On the Risks of Collecting Multidimensional Data Under Local Differential Privacy

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

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

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

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 experimented 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(5): 1126 - 1139, 2023. doi:10.14778/3579075.3579086

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, 28], 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 [53], 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 [45].

A fundamental task under LDP guarantees is frequency estimation [12, 14, 23, 26, 27, 45, 49, 50], 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 [6, 10, 32, 54] and privacy [8, 22, 24, 35] threats to LDP protocols (mainly for frequency estimation), which are discussed in detail in Section 7.

In this paper, we further investigate the privacy threats for the users when the server aims to perform frequency estimation of multiple attributes under LDP guarantees. In this setting [4, 37, 46, 47, 50], 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 \( j \in [d] \). There are \( n \) users \( \mathcal{U} = \{u_1, \ldots, u_n\} \), and each user \( u_i \), for \( i \in [n] \), holds a private tuple \( v^{(i)} = (v^{(i)}_1, v^{(i)}_2, \ldots, v^{(i)}_d) \), in which \( v^{(i)}_j \) represents the value of attribute \( A_j \) in record \( v^{(i)} \). Thus, for each attribute \( A_j \in \mathcal{A} \), for \( j \in [d] \), the aggregator’s goal is to estimate a \( k_j \)-bins histogram.

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 \( v = (v_1, v_2, \ldots, v_d) \), which are described in the following:

- **Splitting (SPL).** This naive solution directly splits the privacy budget \( \epsilon \) by \( d \) attributes and reports all attributes with \( \epsilon/d \)-LDP, thus incurring a high estimation error [5, 37, 47, 50].
- **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 \( \epsilon \)-LDP [5, 37, 47, 50].
- **Random Sampling Plus Fake Data (RS+FD) [4].** 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 state-of-the-art RS+FD solution is precisely to enable users to “hide” the sampled attribute (i.e., \( \epsilon \)-LDP value) by also generating one uniformly random fake data for each non-sampled attribute. Thus, RS+FD creates uncertainty on the server-side.

---

\(^1\)This is a different task of joint distribution estimation under LDP guarantees [38, 56].

\(^2\)For simplicity, we omit the index notation \( v^{(i)} \) and focus on one arbitrary user \( u_i \).
Focusing on the state-of-the-art solutions SMP and RS+FD, first, we empirically demonstrate through extensive experiments that the **SMP solution is vulnerable to re-identification attacks** when collecting users’ multidimensional data several times with $\epsilon$ values commonly used by industry nowadays [13, 39]. For instance, assume 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 $\epsilon$-LDP report. Nevertheless, we show that an adversary who can see every tuple containing (sampled attribute, $\epsilon$-LDP report) 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 characterized, one can leverage well-known attacks (e.g., homogeneity) [11, 30, 33, 40–43].

More specifically, to attack the SMP solution, our adversarial analysis focuses on the reduced "plausible deniability" [16, 53] of using the whole privacy budget $\epsilon$ to report a single attribute out of $d$ ones. In this setting, the adversary has a higher chance to infer the users’ true value for each data collection performed. Consequently, in multiple data collections, the adversary can build partial or even sometimes complete profiles of each user, then using it to perform a re-identification attack. However, this depends on the LDP protocol being used as the encoding and randomization vary across them [12, 50]. In our experiments, we have assessed five widely used LDP protocols for frequency estimation (a.k.a. frequency oracle protocols [51, 52]), namely Generalized Randomized Response (GRR) [26, 27], Optimal Local Hashing (OLH) [50], Subset Selection (SS) [48, 55] and two Unary Encoding (UE) protocols (Basic One-time RAPPOR [23] and Optimal UE [50]). To assess the risks of re-identification we have also considered two privacy models, usual LDP and the relaxed version of LDP developed in [35] (the latter in Appendix C) for measuring re-identification risks.

Secondly, we observe that since the RS+FD solution generates fake data uniformly at random in [4, 46], 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 in hiding the sampled attribute to the aggregator by varying the privacy budget $\epsilon$, 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 on multiple collections. 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 due to chaining errors from incorrectly predicting the sampled attribute and user’s value in multiple collections. 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 [5, 37, 47, 50] and RS+FD [4], providing insightful adversarial analysis to 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" [16, 53] when using the whole privacy budget to report a single attribute.
- We propose three attack models to predict the sampled attribute of users when collecting multidimensional data with the RS+FD solution with about a 2-20 fold increment over a random guess baseline model.
- We show through empirical results that the RS+FD solution can prevent (to some extent) re-identification attacks.
- Finally, we present an adaptation of the RS+FD solution that can serve as a countermeasure to the identified privacy threats while improving both privacy and utility.

**Outline.** In Section 2, we review the LDP privacy model, 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 analyze our results before in Section 5 presenting an improvement of the RS+FD as a countermeasure. Next, we provide a general discussion in Section 6. Finally in Section 7, we review related work before concluding with future perspectives of this work in Section 8.

## 2 PRELIMINARIES

This section briefly reviews the LDP model, state-of-the-art LDP frequency estimation protocols and three solutions for multiple attribute frequency estimation under LDP.

### 2.1 Local Differential Privacy

In this paper, we use LDP (Local Differential Privacy) [18, 28] as the privacy model considered, which is formalized as:

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

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

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 $\epsilon$ controls the privacy-utility trade-off for which lower values of $\epsilon$ result in tighter privacy protection. Similar to central DP, LDP also has several important properties, such as immunity to post-processing and composability [21].

### 2.2 LDP Frequency Estimation 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 $i \in \{k_j\}$, under LDP guarantees.
2.2.1 Generalized Randomized Response. Randomized response (RR) [53] is the classical technique for achieving LDP, which provides “plausible deniability” for individuals responding to embarrassing (binary) questions in a survey. The Generalized RR (GRR) [26, 27] protocol extends RR to the case of \( k_j \geq 2 \) while satisfying \( \varepsilon \)-LDP. Given a value \( v_i \in A_j \), for \( i \in [k_j] \), 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:

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

in which \( y \) is the perturbed value sent to the aggregator. The GRR protocol satisfy \( \varepsilon \)-LDP since \( e^r = e^\varepsilon \). To estimate the normalized frequency of \( v_i \in A_j \), for \( i \in [k_j] \), one counts how many times \( v_i \) is reported, expressed as \( C(v_i) \), and then computes [50]:

\[
f'(v_j) = \frac{C(v_j) - nq}{n(p - q)}, \tag{2}
\]

in which \( n \) is the total number of users. In [50], it was proven that Eq. (2) is an unbiased estimator (i.e., \( E(f'(v_i)) = f(v_i) \)).

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 [50] have proposed Optimal LH (OLH), which selects \( g_j = e^r + 1 \). Given a value \( v_i \in A_j \), for \( i \in [k_j] \), 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\()\) 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:

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

in which \( y \) is the hash function and perturbed value sent to the aggregator. From this, the aggregator can obtain the unbiased estimation of \( v_i \in A_j \), for \( i \in [k_j] \), with Eq. (2) by setting \( p = p' \) and \( q = \frac{1}{g_j} \cdot p' + \left(1 - \frac{1}{g_j}\right) \cdot q' = \frac{1}{g_j} \) [50].

2.2.3 Subset Selection. The main idea of \( \omega \)-Subset Selection (\( \omega \)-SS) [48, 55] is to randomly select \( \omega \) 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 \( i \in [k_j] \), has higher probability of being included in the subset \( \Omega \), compared to other values in \( A_j \setminus \{v_i\} \) that are sampled uniformly at random (without replacement). The optimal subset size \( \omega = \lceil |\Omega| \rceil \) that minimizes the variance is \( \omega = \frac{k_j}{e^r} \) [48, 55].

Given a value \( v_i \in A_j \), for \( i \in [k_j] \), the \( \omega \)-SS protocol starts by initializes an empty subset \( \Omega \). Afterwards, the true value \( v_i \) is added to \( \Omega \) with probability \( p = \frac{e^r}{e^r + k_j - \omega} \). Finally, it adds values to \( \Omega \) as follows [48, 55]:

- If \( v_i \) has been added to \( \Omega \) 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 \( \Omega \);
- If \( v_i \) has not been added to \( \Omega \) 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 \( \Omega \);
- If \( v_i \) has not been added to \( \Omega \) 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 \( \Omega \).

From this, the aggregator can obtain the unbiased estimation of \( v_i \in A_j \), for \( i \in [k_j] \), with Eq. (2) by setting \( p = \frac{e^r}{e^r + k_j - \omega} \) and \( q = \frac{1}{e^r + k_j - \omega} \) [48, 55].

2.2.4 Unary Encoding Protocols. Unary encoding (UE) protocols interpret the user’s input \( v_i \in A_j \), for \( i \in [k_j] \), as a one-hot \( k_j \)-dimensional vector. More specifically, \( B = UE(v_i) \) 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 (SU) [50], which randomizes the bits from \( B \) independently with probabilities:

\[
\forall i \in [k_j] : \quad \Pr[B_i = 1] = \begin{cases} 
    p = \frac{e^r/2}{e^r/2 + 1}, & \text{if } B_i = 1 \\
    q = \frac{1}{e^r/2 + 1}, & \text{if } B_i = 0.
\end{cases} \tag{3}
\]

Afterwards, the client sends \( B \) to the aggregator. More recently, to minimize the variance of the SU protocol, the authors in [50] proposed Optimal UE (OUE), which selects probabilities \( p = \frac{1}{2} \) and \( q = \frac{1}{e^r/2 + 1} \) in Eq. (3) asymmetrically (i.e., \( p \neq q \)). The estimation method used in Eq. (2) applies equally to both SU and OUE protocols, in which both satisfy \( \varepsilon \)-LDP for \( \varepsilon = \ln \frac{(2(1-q))}{(1-p)q} \) [23, 50].

2.3 Multidimensional Frequency Estimation

Let \( n \) be the total number of users, \( d \geq 2 \) be the total number of attributes, \( k = [k_1, k_2, \ldots, k_d] \) be the domain size of each attribute, \( M \) be a local randomizer and \( \varepsilon \) be the privacy budget. Each user holds a tuple \( v = [v_1, v_2, \ldots, v_d] \), (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. Previous works in the local DP setting considered the following approaches [5, 37, 47, 50]:

- **SPL.** On the one hand, due to the sequential composition theorem [21], users can split the privacy budget \( \varepsilon \) over the number of attributes \( d \) and send all randomized values \( y_j \), for \( j \in [d] \), with \( \frac{\varepsilon}{d} \)-LDP to the aggregator (i.e., a tuple \( y = [y_1, y_2, \ldots, y_d] \)). However, this naive SPL solution leads to high estimation error [5, 37, 47, 50].

- **SMP.** Instead of splitting the privacy budget \( \varepsilon \), this state-of-the-art solution allows each user to sample a single attribute \( j \in [d] \) at random and uses all the privacy budget to send it with \( \varepsilon \)-LDP [5, 37, 47, 50]. In this case, each user tells the aggregator which attribute is sampled, and what is the perturbed value for it ensuring \( \varepsilon \)-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 attributes, 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 [4] solution is composed of two steps, namely local randomization and fake data generation. More precisely, each user samples a unique attribute uniformly at random \( j = Uniform([d]) \) and uses an \( \varepsilon \)-LDP protocol to sanitize its value \( v_j \). Next, for each non-sampled
attribute \( i \in [d] \setminus \{ j \} \), the user generates uniform 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 \( y = [y_1, y_2, \ldots, y_d] \)). In this manner, the sampling result is not disclosed to the aggregator, thus increasing the uncertainty. For this reason, to satisfy \( \epsilon \)-LDP, following the parallel composition theorem [21] and the amplification by sampling result [31], RS+FD utilizes an amplified privacy budget \( \epsilon' = \ln(d \cdot (\epsilon^e - 1) + 1) \) for the sampled attribute [4].

With the RS+FD solution, the estimator should remove the bias introduced by the local randomizer \( M \) and uniform fake data. In [4], the authors used GRR and OUE as LDP protocols within the RS+FD solution, which results in RS+FD[GRR], RS+FD[OUE-x] 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 [4, 46].

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

- **RS+FD[GRR]** [4]. For each non-sampled attribute \( i \in [d] \setminus \{ j \} \), the user generates fake data uniformly at random according to the domain size \( k_j \). On the server-side, the unbiased estimator for this protocol is: 
  \[
  \hat{f}(v_i) = \frac{C(v_i) k_j - n(d-1) v_i}{n k_j(p-q)},
  \]
  in which \( C(v_i) \) is the number of times \( v_i \) has been reported, \( p = \frac{\epsilon^e}{e^e k_j - 1} \) and \( q = 1 - \frac{p}{k_j - 1} \).

- **RS+FD[UE-x]** [4]. For each non-sampled attribute \( i \in [d] \setminus \{ j \} \), 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: 
  \[
  \hat{f}(v_i) = \frac{d(C(v_i) - n q)}{n(p-q)},
  \]
  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 [50] protocols.

- **RS+FD[UE-r]** [4]. For each non-sampled attribute \( i \in [d] \setminus \{ j \} \), 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: 
  \[
  \hat{f}(v_i) = \frac{C(v_i) k_j - n q k_j p (p-q)}{n k_j (p-q)},
  \]
  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 [50] protocols.

## 3 SYSTEM OVERVIEW & PRIVACY THREATS

Hereafter, we describe the system and adversary models before presenting our adversarial analyses of SMP and RS+FD.

### 3.1 System Overview

We consider the situation in which a (possibly untrusted) server collects users’ multidimensional data \( d \geq 2 \) for frequency estimation under \( \epsilon \)-LDP guarantees multiple times. Particularly, in each data collection (i.e., survey), the server can select a different number of attributes. 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 \( U \) 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. We assume that the server uses one of the state-of-the-art LDP solutions (e.g., SMP or RS+FD) to collect one random attribute per user. Thus, we do not consider the SPL solution in our attacks as all attributes would be collected at once, thus resulting in a low level of utility [5, 37, 47, 50].

**Adversary model.** Following the LDP assumptions [18, 28], 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 \( D_{BK} \) coming from public available source, such as Census data [1]. 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 \( A_j = [v_y, v_a] \) (e.g., a value “Yes” for an attribute \( A_j \) denoting whether someone cheated on their partner). As long as \( \Pr[M(v_y) = v_y] < 1 \), the user can deny to have \( A_j = v_y \) [16].

The LDP protocols of Section 2.2 are based on RR [53], which provides “plausible deniability” for users’ reports. However, increasing \( \epsilon \) to improve utility of LDP protocols compromises the “plausible deniability” of the users’ reports. Indeed, common \( \epsilon \) values used daily by users in high-scale industrial systems nowadays range from small \( \epsilon \leq 1 \) to high values \( \epsilon \geq 8 \) [13, 39, 44]. 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 \( \epsilon \) 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 setting, in which many surveys are proposed by the server to the same set of users with possibly different number of attributes (e.g., demographic, preference, application usage, . . .), an attacker knowing the tuple (sampled attribute, \( \epsilon \)-LDP report) will be able to profile each user throughout time. Therefore, once a partial or complete profile of the target user is built (see details in Sections 3.2.2, 3.2.3 and 3.2.4), the adversary could use his background knowledge \( D_{BK} \) to possibly re-identify a user within population [30, 33, 40–43], possibly also inferring all other available attributes. The next four subsections analyze the “plausible deniability” of LDP protocols in single and multiple collections, and describes the proposed re-identification attack models, respectively.

#### 3.2.1 Plausible Deniability of LDP protocols.

Given a user’s true value \( v \in A_j \), different LDP protocols \( M \) have different type of
output \( y_i = M(\omega, \epsilon) \) [50]. For instance, UE protocols output unary encoded vectors, \( \omega \) SS outputs a subset \( \Omega \) of \( \omega \) non-encoded values and so on (cf. Section 2.2). Thus, for each user \( u_i \in U \), for \( i \in [n] \), 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 \( \hat{v} = \hat{v} \) over the number of users \( n \): 

\[
\text{ACC}(\%) = 100 \cdot \frac{\sum_{i=1}^{n} I(v_i, \hat{v}_i)}{n},
\]

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 \( v \in A_j \backslash \{v_i\} \), we now describe our attack strategy to each LDP protocol. By the time of completing this paper, we learned about a recent work showing that the expectation of our attacks could be analytically formalized with the Bayes adversary of [22]. We believe this work is complementary to our "plausible deniability" attacking interpretation.

**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 \( \text{ACC}_{GRR}(\%) = 100 \cdot \frac{1}{2} \).

**Plausible Deniability of OLH.** Since the output of \( \omega \)-SS for user \( u_i \) is a binary function \( H_i \) used to hash the user’s value and the hashed value \( h_i = H_i(u) \), 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., \( A_{h_i} = \{ v \in A_j : H_i(u) = v \} \)). Thus, the attacker’s best guess is a random choice \( \hat{v} \in \text{Uniform}(A_{h_i}) \). On expectation [22], one achieves: 

\[
\text{ACC}_{OLH}(\%) = 100 \cdot \frac{1}{2} \max \left\{ \frac{1}{2^n}, 1 \right\}.
\]

**Plausible Deniability of \( \omega \)-SS.** Since the output of \( \omega \)-SS for user \( u_i \) is set \( \Omega \subseteq A_j \), the most likely value after randomization is one within the subset \( \Omega \). Thus, the attacker’s best guess is a random choice \( \hat{v} \in \text{Uniform}(\Omega) \). Selecting \( \omega = \frac{k_j}{\epsilon^{1/4}} \) in \( \omega \)-SS [48, 55], on expectation [22], one achieves: 

\[
\text{ACC}_{\omega-SS}(\%) = 100 \cdot \frac{1}{2} \max \left\{ \frac{1}{2^n}, 1 \right\}.
\]

**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 \( \hat{v} = b_i \); 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 \( \hat{v} \in \text{Uniform} \left( \{ b \in \{ k_j \} : B_j = 1 \} \right) \); and 3) no bit in \( B \) is set to 1, in which the attacker’s best guess is a random choice of the domain \( \hat{v} \in \text{Uniform}(A_j) \). Therefore, on expectation [22], the attacker’s accuracy for UE is: 

\[
\text{ACC}_{UE}(\%) = 100 \cdot \frac{1}{k_j} \cdot \frac{1}{\epsilon^{1/4}} \cdot \frac{e^{1/2} - k_j^{-1}}{e^{1/2} + k_j^{-1}} + \sum_{i=1}^{k_j} \left( \frac{e^{1/2}}{e^{1/2} + 1} \right)^i \cdot \text{Bin} \left( i - 1; k_j^{-1}, 1 - \frac{1}{\epsilon^{1/2}} \right),
\]

in which \( \text{Bin}(\cdot) \) denotes a Binomial distribution with \( k_j^{-1} \) trials, success probability \( 1 - \frac{1}{\epsilon^{1/2}} \), and exactly \( i - 1 \) successes. On the other hand, on expectation [22], the attacker’s accuracy for OUE is: 

\[
\text{ACC}_{OUE}(\%) = 100 \cdot \frac{1}{\epsilon^{1/4}} \cdot \frac{e^{1/2} - k_j^{-1}}{e^{1/2} + k_j^{-1}} + \sum_{i=1}^{k_j} \frac{k_j}{2i} \cdot \text{Bin} \left( i - 1; k_j^{-1}, 1 - \frac{1}{\epsilon^{1/2}} \right).
\]

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

When collecting multidimensional data \( d \geq 2 \) 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 \( \text{ACC}_{FO} \) depends on \( \epsilon \) 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 \#surveys = \( d \), denoted as \( \text{ACC}^{NU}_{FO} \), now follows:

\[
\text{ACC}^{NU}_{FO}(\%) = 100 \cdot \frac{d}{d + 1} \cdot \frac{\text{ACC}_{FO}(\epsilon, k_j)}{d}.
\]

Since each survey is independent and samples without replacement, Eq. (4) represents the expected probability of accurately profiling users with exactly \( d \) attributes.

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

On the other hand, when collecting multidimensional data \( d \geq 2 \) with the SMP solution multiple times, the server can allow users to sample attributes with replacements in each data collection. In case of a repeated attribute, the user can report the previous randomized value (a.k.a. memoization [5, 14, 23]). This way, users will have a non-uniform privacy metric. Depending on the LDP protocol, the expected ACC with non-uniform privacy metric after \#surveys = \( d \), denoted as \( \text{ACC}^{NU}_{FO} \), now follows:

\[
\text{ACC}^{NU}_{FO}(\%) = 100 \cdot \frac{d}{d + 1} \cdot \frac{\text{ACC}_{FO}(\epsilon, k_j)}{d}.
\]

Since each survey is independent but attributes are sampled with replacement, Eq. (5) denotes the overall adversary’s accuracy only considering users that report a different attribute in each survey (i.e., of accurately profiling users with exactly \( d \) attributes). Thus, in this setting, users can also end-up with partial profiles.

**Analytical analysis of expected ACC.** In Fig. 1, we illustrate the expected \( \text{ACC}^{NU}_{FO} \) following Eq. (4) and the \( \text{ACC}^{NU}_{FO} \) following Eq. (5) of each LDP protocol with the following parameters (taken from Section 4): \( \epsilon = [1, 2, \ldots, 9, 10] \), \( d = 3 \), \( k = [74, 7, 16] \), and \#surveys = \( d \). From Fig. 1 (a), one can notice that GRR, \( \omega \)-SS and UE have the highest attacker’s accuracy, which would enable an adversary to accurately infer a complete profile after \#surveys = \( d \). 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 \( \frac{d!}{d^d} \).

Note that the expected ACC in both Eqs. (4) and (5) decreases with the \#surveys since the probability of accurately inferring the users’ true value is independent in each survey.

### 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 \( D_{BK} \) to re-identify the target user. The latter PK-RI model considers that the attacker only has access to a subset \( D_{PK} \subseteq D_{BK} \) 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 protocol considered.

In particular, after \#surveys, the attacker will have a profile \( y_i \) of at most \#surveys sanitized values for the target user \( u_i \). 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 $R$, which takes as input the sanitized profile $y_i$ and the background knowledge $D_{BK}$ (or $D_{PK}$ for PK-RD), and outputs a score $c_i \in \mathbb{R}$. More precisely, the score $c_i$ measures the distance between the target $y_i$ and all samples $r \in D_{BK}$. Since the LDP protocols from Section 2.2 do not have a notion of "distance" when randomizing a value, when an attribute in $y_i$ is not $r$, the distance is 1 and 0 otherwise. A smaller distance between $y_i$ and a profile in $D_{BK}$ indicates that it is highly likely that $y_i$ has been re-identified through the uniqueness combination of $\epsilon$ surveys attributes [30, 33, 40–43]. Finally, a decision algorithm $G$ takes as input the computed distances and outputs a list of top-$k$ possible profiles (or IDs) in $D_{BK}$ that corresponds to the target user $u_i \in U$. 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 $n$: 

$$\text{RID-ACC}(\%) = 100 \cdot \frac{\sum_{i=1}^{n} 1_{u_{id} = \hat{u}_{id}}} {n},$$

in which $f(u_{id}, \hat{u}_{id}) = 1$ if $u_{id} = \hat{u}_{id}$ and 0 otherwise. The attacker’s RID-ACC depends on the accuracy of partially or completely profiling the target user (i.e., as measured by Eqs. (4) and (5)) and the "uniqueness" of users with respect to the collected attributes (unknown by the server) and in the background knowledge $D_{BK}$.

### 3.3 Attacking RS+FD: Uncovering the Sampled Attribute ($\rightarrow$ SMP)

Because the objective of the RS+FD solution is to hide the LDP value among fake data [4], discovering the sampled attribute of each user would convert RS+FD into the SMP solution again. Even more, unlike SMP (and SPL), RS+FD utilizes an amplified $\epsilon' > \epsilon$, which decreases the "plausible deniability" of the user’s report (cf. Section 3.2.1) and could thus 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 $u \in U$ whose sampled attribute is $t \in [d]$ produces an RS+FD’s output tuple as $y = [y_1, y_2, \ldots, y_d]$. In this situation, the baseline classification model is just a random guess $\hat{t} = \text{Uniform}([d])$. In addition, we propose a classifier learning setting in which an attacker aims to train a classifier over a learning dataset $D_l = \{(y_i, t_i) \mid i \in [r]\}$ of $r$ rows and $d = d + 1$ columns. That is, for each user $u_i$, $y_i$ 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 $[d]$). 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 $D_l$, which depends on the attack model. In all these settings, we assume that the attacker has the knowledge of the privacy budget $\epsilon$, 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 $\hat{t} = \hat{t}$ over the number of users in the testing dataset $n_t$:

$$\text{AIF-ACC}(\%) = 100 \cdot \frac{\sum_{i=1}^{n_t} 1_{\hat{t} = \hat{t}}} {n_t},$$

in which $f(t, \hat{t}) = 1$ 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 $u \in U$ and after aggregating users’ LDP data, an attacker could use the estimated frequencies $\hat{f} = [\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_d]$ to generate $s$ synthetic profiles $\hat{s}_i = [s_1, s_2, \ldots, s_d]$, for $i \in [s]$, 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 $D_l$. 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 $D_t$ is composed of all the real RS+FD’s sanitized tuples $y$ of users $u \in U$, and the objective is to accurately classify their sampled attribute $t \in [d]$.

#### 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 $U_{pk} \subset U$. 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 $D_l$ depends on the number of (compromised) profiles $n_{pk}$ the attacker has access to and the testing set $D_t$ has $n - n_{pk}$ sanitized tuples $y$ of users $u \in U \setminus U_{pk}$, in which the objective is to accurately classify their sampled attribute $t \in [d]$.

#### 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 $U_{pk} \subset U$ with $s$ synthetic profiles. In this hybrid model (HM), the learning set $D_l$ 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 $D_t$ has $n - n_{pk}$ sanitized tuples $y$ of users $u \in U \setminus U_{pk}$, and the goal is to accurately classify their sampled attribute $t \in [d]$.

### 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.

---

3If $U_{pk} \subseteq U$, this will correspond to a full-knowledge model in which the adversary has knowledge of all users’ sampled attribute (i.e., SMP solution).
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. 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 \( k = [92, 25, 5, 2, 2, 9, 4, 5, 5, 4, 2, 18, 2, 2, 3, 9, 3, 6] \).
- 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 \( k = [74, 7, 16, 7, 14, 6, 5, 2, 41, 2] \), 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, \( \omega \)-SS, SUE and OUE.

Privacy protection. We vary the privacy budget in the interval \( \epsilon = [1, 2, \ldots, 9, 10] \), which corresponds to values used by industry nowadays [13, 39] and experiments found in the LDP attacking literature with single [8, 22, 35] and multiple [24] collections.

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

Baseline. For each top-\( k \), the baseline re-identification model follows top-\( k \) random guesses (i.e., \( \tilde{u}_d = \text{Uniform}([n]) \)) without replacement with expected RID-ACC: top-\( k \)-n.

Experimental evaluation. We set \#surveys = 5, in which each survey \( s \in \{1, \ldots, 5\} \), has a different number of attributes \( d_s = \text{Uniform}[\frac{d}{2}, \ldots, d] \) (i.e., with at least \( \frac{d}{2} \) attributes). The attributes are also selected at random per survey. Due to space constraints, we only present here the experiments with the FK-RI model (cf. Section 3.2.4), considering the \( d \)-dimensional dataset as background knowledge \( D_{BK} \), and with the uniform privacy metric setting from Section 3.2.2. Finally, we measure the attacker’s RID-ACC after \#surveys = 2, which results in the inferred profile of each user having respectively 2, 3, 4 or 5 attributes, to be used for the re-identification attack.

Results. Fig. 2 illustrates the attacker’s RID-ACC metric on the Adult dataset for top-\( k \) re-identification using the SMP solution, the FK-RI model with uniform \( \epsilon \)-LDP privacy metric across users, by varying the LDP protocol and the number of surveys. Additional results with all LDP protocols, Adult and ACSEmployement datasets, FK-RI and PK-RI models, uniform and non-uniform privacy metric settings as well as with the relaxed LDP metric of [35] are presented in Appendix C.

Analysis. In general, the experimental results of Fig. 2 match the numerical results of the expected values from Fig. 1. From Fig. 2, one can observe that our re-identification attacks present significant improvement over a random baseline model that has RID-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 for GRR (and \( \omega \)-SS) and SUE after about \#surveys \( \geq 4 \), with at most \( \sim 10 \% \) of RID-ACC. In comparison, both OUE and OLH protocols have about 10x less RID-ACC, (i.e., at most \( \sim 1 \% \) of RID-ACC for top-1). On the other hand, when there is a set of top-10 profiles, the adversary achieves RID-ACC \( \geq 2.5 \% \) for GRR (and \( \omega \)-SS) after only 2 surveys with an upper bound of about 33% of RID-ACC after 5 surveys. Though with slightly smaller RID-ACC, the SUE protocol also achieves about 28% of RID-ACC after 5 surveys, and both OUE and OLH are upper bounded by about 5% of RID-ACC. Although the user is not uniquely re-identified, this still represents a threat due to the possibility of performing, e.g., homogeneity attacks [11, 30, 33].

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 of Appendix C). Additionally, the uniform privacy metric setting require the users to always sample a new attribute, increasing the privacy leakage. In a more realistic scenario, the non-uniform privacy metric setting minimizes the RID-ACC (see Fig. 11 of Appendix C) as already shown in Fig. 1. Furthermore, the FK-RI model allows the attacker to use the whole background knowledge \( D_{BK} \) to match the inferred profiles. For instance, the attacker’s RID-ACC metric decreased by almost half when considering the FK-RI model (cf. Fig. 10 of Appendix C) since there are fewer attributes as background information to use for the matching algorithm \( R \) (see Section 3.2.4). Lastly, we used the same dataset for private data collection and as (partial) 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 (\( \rightarrow \) 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 \( i = \text{Uniform}([d]) \) with expected AIF-ACC: 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
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] [4] 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 $O_{HK}$ 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 $\epsilon$-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. 14 in Appendix D, 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-10 group the attacker’s RID-ACC has meaningful improvement over the Baseline model. These results with the RS+FD[GRR] protocol indicate that RS+FD is already (to some extent) a countermeasure to re-identification attacks, except for RS+FD[SUE-z] in which the attacker can predict the attribute with high confidence with high $\epsilon$.

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 [4] 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 [1] 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 \( v = [v_1, v_2, \ldots, v_d] \), the domain size of attributes \( k = [k_1, k_2, \ldots, k_d] \), the attributes' prior distributions \( f = [f_1, f_2, \ldots, f_d] \) (transmitted by the server in advance), the privacy parameter \( \epsilon \) and a local randomizer \( M \). The output is a tuple \( y = [y_1, y_2, \ldots, y_d] \) of values (LDP and fake). In Alg. 1, line 6, Sample means a random sample is generated following prior \( f_i \) of the attribute \( i \in [d] \setminus \{j\} \).

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

Input: tuple \( v = [v_1, v_2, \ldots, v_d] \), domain size of attributes \( k = [k_1, k_2, \ldots, k_d] \), prior distribution of attributes \( f = [f_1, f_2, \ldots, f_d] \), privacy parameter \( \epsilon \) and local randomizer \( M \).
Output: sanitized tuple \( y = [y_1, y_2, \ldots, y_d] \).
1. \( \epsilon' = \ln (d \cdot (e^\epsilon - 1) + 1) \) \# Amplification by sampling [31]
2. \( y_j \leftarrow \text{Uniform}([d]) \) \# Selection of attribute to sanitize
3. \( y_j \leftarrow \text{Encode}(y_j) \) \# Encode (if needed)
4. \( v_i \leftarrow \text{Sample}(1, \ldots, k_i), f_i \) \# For each non-sampled attributes
5. for \( i \in [d] \setminus \{j\} \) do
6. \( y_j \leftarrow \text{Sample}(\{1, \ldots, k_j\}, f_j) \) \# Generate fake data
7. end for
8. return: \( y = [y_1, y_2, \ldots, y_d] \) \# Sanitized tuple

Server-Side. The aggregator performs multiple frequency estimation on the collected data by removing bias introduced by the local randomizer \( M \) and fake data. The new estimators of using RS+RFD with GRR or UE-based protocols (e.g., SUE [23] or OUE [50]) as local randomizer \( M \) in Alg. 1 is presented in the following. For each attribute \( i \in [d] \), the aggregator estimates \( \hat{f}(v_i) \) for the frequency of each value \( i \in [k_j] \) as:

- **RS+RFD[GRR].** The RS+RFD[GRR] estimator is:
  \[
  \hat{f}_{\text{GRR}}(v_i) = \frac{dC(v_i) - n (q + (d - 1)\hat{f}_j(v_i))}{n(p - q)},
  \]
  in which \( C(v_i) \) is the number of times \( v_i \) has been reported, \( \hat{f}_j(v_i) \) is the prior distribution of value \( v_i \in A_j \), \( \epsilon' = \ln (d \cdot (e^{\epsilon' - 1}) + 1) \), \( p = \frac{e^{\epsilon'}}{e^{\epsilon'} + 1} \), and \( q = \frac{1 - p}{1 + p} \). The probability tree of the RS+RFD[GRR] protocol, the proof that the estimator in Eq. (6) is unbiased and its variance computation are provided in Appendix A.

- **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 \( i \in [d] \setminus \{j\} \), the user generates fake data by applying an UE protocol to encoded random data following prior distribution \( f_i \). The RS+RFD[UE-r] estimator is:
  \[
  \hat{f}_{\text{UE-r}}(v_i) = \frac{dC(v_i) - n (q + (p - q)(d - 1)\hat{f}_j(v_i) + q(d - 1))}{n(p - q)},
  \]
  in which \( C(v_i) \) is the number of times \( v_i \) has been reported, \( \epsilon' = \ln (d \cdot (e^{\epsilon' - 1}) + 1) \) and \( \hat{f}_j(v_i) \) is the prior distribution of value \( v_i \in A_j \). Parameters \( p \) and \( q \) can be selected following the SUE [23] protocol (\( p = \frac{e^{\epsilon'}}{e^{\epsilon'} + 1} \) and \( q = \frac{1 - p}{1 + p} \)) or OUE [50] 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. (7) is unbiased and its variance calculation is provided in Appendix B.

Privacy analysis. Similar to the RS+FD solution [4], let \( M \) be any existing LDP mechanism, Alg. 1 satisfies \( \epsilon \)-LDP, in a way that \( \epsilon' = \ln (d \cdot (e^{\epsilon' - 1}) + 1) \), in which \( d \) is the number of attributes.

Limitations. Besides known limits of the RS+FD solution [4, 46], RS+RFD adds a limitation on being dependent on the underlying prior distributions \( f \) to generate realistic fake data. Yet, many demographic attributes have Census data [1] 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 ACSEmployement dataset described in Section 4.1.

Prior distribution. To simulate “Correct” prior distributions \( f = [f_1, f_2, \ldots, f_d] \) to be used to generate realistic fake data with RS+RFD, we perturb the real frequency of each attribute \( j \in [d] \) with the standard Laplace mechanism [19–21] in centralized DP satisfying \( \epsilon = 0.1/d \) (i.e., split \( \epsilon = 0.1 \) by \( d \) attributes). In addition, to simulate an “incorrect” scenario in which prior distributions are wrongly specified, we use Dirichlet distributions with parameter 1.

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 [4] 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 [4], we vary \( \epsilon \) in the interval \( \epsilon = [\ln (2), \ln (3), \ldots, \ln (7)] \) and measure the quality of the estimated frequencies with the averaged mean squared error metric:

\[
\text{MSE}_{\text{avg}} = \frac{1}{d} \sum_{j \in [d]} \frac{1}{|A_j|} \sum_{v \in A_j} \left( \hat{f}(v) - f(v) \right)^2.
\]

Results. Fig. 5 illustrates for all methods the MSEavg metric (y-axis) according to the privacy parameter \( \epsilon \) (x-axis) for both “Correct” and “Incorrect” priors. Additional empirical and analytical results with the Adult dataset are provided in Appendix E.

Analysis. For the “Correct” prior, one can observe that the MSEavg metric of our proposed RS+RFD protocols consistently and significantly outperform 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. Indeed, even with "Incorrect" priors, our RS+RFD protocols still outperform the RS+FD protocols, with the exception of RS+RFD[OUE-r] with similar utility RS+RD[OUE-r] in low privacy regimes. 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.

In brief, we identified and evaluated empirically two threats to users’ privacy when collecting multidimensional data with the state-of-the-art solutions SMP and RS+FD, namely re-identification attack and inference of the sampled attribute. These threats are generic to any LDP protocol and can be modelled by extending the “plausible deniability” attack analysis of Section 3.2.1. Hereafter, we summarize the key findings that can be used by practitioners and help substantiate the main claims of this paper.

Regarding the SMP solution, in our experiments, the GRR and ω-SS protocol had the highest RID-ACC as the probability of accurately inferring the user’s full profile was higher with relatively small $k_j$ values (see also Fig. 1). With other protocols, such as OLH and OUE, which are the current state-of-the-art for preserving utility [50], the adversary cannot accurately infer the profile of users when using $\epsilon$-LDP as privacy model, which leads to lower re-identification risks (see Fig. 2 (c) and (d)). On the other hand, as shown in Appendix C, when using the relaxed version of LDP from [35], the RID-ACC increases considerably for both OLH and OUE protocols. Though we only experimented with #surveys ≤ 5, we believe that more data collections can lead to higher RID-ACC as long as the profile is accurately inferred. Yet, under standard sequential composition [21], the overall privacy loss is excessive when using high values for $\epsilon$, but we have also considered them due to their use in practical deployments [44, 45] and similar experiments found in [24] (though with higher #surveys ∈ {7, 30, 90, 180}).

On the other hand, when using the RS+FD to "hide" the sampled attribute, the utility-oriented protocol RS+FD[UE-z] has the highest AIF-ACC due to generating fake data with zero-vectors and we recommend not using it in practice. Even with the RS+FD[GRR] or RS+FD[UE-r] protocols the attacker’s AIF-ACC is considerably greater than a random guess. Yet, since there are chained errors in multiple collections on accurately predicting the sampled attribute and on inferring the user’s value, the RS+FD considerably minimizes the risks of re-identification presented by the SMP solution.

Overall, though some LDP protocols minimized the RID-ACC or AIF-ACC in our experiments (see the main body and Appendices C, D and E), they did not fully mitigate the risks when increasing $\epsilon$ as done in practice to get more accurate estimates. This means they still allow a small portion of users to leak more information than others and corroborate with DP consensus of using $\epsilon$ ≤ 1.

Therefore, considering the setting described in Section 3.1, the overall recommendation when using the SMP solution is to select: the standard $\epsilon$-LDP as privacy model, the OUE and/or OLH protocols (depending on $k_j$ due to communication costs [50]), the non-uniform privacy metric setting (i.e., allowing users to sample with replacement and enforce memoization [5, 14, 23]) and to keep $\epsilon$ ≤ 1. On the other hand, when using the RS+FD solution, even when no prior is available, we highly recommend the proposed version in this paper, i.e., RS+RFD with non-uniform fake data.

### 6 DISCUSSION

In brief, we identified and evaluated empirically two threats to users’ privacy when collecting multidimensional data with the state-of-the-art solutions SMP and RS+FD, namely re-identification attack and inference of the sampled attribute. These threats are generic to any LDP protocol and can be modelled by extending the “plausible deniability” attack analysis of Section 3.2.1. Hereafter, we summarize the key findings that can be used by practitioners and help substantiate the main claims of this paper.

Regarding the SMP solution, in our experiments, the GRR and ω-SS protocol had the highest RID-ACC as the probability of accurately estimating with the RS+RFD and RS+FD solutions.

![Figure 5: Averaged MSE metric varying $\epsilon$ for (a) "Correct" and (b) "Incorrect" priors for multidimensional frequency estimation with the RS+RFD and RS+FD solutions.](image)

5.2.3 Uncovering the Sampled Attribute of the RS+RFD Solution (→ SMP). This section follows similar parameters (dataset, $\epsilon$ 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 ACSEmployment 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] with “Correct” priors), varying $\epsilon$, the number of synthetic profiles $s$ and the number of compromised profiles $n_{pk}$. Further results with “Incorrect” priors are in Appendix E.

**Analysis.** We highlight that the non-stability in the plots of Fig. 6 is due to different sources of randomness: $\epsilon = 0.1$-DP for “Correct” prior distributions $\tilde{f}$, $\epsilon$-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.

### 7 RELATED WORK

The literature on the local DP model has largely explored the issue of improving the utility of LDP protocols [4, 5, 14, 18, 23, 26, 27, 37, 45–47, 50–52]. Recently, a few works have started to design attacks on LDP protocols. Some authors focused on maliciously modifying the estimated statistic on the server through targeted or untargeted attacks [6, 10, 32, 54]. To counter such kinds of attacks, some works [3, 29] 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 [8, 22, 24, 35] and in this work. For instance, Chatzikokolakis et al. [8] proposed the Bayes security measure to quantify the expected gain over a random guess of an adversary that observes a report of the RR protocol. Similar to [8], this paper provides a “plausible
deniability” attacking interpretation of five state-of-the-art LDP protocols to infer the user’s true value by observing an LDP report. In an independent and concurrent work, Gursoy et al. [22] proposed a formalized Bayes adversary for the same attack, which was referenced in Section 3.2.1 to give the expected accuracy of our analyses. Besides, we extended our attack to multiple collections of multidimensional data in Sections 3.2.2 and 3.2.3, which were proposed to account for the consequent risks of re-identification [25, 34–36] in Section 3.2.4. Re-identification risks in the LDP model for single-frequency estimation were first investigated by Murakami and Takahashi [35]. However, different from [35] that focused on a single attribute (e.g., location traces), our work considers multiple attributes being collected multiple times. Regarding multiple collections, Gadotti et al. [24] introduced pool inference attacks to LDP protocols for single-frequency estimation in a way that an adversary can infer the user’s preferred pool (e.g., skin tone used in emojis).

On the other hand, Arcoley et al. [4] introduced the RS+FD solution focusing only on the utility of the protocols, which was also later studied in [46]. In this work, we are the first to propose three attack models to the RS+FD solution, showing it is possible to distinguish the ε-LDP report from fake data. Consequently, in multiple collections, we also show that RS+FD is still subject to (reduced) re-identification risks. We thus proposed an improvement of the RS+FD solution that generates non-uniform fake data (i.e., RS+RFD of Section 5.1) and can serve as a countermeasure solution.

8 CONCLUSION AND PERSPECTIVES

In this paper, we 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 [4]. On the one hand, we presented inference attacks based on “plausible deniability” [53] of five widely used LDP protocols (i.e., GRR [26, 27], OLH [50], ω-SS [48, 55], RAPPRO [23] and OUE [50]) under multiple collections following the SMP solution. We introduce a refined version of the RS+FD solution, called RS+RFD 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 attack.

Though we identified and investigated two privacy threats for LDP protocols for multidimensional data in single and multiple data collections, these are not unique and we believe that our work opens new avenues of research in this direction. For future work, we suggest and aim to formalize the re-identification risks considering different LDP and d-privacy [2, 7, 49] protocols, the number of collections, the number of attributes and the “uniqueness” of users in a given dataset. Such a formalization will allow to design other countermeasure solutions beyond RS+FD [4] and our RS+RFD.

ACKNOWLEDGMENTS

The authors deeply thank the anonymous reviewers for their insightful suggestions. This work was partially supported by the ERC project HYPATIA with grant agreement N° 835294 and by the EIPHI-BFC Graduate School (contract “ANR-17-EURE-0002”). Sébastien Gamb is supported by the Canada Research Chair program as well as a Discovery Grant from NSERC. All computations were performed on the “Mésocentre de Calcul de Franche-Comté”.

Figure 6: Attacker’s AIF-ACC on the ACSEmployment 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] with “Correct” priors), varying ϵ, the number of synthetic profiles s the attacker generates and the number of compromised profiles npk the attacker has access to.
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. (6) is an unbiased estimation of $f(v_i)$ for any value $v_i \in A_j$.

Proof.

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

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

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

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

$= \frac{n}{d} \left(f(v_i)(p - q) + q + (d - 1)\tilde{f}(v_i)\right).$

Therefore,

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

□

Theorem 2. The variance of the estimation in Eq. (6) is:

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

$y = \frac{1}{d} \left(q + f(v_i)(p - q) + (d - 1)\tilde{f}(v_i)\right).$

Proof. Thanks to Eq. (6) we have

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

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,

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

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

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

□

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).

Theorem 3. For $j \in [1,d]$, the estimation result $\hat{f}_{UE-R}(v_i)$ in Eq. (7) is an unbiased estimation of $f(v_i)$ for any value $v_i \in A_j$.

Proof.

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

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

On expectation, the number of times that $v_i$ is reported is:
This section provides additional results for the re-identification risks of the PIE (Personal Information Entropy) model with uniform privacy and if $U \subseteq \Psi$ and $\alpha \in \mathbb{R}_{\geq 0}$. If an obfuscation mechanism $M$ provides ($U, \alpha$)-PIE privacy and if $U$ is uniformly distributed (i.e., one data item), then

$$\sup_{p \in \Psi} I(U; Y) \leq \alpha \text{(bits)}. \quad (10)$$

Since the inequality in Eq. (10) holds for any distribution $p \in \Psi$, the PIE is upper bounded by $\alpha$ irrespective of the adversary’s background knowledge [35]. The parameter $\alpha$ plays a role similar to the privacy budget $\epsilon$ 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 [35]). Let $U \subseteq \Psi$ and $\alpha \in \mathbb{R}_{\geq 0}$. If an obfuscation mechanism $M$ provides $\epsilon$-LDP, then it provides ($U, \alpha$)-PIE privacy for any $U \subseteq \Psi$ such that $|U| = n$, where

$$\alpha = \min \left\{ \epsilon \log_2(e), \epsilon^2 \log_2(e), \log_2(n), \log_2(k_f) \right\}. \quad (12)$$

As stated in [35], Proposition 1 holds for any LDP protocol.

### C.2 Experiments

We follow a similar experimental evaluation to the one described in Section 4.2 and we vary the following:

- **Dataset.** Adult [17] and ACSEmployement [15] datasets.
- **LDP protocol.** GRR [26, 27], OLH [50], $\omega$-SS [48, 55], SUE (a.k.a. Basic One-Time RAPPOR [23]) and OUE [50].
- **Re-identification model.** Full knowledge (FK-RI) and partial knowledge (PK-RI) models (cf. Section 3.2.4). The background knowledge $D_{BK}$ and $D_{PK} \subseteq D_{BK}$ (a random subset with at least $\frac{d}{2}$ attributes) are the own Adult and ACSEmployement datasets.
- **Privacy metric.** Local differential privacy (with $\epsilon = [1, 2, \ldots, 9, 10]$) and $\alpha$-PIE privacy [35] (with $\beta_{U|S} = [0.95, 0.9, \ldots, 0.55, 0.5]$, i.e., from tighter privacy regimes to lower ones). In this case, following Eq. (12) and [35, Proposition 9], when $k_f$ (attribute’s domain size) is small, one does not need to use an LDP protocol (i.e., $y = o$).
- **Privacy metric across users.** Uniform (cf. Section 3.2.2) and non-uniform (cf. Section 3.2.3) privacy metrics.

First, within the same setting of Fig. 2 (Adult dataset), Fig. 9 (ACSEmployment dataset) illustrates the attacker’s RID-ACC for top-k re-identification on using the SMP solution, the FK-RI model with uniform $\epsilon$-LDP privacy metric across users, by varying the LDP protocol and the number of surveys. One can notice that both Figs. 2 and 9 follow similar patterns for all LDP protocols with only different upper bounds for the RID-ACC metric. Therefore, results for the FK-RI model, $\alpha$-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.

Secondly, to complement the results of Fig. 2 (FK-RI model), Fig. 10 (PK-RI model) illustrates the attacker’s RID-ACC for top-k
re-identification on using the SMP solution, the Adult dataset, with uniform $\epsilon$-LDP privacy metric across users, by varying the LDP protocol and the number of surveys. From Figs. 2 and 10, one can notice that the PK-RI model minimized the RID-ACC by about 50% for all LDP protocols, which represents a more realistic scenario for the background knowledge of adversaries in real-life.

Since Figs. 2 and 10 present results for the uniform $\epsilon$-LDP privacy metric setting, Fig. 11 illustrates the attacker’s RID-ACC on the Adult dataset for top-k re-identification on using the SMP solution, the full knowledge FK-RI model (left-side plots) and partial knowledge PK-RI model (right-side plots) with non-uniform $\epsilon$-LDP privacy metric across users, and by varying the LDP protocol and the number of surveys (i.e., data collections). With a uniform privacy metric, all users report a different attribute per survey, which can lead 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 can bound the re-identification risks. For these reasons, the RID-ACC of Fig. 11 decreased by about 40% in comparison with the RID-ACC of Figs. 2 and 10.

Focusing now on the $\alpha$-PIE metric, Figs. 12 (uniform setting) and 13 (non-uniform setting) illustrate the attacker’s RID-ACC on the Adult dataset for top-k re-identification on using the SMP solution, the full knowledge FK-RI model (left-side plots) and partial knowledge PK-RI model (right-side plots), and by varying the LDP protocol and the number of surveys (i.e., data collections). As one can note, even using a high Bayes error probability as $\beta_{U|S} = 0.95$ for each attribute and collection, the $(U, \alpha)$-PIE privacy metric leads to higher attacker’s RID-ACC in comparison with Figs. 2, 10 and 11 in which $\epsilon = 1$. In fact, these higher RID-ACC when using $\beta_{U|S}$ can be explained due to not using a local randomizer when $k_j$ is small [35, Proposition 9] (the case for several attributes in the Adult dataset), with more difference for other protocols such as SUE, OUE and OLH. For this reason, the re-identification rates of all LDP protocols follow similar behaviors on Figs. 12 and 13.

Overall, 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 (cf. Sections 3.2.2 and 3.2.3) such that the profile can be found in the background knowledge DK (or DPK). Besides, the results agree with intuitive expectations, as increasing $\epsilon$ (or decreasing $\beta_{U|S}$) also increases the privacy leakage, thus leading to higher re-identification rates. Moreover, following the $\epsilon$-LDP metric, the lowest attacker’s RID-ACC were achieved by both OLH and OUE protocols [50], as they have an expected upper bound for the attacker’s ACCDP = $1/2$ [22] (cf. Section 3.2.1) per data collection. Besides, the $\omega$-SS protocol follows similar re-identification rates as the GRR protocol (also shown analytically in Fig. 1), which is the highest among all LDP protocols. On the other hand, the re-identification rates of all LDP protocols follow similar behaviors when using the $\alpha$-PIE privacy model on Figs. 12 and 13.

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 [4] 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, $n = 45,222$ and $k = \{74, 16, 7, 14, 6, 5, 2, 41, 2\}$) and Nursery ($d = 9$ attributes, $n = 12,959$ and $k = \{3, 5, 4, 3, 2, 3, 3, 5\}$) datasets from the UCI ML repository [17].
- **LDP protocol within RS+FD.** All protocols from Section 2.3.2, namely, RS+FD[GR], 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. 14 and 15 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[GR], RS+FD[SUE-z], RS+FD[OUE-z], RS+FD[SUE-r] and RS+FD[OUE-r]), varying $\epsilon$, the number of synthetic profiles $s$ and the number of compromised profiles $n_p$.

Similar to the results with the ACSEmployment [15] dataset in Fig. 3, one can notice in Figs. 14 and 15 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[GR] 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 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 14). Lastly, increasing the number of synthetic profiles the attacker generates $s$ and/or the number of compromised profiles the attacker has access to $n_p$, had few influence with the Adult dataset in Fig. 14. Conversely, both ACSEmployment (Fig. 3) and Nursery (Fig. 15) datasets showed sensitivity to a change in both parameters, especially with $n_p$ in the PK model.

E ADDITIONAL RESULTS FOR SECTION 5.2

This section provides additional results for the utility and privacy analysis of our proposed RS+RFD solution.

E.1 Multidimensional Frequency Estimation

We compare the utility of our RS+RFD and RS+FD [4] solutions for frequency estimation of multiple attributes following the experimental evaluation described in Section 5.2.2 by using:

- **Dataset.** We use the Adult [17] ($d = 10$ attributes, $n = 45,222$ and $k = \{74, 16, 7, 14, 6, 5, 2, 41, 2\}$).
**LDP protocol within RS+FD.** RS+FD[GRR], RS+FD[SUE-r] and RS+FD[OUE-r], described in Section 2.3.2.

**LDP protocol within RS+RFD.** RS+RFD[GRR], RS+RFD[SUE-r] and RS+RFD[OUE-r], described in Section 5.1.

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

- **“Incorrect” priors.** To simulate “Incorrect” scenarios where prior distributions are wrongly specified, we use the following distributions: Dirichlet distributions (DIR) with parameter 1, Zipf distributions (ZIPF) with parameter \( s = 1.01 \) and Exponential distributions (EXP) with \( \lambda = 1 \). For both ZIPF and EXP distributions, one hundred thousand samples are generated, and for each attribute \( j \in [d] \), we reconstruct the histogram with \( k_j \) buckets.

- **Analytical analysis.** We use the approximate variance to measure the utility loss of our protocols by setting \( f(a_j) = 0 \) in Eq. (8) for RS+RFD[GRR] and in Eq. 9 for RS+RFD[OUE-r].

- **Evaluation metrics.** To compare with [4], we vary \( \epsilon \) 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: \( \text{MSE}_{\text{avg}} = \frac{1}{d} \sum_{j \in [d]} \frac{1}{|A_j|} \sum_{v \in A_j} (f(v) - \hat{f}(v))^2 \).

Fig. 16 illustrates for all methods analytical (averaged approximate variance values) and experimental (averaged MSE metric) results varying \( \epsilon \) for “Correct” (a and b plots) and “Incorrect” (c to h) priors for multidimensional frequency estimation with the RS+RFD and RS+FD solutions with the Adult dataset.

For all plots of Fig. 16, one can notice that the experimental evaluations are consistent with the numerical evaluation of the variance for each protocol. Moreover, for “Correct” priors, similar to Fig. 5, one can observe that the \( \text{MSE}_{\text{avg}} \) metric of our proposed RS+RFD protocols consistently and considerably 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. With “Incorrect” priors (e.g., DIR), our RS+RFD protocols still outperform the RS+FD protocols, with the exception of RS+RFD[OUE-r] with similar utility to RS+FD[OUE-r] in low privacy regimes (similar to Fig. 5). Last, with both “Incorrect” priors following ZIPF and EXP distributions, the MSE$_{avg}$ (approximate variance, respectively) improved considerably in comparison with RS+FD protocols following uniform distribution, especially for our RS+RFD[GRR] protocol.

E.2 Inference of Sampled Attribute Attack

Next, we compare the robustness of our RS+RFD protocols against the inference of the sampled attribute attack following the experimental evaluation described in Section 5.2.3 by using:

- **Classifier.** We use the state-of-the-art XGBoost [9] algorithm to predict the sampled attribute of users in a multi-class classification framework (i.e., d attributes) with default parameters.
- **Dataset.** We use the ACSEmployement [15] (d = 18 attributes, n = 10,336 and k = [92, 25, 5, 2, 2, 9, 4, 5, 4, 2, 18, 2, 2, 3, 9, 3, 6]) dataset.
- **LDP protocol within RS+RFD.** RS+RFD[GRR], RS+RFD[SUE-r] and RS+RFD[OUE-r], described in Section 5.1.
- **“Correct” priors.** To simulate “Correct” prior distributions $\mathbf{f} = [f_1, f_2, \ldots, f_d]$ to be used to generate realistic fake data with RS+RFD, we perturb the real frequency of each attribute $j \in [d]$ with the standard Laplace mechanism [19–21] in centralized DP satisfying $\epsilon = 0.1/d$ (i.e., split $\epsilon = 0.1$ by $d$ attributes).
- **“Incorrect” priors.** To simulate “Incorrect” scenarios where prior distributions are wrongly specified, we use the following distributions: Dirichlet distributions (DIR)
Figure 11: Attacker’s re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification on using the SMP solution, the full knowledge FK-RI model (left-side plots) and partial knowledge PK-RI model (right-side plots) with non-uniform $\epsilon$-LDP privacy metric across users, and by varying the LDP protocol and the number of surveys (i.e., data collections).
Figure 12: Attacker’s re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification on using the SMP RAPPOR [23] protocol.

(a) FK-RI risk of the GRR [26, 27] protocol.

(b) PK-RI risk of the GRR [26, 27] protocol.

(c) FK-RI risk of the $\omega$-SS [48, 55] protocol.

(d) PK-RI risk of the $\omega$-SS [48, 55] protocol.

(e) FK-RI risk of the OLH [50] protocol.

(f) PK-RI risk of the OLH [50] protocol.

(g) FK-RI risk of the OUE [50] protocol.

(h) PK-RI risk of the OUE [50] protocol.

(i) FK-RI risk of the SUE (a.k.a. RAPPOR) [23] protocol.

(j) PK-RI risk of the SUE (a.k.a. RAPPOR) [23] protocol.

Figure 12: Attacker’s re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification on using the SMP solution, the full knowledge FK-RI model (left-side plots) and partial knowledge PK-RI model (right-side plots) with uniform $\alpha$-PIE privacy metric across users, and by varying the LDP protocol and the number of surveys (i.e., data collections).
Figure 13: Attacker’s re-identification accuracy (RID-ACC) on the Adult dataset for top-k re-identification on using the SMP solution, the full knowledge FK-RI model (left-side plots) and partial knowledge PK-RI model (right-side plots) with non-uniform $\alpha$-PIE privacy metric across users, and by varying the LDP protocol and the number of surveys (i.e., data collections).
with parameter 1, Zipf distributions (ZIPF) with parameter \( s = 1.01 \) and Exponential distributions (EXP) with \( \lambda = 1 \).
For both ZIPF and EXP distributions, one hundred thousand samples are generated, and for each attribute \( j \in [d] \), we reconstruct the histogram with \( k_j \) buckets.

- **Attribute inference model.** All three protocols are evaluated with the No Knowledge (NK) attack model of Section 3.3.

Fig. 14 illustrates the attacker’s AIF-ACC metric on the ACSeMployement dataset with the NK attack model and our three protocols.
(b) PK model with RS+FD[GRR] protocol.

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

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

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

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

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

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

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

Figure 15: 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 $\epsilon$, the number of synthetic profiles $s$ the attacker generates and the number of compromised profiles $n_{pk}$ the attacker has access to.

(i.e., RS+FD[GRR], RS+FD[SUE-r] and RS+FD[OUE-r] with “Incorrect” priors), varying $\epsilon$, the number of synthetic profiles $s$ the attacker generates and the “Incorrect” prior distribution (i.e., DIR, ZIPF and EXP). We remark that the non-stability in the plots of Fig. 17 is due to different sources of randomness: different random “Incorrect” distributions $f$, $\epsilon$-LDP randomization, fake data generation and the XGBoost algorithm.
Figure 16: Analytical (averaged approximate variance values) and experimental (averaged MSE metric) results varying $\epsilon$ for "Correct" (a and b plots) and "Incorrect" (c to h) priors for multidimensional frequency estimation with the RS+RFD and RS+FD solutions with the Adult dataset.
Figure 17: Attacker’s AIF-ACC on the ACSEmployment dataset with the No Knowledge (NK) attack model and our three protocols (i.e., RS+RFD[GRR], RS+RFD[SUE-r] and RS+RFD[OUE-r] with “Incorrect” priors), varying \( \epsilon \), the number of synthetic profiles \( s \) the attacker generates and the prior distribution (i.e., DIR, ZIPF and EXP).

In Fig. 17, for both ZIPF and EXP distributions, one can notice that our RS+RFD protocols considerably decrease the attacker’s AIF-ACC when comparing with their respective RS+FD version in Fig. 3. More specifically, the RS+FD[GRR] and RS+FD[UE-r] protocols presented at most \( \sim 25\% \) of AIF-ACC within the NK attack model (see Fig. 3). On the other hand, our RS+RFD[GRR] and RS+RFD[UE-r] protocols with both ZIPF and EXP “Incorrect” priors presented at most \( \sim 8\% \) of AIF-ACC (cf. Fig. 17) within the same NK attack model, i.e., a 3x decrease factor. Overall, the lowest performance was achieved when using DIR “Incorrect” priors, which showed small reductions compared to the NK curves in Fig. 3 (i.e., RS+FD with prior uniforms).