Digital Communications and Networks(DCN)

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/dcan

Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates☆

Héber H. Arcolezi*[a,b], Jean-François Couchot[b], Bechara Al Bouna[c], Xiaokui Xiao[d]

[a] *Inria and École Polytechnique (IPP), Palaiseau, France*
[b] *Femto-ST Institute, Univ. Bourg. Franche-Comté, UBFC, CNRS, Belfort, France*
[c] *TICKET Lab., Antonine University Hadat-Baabda, Baabda, Lebanon*
[d] *School of Computing, National University of Singapore, Singapore, Singapore*

Abstract

This paper investigates the problem of collecting multidimensional data throughout time (i.e., longitudinal studies) for the fundamental task of frequency estimation under Local Differential Privacy (LDP) guarantees. Contrary to frequency estimation of a single attribute, the multidimensional aspect demands particular attention to the privacy budget. Besides, when collecting user statistics longitudinally, privacy progressively degrades. Indeed, the "multiple" settings in combination (i.e., many attributes and several collections throughout time) impose several challenges, for which this paper proposes the first solution for frequency estimates under LDP. To tackle these issues, we extend the analysis of three state-of-the-art LDP protocols (Generalized Randomized Response – GRR, Optimized Unary Encoding – OUE, and Symmetric Unary Encoding – SUE) for both longitudinal and multidimensional data collections. While the known literature uses OUE and SUE for two rounds of sanitization (a.k.a. memoization), i.e., L-OUE and L-SUE, respectively, we analytically and experimentally show that starting with OUE and then with SUE provides higher data utility (i.e., L-OSUE). Also, for attributes with small domain sizes, we propose Longitudinal GRR (L-GRR), which provides higher utility than the other protocols based on unary encoding. Last, we also propose a new solution named Adaptive LDP for LOngitudinal and Multidimensional FREquency Estimates (ALLOMFREE), which randomly samples a single attribute to be sent with the whole privacy budget and adaptively selects the optimal protocol, i.e., either L-GRR or L-OSUE. As shown in the results, ALLOMFREE consistently and considerably outperforms the state-of-the-art L-SUE and L-OUE protocols in the quality of the frequency estimates.

KEYWORDS: Local differential privacy, Discrete distribution estimation, Frequency estimation, Multidimensional data, Longitudinal studies.

1. Introduction

1.1. Background

In recent years, Differential Privacy (DP) [1, 2] has been increasingly accepted as the current standard for data privacy [3, 4, 5, 6]. In the centralized model of DP, a trusted curator has access to the entire raw data of users (e.g., the Census Bureau [7, 8]). By "trusted", we mean that curators do not misuse or leak private information of individuals. However, this assumption does not always hold in real life, e.g., data breaches are all too common [9].

To preserve privacy at the user-side, an alternative approach, namely, Local Differential Privacy (LDP), was initially formalized in [10]. With LDP, rather than trust a data curator to have the raw data and sanitize it to output queries, each user applies a DP mechanism to their data before transmitting it to the data collector server. The local DP model allows collecting data in unprecedented ways and, therefore, it has been widely adopted by industry (e.g., Google Chrome browser [11], Microsoft windows 10 operation system [12], Apple iOS and macOS [13]).

1.2. Motivation and problem statement

When collecting data in practice, one is often interested in multiple attributes of a population, i.e., *multidimensional data*. For instance, in crowd-sourcing applications, the server may collect both demographic

☆Final version accepted in the journal Digital Communications and Networks (soon to be updated with DOI).

*Corresponding author (email: heber.hwang-arcolezi@inria.fr).

information (e.g., gender, nationality) and user habits in order to develop personalized solutions for specific groups. In addition, one generally aims to collect data from the same users throughout time (i.e., *longitudinal* studies), which is essential in many situations [11, 12]. For example, the fact that two medical acts identified at a different time have been performed on the same patient, or two different patients mean treatment in the first case or two isolated acts in the second.

So, in this paper, we focus on the problem of private frequency (or histogram) estimation of multiple attributes throughout time with LDP. Frequency estimation is a primary objective of LDP, in which the data collector (a.k.a. the aggregator) decodes all the privatized data of the users and then estimates the number of users for each possible value. More formally, we assume there are d attributes $A = \{A_1, A_2, ..., A_d\}$, where each attribute A_j with a discrete domain has a specific number of value $k_j = |A_j|$. Each user u_i for $i \in \{1, 2, ..., n\}$ has a tuple $\mathbf{v}^{(i)} = (v_1^{(i)}, v_2^{(i)}, ..., v_d^{(i)})$, where $v_j^{(i)}$ represents the value of attribute A_j in record $\mathbf{v}^{(i)}$. Thus, for each attribute A_j at time $t \in [1, \tau]$, the aggregator's goal is to estimate a k_j-bins histogram, including the frequency of all values in A_j.

Indeed, in both longitudinal and multidimensional settings, one needs to consider the allocation of the privacy budget, which can grow extremely quickly due to the composition theorem [3]. However, on the one hand, most academic literature on frequency estimation [14, 15, 16, 17, 18, 19, 20, 21, 22] focuses on a single data collection (i.e., non-longitudinal studies). On the other hand, the studies for collecting multidimensional data with LDP mainly focus on other complex tasks (e.g., analytical/range queries [23, 24, 25, 26], estimating marginals [27, 28, 29, 30, 31]) or numerical data only (e.g., [32, 33, 34, 35]).

1.3. Summary of contributions

In this paper, we extend the analysis of three state-of-the-art LDP protocols, namely, Generalized Randomized Response (GRR) [18], Optimized Unary Encoding (OUE) [14], and Symmetric Unary Encoding (SUE) [11] for both longitudinal and multidimensional frequency estimates. On the one hand, for all three protocols, we theoretically prove that randomly sampling a single attribute per user improves data utility, which is an extension of common results in the LDP literature [36, 24, 37, 29, 38].

On the other hand, in the literature, both SUE and OUE protocols have been extended (and also applied [39, 40]) to longitudinal studies based on the concept of *memoization* [11, 12], i.e., L-SUE and L-OUE, respectively. However, we numerically and experimentally show that combining both protocols provides higher data utility, i.e., starting with OUE and then with SUE (L-OSUE) optimizes data utility better than using SUE or OUE twice. In addition, we also extend GRR for longitudinal studies (i.e., L-GRR), which provides higher data utility than the other protocols based on unary encoding for attributes with a small domain size.

Lastly, in a multidimensional setting having different domain sizes for each attribute, a dynamic selection of longitudinal LDP protocols is preferred. Therefore, we propose a new solution named Adaptive LDP for LOngitudinal and Multidimensional FREquency Estimates (ALLOMFREE), which combines all the aforementioned results. More specifically, ALLOMFREE randomly samples a single attribute to be sent with the whole privacy budget and adaptively selects the optimal protocol, i.e., either L-GRR or L-OSUE. To validate our proposal, we conduct a comprehensive and extensive set of experiments on four real-world open datasets. Under the same privacy guarantee, results show that ALLOMFREE consistently and considerably outperforms the state-of-the-art L-SUE and L-OUE protocols in the quality of the frequency estimates.

The remainder of this paper is organized as follows. In Section 2, we review the privacy notion in consideration, i.e., LDP and the protocols. In Section 3, we extend the analysis of GRR, OUE, and SUE to multidimensional data collections. In Section 4 we present the *memoization*-based framework for longitudinal data collections, the extension and analysis of longitudinal GRR and the longitudinal UE-based protocols and the numerical evaluation of their performance, and we present our ALLOMFREE solution. In Section 5, we present experimental results and discuss our results. In Section 6 we review the related work. Lastly, in Section 7, we present the concluding remarks and future directions.

2. Theoretical background

In this section, we briefly present the concept of privacy considered in this work, that is, LDP, and the LDP protocols we will apply in this paper.

2.1. LDP

Local differential privacy, initially formalized in [10], protects an individual's privacy during the data collection process. A formal definition of LDP is given as follows:

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

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

Similar to the centralized model of DP, LDP also enjoys several important properties, e.g., immunity to post-processing ($F(\mathcal{A})$ is ϵ-LDP for any function F) and composability [3]. That is, combining the results from d locally differentially private protocols also satisfy LDP. If these protocols are applied separately in disjointed subsets of the dataset, $\epsilon = max(\epsilon_1\text{-}, \dots, \epsilon_d)$-LDP (parallel composition). On the other hand, if these protocols are sequentially applied to the same dataset, $\epsilon = \sum_{i=1}^{d} \epsilon_i$-LDP (sequential composition).

2.2. LDP protocols

Randomized Response (RR), a surveying technique proposed by Warner [41], has been the building block for many LDP protocols. Let $A_j = \{v_1, v_2, ..., v_{k_j}\}$ be a set of $k_j = |A_j|$ values of a given attribute and let ϵ be the privacy budget, we review three state-of-the-art LDP mechanisms for single-frequency estimation (a.k.a. frequency oracles) that will be used in this paper.

2.2.1. GRR

The k-Ary RR [18] mechanism extends RR to the case of $k_j \geq 2$ and is also referred to as direct encoding [14] or Generalized RR (GRR) [42, 43, 29]. Throughout this paper, we use the term GRR for this LDP protocol. Given a value $v \in A_j$, *GRR(v)* outputs the true value with probability p, and any other value $v' \in A_j$ such that $v' \neq v$ with probability $1 - p$. More formally, the perturbation function is defined as:

$$\forall y \in A_j \, \Pr[\mathcal{A}_{GRR(\epsilon)}(v) = y] = \begin{cases} p = \frac{e^{\epsilon}}{e^{\epsilon}+k_j-1}, & \text{if } y = v \\ q = \frac{1}{e^{\epsilon}+k_j-1}, & \text{if } y \neq v \end{cases}$$

This satisfies ϵ-LDP since $\frac{p}{q} = e^{\epsilon}$. On expectation, the number of times that a value v_i is reported, N_i, for $i \in [1, k_j]$, is given by:

$$\mathbb{E}[N_i] = nf(v_i)p + n(1 - f(v_i))q$$

in which N_i is the number of times the value v_i has been reported, $f(v_i)$ is the real frequency of value v_i, and n is the total number of users. This immediately provides the normalized estimation $\hat{f}(v_i)$ that each value v_i occurs as [18, 14, 11]:

$$\hat{f}(v_i) = \frac{N_i - nq}{n(p - q)} \tag{1}$$

In [14], the authors prove that $\hat{f}(v_i)$ in Eq. (1) is an unbiased estimation of the true frequency $f(v_i)$, and the variance of this estimation is $Var[\hat{f}(v_i)] = \frac{q(1-q)}{n(p-q)^2} + \frac{f(v_i)(1-p-q)}{n(p-q)}$. In the case of small $f(v_i) \sim 0$, this variance is dominated by the first term, which gives the *approximate* variance as [14]:

$$Var^*[\hat{f}(v_i)] = \frac{q(1 - q)}{n(p - q)^2} \tag{2}$$

Since the estimation in Eq. (1) is unbiased, its variance $Var[\hat{f}(v_i)]$ is equal to the Mean Squared Error (MSE), which is commonly used as an accuracy metric (e.g., cf. [43, 35]) and also adopted in this paper.

Replacing $p = \frac{e^{\epsilon}}{e^{\epsilon}+k_j-1}$ and $q = \frac{1}{e^{\epsilon}+k_j-1}$ into Eq. (2), the GRR variance is calculated as:

$$Var^*[\hat{f}_{GRR}(v_i)] = \frac{e^{\epsilon} + k_j - 2}{n(e^{\epsilon} - 1)^2} \tag{3}$$

2.2.2. Unary encoding-based

Protocols based on Unary Encoding (UE) consist of transforming a value v into a binary representation of it. So, first, for a given value v, $B = UE(v)$, where $B = [0, 0, ..., 1, 0, ...0]$, a k_j-bit array where only the v-th position is set to one. Next, the bits i, for $i \in [1, k_j]$, from B are flipped, depending on parameters p and q, to generate a sanitized vector B', in which:

$$\Pr[B'_i = 1] = \begin{cases} p, & \text{if } B_= 1 \\ q, & \text{if } B_= 0 \end{cases}$$

The proof that the UE-based protocols satisfy ϵ-LDP for

$$\epsilon = ln\left(\frac{p(1 - q)}{(1 - p)q}\right) \tag{4}$$

is known in the literature and can be found in [11, 14]. In [14] the authors presented two ways for selecting probabilities p and q, which determines the protocol variance. One well-known UE-based protocol is the basic one-time RAPPOR [11], referred to as Symmetric UE (SUE), which selects $p = \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}$ and $q = \frac{1}{e^{\epsilon/2}+1}$, where $p+q = 1$ (symmetric). The estimated frequency $\hat{f}(v_i)$ that a value v_i occurs for $i \in [1, k_j]$ is also calculated using Eq. (1). Replacing $p = \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}$ and $q = \frac{1}{e^{\epsilon/2}+1}$ into Eq. (2), the SUE variance is calculated as [11]:

$$Var^*[\hat{f}_{SUE}(v_i)] = \frac{e^{\epsilon/2}}{n(e^{\epsilon/2} - 1)^2} \tag{5}$$

Moreover, rather than select p and q to be symmetric, Wang *et al.* [14] proposed Optimized UE (OUE), which selects parameters $p = \frac{1}{2}$ and $q = \frac{1}{e^{\epsilon}+1}$ that minimize the variance of UE-based protocols while still satisfying ϵ-LDP. Similarly, the estimation method used in Eq. (1) equally applies to OUE. Replacing $p = \frac{1}{2}$ and $q = \frac{1}{e^{\epsilon}+1}$ into Eq. (2), the OUE variance is calculated as [14]:

$$Var^*[\hat{f}_{OUE}(v_i)] = \frac{4e^{\epsilon}}{n(e^{\epsilon} - 1)^2} \tag{6}$$

3. Multidimensional frequency estimates with LDP

In the literature, few work for collecting multidimensional data with LDP is based on random sampling (i.e., dividing users in groups) [32, 33, 34, 35, 14, 38]. This technique reduces both dimensionality and communication costs, which will also be the focus of this paper. Let $d \geq 2$ be the total number of attributes, $\mathbf{k} = [k_1, k_2, ..., k_d]$ be the domain size of each attribute, n be the number of users, and ϵ be the privacy budget. An intuitive solution (*Spl*) is to split the privacy budget, i.e., assigning ϵ/d for each attribute. The other solution (*Smp*) is based on uniformly sampling (without replacement) only r attribute(s) out of d possible ones, i.e., assigning ϵ/r per attribute. Notice that both solutions satisfy ϵ-LDP according to the sequential composition theorem [3].

For the first case, *Spl*, the variances (σ_1^2) of GRR, SUE, and OUE are respectively:

$$\begin{aligned}\sigma^2_{1,GRR} &= \frac{e^{\epsilon/d}+k_j-2}{n(e^{\epsilon/d}-1)^2}\\ \sigma^2_{1,SUE} &= \frac{e^{\epsilon/2d}}{n(e^{\epsilon/2d}-1)^2}\\ \sigma^2_{1,OUE} &= \frac{4e^{\epsilon/d}}{n(e^{\epsilon/d}-1)^2}\end{aligned} \tag{7}$$

For the second case, *Smp*, the number of users per attribute is reduced to nr/d. Thus, the variances (σ_2^2) of GRR, SUE, and OUE are, respectively:

$$\begin{aligned}\sigma^2_{2,GRR} &= \frac{d(e^{\epsilon/r}+k_j-2)}{nr(e^{\epsilon/r}-1)^2}\\ \sigma^2_{2,SUE} &= \frac{d(e^{\epsilon/2r})}{nr(e^{\epsilon/2r}-1)^2}\\ \sigma^2_{2,OUE} &= \frac{d(4e^{\epsilon/r})}{nr(e^{\epsilon/r}-1)^2}\end{aligned} \tag{8}$$

Notice that if $r = d$ in Eq. (8), one achieves Eq. (7). Practically, the objective is reduced to finding r, which minimizes σ_2^2 for each protocol. In this way, to find the optimal r for each protocol, we first multiply each σ_2^2 in Eq. (8) by ϵ. Without losing generality, minimizing $\sigma^2_{2,GRR}$, $\sigma^2_{2,SUE}$, and $\sigma^2_{2,OUE}$ is equivalent to minimizing $\frac{\epsilon e^{\epsilon/r}}{r(e^{\epsilon/r}-1)^2}$, $\frac{\epsilon e^{\epsilon/2r}}{r(e^{\epsilon/2r}-1)^2}$, and $\frac{\epsilon e^{\epsilon/r}}{r(e^{\epsilon/r}-1)^2}$, respectively. Hence, let $x = r/\epsilon$ be the independent variable, $\sigma^2_{2,GRR}$ and $\sigma^2_{2,OUE}$ can be rewritten as $y_1 = \frac{1}{x}\cdot\frac{e^{1/x}}{(e^{1/x}-1)^2}$, and $\sigma^2_{2,SUE}$ can be rewritten as $y_2 = \frac{1}{x}\cdot\frac{e^{1/2x}}{(e^{1/2x}-1)^2}$ as functions over x. It is not hard to prove that both y_1 and y_2 are increasing functions w.r.t. x. Therefore, the minimum and optimal number of attributes per user is $r = 1$ for all three protocols. We highlight that this is a common result in the LDP literature obtained for different protocols and contexts [32, 33, 35, 14, 24, 37, 36, 44].

Therefore, in this paper, we adopt the multidimensional setting *Smp* with $r = 1$. In this setting, users tell the data collector whose attribute is sampled, and its perturbed value ensures ϵ-LDP by applying either GRR or UE-based protocols; the data analyst server would not receive any information about the remaining $d-1$ attributes.

4. Longitudinal frequency estimates with LDP

In this section, we first present the *memoization*-based framework for longitudinal data collections. Next, we present the analysis of longitudinal GRR and longitudinal UE-based protocols. Lastly, we numerically evaluate the extended longitudinal protocols and propose our ALLOMFREE solution.

4.1. Memoization-based data collection with LDP

In the literature, many studies focus on how to collect and analyze categorical data longitudinally based on *memoization* [11, 12, 36]. The key idea behind memoization is using two sanitization processes. The first round (RR_1) replaces the real value B with a sanitized one B' with a higher epsilon (ϵ_∞). Whenever one intends to report B, B' shall be reused to produce other sanitized versions B'' with lower epsilon values. Notice that the second sanitization (RR_2) is a *must* to avoid "averaging attacks", in which adversaries can reconstruct the true value from multiple sanitized versions of it. This technique allows achieving privacy over time with an upper bound value of ϵ_∞-LDP.

Let $A_j = \{v_1, v_2, ..., v_{k_j}\}$ be a set of $k_j = |A_j|$ values of a given attribute and let ϵ be the privacy budget. In this paper, for both RR_1 and RR_2 steps, we will apply either GRR, SUE, or OUE. The unbiased estimator in Eq. (1) for the frequency $f(v_i)$ of each value v_i for $i \in [1, k_j]$ is now extended to:

$$\hat{f}_L(v_i) = \frac{\frac{N_i - nq_2}{(p_2-q_2)} - nq_1}{n(p_1-q_1)} = \frac{N_i - nq_1(p_2-q_2) - nq_2}{n(p_1-q_1)(p_2-q_2)} \tag{9}$$

in which N_i is the number of times the value v_i has been reported, n is the total number of users, p_1 and q_1 are the parameters used by an LDP protocol for RR_1, and p_2 and q_2 are the parameters used by an LDP protocol for RR_2. Eq. (9) is the result of using the unbiased estimator of Eq. (1) with two rounds of sanitization.

Theorem 1. *The estimation result $\hat{f}_L(v_i)$ in Eq. (9) is an unbiased estimation of $f(v_i)$ for any value $v_i \in A_j$.*

Proof.

$$\begin{aligned}\mathbb{E}[\hat{f}_L(v_i)] &= \mathbb{E}\left[\frac{N_i - nq_1(p_2-q_2) - nq_2}{n(p_1-q_1)(p_2-q_2)}\right]\\ &= \frac{\mathbb{E}[Ni] - nq_1(p_2-q_2) - nq_2}{n(p_1-q_1)(p_2-q_2)}\end{aligned}$$

Let us focus on

$$\begin{aligned}\mathbb{E}[N_i] = {} & nf(v_i)\,(p_1p_2 + q_2\,(1-p_1))\\ & + n\,(1-f(v_i))\,(p_2q_1 + q_2\,(1-q_1))\end{aligned}$$

Thus,

$$\mathbb{E}[\hat{f}_L(v_i)] = \frac{nf(v_i)\,(p_1p_2 + q_2\,(1-p_1)) - nq_1\,(p_2 - q_2) - nq_2}{n\,(p_1 - q_1)\,(p_2 - q_2)} + \frac{(-f(v_i)n + n)\,(p_2q_1 + q_2\,(1-q_1))}{n\,(p_1 - q_1)\,(p_2 - q_2)} = f(v_i)$$

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

$$Var[\hat{f}_L(v_i)] = \frac{\gamma(1-\gamma)}{n(p_1 - q_1)^2(p_2 - q_2)^2},\ where$$
$$\gamma = f(v_i)\,(2p_1p_2 - 2p_1q_2 + 2q_2 - 1) + p_2q_1 + q_2(1 - q_1) \tag{10}$$

Proof. Thanks to Eq. (9), we have

$$Var\left[\hat{f}_L(v_i)\right] = \frac{Var[N_i]}{n^2(p_1 - q_1)^2(p_2 - q_2)^2}$$

Since N_i is the number of times the value v_i is observed, it can be defined as $N_i = \sum_{z=1}^{n} X_z$, where X_z is equal to 1 if the user z, $1 \leq z \leq n$ reports value v_i, and 0 otherwise. We thus have $Var[N_i] = \sum_{z=1}^{n} Var[X_z] = nVar[X]$. Since all the users are independent,

$$\Pr[X = 1] = P[X^2 = 1] = f(v_i)\,(2p_1p_2 - 2p_1q_2 + 2q_2 - 1) + p_2q_1 + q_2(1 - q_1) = \gamma$$

We thus have $Var[X] = \gamma - \gamma^2 = \gamma(1 - \gamma)$ and, finally,

$$Var[\hat{f}_L(v_i)] = \frac{\gamma(1-\gamma)}{n(p_1 - q_1)^2(p_2 - q_2)^2}$$

In this work, we will use the *approximate variance*, in which $f(v_i) = 0$ in Eq. (10), which gives:

$$Var^*\left[\hat{f}_L(v_i)\right] = \frac{(p_2q_1 - q_2\,(q_1 - 1))\,(-p_2q_1 + q_2\,(q_1 - 1) + 1)}{n(p_1 - q_1)^2(p_2 - q_2)^2} \tag{11}$$

4.2. Longitudinal GRR (L-GRR): definition and ϵ-LDP study

Let $V = \{v_1, v_2, ..., v_{k_j}\}$ be a set of k_j values of a given attribute and let v_i be the real value. We now describe an extension of GRR for longitudinal studies; we refer to this protocol as L-GRR for the rest of this paper. First, $Encode(v_i) = v_i$ (direct encoding). Next, there are two rounds of sanitization, RR_1 and RR_2 applying GRR, as described in the following equations.

1. $RR_1[GRR]$: Memoize a value B' such that

$$B' = \begin{cases} v_i, & \text{with probability } p_1 \\ v_{k \neq v_i}, & \text{with probability } q_1 = \frac{1-p_1}{k_j - 1} \end{cases}$$

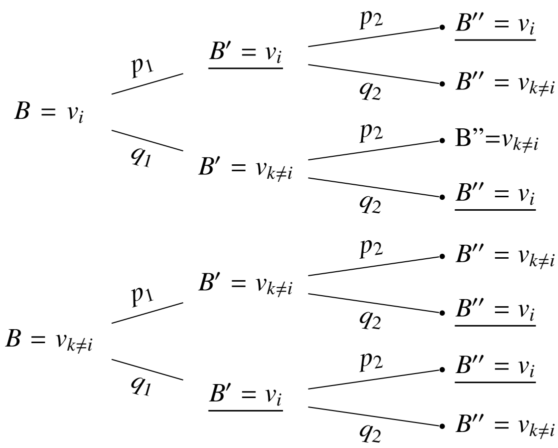

Fig. 1: Probability trees for two rounds of sanitization using GRR (L-GRR).

in which p_1 and q_1 control the level of longitudinal ϵ_∞-LDP. The value B' shall be reused as the basis for all future reports on the real value v_i.

2. $RR_2[GRR]$: Generate a reporting B'' such that

$$B'' = \begin{cases} B', & \text{with probability } p_2 \\ v_{k \neq B'}, & \text{with probability } q_2 = \frac{1-p_2}{k_j - 1} \end{cases}$$

in which B'' is the report to be sent to the server.

Visually, Fig. 1 illustrates the probability tree of the L-GRR protocol. In the first round of sanitization, RR_1, our proposed L-GRR applies GRR with $p_1 = \Pr[B' = v_i | B = v_i] = \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty} + k_j - 1}$ and $q_1 = \Pr[B' = v_i | B = v_{k\neq i}] = \frac{1-p_1}{k_j - 1} = \frac{1}{e^{\epsilon_\infty} + k_j - 1}$ (underlined in the middle of Fig. 1), where $k_j = |A_j|$. As discussed in subsection 2.2.1, this *permanent* memoization satisfies ϵ_∞-LDP since $\frac{p_1}{q_1} = e^{\epsilon_\infty}$, which is the upper bound.

On the other hand, with a single collection of data, the attacker's knowledge of v_i comes only from B'', which is generated using two randomization steps with GRR. This provides a higher level of privacy protection [11]. From Fig. 1, we can obtain the following conditional probabilities:

$$\Pr[B''|B] = \begin{cases} \Pr[B'' = v_i | B = v_i] = p_1p_2 + q_1q_2 \\ \Pr[B'' = v_{k\neq i} | B = v_i] = p_1q_2 + q_1p_2 \\ \Pr[B'' = v_i | B = v_{k\neq i}] = p_1q_2 + q_1p_2 \\ \Pr[B'' = v_{k\neq i} | B = v_{k\neq i}] = p_1p_2 + q_1q_2 \end{cases}$$

Let $p_s = \Pr[B'' = v_i | B = v_i]$ and $q_s = \Pr[B'' = v_i | B = v_{k\neq i}]$ (underlined in the far right of Fig. 1), with the second round of sanitization, $RR_2[GRR]$, our proposed L-GRR protocol satisfies ϵ_1-LDP since $\frac{p_s}{q_s} = e^{\epsilon_1}$. Notice that ϵ_1 corresponds to a single report (lower bound) and its extension to infinity reports is limited by ϵ_∞ (upper bound) since $RR_2[GRR]$ uses as input the output of $RR_1[GRR]$. More specifically, the calculus of ϵ_1 for L-GRR is:

$$\epsilon_1 = \ln\left(\frac{p_1 p_2 + q_1 q_2}{p_1 q_2 + q_1 p_2}\right) \quad (12)$$

in which $p_1 = \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+k_j-1}$, $q_1 = \frac{1-p_1}{k_j-1}$, and both p_2 and q_2 are selectable according to ϵ_∞, ϵ_1, and k_j, calculated as:

$$p_2 = \frac{e^{\epsilon_1+\epsilon_\infty} - 1}{-k_j e^{\epsilon_1} + \left(k_j - 1\right) e^{\epsilon_\infty} + e^{\epsilon_1} + e^{\epsilon_1+\epsilon_\infty} - 1}$$
$$q_2 = \frac{1 - p_2}{k_j - 1} \quad (13)$$

The estimated frequency $\hat{f}_L(v_i)$ that a value v_i occurs for $i \in [1, k_j]$ is calculated using Eq. (9). Lastly, one can calculate the L-GRR approximate variance by replacing the resulting p_1, q_1, p_2, q_2 parameters into Eq. (11).

4.3. Longitudinal UE (L-UE): definition and ϵ-LDP study

We now describe the UE-based protocol for longitudinal studies. We refer to this protocol as L-UE for the rest of this paper. Let $V = \{v_1, v_2, ..., v_{k_j}\}$ be a set of k_j values of a given attribute and let v_i be the real value. First, $Encode(v_i) = B$ (unary encoding), where $B = [0, 0, ..., 1, 0, ...0]$, a k_j-bit array where only the v-th position is set to one. Next, there are two rounds of sanitization, RR_1 and RR_2, which apply the UE-based protocols, described as follows.

1. $RR_1[UE]$: For each bit i, $1 \leq i \leq k_j$ in B, memoize a value B' such that

$$\Pr[B'_i = 1] = \begin{cases} p_1, & \text{if } B_i = 1 \\ q_1, & \text{if } B_i = 0 \end{cases}$$

   in which p_1 and q_1 control the level of longitudinal ϵ_∞-LDP. The value B' shall be reused as the basis for all future reports on the real value v_i.

2. $RR_2[UE]$: For each bit i, $1 \leq i \leq k_j$ in B', generate a reporting B'' that

$$\Pr[B''_i = 1] = \begin{cases} p_2, & \text{if } B'_i = 1 \\ q_2, & \text{if } B'_i = 0 \end{cases}$$

   in which B'' is the report to be sent to the server.

Visually, Fig. 2 illustrates the probability tree of the L-UE protocol. **One natural question emerges: how to select the parameters $\{p_1, q_1, p_2, q_2\}$ in order to optimize the utility of this L-UE protocol?** One can see $RR_1[UE]$ as a *permanent* sanitization and $RR_2[UE]$ as a 'small' perturbation to avoid averaging attacks and keep privacy over time.

Based on SUE and OUE, we are then left with four options: two popular solutions that strictly use only OUE or SUE parameters in both sanitization steps and

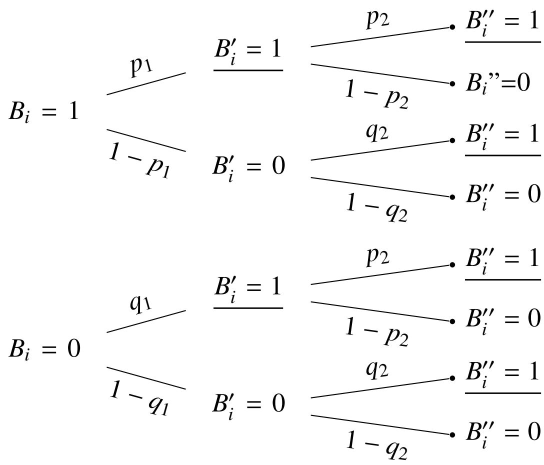

Fig. 2: Probability trees for two rounds of sanitization using UE (L-UE).

two proposed settings that combine both OUE and SUE. These four L-UE protocols are summarized below:

I both sanitizations with OUE (L-OUE);

II both sanitizations with SUE (L-SUE);

III starting with OUE and then with SUE (L-OSUE);

IV starting with SUE and then with OUE (L-SOUE);

in which L-SUE is the well-known Basic-RAPPOR protocol [11], L-OUE is the state-of-the-art OUE protocol [14] with memoization, and both L-OSUE and L-SOUE are proposed in this paper.

As presented in [14], the OUE variance in Eq. (6) is smaller than the SUE variance in Eq. (5) and, therefore, the former can provide higher utility than the latter for RR_1. On the other hand, we argue that OUE might be too strict for RR_2 since the parameter $p_2 = 1/2$ is constant. Thus, we hypothesize that option III (i.e., L-OSUE) is the most suitable one. Without losing generality, **the following analyses are done only for L-OSUE**, which can be easily extended to any of the other combinations.

In the first round of sanitization, RR_1, our solution L-OSUE applies OUE with $p_1 = \Pr[B'_i = 1 | B_i = 1] = \frac{1}{2}$ and $q_1 = \Pr[B'_i = 1 | B_i = 0] = \frac{1}{e^{\epsilon_\infty}+1}$ (underlined in the middle of Fig. 2). As discussed in Section 2.2.2, this *permanent* memoization satisfies ϵ_∞-LDP since $\frac{p_1(1-q_1)}{(1-p_1)q_1} = e^{\epsilon_\infty}$, which is the upper bound.

Following the same development as for L-GRR, on the other hand, with a single collection of data, the attacker's knowledge of $B = UE(v)$ comes only from B'', which is generated using two randomization steps with OUE and SUE, respectively. This provides a higher level of privacy protection [11]. From Fig. 2, we can obtain the following conditional probabilities according to each bit $i \in [1, k_j]$:

$$\Pr[B_i''|B_i] = \begin{cases} \Pr[B_i'' = 1|B_i = 1] = p_1p_2 + (1-p_1)q_2 \\ \Pr[B_i'' = 0|B_i = 1] = p_1(1-p_2) + (1-p_1)(1-q_2) \\ \Pr[B_i'' = 1|B_i = 0] = q_1p_2 + (1-q_1)q_2 \\ \Pr[B_i'' = 0|B_i = 0] = q_1(1-p_2) + (1-q_1)(1-q_2) \end{cases}$$

Let $p_s = \Pr[B_i'' = 1|B_i = 1]$ and $q_s = \Pr[B_i'' = 1|B_i = 0]$ (underlined in far right of Fig. 2), with the second round of sanitization, $RR_2[SUE]$, our proposed L-OSUE protocol satisfies ϵ_1-LDP since $\frac{p_s(1-q_s)}{(1-p_s)q_s} = e^{\epsilon_1}$. Notice that ϵ_1 corresponds to a single report (lower bound) and its extension to infinity reports is limited by ϵ_∞ (upper bound) since $RR_2[SUE]$ uses as input the output of $RR_1[OUE]$. More specifically, the calculus of ϵ_1 for L-OSUE (or L-UE protocols in general) is:

$$\epsilon_1 = \ln\left(\frac{(p_1p_2 - q_2(p_1-1))(p_2q_1 - q_2(q_1-1)-1)}{(p_2q_1 - q_2(q_1-1))(p_1p_2 - q_2(p_1-1)-1)}\right) \tag{14}$$

in which, for L-OSUE, we have $p_1 = \frac{1}{2}$, $q_1 = \frac{1}{e^{\epsilon_\infty}+1}$, and both p_2 and q_2 are symmetric ($p_2 + q_2 = 1$) and selectable according to ϵ_∞ and ϵ_1, calculated as:

$$p_2 = \frac{1 - e^{\epsilon_1+\epsilon_\infty}}{e^{\epsilon_1} - e^{\epsilon_\infty} - e^{\epsilon_1+\epsilon_\infty} + 1} \qquad q_2 = 1 - p_2 \tag{15}$$

Similarly, the estimated frequency $\hat{f}_L(v_i)$ that a value v_i occurs for $i \in [1, k_j]$ is calculated using Eq. (9). Lastly, one can calculate the L-OSUE (or L-UE protocols in general) approximate variance by replacing the resulting p_1, q_1, p_2, q_2 parameters into Eq. (11).

4.4. Numerical evaluation of L-GRR and L-UE protocols

In this subsection, we evaluate numerically the approximate variance of all developed longitudinal protocols, namely, L-GRR, and the four UE-based options, namely, L-OUE, L-SUE, L-OSUE, and L-SOUE, respectively. As aforementioned, once both ϵ_∞ and ϵ_1 privacy guarantees are defined, one can obtain parameters p_1 and q_1 depending on ϵ_∞, and parameters p_2 and q_2 depending on both ϵ_∞ and ϵ_1 (and the domain size k_j for L-GRR), as given in Eq. (13) for L-GRR and in Eq. (15) for L-OSUE.

Next, once the parameters $\{p_1, q_1, p_2, q_2\}$ are computed, one can calculate the approximate variance with Eq. (11) for each protocol. In other words, following our proposal, one has to set both the upper (ϵ_∞) and lower (ϵ_1) bounds of the privacy guarantees. For example, let $\epsilon_\infty = 2$, one might want the first ϵ_1-LDP report to have high privacy such as $\epsilon_1 = 0.1$, i.e., $\epsilon_1 = 0.05\epsilon_\infty$ (**we will use this percentage notation to set up the privacy guarantees**).

Table 1 exhibits the numerical values of the approximate variance using Eq. (11) for all longitudinal protocols with $n = 10000$, $\epsilon_\infty = [0.5, 1.0, 2.0, 4.0]$ (as in [14]), and $\epsilon_1 = \{0.6\epsilon_\infty, 0.5\epsilon_\infty, 0.4\epsilon_\infty, 0.3\epsilon_\infty, 0.2\epsilon_\infty, 0.1\epsilon_\infty\}$. For values of ϵ_1 higher than $0.6\epsilon_\infty$, neither L-OUE nor L-SOUE could satisfy some values of ϵ_1 because of the constant $p_2 = 1/2$ in RR_2. However, it is not desirable to have higher values of ϵ_1 and, thus, we do not consider values above $0.6\epsilon_\infty$ in our analysis. Besides, Table 2 exhibits the numerical values for the non-longitudinal GRR, OUE, and SUE protocols, which allow evaluating how utility degrades with a second step of sanitization.

From Table 1, one can notice that L-GRR presents the smallest variance values for binary attributes (i.e., when $k_j = 2$). On the other hand, L-GRR is also most sensitive to changes in privacy parameters ϵ_∞ and ϵ_1 when k_j is large, which shows a much higher variance than when using a non-longitudinal GRR, as shown in Table 2. Similar to the non-longitudinal GRR, this increase in the variance is due to the number of values k_j, which decreases the probability p of reporting the true value. With two rounds of sanitization, it further deteriorates the accuracy of the L-GRR protocol that gets extremely high values, e.g., see L-GRR($k_j = 2^{10}$). Interestingly, when $k_j = 2$ in Table 1, the variance of L-GRR with $\epsilon_1 = 0.5\epsilon_\infty$ is a lagged version of the variance values given by the non-longitudinal GRR in Table 2. This effect is also observed for both the L-SUE (cf. SUE in Table 2) and L-OSUE (cf. OUE in Table 2) protocols, which use symmetric probabilities on RR_2 (i.e., $p_2 + q_2 = 1$). We highlight these values in **bold font**. However, for L-GRR, this is not true for other values of k_j, the further analysis of which is beyond the scope of this paper.

On the other hand, the L-UE protocols avoid having a variance that depends on k_j by encoding the value into the unary representation, which results in a constant variance regardless of the size of the attribute. To complement the results of Table 1, Fig. 3 illustrates the numerical values of the approximate variance for the L-UE protocols with $\epsilon_1 = \{0.3\epsilon_\infty, 0.6\epsilon_\infty\}$. With the four options I-IV analyzed, on the high privacy regimes, L-OSUE and L-SUE have similar performance while *always* favoring the proposed L-OSUE. On lower privacy regimes, our proposed protocols L-SOUE and L-OSUE have similar performance, which outperform both the L-OUE and L-SUE protocols. As shown in our experiments, the L-OUE protocol has the worst performance among the four options analyzed, with the exception of high values for ϵ_∞ (see the plot on the bottom of Fig. 3), when it has performance superior or similar to that of L-SUE. Indeed, for L-OUE, selecting $p_2 = 1/2$ for the second sanitization step is too strict, which results in higher variance values. Therefore, by comparing the approximate variances,

Privacy Guarantees		L-GRR			L-UE			
		$k_j = 2$	$k_j = 32$	$k_j = 2^{10}$	L-OSUE	L-SUE	L-SOUE	L-OUE
$\epsilon_1 = 0.6\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.30$	0.001103	0.980969	26706	0.004411	0.004436	0.005306	0.005549
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.60$	0.000270	0.125036	3153	0.001078	0.001103	0.001234	0.001347
	$\epsilon_\infty = 2.0, \epsilon_1 = 1.20$	0.000062	0.006327	117	0.000247	0.000270	0.000264	0.000310
	$\epsilon_\infty = 4.0, \epsilon_1 = 2.40$	0.000011	0.000078	0.25903	0.000044	0.000062	0.000045	0.000057
$\epsilon_1 = 0.5\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.25$	0.001592	2.088372	60218	0.006367	0.006392	0.007336	0.007611
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.50$	**0.000392**	0.268074	7198	**0.001567**	**0.001592**	0.001740	0.001872
	$\epsilon_\infty = 2.0, \epsilon_1 = 1.00$	**0.000092**	0.013926	281	**0.000368**	**0.000392**	0.000389	0.000447
	$\epsilon_\infty = 4.0, \epsilon_1 = 2.00$	**0.000018**	0.000188	0.74088	**0.000072**	**0.000092**	0.000073	0.000092
$\epsilon_1 = 0.4\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.20$	0.002492	4.530779	135874	0.009967	0.009992	0.011012	0.011324
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.40$	0.000617	0.586823	16443	0.002467	0.002492	0.002658	0.002812
	$\epsilon_\infty = 2.0, \epsilon_1 = 0.80$	0.000148	0.031552	673	0.000593	0.000617	0.000617	0.000690
	$\epsilon_\infty = 4.0, \epsilon_1 = 1.60$	0.000032	0.000484	2.12772	0.000127	0.000148	0.000128	0.000156
$\epsilon_1 = 0.3\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.15$	0.004436	10	329836	0.017744	0.017769	0.018863	0.019214
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.30$	0.001103	1.398568	40412	0.004411	0.004436	0.004620	0.004799
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.60$	0.000270	0.078202	1737	0.001078	0.001103	0.001106	0.001198
	$\epsilon_\infty = 2.0, \epsilon_1 = 1.20$	0.000062	0.001389	6	0.000247	0.000270	0.000248	0.000291
$\epsilon_1 = 0.2\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.10$	0.009992	30	972656	0.039967	0.039992	0.041148	0.041536
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.20$	0.002492	4.080052	120651	0.009967	0.009992	0.010190	0.010394
	$\epsilon_\infty = 2.0, \epsilon_1 = 0.40$	0.000617	0.237925	5443	0.002467	0.002492	0.002498	0.002610
	$\epsilon_\infty = 4.0, \epsilon_1 = 0.80$	0.000148	0.004939	24	0.000593	0.000617	0.000595	0.000659
$\epsilon_1 = 0.1\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.05$	0.039992	154	4941829	0.159967	0.159992	0.161191	0.161608
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.10$	0.009992	20	620584	0.039967	0.039992	0.040201	0.040424
	$\epsilon_\infty = 2.0, \epsilon_1 = 0.20$	0.002492	1.255550	29356	0.009967	0.009992	0.010000	0.010130
	$\epsilon_\infty = 4.0, \epsilon_1 = 0.40$	0.000617	0.030494	156	0.002467	0.002492	0.002469	0.002560

Table 1: Numerical values of Eq. (11) (i.e., $Var^*[\hat{f}_L(v_i)]$) for L-GRR and L-UE protocols with different ϵ_∞ and ϵ_1 privacy guarantees, following $\epsilon_1 = \{0.6\epsilon_\infty, 0.5\epsilon_\infty, 0.4\epsilon_\infty, 0.3\epsilon_\infty, 0.2\epsilon_\infty, 0.1\epsilon_\infty\}$, respectively.

ϵ_∞	GRR($k_j = 2$)	GRR($k_j = 32$)	GRR($k_j = 2^{10}$)	OUE	SUE
$\epsilon_\infty = 0.5$	**0.000392**	0.007520	0.243240	**0.001567**	**0.001592**
$\epsilon_\infty = 1.0$	**0.000092**	0.001108	0.034707	**0.000368**	**0.000392**
$\epsilon_\infty = 2.0$	**0.000018**	0.000092	0.002522	**0.000072**	**0.000092**
$\epsilon_\infty = 4.0$	0.000002	0.000003	0.000037	0.000008	0.000018

Table 2: Numerical values of $Var^*[\hat{f}(v_i)]$ for the non-longitudinal GRR, OUE, and SUE protocols with different ϵ_∞ privacy guarantees.

the best option for L-UE protocols, in terms of utility, is to start with OUE and then with SUE as we propose in this paper, i.e., L-OSUE.

4.5. The ALLOMFREE algorithm

Let $A = \{A_1, A_2, ..., A_d\}$ be a set of d attributes with the domain size $\mathbf{k} = [k_1, k_2, ..., k_d]$, $\mathbb{A} = \{L\text{-}GRR, L\text{-}OSUE\}$ be a set of optimal longitudinal LDP protocols, and ϵ_∞ and ϵ_1 be the longitudinal and *single-report* privacy guarantees, respectively. Each user u_i, for $1 \leq i \leq n$, holds a tuple $\mathbf{v}^{(i)} = (v_1^{(i)}, v_2^{(i)}, ..., v_d^{(i)})$, i.e., a private value per attribute. From now on, we will simply omit the index notation $\mathbf{v}^{(i)}$ and use $\mathbf{v}$ in the analysis as we focus on one arbitrary user u_i here. For each attribute $j \in [1, d]$ (we slightly abuse the notation and use j for A_j) at time $t \in [1, \tau]$, the aggregator aims to estimate the frequencies of each value $v \in A_j$.

Client-Side. In a multidimensional setting with different domain sizes for each attribute, a dynamic selection of longitudinal LDP protocols is preferred. As mentioned in Section 3, we propose that each user randomly sample $r = Uniform(1, 2, ..., d)$ to select a single attribute A_r. Given k_r (the domain size), ϵ_∞, and ϵ_1, one calculates the parameters $fp_{L-GRR} = \{p_1, q_1, p_2, q_2\}$ and $fp_{L-OSUE} = \{p_1, q_1, p_2, q_2\}$, for L-GRR and L-OSUE, respectively (cf. Eq. (13) and Eq. (15)). Next, with fp_{L-GRR} and fp_{L-OSUE}, one calculates the approximate variances $Var^*[\hat{f}_{L(L\text{-}GRR)}]$ for L-

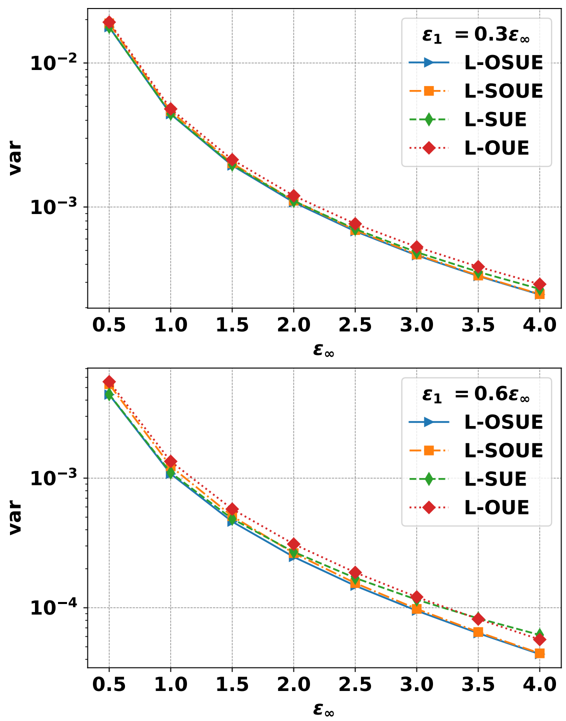

Fig. 3: Numerical values of $Var^*[\hat{f}_L(v_i)]$ for L-UE protocols with $\epsilon_1 = 0.3 \cdot \epsilon_\infty$ (plot on the top) and with $\epsilon_1 = 0.6 \cdot \epsilon_\infty$ (plot on the bottom).

GRR and $Var^*[\hat{f}_{L_{(L\text{-}OSUE)}}]$ for L-OSUE with Eq. (11). Lastly, to select L-GRR as the local randomizer, we are then left to evaluate if $Var^*[\hat{f}_{L_{(L\text{-}GRR)}}] \leq Var^*[\hat{f}_{L_{(L\text{-}OSUE)}}]$. Therefore, the first round of sanitization ensures a *permanent memoization* B' that is always used for the second round of sanitization to generate B'' each time $t \in [1, \tau]$ the user will report the real value B. We call our solution Adaptive LDP for LOngitudinal and Multidimensional FREquency Estimates (ALLOMFREE), which is summarized in Algorithm 1 as a pseudocode.

The intuition of ALLOMFREE is as follows. By requiring each user to submit only 1 attribute with the whole privacy budget, it reduces both the variance incurred as well as the communication cost. Also, since we develop the calculus of the approximate variance in Eq. (11) for the proposed longitudinal protocols (L-GRR and L-OSUE), ALLOMFREE can adaptively select the protocol with a smaller variance value to optimize the data utility. Therefore, ALLOMFREE utilizes optimal solutions for both multidimensional and longitudinal data collection settings developed in Sections 3 and 4 of this paper, respectively.

Server-Side. On the server-side, for each attribute $j \in [1, d]$ at time $t \in [1, \tau]$, the estimated frequency $\hat{f}_L(v_i)$ that a value v_i occurs for $i \in [1, k_j]$ is calculated using Eq. (9).

Privacy analysis. On the one hand, according to the analysis in subsections 4.2 and 4.3, Alg. 1 satisfies ϵ-LDP with upper ϵ_∞ (infinity reports) and lower ϵ_1 (a single report) bounds as it uses either L-GRR or L-OSUE to sanitize a single attribute per user. **Notice that, to ensure the users' privacy over time and to avoid the sequential composition theorem [3], each user must always report the same unique attribute** A_r. In addition, the privacy of a user decreases gracefully according to the number of LDP reports $t \leq \tau$ that an adversary has gained access to, which is calculated as [45, 36]:

$$\epsilon_t = \ln\left(\frac{e^{\epsilon_\infty + t\epsilon_1} + 1}{e^{\epsilon_\infty} + e^{t\epsilon_1}}\right) \leq \min\{\epsilon_\infty, t\epsilon_1\} \tag{16}$$

Limitations. Similar to other sampling-based methods for collecting multidimensional data under LDP [34, 32, 33, 35], our ALLOMFREE algorithm also entails a *sampling error*, which is due to observing a sample instead of the entire population. In addition, concerning the privacy guarantees, the memoization step of ALLOMFREE is certainly effective for longitudinal privacy in the cases where the true client's data does not vary (static) or vary very slowly or in an uncorrelated manner [11]. In many application scenarios, gender, age range, nationality, and other demographic data are generally static or hardly ever vary. On the other hand, for dynamic attributes such as the location or the time spent in the application, this is not the case. Therefore, for each different value, a new memoized value would be generated, thus accumulating the privacy budget ϵ_∞ by the sequential composition theorem [3].

5. Experimental results

In this section, we present the setup of our experiments and the results with real-world data.

5.1. Setup of experiments

The main goal of our experiments is to evaluate the proposed longitudinal LDP protocols on multidimensional frequency estimates a single time, i.e., satisfying ϵ_1-LDP (as in [11, 40, 39], for example).

Environment. All algorithms are implemented in Python 3.8.8 with NumPy 1.19.5 and Numba 0.53.1 libraries. The codes we develop and use for all experiments are available in a Github repository[1]. In all experiments, we report average results over 100 runs as LDP algorithms are randomized.

Methods evaluated. We consider for evaluation the following solutions and protocols:

- Solution *Smp* (cf. Section 3), which randomly samples a single attribute to be sent with the whole privacy budget. We will experiment with the state-of-the-art protocols, namely, L-SUE and L-OUE, and with our extended protocols L-OSUE and L-SOUE;
- Our ALLOMFREE solution (cf. Alg. 1), which also randomly samples a single attribute to be sent with the whole privacy budget but adaptively select the optimal protocol, i.e., either L-GRR or L-OSUE.

Experimental evaluation and metrics. We vary the longitudinal privacy parameter in the range $\epsilon_\infty = [0.5, 1, ..., 3.5, 4]$ with $\epsilon_1 = [0.3\epsilon_\infty, 0.6\epsilon_\infty]$ to compare our experimental results with numerical ones from subsection 4.4. Notice that this range of privacy guarantees is commonly used in the literature for multidimensional data (e.g., in [33] the range is $\epsilon = [0.5, ..., 4]$ and in [35] the range is $\epsilon = [0.1, ..., 10]$).

To evaluate our results, we use the MSE metric averaged per the number of attributes d **in a single data collection** $\tau = 1$**, i.e., with** ϵ_1**-LDP**. Thus, for each attribute j, we compute for each value $v_i \in A_j$ the estimated frequency $\hat{f}(v_i)$ and the real one $f(v_i)$ and calculate their differences. More precisely,

$$MSE_{avg} = \frac{1}{\tau} \sum_{t \in [1,\tau]} \frac{1}{d} \sum_{j \in [1,d]} \frac{1}{|A_j|} \sum_{v \in A_j} (f(v_i) - \hat{f}(v_i))^2$$

Datasets. For the ease of reproducibility, we conduct our experiments on four multidimensional open datasets.

[1] `https://github.com/hharcolezi/ldp-protocols-mobility-cdrs`

Algorithm 1 User-side algorithm of ALLOMFREE.

```
Input : v = [v_1, v_2, ..., v_d], k = [k_1, k_2, ..., k_d], A = {L-GRR, L-OSUE}, ε_∞, ε_1, number of reports τ.
r ← Uniform({1, 2, ..., d})                                          ▹ Select attribute only once
B ← Encode(v_r)                                                      ▹ Encode (if needed)
fp_{L-GRR} ← p_1 = e^{ε_∞}/(e^{ε_∞}+k_r−1), q_1 = (1−p_1)/(k_r−1),
    p_2 = (e^{ε_1+ε_∞}−1)/(−k_r e^{ε_1}+(k_r−1)e^{ε_∞}+e^{ε_1}+e^{ε_1+ε_∞}−1), q_2 = (1−p_2)/(k_r−1)
                                                                     ▹ Get p_2 and q_2 with Eq. (12)
fp_{L-OSUE} ← p_1 = 1/2, q_1 = 1/(e^{ε_∞}+1),
    p_2 = (1−e^{ε_1+ε_∞})/(e^{ε_1}−e^{ε_∞}−e^{ε_1+ε_∞}+1), q_2 = 1 − p_2
                                                                     ▹ Get p_2 and q_2 with Eq. (14)
if Var*[f̂_{L(L-GRR)}](fp_{L-GRR}) ≤ Var*[f̂_{L(L-OSUE)}](fp_{L-OSUE}) :   ▹ Check variances with Eq. (11)
    𝒜 ← L-GRR                                                        ▹ Select L-GRR as local randomizer
else
    𝒜 ← L-OSUE                                                       ▹ Select L-OSUE as local randomizer
B′ ← 𝒜(B, p_1, q_1, k_r)                 ▹ First round of sanitization (permanent memoization)
for t ∈ [1, τ] do
    B″ = 𝒜(B′, p_2, q_2, k_r)                                        ▹ Second round of sanitization
end for
send : (t, ⟨r, B″⟩) for t ∈ [1, τ]
```

- *Nursery.* A dataset from the UCI machine learning repository [46] with $d = 9$ categorical attributes and $n = 12960$ samples. The domain size of each attribute is $\mathbf{k} = [3, 5, 4, 4, 3, 2, 3, 3, 5]$, respectively.
- *Adult.* A dataset from the UCI machine learning repository [46] with $d = 9$ categorical attributes and $n = 45222$ samples after cleaning the data. The domain size of each attribute is $\mathbf{k} = [7, 16, 7, 14, 6, 5, 2, 41, 2]$, respectively.
- *MS-FIMU.* An open dataset from [47] with $d = 6$ categorical attributes and $n = 88935$ samples. The domain size of each attribute is $\mathbf{k} = [3, 3, 8, 12, 37, 11]$, respectively.
- *Census-Income.* A dataset from the UCI machine learning repository [46] with $d = 33$ categorical attributes and $n = 299285$ samples. The domain size of each attribute is $\mathbf{k} = [9, 52, 47, 17, 3, ..., 43, 43, 43, 5, 3, 3, 3, 2]$, respectively.

5.2. *Results*

Our experiments were conducted on four real-world datasets with varied parameters for n, d, and $\mathbf{k}$, which allowed evaluating our solutions more practically. Fig. 4 (*Nursery*), Fig. 5 (*Adult*), Fig. 6 (*MS-FIMU*), and Fig. 7 (*Census-Income*) illustrate for all the evaluated protocols, the averaged MSE_{avg} (y-axis) according to the longitudinal privacy parameter ϵ_∞ (x-axis) with $\epsilon_1 = 0.3\epsilon_\infty$ (plot on the top) and with $\epsilon_1 = 0.6\epsilon_\infty$ (plot on the bottom), respectively.

As one can notice in the results, for all datasets, ALLOMFREE consistently and considerably outperforms the state-of-the-art protocols, namely, L-SUE (a.k.a. Basic-RAPPOR) [11] and L-OUE (that uses OUE [14] twice). Indeed, the difference between the performances of ALLOMFREE and the other longitudinal LDP protocols increases proportionally according to the privacy guarantees, i.e., for high ϵ_∞ and ϵ_1 values, the gap is bigger. This is first because in all datasets there are attribute(s) with a small domain size (e.g., $k_j = 2$ or $k_j = 3$), in which L-GRR can provide smaller variance values than the L-UE protocols (cf. subsection 4.4). Secondly, by adequately selecting the probabilities p_1, q_1, p_2, q_2 for the L-UE protocol (i.e., L-OSUE) also optimizes data utility. Thus, since there is a way to measure the approximate variance of the extended protocols (i.e., Eq. (11)), given the sampled attribute, ALLOMFREE adaptively selects one of the optimized protocol (i.e., L-GRR or L-OSUE) whose smaller variance improves the data utility.

In addition, among the L-UE protocols applied individually, the experimental results with multidimensional data approximate the numerical results with a single attribute from subsection 4.4. For instance, the proposed L-OSUE provides similar or better performance than L-SUE while always outperforming L-OUE. Besides, L-SOUE always outperforms L-OUE too, achieving performance similar to those of L-OSUE and L-SUE in low privacy regimes (i.e., high ϵ values). As we have already shown in subsection 4.4, even though OUE has better utility than SUE for one-time collection [14], applying OUE twice does not provide higher utility.

To complement the results of Figs. 4 – 7, Table 3 ($\epsilon_1 = 0.3\epsilon_\infty$) and Table 4 ($\epsilon_1 = 0.6\epsilon_\infty$) exhibit all datasets and ϵ_∞ guarantees the following utility metrics:

$$\mathcal{U}_{\textit{L-SUE}} = \frac{MSE_{avg_{(\textit{L-SUE})}} - MSE_{avg_{(\textit{ALLOMFREE})}}}{MSE_{avg_{(\textit{L-SUE})}}}$$
$$\mathcal{U}_{\textit{L-OUE}} = \frac{MSE_{avg_{(\textit{L-OUE})}} - MSE_{avg_{(\textit{ALLOMFREE})}}}{MSE_{avg_{(\textit{L-OUE})}}} \tag{17}$$

in which $\mathcal{U}_{\textit{L-SUE}}$ and $\mathcal{U}_{\textit{L-OUE}}$ represent the accuracy gain of ALLOMFREE over the state-of-the-art L-SUE and L-OUE protocols, respectively.

From Tables 3 and 4, one can notice that ALLOMFREE considerably improves the quality of the fre-

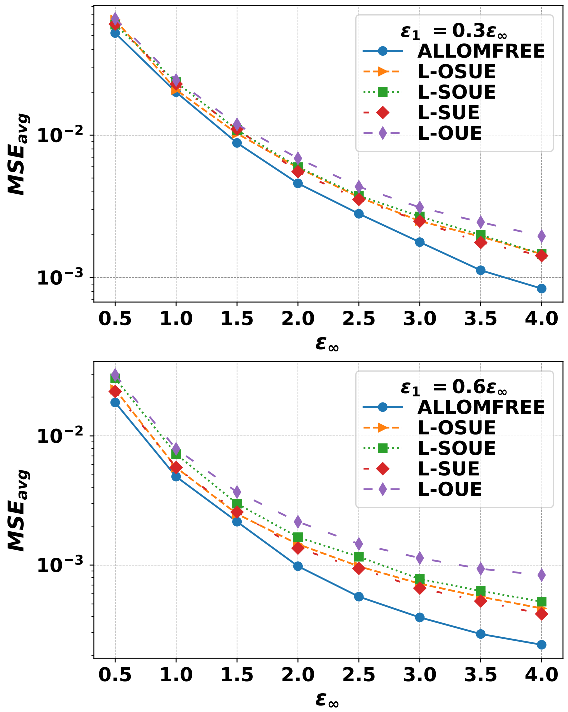

Fig. 4: Averaged MSE varying ϵ_∞ with $\epsilon_1 = 0.3\epsilon_\infty$ (plot on the top) and with $\epsilon_1 = 0.6\epsilon_\infty$ (plot on the bottom) on the *Nursery* dataset.

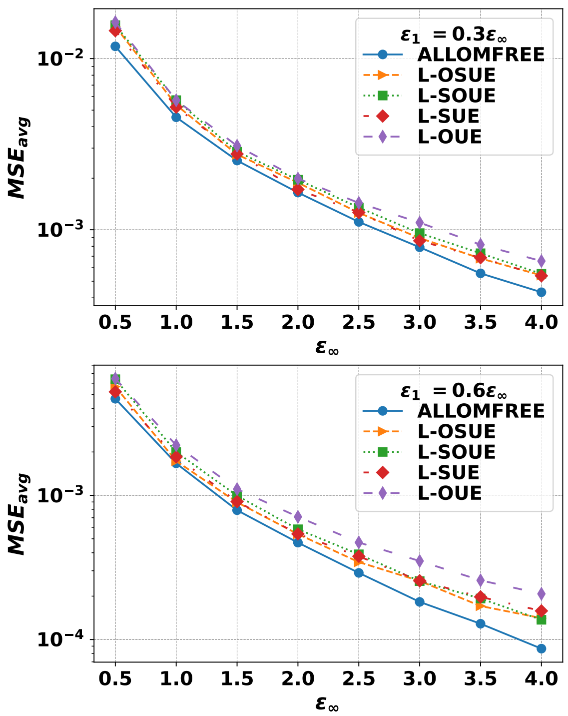

Fig. 5: Averaged MSE varying ϵ_∞ with $\epsilon_1 = 0.3\epsilon_\infty$ (plot on the top) and with $\epsilon_1 = 0.6\epsilon_\infty$ (plot on the bottom) on the *Adult* dataset.

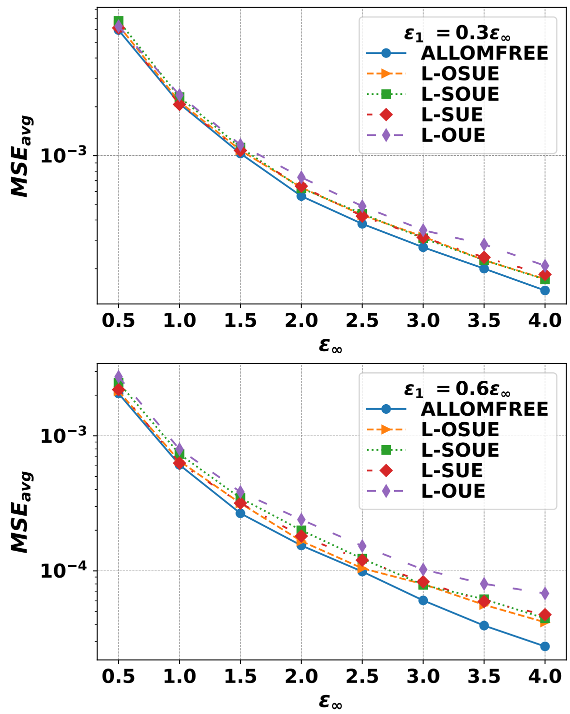

Fig. 6: Averaged MSE varying ϵ_∞ with $\epsilon_1 = 0.3\epsilon_\infty$ (plot on the top) and with $\epsilon_1 = 0.6\epsilon_\infty$ (plot on the bottom) on the *MS-FIMU* dataset.

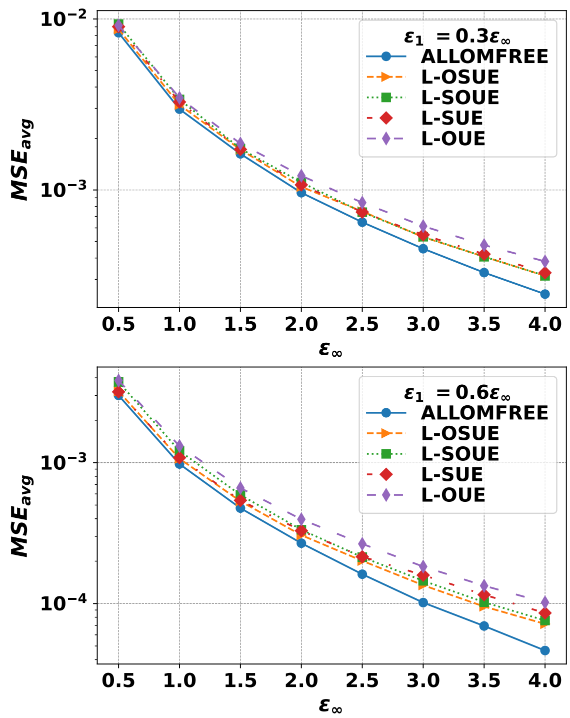

Fig. 7: Averaged MSE varying ϵ_∞ with $\epsilon_1 = 0.3\epsilon_\infty$ (plot on the top) and with $\epsilon_1 = 0.6\epsilon_\infty$ (plot on the bottom) on the *Census-Income* dataset.

ϵ_∞	*Nursery*		*Adult*		*MS-FIMU*		*Census-Income*	
	$\mathcal{U}_{L\text{-}SUE}$	$\mathcal{U}_{L\text{-}OUE}$	$\mathcal{U}_{L\text{-}SUE}$	$\mathcal{U}_{L\text{-}OUE}$	$\mathcal{U}_{L\text{-}SUE}$	$\mathcal{U}_{L\text{-}OUE}$	$\mathcal{U}_{L\text{-}SUE}$	$\mathcal{U}_{L\text{-}OUE}$
0.5	13.51	20.63	19.03	27.73	3.03	5.43	7.84	9.48
1.0	12.36	17.75	12.77	20.44	1.01	11.57	9.21	14.08
1.5	19.95	25.86	8.47	18.01	4.13	11.55	5.82	12.92
2.0	17.18	33.24	4.11	17.16	13.22	23.44	10.06	20.41
2.5	20.70	35.40	11.93	22.54	10.41	22.25	12.77	23.15
3.0	28.69	42.98	8.35	28.22	13.07	21.56	17.07	26.21
3.5	36.19	54.02	18.97	32.02	14.78	29.10	22.02	30.96
4.0	41.24	57.16	19.81	34.25	20.38	29.64	24.99	35.60
Mean	23.73	35.88	12.93	25.05	10.00	19.32	13.72	21.60

Table 3: Accuracy gain of ALLOMFREE over the state-of-the-art L-SUE and L-OUE protocols for all datasets with $\epsilon_1 = 0.3\epsilon_\infty$, measured with the $\mathcal{U}_{L\text{-}SUE}$ and $\mathcal{U}_{L\text{-}OUE}$ metrics expressed in %.

ϵ_∞	*Nursery*		*Adult*		*MS-FIMU*		*Census-Income*	
	$\mathcal{U}_{L\text{-}SUE}$	$\mathcal{U}_{L\text{-}OUE}$	$\mathcal{U}_{L\text{-}SUE}$	$\mathcal{U}_{L\text{-}OUE}$	$\mathcal{U}_{L\text{-}SUE}$	$\mathcal{U}_{L\text{-}OUE}$	$\mathcal{U}_{L\text{-}SUE}$	$\mathcal{U}_{L\text{-}OUE}$
0.5	17.82	38.84	10.42	27.46	6.41	24.79	5.65	21.61
1.0	14.99	38.97	9.83	25.14	2.97	23.32	9.79	25.46
1.5	15.88	41.05	12.90	28.59	16.00	30.52	11.88	28.05
2.0	27.52	54.69	12.95	33.78	14.81	35.65	18.45	32.31
2.5	39.59	60.96	23.28	38.50	17.71	35.34	24.89	39.11
3.0	40.64	65.32	28.59	47.95	27.26	40.97	36.12	44.48
3.5	44.39	68.73	34.85	50.00	33.69	50.94	40.01	48.18
4.0	42.24	71.13	45.26	58.33	41.83	59.47	45.85	54.44
Mean	30.38	54.96	22.26	38.72	20.08	37.62	24.08	36.70

Table 4: Accuracy gain of ALLOMFREE over the state-of-the-art L-SUE and L-OUE protocols for all datasets with $\epsilon_1 = 0.6\epsilon_\infty$, measured with the $\mathcal{U}_{L\text{-}SUE}$ and $\mathcal{U}_{L\text{-}OUE}$ metrics expressed in %.

quency estimates in comparison with the state-of-the-art L-SUE and L-OUE protocols. On average, ALLOMFREE improves the results of L-SUE at least 10% with the *MS-FIMU* dataset in Table 3 and at most 30.38% with the *Nursery* dataset in Table 4 for the privacy guarantees ϵ_∞ and ϵ_1 analyzed. Similarly, on average, ALLOMFREE improves the results of L-OUE at least 19.32% with the *MS-FIMU* dataset in Table 3 and at most 54.96% with the *Nursery* dataset in Table 4. The highest gain of accuracy was about $\sim 71\%$, achieved with the *Nursery* dataset when $\epsilon_\infty = 4$ in Table 4 in comparison with the L-OUE protocol. Finally, as one can note, with higher values of ϵ_1, ALLOMFREE will provide much higher utility than the other protocols.

6. Related work

In recent times, there have been several studies on the local DP setting in both academia [16, 33, 32, 10, 14, 20, 19, 48, 49, 18, 35, 45, 15, 50] and practical deployment [11, 12, 13, 51]. The local DP model does not rely on collecting raw data anymore, which has a clear connection with the concept of randomized response [41]. Among many other complex tasks (e.g., heavy hitter estimation [48, 37, 44], machine learning [52, 53], frequent itemset mining [42, 54]), frequency estimation is a fundamental primitive in LDP and has received considerable attention for a single attribute [15, 16, 19, 35, 14, 18, 20, 11, 12, 39, 55, 21, 22, 17].

However, most studies for collecting multidimensional data with LDP mainly focused on numerical data [49] (e.g., [32, 33, 34, 35]) or other complex tasks with categorical data (e.g., marginal estimation [27, 28, 29, 30, 31], analytical/range queries [24, 23, 25, 26]). Our ALLOMFREE solution is based on the multidimensional *Smp* solution, which randomly samples a single attribute per user only, minimizing the variance of the estimation and the communication cost. A recent study [50] proposes the Random Sampling plus Fake Data (RS+FD) solution for multidimensional data, in which the user samples a single attribute, but also generates fake data for all non-sampled attributes. The RS+FD solution creates uncertainty in the view of the aggregator while achieving similar data utility as the *Smp* solution. An interesting direction would be to extend ALLOMFREE to add fake data for non-sampled attributes too.

Besides, most academic literature on frequency estimation focuses on single data collection. To address longitudinal data collections, in [11, 12], the authors proposed LDP protocols based on two rounds of sanitization, i.e., *memoization*, which was also adopted in this paper. In the literature, some studies [39, 40] applied L-SUE (a.k.a. Basic-RAPPOR [11]) and L-OUE (i.e., OUE [14] with memoization) for longitudinal frequency estimates. However, rather than strictly using only SUE or OUE, we prove that the optimal combination is to start with OUE and then with SUE (i.e., L-OSUE). The privacy guarantees of chaining two LDP protocols has been further studied in [45, 36], which results in Eq. (16). Indeed, combining "multiple" settings (i.e., many attributes and several collections throughout time) imposes several challenges, for which this paper proposes the first solution named ALLOMFREE under LDP.

7. Conclusion

This paper investigates the problem of collecting multidimensional data throughout time for the fundamental task of frequency estimation under LDP guarantees. We extend and analyze three state-of-the-art LDP protocols, namely, GRR [18], OUE [14], and SUE [11], and propose an optimized solution, namely, ALLOMFREE, which randomly samples one attribute per user and adaptively selects a protocol with a lower variance (i.e., L-GRR or L-OSUE) in order to improve data utility. Through experimental validations, we demonstrate the advantages of ALLOMFREE over the state-of-the-art protocols L-SUE [11] and L-OUE [14] by using four real-world datasets, with the gain of accuracy on average ranging from 10% up to 55% for the analyzed range of ϵ_∞ and ϵ_1 privacy guarantees. For future work, we suggest and intend to improve the frequency estimates through post-processing techniques [56, 43] and to design LDP protocols for longitudinal and multidimensional studies considering both numerical and categorical data.

Acknowledgements

This work was supported by the EIPHI-BFC Graduate School (contract "ANR-17-EURE-0002") and by the Region of Bourgogne Franche-Comté CADRAN Project. The work of Héber H. Arcolezi was partially supported by the European Research Council (ERC) project HYPATIA under the European Union's Horizon 2020 research and innovation programme. Grant agreement n. 835294. All computations have been performed on the "Mésocentre de Calcul de Franche-Comté".

References

[1] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography, Springer Berlin Heidelberg, 2006, pp. 265–284. doi:10.1007/11681878₁4.

[2] C. Dwork, Differential privacy, in: Automata, Languages and Programming, Springer Berlin Heidelberg, 2006, pp. 1–12. doi:10.1007/11787006₁.

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

[4] A. Aktay, S. Bavadekar, G. Cossoul, J. Davis, D. Desfontaines, A. Fabrikant, E. Gabrilovich, K. Gadepalli, B. Gipson, M. Guevara, et al., Google COVID-19 community mobility reports: anonymization process description (version 1.1), arXiv preprint arXiv:2004.04145 (2020).

[5] R. Rogers, S. Subramaniam, S. Peng, D. Durfee, S. Lee, S. K. Kancha, S. Sahay, P. Ahammad, Linkedin's audience engagements API: A privacy preserving data analytics system at scale, Journal of Privacy and Confidentiality 11 (3) (2021) 1–27. doi:10.29012/jpc.782.

[6] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, CCS '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 308–318. doi:10.1145/2976749.2978318.

[7] J. M. Abowd, The U.S. census bureau adopts differential privacy, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018. doi:10.1145/3219819.3226070.

[8] S. Garfinkel, Implementing differential privacy for the 2020 census, USENIX Association, 2021.

[9] D. McCandless, T. Evans, M. Quick, E. Hollowood, C. Miles, D. Hampson, D. Geere, World's biggest data breaches & hacks. https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/, 2021 (accessed 11 march 2021).

[10] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, A. Smith, What can we learn privately?, in: 2008 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE, 2008. doi:10.1109/focs.2008.27.

[11] U. Erlingsson, V. Pihur, A. Korolova, RAPPOR: Randomized aggregatable privacy-preserving ordinal response, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, ACM, New York, NY, USA, 2014, pp. 1054–1067. doi:10.1145/2660267.2660348.

[12] B. Ding, J. Kulkarni, S. Yekhanin, Collecting telemetry data privately, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 3571–3580.

[13] Apple Differential Privacy Team, Learning with privacy at scale. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf, 2017 (accessed 11 march 2021).

[14] T. Wang, J. Blocki, N. Li, S. Jha, Locally differentially private protocols for frequency estimation, in: 26th USENIX Security Symposium (USENIX Security 17), USENIX Association, Vancouver, BC, 2017, pp. 729–745.

[15] G. Cormode, S. Maddock, C. Maple, Frequency estimation under local differential privacy, Proceedings of the VLDB Endowment 14 (11) (2021) 2046–2058. doi:10.14778/3476249.3476261.

[16] T. Murakami, Y. Kawamoto, Utility-Optimized local differential privacy mechanisms for distribution estimation, in: 28th USENIX Security Symposium (USENIX Security 19), USENIX Association, Santa Clara, CA, 2019, pp. 1877–1894.

[17] S. Wang, Y. Nie, P. Wang, H. Xu, W. Yang, L. Huang, Local private ordinal data distribution estimation, in: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, IEEE, 2017. doi:10.1109/infocom.2017.8056977.

[18] P. Kairouz, K. Bonawitz, D. Ramage, Discrete distribution estimation under local privacy, in: International Conference on Machine Learning, PMLR, 2016, pp. 2436–2444.

[19] J. Acharya, Z. Sun, H. Zhang, Hadamard response: Estimating distributions privately, efficiently, and with little communication, in: K. Chaudhuri, M. Sugiyama (Eds.), Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, Vol. 89 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 1120–1129.

[20] M. Alvim, K. Chatzikokolakis, C. Palamidessi, A. Pazii, Invited paper: Local differential privacy on metric spaces: Optimizing the trade-off with utility, in: 2018 IEEE 31st Computer Security Foundations Symposium (CSF), IEEE, 2018. doi:10.1109/csf.2018.00026.

[21] D. Zhao, H. Chen, S. Zhao, X. Zhang, C. Li, R. Liu, Local differential privacy with k-anonymous for frequency estimation, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019. doi:10.1109/bigdata47090.2019.9006022.

[22] Z. Li, T. Wang, M. Lopuhaä-Zwakenberg, N. Li, B. Škoric, Estimating numerical distributions under local differential privacy, in: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, ACM, 2020. doi:10.1145/3318464.3389700.

[23] M. Xu, B. Ding, T. Wang, J. Zhou, Collecting and analyzing data jointly from multiple services under local differential privacy, Proceedings of the VLDB Endowment 13 (12) (2020) 2760–2772. doi:10.14778/3407790.3407859.

[24] J. Yang, T. Wang, N. Li, X. Cheng, S. Su, Answering multi-dimensional range queries under local differential privacy, Proc. VLDB Endow. 14 (3) (2020) 378–390. doi:10.14778/3430915.3430927.

[25] X. Gu, M. Li, Y. Cao, L. Xiong, Supporting both range queries and frequency estimation with local differential privacy, in: 2019 IEEE Conference on Communications and Network Security (CNS), IEEE, 2019. doi:10.1109/cns.2019.8802778.

[26] G. Cormode, T. Kulkarni, D. Srivastava, Answering range queries under local differential privacy, Proceedings of the VLDB Endowment 12 (10) (2019) 1126–1138. doi:10.14778/3339490.3339496.

[27] Z. Shen, Z. Xia, P. Yu, PLDP: Personalized local differential privacy for multidimensional data aggregation, Security and Communication Networks 2021 (2021) 1–13. doi:10.1155/2021/6684179.

[28] F. Peng, S. Tang, B. Zhao, Y. Liu, A privacy-preserving data aggregation of mobile crowdsensing based on local differential privacy, in: Proceedings of the ACM Turing Celebration Conference - China, ACM, 2019. doi:10.1145/3321408.3321602.

[29] Z. Zhang, T. Wang, N. Li, S. He, J. Chen, CALM: Consistent adaptive local marginal for marginal release under

local differential privacy, Proceedings of the ACM Conference on Computer and Communications Security (2018) 212–229doi:10.1145/3243734.3243742.

[30] X. Ren, C.-m. Yu, W. Yu, S. Yang, S. Member, X. Yang, J. A. Mccann, P. S. Yu, L. Fellow, LoPub : High-Dimensional Crowdsourced Data 13 (9) (2018) 2151–2166. doi:10.1109/TIFS.2018.2812146.

[31] G. Fanti, V. Pihur, Ú. Erlingsson, Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries, Proceedings on Privacy Enhancing Technologies 2016 (3) (2016) 41–61. doi:10.1515/popets-2016-0015.

[32] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, J. Shin, Collecting and analyzing data from smart device users with local differential privacy, arXiv preprint arXiv:1606.05053 (2016).

[33] N. Wang, X. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, G. Yu, Collecting and analyzing multidimensional data with local differential privacy, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019. doi:10.1109/icde.2019.00063.

[34] J. C. Duchi, M. I. Jordan, M. J. Wainwright, Minimax optimal procedures for locally private estimation, Journal of the American Statistical Association 113 (521) (2018) 182–201. doi:10.1080/01621459.2017.1389735.

[35] T. Wang, J. Zhao, Z. Hu, X. Yang, X. Ren, K.-Y. Lam, Local differential privacy for data collection and analysis, Neurocomputing 426 (2021) 114–133. doi:10.1016/j.neucom.2020.09.073.

[36] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, S. Song, K. Talwar, A. Thakurta, Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation, arXiv preprint arXiv:2001.03618 (2020).

[37] T. Wang, N. Li, S. Jha, Locally differentially private heavy hitter identification, IEEE Transactions on Dependable and Secure Computing 18 (2) (2021) 982–993. doi:10.1109/tdsc.2019.2927695.

[38] H. H. Arcolezi, J.-F. Couchot, B. A. Bouna, X. Xiao, Longitudinal collection and analysis of mobile phone data with local differential privacy, in: M. Friedewald, S. Schiffner, S. Krenn (Eds.), Privacy and Identity Management, Springer International Publishing, Cham, 2021, pp. 40–57. doi:10.1007/978-3-030-72465-8$_3$.

[39] J. W. Kim, D.-H. Kim, B. Jang, Application of local differential privacy to collection of indoor positioning data, IEEE Access 6 (2018) 4276–4286. doi:10.1109/access.2018.2791588.

[40] I. D. C. Vidal, A. L. da Costa Mendonça, F. Rousseau, J. D. C. Machado, ProTECting: An application of local differential privacy for IoT at the edge in smart home scenarios, in: Anais XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2020), Sociedade Brasileira de Computação, 2020. doi:10.5753/sbrc.2020.12308.

[41] S. L. Warner, Randomized response: A survey technique for eliminating evasive answer bias, Journal of the American Statistical Association 60 (309) (1965) 63–69. doi:10.1080/01621459.1965.10480775.

[42] T. Wang, N. Li, S. Jha, Locally differentially private frequent itemset mining, in: 2018 IEEE Symposium on Security and Privacy (SP), IEEE, 2018. doi:10.1109/sp.2018.00035.

[43] T. Wang, M. Lopuhaa-Zwakenberg, Z. Li, B. Skoric, N. Li, Locally differentially private frequency estimation with consistency, in: Proceedings 2020 Network and Distributed System Security Symposium, Internet Society, 2020. doi:10.14722/ndss.2020.24157.

[44] R. Bassily, K. Nissim, U. Stemmer, A. Thakurta, Practical locally private heavy hitters, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 2285–2293.

[45] M. Naor, N. Vexler, Can Two Walk Together: Privacy Enhancing Methods and Preventing Tracking of Users, in: A. Roth (Ed.), 1st Symposium on Foundations of Responsible Computing (FORC 2020), Vol. 156 of Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2020, pp. 4:1–4:20. doi:10.4230/LIPIcs.FORC.2020.4.

[46] D. Dua, C. Graff, UCI machine learning repository. http://archive.ics.uci.edu/ml, 2017 (accessed 11 march 2021).

[47] H. H. Arcolezi, J.-F. Couchot, O. Baala, J.-M. Contet, B. A. Bouna, X. Xiao, Mobility modeling through mobile data: generating an optimized and open dataset respecting privacy, in: 2020 International Wireless Communications and Mobile Computing (IWCMC), IEEE, 2020. doi:10.1109/iwcmc48107.2020.9148138.

[48] R. Bassily, A. Smith, Local, private, efficient protocols for succinct histograms, in: Proceedings of the forty-seventh annual ACM symposium on Theory of Computing, ACM, 2015. doi:10.1145/2746539.2746632.

[49] X. Xiong, S. Liu, D. Li, Z. Cai, X. Niu, A comprehensive survey on local differential privacy, Security and Communication Networks 2020 (2020) 1–29. doi:10.1155/2020/8829523.

[50] H. H. Arcolezi, J.-F. Couchot, B. Al Bouna, X. Xiao, Random sampling plus fake data: Multidimensional frequency estimates with local differential privacy, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, ACM, 2021, pp. 47–57. doi:10.1145/3459637.3482467.

[51] S. Kessler, J. Hoff, J.-C. Freytag, SAP HANA goes private, Proceedings of the VLDB Endowment 12 (12) (2019) 1998–2009. doi:10.14778/3352063.3352119.

[52] P. C. Mahawaga Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, M. Atiquzzaman, Local differential privacy for deep learning, IEEE Internet of Things Journal 7 (7) (2020) 5827–5842. doi:10.1109/JIOT.2019.2952146.

[53] X. Zhou, J. Tan, Local differential privacy for bayesian optimization, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 11152–11159.

[54] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, K. Ren, Heavy hitter estimation over set-valued data with local differential privacy, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2016. doi:10.1145/2976749.2978409.

[55] H. H. Arcolezi, J.-F. Couchot, S. Cerna, C. Guyeux, G. Royer, B. A. Bouna, X. Xiao, Forecasting the number of firefighter interventions per region with local-differential-privacy-based data, Computers & Security 96 (2020) 101888. doi:10.1016/j.cose.2020.101888.

[56] E. ElSalamouny, C. Palamidessi, Generalized iterative bayesian update and applications to mechanisms for privacy protection, in: 2020 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2020. doi:10.1109/eurosp48549.2020.00038.