[12] Graham Cormode, Samuel Maddock, and Carsten Maple. 2021. Frequency estimation under local differential privacy. *Proceedings of the VLDB Endowment* 14, 11 (July 2021), 2046–2058. https://doi.org/10.14778/3476249.3476261

[13] Damien Desfontaines. 2021. A list of real-world uses of differential privacy. Available online: https://desfontain.es/privacy/real-world-differential-privacy.html (accessed on 27 May 2022).

[14] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting Telemetry Data Privately. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3571–3580.

[15] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021).

[16] Josep Domingo-Ferrer and Jordi Soria-Comas. 2018. Connecting randomized response, post-randomization, differential privacy and t-closeness via deniability and permutation. *arXiv preprint arXiv:1803.02139* (2018).

[17] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[18] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. 2013. Local Privacy and Statistical Minimax Rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE. https://doi.org/10.1109/focs.2013.53

[19] Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming*, Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–12.

[20] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*. Springer Berlin Heidelberg, 265–284. https://doi.org/10.1007/11681878_14

[21] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

[22] M. Emre Gursoy, Ling Liu, Ka-Ho Chow, Stacey Truex, and Wenqi Wei. 2022. An Adversarial Approach to Protocol Analysis and Selection in Local Differential Privacy. *IEEE Transactions on Information Forensics and Security* (2022), 1–1. https://doi.org/10.1109/TIFS.2022.3170242

[23] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (Scottsdale, Arizona, USA). ACM, New York, NY, USA, 1054–1067. https://doi.org/10.1145/2660267.2660348

[24] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2014. De-anonymization attack on geolocated data. *J. Comput. System Sci.* 80, 8 (2014), 1597–1614. https://doi.org/10.1016/j.jcss.2014.04.024

[25] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. 2016. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*. PMLR, 2436–2444.

[26] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2016. Extremal mechanisms for local differential privacy. *The Journal of Machine Learning Research* 17, 1 (2016), 492–542.

[27] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2008. What Can We Learn Privately?. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE. https://doi.org/10.1109/focs.2008.27

[28] Fumiyuki Kato, Yang Cao, and Masatoshi Yoshikawa. 2021. Preventing Manipulation Attack in Local Differential Privacy Using Verifiable Randomization Mechanism. In *Data and Applications Security and Privacy XXXV*. Springer International Publishing, 43–60. https://doi.org/10.1007/978-3-030-81242-3_3

[29] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE. https://doi.org/10.1109/icde.2007.367856

[30] Ninghui Li, Wahbeh Qardaji, and Dong Su. 2012. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security - ASIACCS '12*. ACM Press. https://doi.org/10.1145/2414456.2414474

[31] Xiaoguang Li, Neil Zhenqiang Gong, Ninghui Li, Wenhai Sun, and Hui Li. 2022. Fine-grained Poisoning Attacks to Local Differential Privacy Protocols for Mean and Variance Estimation. *arXiv preprint arXiv:2205.11782* (2022).

[32] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. 2006. L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*. IEEE. https://doi.org/10.1109/icde.2006.1

[33] Takao Murakami, Atsunori Kanemura, and Hideitsu Hino. 2017. Group Sparsity Tensor Factorization for Re-Identification of Open Mobility Traces. *IEEE Transactions on Information Forensics and Security* 12, 3 (2017), 689–704. https://doi.org/10.1109/TIFS.2016.2631952

[34] Takao Murakami and Kenta Takahashi. 2021. Toward Evaluating Re-identification Risks in the Local Privacy Model. *Transactions on Data Privacy* 14 (2021), 79–116. Issue 3.

[35] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 111–125. https://doi.org/10.1109/SP.2008.33

[36] Thông T Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. 2016. Collecting and analyzing data from smart device users with local differential privacy. *arXiv preprint arXiv:1606.05053* (2016).

[37] Xuebin Ren, Chia-mu Yu, Weiren Yu, Shusen Yang, Senior Member, Xinyu Yang, Julie A Mccann, Philip S Yu, and Life Fellow. 2018. LoPub : High-Dimensional Crowdsourced Data. 13, 9 (2018), 2151–2166. https://doi.org/10.1109/TIFS.2018.2812146

[38] Ryan Rogers, Subbu Subramaniam, Sean Peng, David Durfee, Seunghyun Lee, Santosh Kumar Kancha, Shraddha Sahay, and Parvez Ahammad. 2021. LinkedIn's Audience Engagements API: A Privacy Preserving Data Analytics System at Scale. *Journal of Privacy and Confidentiality* 11, 3 (Dec. 2021). https://doi.org/10.29012/jpc.782

[39] Pierangela Samarati and Latanya Sweeney. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. (1998).

[40] Latanya Sweeney. 2002. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (Oct. 2002), 557–570. https://doi.org/10.1142/s0218488502001648

[41] Latanya Sweeney. 2015. Only you, your doctor, and many others may know. *Technology Science* 2015092903, 9 (2015), 29.

[42] Apple Differential Privacy Team. 2017. Learning with privacy at scale. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf. Online; accessed 11 December 2021.

[43] Gatha Varma, Ritu Chauhan, and Dhananjay Singh. 2022. Sarve: synthetic data and local differential privacy for private frequency estimation. *Cybersecurity* 5, 1 (Aug. 2022). https://doi.org/10.1186/s42400-022-00129-6

[44] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. 2019. Collecting and Analyzing Multidimensional Data with Local Differential Privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE. https://doi.org/10.1109/icde.2019.00063

[45] Shaowei Wang, Liusheng Huang, Pengzhan Wang, Yiwen Nie, Hongli Xu, Wei Yang, Xiang-Yang Li, and Chunming Qiao. 2016. Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*

(2016).

[46] Shaowei Wang, Yiwen Nie, Pengzhan Wang, Hongli Xu, Wei Yang, and Liusheng Huang. 2017. Local private ordinal data distribution estimation. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. IEEE. https://doi.org/10.1109/infocom.2017.8056977

[47] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally Differentially Private Protocols for Frequency Estimation. In *26th USENIX Security Symposium (USENIX Security 17)*. USENIX Association, Vancouver, BC, 729–745.

[48] Tianhao Wang, Ninghui Li, and Somesh Jha. 2018. Locally Differentially Private Frequent Itemset Mining. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE. https://doi.org/10.1109/sp.2018.00035

[49] Tianhao Wang, Milan Lopuhaa-Zwakenberg, Zitao Li, Boris Skoric, and Ninghui Li. 2020. Locally Differentially Private Frequency Estimation with Consistency. In *Proceedings 2020 Network and Distributed System Security Symposium*. Internet Society. https://doi.org/10.14722/ndss.2020.24157

[50] Stanley L. Warner. 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *J. Amer. Statist. Assoc.* 60, 309 (March 1965), 63–69. https://doi.org/10.1080/01621459.1965.10480775

[51] Yongji Wu, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2022. Poisoning Attacks to Local Differential Privacy Protocols for Key-Value Data. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 519–536. https://www.usenix.org/conference/usenixsecurity22/presentation/wu-yongji

[52] Min Ye and Alexander Barg. 2018. Optimal Schemes for Discrete Distribution Estimation Under Locally Differential Privacy. *IEEE Transactions on Information Theory* 64, 8 (2018), 5662–5676. https://doi.org/10.1109/TIT.2018.2809790

[53] Zhikun Zhang, Tianhao Wang, Ninghui Li, Shibo He, and Jiming Chen. 2018. CALM: Consistent adaptive local marginal for marginal release under local differential privacy. *Proceedings of the ACM Conference on Computer and Communications Security* (2018), 212–229. https://doi.org/10.1145/3243734.3243742

A RS+RFD WITH GRR

Visually, Fig. 7 illustrates the probability tree of the RS+RFD[GRR] protocol (cf. Section 2.3.2).

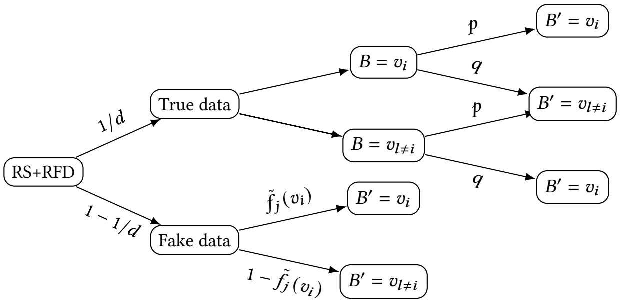


Figure 7: Probability tree for the RS+FD[GRR] protocol.

Theorem 1. *For $j \in [1, d]$, the estimation result $\hat{f}_{GRR}(v_i)$ in Eq. (9) is an unbiased estimation of $f(v_i)$ for any value $v_i \in A_j$.*

Proof.

$$\mathbb{E}\left[\hat{f}_{\text{GRR}}(v_i)\right] = \mathbb{E}\left[\frac{dC(v_i) - n\left(q + (d-1)\tilde{f}_j(v_i)\right)}{n(p-q)}\right]$$
$$= \frac{d}{n(p-q)}\mathbb{E}\left[C(v_i)\right] - \frac{(d-1)\tilde{f}_j(v_i) + q}{(p-q)}.$$

On expectation, the number of times that v_i is reported is:

$$\mathbb{E}\left[C(v_i)\right] = \frac{1}{d}\left(pnf(v_i) + q(n - nf(v_i))\right) + n\frac{(d-1)\tilde{f}_j(v_i)}{d}$$
$$= \frac{n}{d}\left(f(v_i)(p-q) + q + (d-1)\tilde{f}_j(v_i)\right).$$

Therefore,

$$\mathbb{E}\left[\hat{f}_{\text{GRR}}(v_i)\right] = f(v_i).$$

□

Theorem 2. *The variance of the estimation in Eq. (9) is:*

$$\text{VAR}\left[\hat{f}_{GRR}(v_i)\right] = \frac{d^2\gamma(1-\gamma)}{n(p-q)^2}, \textit{ where}$$
$$\gamma = \frac{1}{d}\left(q + f(v_i)(p-q) + (d-1)\tilde{f}_j(v_i)\right). \tag{11}$$

Proof. Thanks to Eq. (9) we have

$$\text{VAR}\left[\hat{f}_{\text{GRR}}(v_i)\right] = \frac{\text{VAR}\left[C(v_i)\right] d^2}{n^2(p-q)^2}.$$

Since $C(v_i)$ is the number of times value v_i is observed, it can be defined as $C(v_i) = \sum_{u=1}^{n} X_u$ where X_u is equal to 1 if the user u, $1 \leq u \leq n$ reports value v_i, and 0 otherwise. We thus have $\text{VAR}\left[C(v_i)\right] = \sum_{u=1}^{n} \text{VAR}\left[X_u\right] = n\,\text{VAR}\left[X\right]$, since all the users are independent. According to the probability tree in Fig. 7,

$$\Pr\left[X = 1\right] = \Pr\left[X^2 = 1\right] = \gamma = \frac{1}{d}\left(q + f(v_i)(p-q) + (d-1)\tilde{f}_j(v_i)\right).$$

We thus have $\text{VAR}\left[X\right] = \gamma - \gamma^2 = \gamma(1-\gamma)$ and, finally,

$$\text{VAR}\left[\hat{f}_{\text{GRR}}(v_i)\right] = \frac{d^2\gamma(1-\gamma)}{n(p-q)^2}.$$

□

B RS+RFD WITH UE-R PROTOCOLS

Visually, Fig. 8 illustrates the probability tree of the RS+RFD[UE-r] protocol (cf. Section 2.3.2).

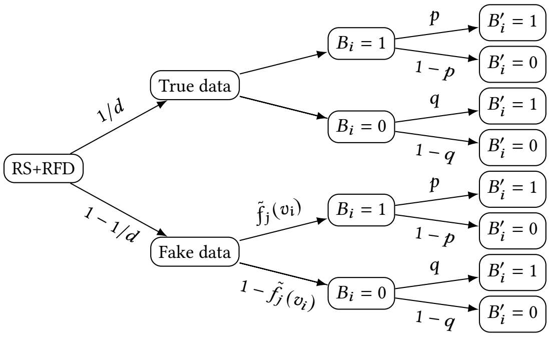


Figure 8: Probability tree for the RS+FD[UE-r] protocol.

Theorem 3. *For $j \in [1, d]$, the estimation result $\hat{f}_{UE\text{-}R}(v_i)$ in Eq. (10) is an unbiased estimation of $f(v_i)$ for any value $v_i \in A_j$.*

Proof.

$$\mathbb{E}\left[\hat{f}_{\text{UE-R}}(v_i)\right] = \mathbb{E}\left[\frac{dC(v_i) - n\left(q + (p-q)(d-1)\tilde{f}_j(v_i) + q(d-1)\right)}{n(p-q)}\right]$$
$$= \frac{d}{n(p-q)}\mathbb{E}\left[C(v_i)\right] - \frac{(p-q)(d-1)\tilde{f}_j(v_i) + q + q(d-1)}{(p-q)}.$$

On expectation, the number of times that v_i is reported is:

$$\mathbb{E}\left[C(v_i)\right] = \frac{n}{d}\left(f(v_i)(p-q)+q\right) + \frac{n(d-1)}{d}\left(\tilde{f}_j(v_i)(p-q)+q\right).$$

Therefore,

$$\mathbb{E}\left[\hat{f}_{\text{UE-R}}(v_i)\right] = f(v_i).$$

□

Theorem 4. *The variance of the estimation in Eq.* (10) *is:*

$$\text{VAR}\left[\hat{f}_{UE\text{-}R}(v_i)\right] = \frac{d^2\gamma(1-\gamma)}{n(p-q)^2}, \textit{ where}$$
$$\gamma = \frac{1}{d}\left(f(v_i)(p-q)+q+(d-1)\left(\tilde{f}_j(v_i)(p-q)+q\right)\right). \tag{12}$$

The proof of Theorem 4 follows the proof of Theorem 2 and is omitted here. In this case, γ follows the probability tree in Fig. 8.

C ADDITIONAL RESULTS FOR SECTION 4.2

This section provides additional results for the risks of re-identification on collecting multidimensional data with the SMP solution. We follow the experimental evaluation described in Section 4.2 and we vary the following:

- **Dataset.** Adult [17] and ACSEmployement [15] datasets.
- **Frequency oracle.** GRR [25, 26], OLH [47], ω-SS [45, 52], SUE (a.k.a. Basic One-Time RAPPOR [23]) and OUE [47].
- **Re-identification model.** Full knowledge (FK-RI) and partial knowledge (PK-RI) models (cf. Section 3.2.4).
- **Privacy metric.** Local differential privacy (with ϵ) and α-PIE privacy [34] (with $\beta_{U|S}$).
- **Privacy metric across users.** Uniform (cf. Section 3.2.2) and non-uniform (cf. Section 3.2.3) privacy metrics.

First, with the ACSEmployement dataset, Fig. 9 illustrates the attacker's RID-ACC for top-k re-identification on using the SMP solution, the FK-RI model with uniform ϵ-LDP privacy metric across users, by varying the frequency oracle and the number of surveys. Secondly, to complement the results of Fig. 2 (i.e., GRR's risks of re-identification), Fig. 10 illustrates the attacker's re-identification accuracy (RID-ACC) metric on the Adult dataset for top-k re-identification on using the SMP solution with the GRR protocol by varying the **non-uniform** privacy metric, the attack model and the number of surveys. Lastly, Figs. 11 to 14, illustrates for all remaining frequency oracles, the attacker's RID-ACC on the Adult dataset by varying the re-identification model and the privacy metrics (uniform and non-uniform). Results for the PK-RI model, α-PIE and non-uniform privacy metrics were omitted for the ACSEmployement dataset as they follow similar results achieved with the Adult dataset where the difference is mainly in the attacker's RID-ACC upper bound.

From Fig. 9 (all plots) and Figs. 2, 10–14 (left-side plots with ϵ-LDP metric), one can note that the attackers' RID-ACC follow similar behavior as the analytical results in Fig. 1. Indeed, the risks of re-identification deeply depend on how well the attacker can profile each user on multiple collections such that the profile can be found on Census-based background knowledge $\mathbb{D}_{BK}$. With uniform privacy metrics, all users report a different attribute per survey, which leads to higher re-identification rates. On the other hand, if users sample attributes with replacement, there is a lower probability of selecting different attributes for each survey, which bounds the re-identification risks. Besides, the results agree with intuitive expectations, as increasing ϵ also increases the privacy leakage, thus leading to higher re-identification rates.

Moreover, following the ϵ-LDP metric, the lowest attacker's RID-ACC were achieved by both OLH and OUE protocols [47], as they have an upper bound for the attacker's $ACC_{FO} = 1/2$ [22] (cf. Section 3.2.1) per data collection. On the other hand, the ω-SS protocol follows similar re-identification rates as the GRR protocol (also shown analytically in Fig. 1), which is the highest among all frequency oracles. In addition, as shown in [34, Proposition 9], the PIE privacy metric does not require a local randomizer when k_j (the domain size) is small. Thus, as the Adult dataset has several attributes with a small domain size (e.g., two binary attributes), for all frequency oracles, the re-identification rates on right-side plots of Figs. 2, 10–14 follow similar behaviors. Indeed, following the α-PIE metric, the re-identification rates were already perceptible even with the highest Bayes error probability (tighter privacy) that we experimented with, i.e., $\beta_{U|S} = 0.95$.

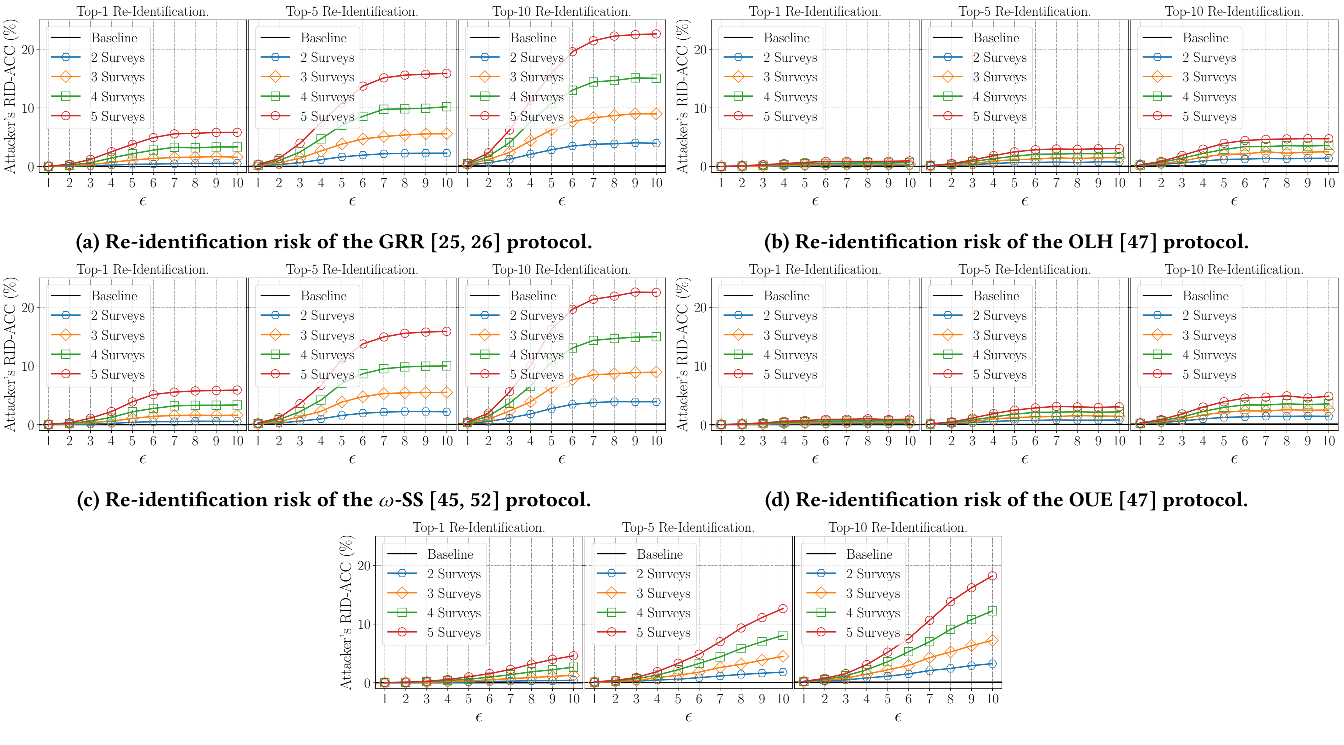
D ADDITIONAL RESULTS FOR SECTION 4.3

This section provides additional results for the inference of the sampled attribute on collecting multidimensional data with the RS+FD [5] solution. We use the state-of-the-art XGBoost [9] algorithm to predict the sampled attribute of users in a multiclass classification framework (i.e., d attributes) with default parameters. We follow the experimental evaluation described in Section 4.3 and we vary the following:

- **Dataset.** We use the Adult ($d = 10$ attributes, $n = 45,222$, $\mathbf{k} = [74, 7, 16, 7, 14, 6, 5, 2, 41, 2]$) and Nursery ($d = 9$ attributes, $n = 12,959$, $\mathbf{k} = [3, 5, 4, 4, 3, 2, 3, 3, 5]$) datasets from the UCI ML repository [17].
- **Frequency oracle within RS+FD.** All protocols from Section 2.3.2, namely, RS+FD[GRR], RS+FD[SUE-z], RS+FD[SUE-r], RS+FD[OUE-z] and RS+FD[OUE-r].
- **Attribute inference model.** All five protocols are evaluated with the three attack models of Section 3.3, namely, No Knowledge (NK), Partial-Knowledge (PK) and Hybrid Model (HM).

Figs. 15 and 16 illustrates the attacker's attribute inference accuracy (AIF-ACC) metric on the Adult and Nursery datasets, respectively, with the three attack models (i.e., NK, PK and HM) and all five protocols (i.e., RS+FD[GRR], RS+FD[SUE-z], RS+FD[OUE-z], RS+FD[SUE-r] and RS+FD[OUE-r]), varying ϵ, the number of synthetic profiles s and the number of compromised profiles n_{pk}.

Similar to the results with the ACSEmployement [15] dataset in Fig. 3, one can notice in Figs. 15 and 16 that the proposed attack models, namely, NK, PK and HM present significant increments in the attacker's AIF-ACC over the Baseline model. More precisely, with the Adult dataset, there is about a 1.3-10 fold increment over a random Baseline model with our NK, PK and HM models. On the one hand, there is about a 0.1-10 fold increment with the Nursery dataset. More precisely, the attack models with both RS+FD[GRR]



(a) Re-identification risk of the GRR [25, 26] protocol.

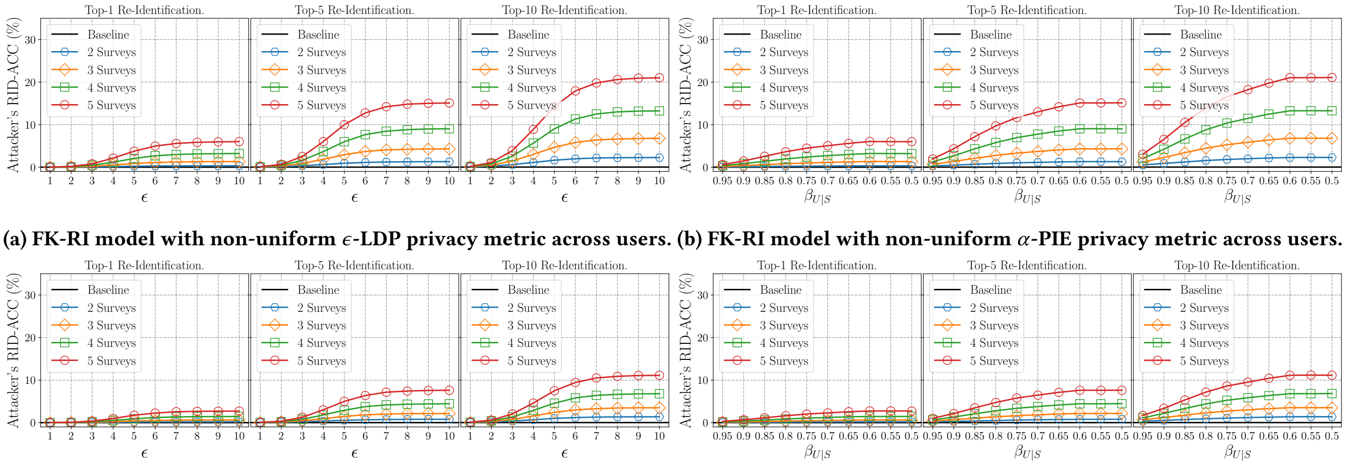
(b) Re-identification risk of the OLH [47] protocol.

(c) Re-identification risk of the ω-SS [45, 52] protocol.

(d) Re-identification risk of the OUE [47] protocol.

(e) Re-identification risk of the SUE (a.k.a. Basic One-Time RAPPOR [23]) protocol.

Figure 9: Attacker's re-identification accuracy (RID-ACC) on the ACSEmployement dataset for top-k re-identification on using the SMP solution, the full knowledge FK-RI model with uniform ϵ-LDP privacy metric across users, and by varying the frequency oracle and the number of surveys (i.e., data collections).

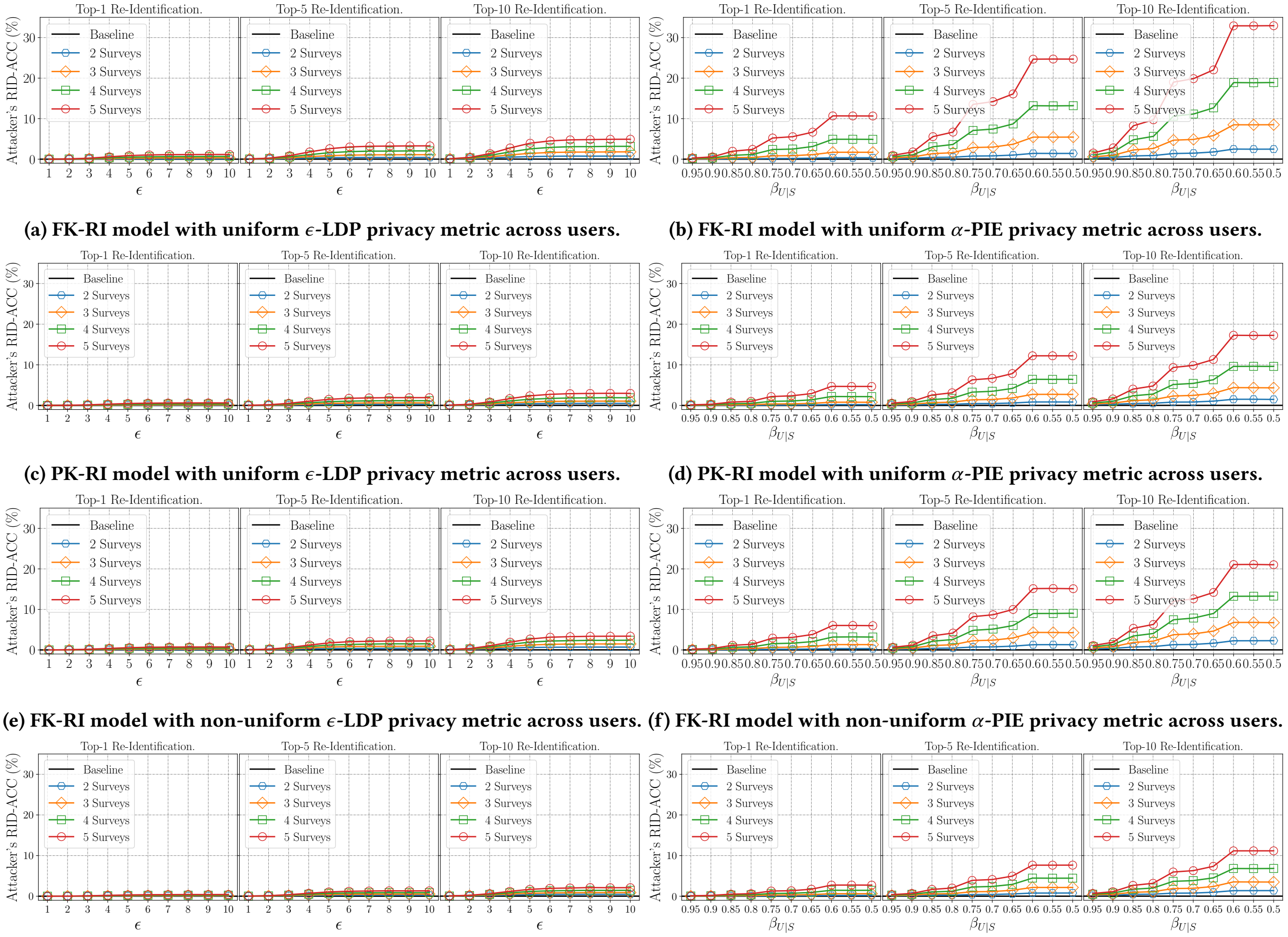


(a) FK-RI model with non-uniform ϵ-LDP privacy metric across users. **(b) FK-RI model with non-uniform α-PIE privacy metric across users.**

(c) PK-RI model with non-uniform ϵ-LDP privacy metric across users. **(d) PK-RI model with non-uniform α-PIE privacy metric across users.**

Figure 10: Attacker's re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification on using the SMP solution with the GRR [25, 26] protocol by varying the non-uniform privacy metric (i.e., ϵ for LDP and $\beta_{U|S}$ for α-PIE), the attack model (i.e., full knowledge – FK-RI, partial knowledge – PK-RI) and the number of surveys (i.e., data collections).

and RS+FD[UE-r] protocols did not provide a meaningful increment over the Baseline model in the Nursery dataset. The reason behind this is that the attributes follow uniform-like distributions. Thus, since fake data are also generated uniformly at random with the



(a) FK-RI model with uniform ϵ-LDP privacy metric across users. **(b) FK-RI model with uniform α-PIE privacy metric across users.**

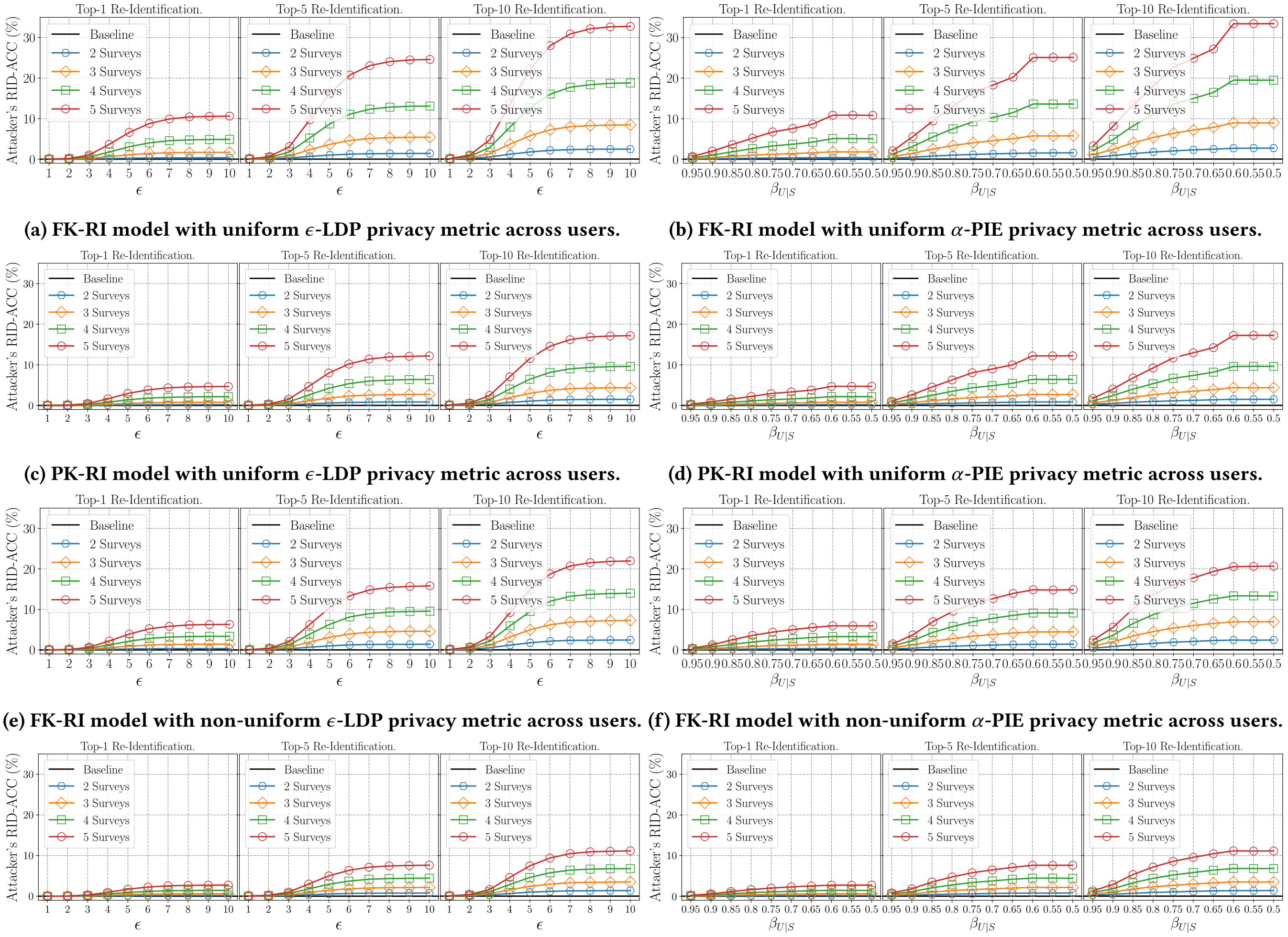
(c) PK-RI model with uniform ϵ-LDP privacy metric across users. **(d) PK-RI model with uniform α-PIE privacy metric across users.**

(e) FK-RI model with non-uniform ϵ-LDP privacy metric across users. **(f) FK-RI model with non-uniform α-PIE privacy metric across users.**

(g) PK-RI model with non-uniform ϵ-LDP privacy metric across users. **(h) PK-RI model with non-uniform α-PIE privacy metric across users.**

Figure 11: Attacker's re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification on using the SMP solution with the OLH [47] protocol by varying the attack model (i.e., full knowledge – FK-RI, partial knowledge – PK-RI), the number of surveys (i.e., data collections) and the privacy metric (i.e., ϵ for LDP and $\beta_{U|S}$ for α-PIE) in both uniform and non-uniform settings.

RS+FD solution, the classifier is not able to distinguish between real and fake data when predicting the sampled attribute. Yet, the attacker's AIF-ACC also achieves about 100% with RS+FD[SUE-z] with all three datasets (see Figs. 3 and. 15). Lastly, increasing the number of synthetic profiles the attacker generates s and/or the number of compromised profiles the attacker has access to n_{pk}, had few influence with the Adult dataset in Fig. 15. Conversely, both ACSEmployement (Fig. 3) and Nursery (Fig. 16) datasets showed sensitivity to a change in both parameters, especially with n_{pk} in the PK model.



(a) FK-RI model with uniform ϵ-LDP privacy metric across users.

(b) FK-RI model with uniform α-PIE privacy metric across users.

(c) PK-RI model with uniform ϵ-LDP privacy metric across users.

(d) PK-RI model with uniform α-PIE privacy metric across users.

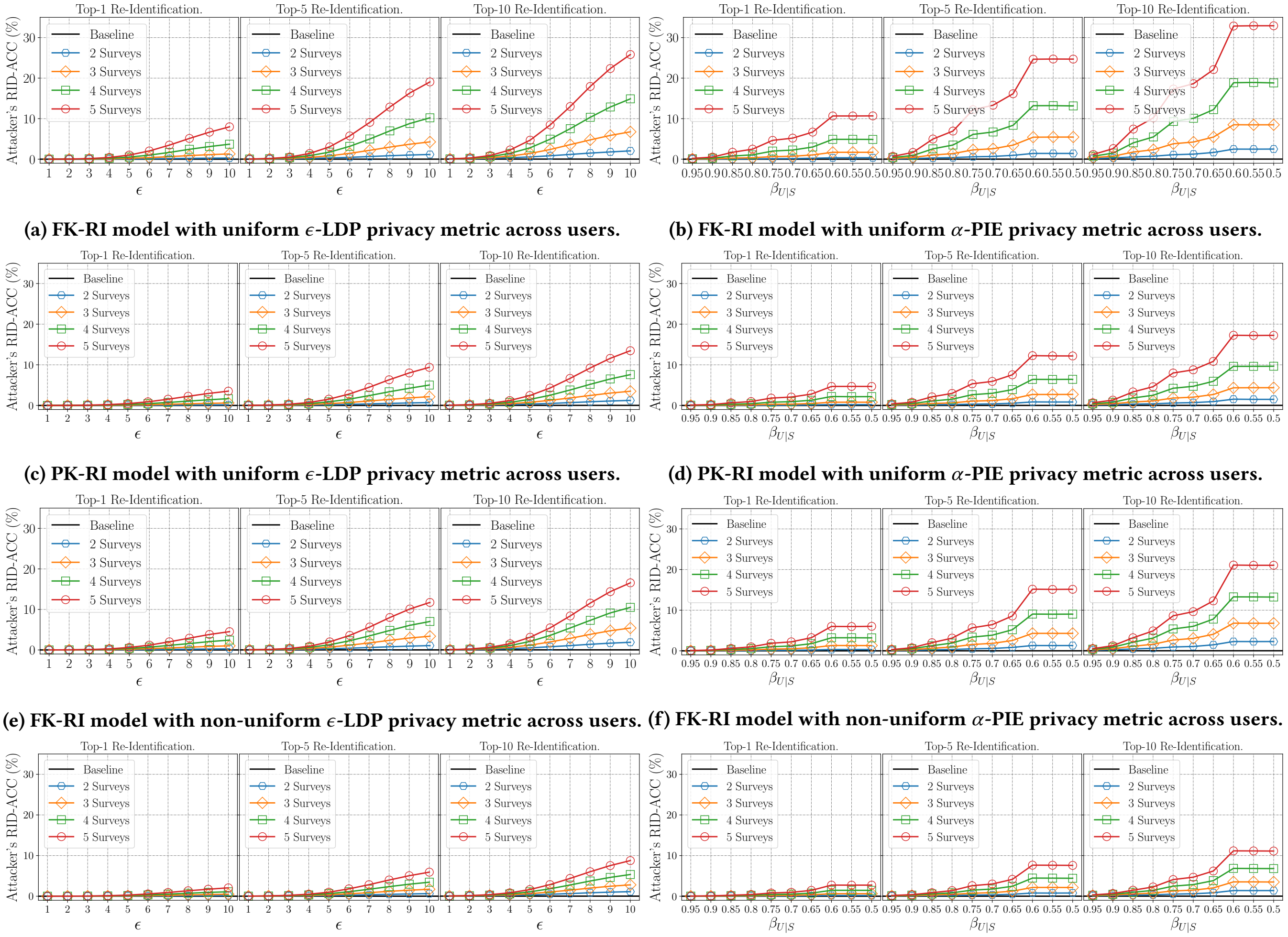
(e) FK-RI model with non-uniform ϵ-LDP privacy metric across users.

(f) FK-RI model with non-uniform α-PIE privacy metric across users.

(g) PK-RI model with non-uniform ϵ-LDP privacy metric across users.

(h) PK-RI model with non-uniform α-PIE privacy metric across users.

Figure 12: Attacker's re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification on using the SMP solution with the ω-SS [45, 52] protocol by varying the attack model (i.e., full knowledge – FK-RI, partial knowledge – PK-RI), the number of surveys (i.e., data collections) and the privacy metric (i.e., ϵ for LDP and $\beta_{U|S}$ for α-PIE) in both uniform and non-uniform settings.



(a) FK-RI model with uniform ϵ-LDP privacy metric across users.

(b) FK-RI model with uniform α-PIE privacy metric across users.

(c) PK-RI model with uniform ϵ-LDP privacy metric across users.

(d) PK-RI model with uniform α-PIE privacy metric across users.

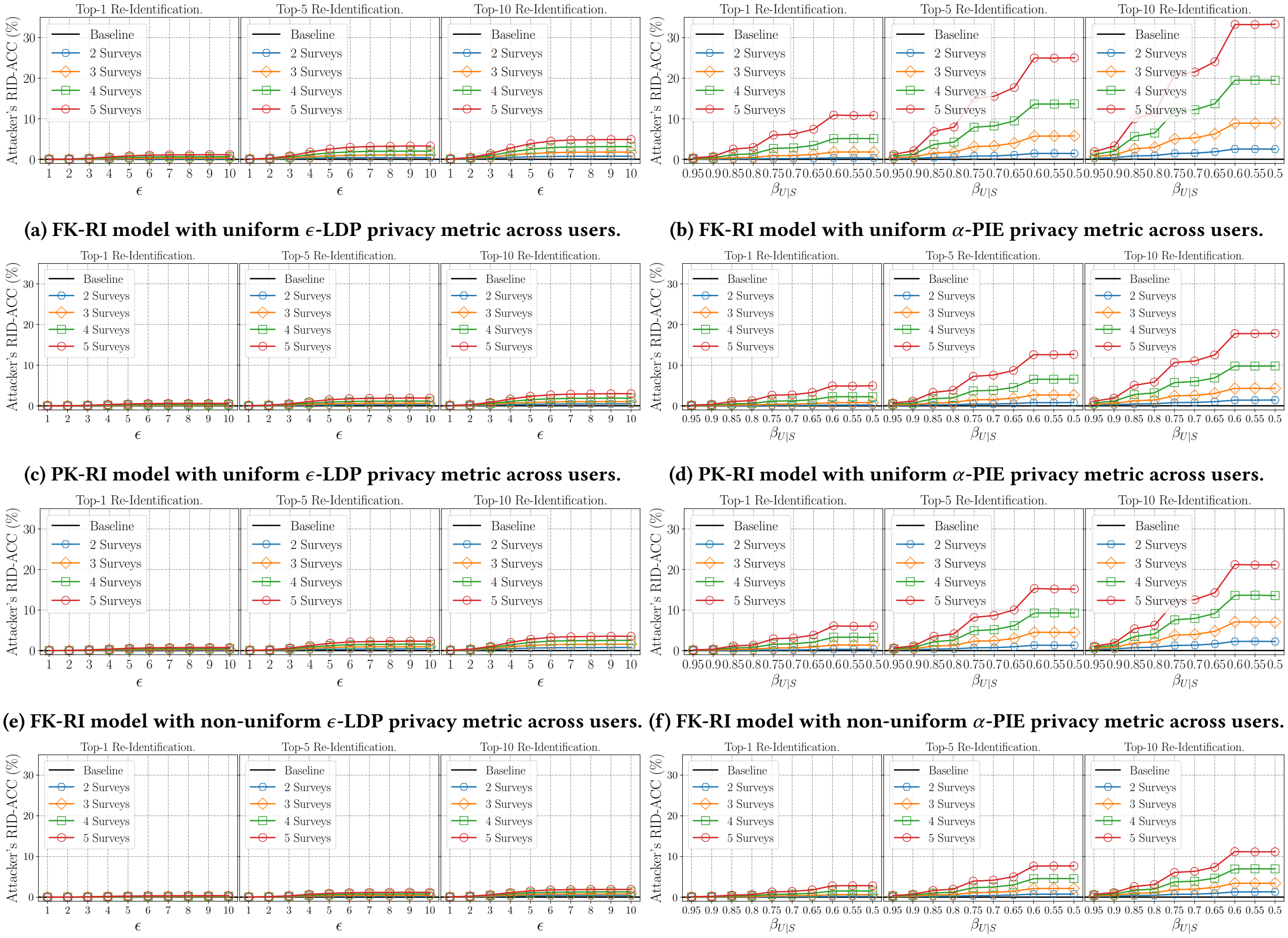
(e) FK-RI model with non-uniform ϵ-LDP privacy metric across users.

(f) FK-RI model with non-uniform α-PIE privacy metric across users.

(g) PK-RI model with non-uniform ϵ-LDP privacy metric across users.

(h) PK-RI model with non-uniform α-PIE privacy metric across users.

Figure 13: Attacker's re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification on using the SMP solution with the SUE [23] protocol by varying the attack model (i.e., full knowledge – FK-RI, partial knowledge – PK-RI), the number of surveys (i.e., data collections) and the privacy metric (i.e., ϵ for LDP and $\beta_{U|S}$ for α-PIE) in both uniform and non-uniform settings.



(a) FK-RI model with uniform ϵ-LDP privacy metric across users.

(b) FK-RI model with uniform α-PIE privacy metric across users.

(c) PK-RI model with uniform ϵ-LDP privacy metric across users.

(d) PK-RI model with uniform α-PIE privacy metric across users.

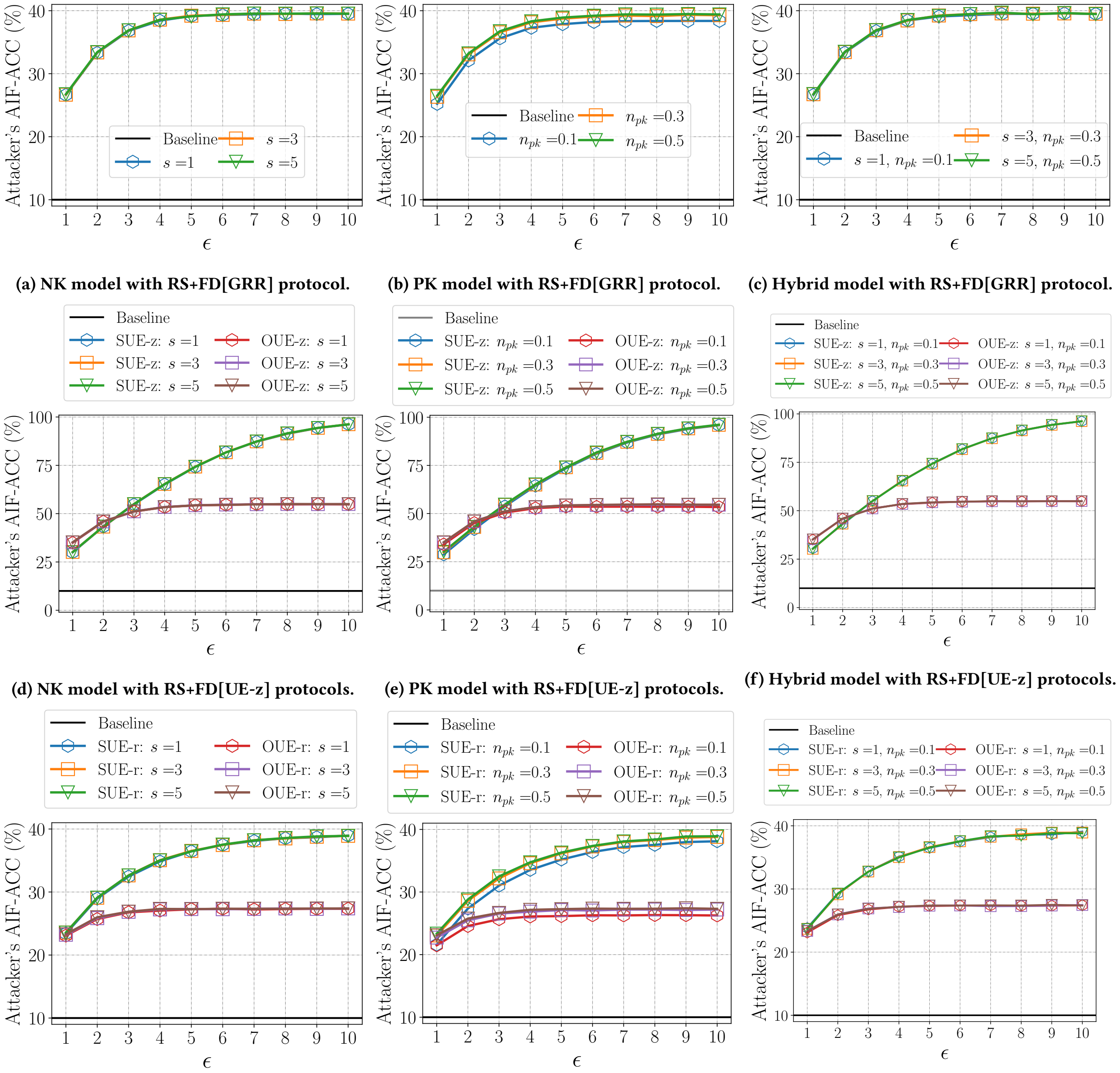
(e) FK-RI model with non-uniform ϵ-LDP privacy metric across users.

(f) FK-RI model with non-uniform α-PIE privacy metric across users.

(g) PK-RI model with non-uniform ϵ-LDP privacy metric across users.

(h) PK-RI model with non-uniform α-PIE privacy metric across users.

Figure 14: Attacker's re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification on using the SMP solution with the OUE [47] protocol by varying the attack model (i.e., full knowledge – FK-RI, partial knowledge – PK-RI), the number of surveys (i.e., data collections) and the privacy metric (i.e., ϵ for LDP and $\beta_{U|S}$ for α-PIE) in both uniform and non-uniform settings.



(a) NK model with RS+FD[GRR] protocol.

(b) PK model with RS+FD[GRR] protocol.

(c) Hybrid model with RS+FD[GRR] protocol.

(d) NK model with RS+FD[UE-z] protocols.

(e) PK model with RS+FD[UE-z] protocols.

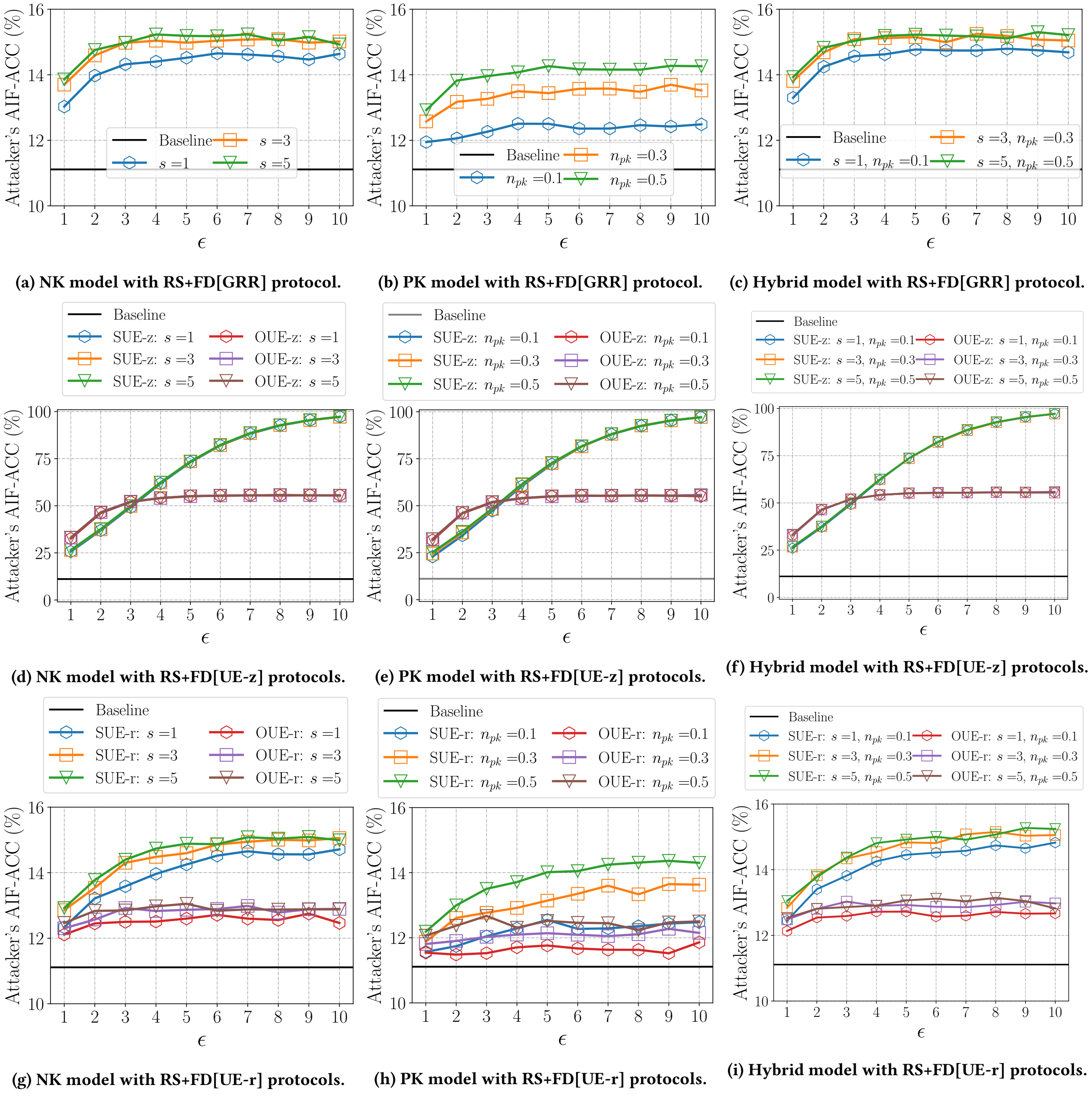
(f) Hybrid model with RS+FD[UE-z] protocols.

(g) NK model with RS+FD[UE-r] protocols.

(h) PK model with RS+FD[UE-r] protocols.

(i) Hybrid model with RS+FD[UE-r] protocols.

Figure 15: Attacker's attribute inference accuracy (AIF-ACC) on the Adult dataset with three attack models (i.e., NK, PK and hybrid) and five protocols (i.e., RS+FD[GRR], RS+FD[SUE-z], RS+FD[OUE-z], RS+FD[SUE-r] and RS+FD[OUE-r]), varying ϵ, the number of synthetic profiles s the attacker generates and the number of compromised profiles n_{pk} the attacker has access to.



(a) NK model with RS+FD[GRR] protocol. **(b) PK model with RS+FD[GRR] protocol.** **(c) Hybrid model with RS+FD[GRR] protocol.**

(d) NK model with RS+FD[UE-z] protocols. **(e) PK model with RS+FD[UE-z] protocols.** **(f) Hybrid model with RS+FD[UE-z] protocols.**

(g) NK model with RS+FD[UE-r] protocols. **(h) PK model with RS+FD[UE-r] protocols.** **(i) Hybrid model with RS+FD[UE-r] protocols.**

Figure 16: Attacker's attribute inference accuracy (AIF-ACC) on the Nursery dataset with three attack models (i.e., NK, PK and hybrid) and five protocols (i.e., RS+FD[GRR], RS+FD[SUE-z], RS+FD[OUE-z], RS+FD[SUE-r] and RS+FD[OUE-r]), varying ϵ, the number of synthetic profiles s the attacker generates and the number of compromised profiles n_{pk} the attacker has access to.