Front.Comput.Sci.
DOI

REVIEW ARTICLE

Locally Differentially Private Frequency Distribution Estimation with Relative Error Optimization

Ning Wang [1], Yifei Liu[1], Zhigang Wang(✉)[1], Zhiqiang Wei[1], Ruichun Tang[1], Peng Tang[2], Ge Yu[3]

1 Ocean University of China, Qingdao 266404, China
2 Shandong University, Qingdao 266237, China
3 Northeastern University, Shenyang 110819, China

Abstract Collecting crowdsourced multidimensional data has numerous real-world applications in crowd sensing systems, while estimating frequency distribution among dimension values, is the most important underlying requirement. However, values from users usually involve personal information, and hence the data collection inevitably raises privacy concerns. Local differential privacy (LDP) has been established as a strong privacy standard for safety data collection. Existing works for LDP-compliant frequency distribution estimation focus on improving result utility by optimizing the traditional absolute error metric, in which the noise scale injected into frequencies is independent of the true values. That makes the frequencies with smaller values dominated by noise and leads to inferior data utility. Considering that different frequencies have different abilities of resisting noise, we employ the relative error metric as the optimization goal and propose a novel LDP-compliant iterative framework *iterUA* to gradually refine the estimation. Its core technique is an adaptive user allocation approach, which allocates more users to the dimensions with smaller frequencies, so as to reduce the scale of noise added. Experiments over real datasets confirm the effectiveness of our methods.

Keywords frequency distribution, local differential privacy, relative error

Received month dd, yyyy; accepted month dd, yyyy

E-mail: Ning Wang, wangning8687@ouc.edu.cn; Yifei Liu, liuyifei@stu.ouc.edu.cn; Zhigang Wang, wangzhigang@ouc.edu.cn; Zhiqiang Wei, weizhiqiang@ouc.edu.cn; Ruichun Tang, tangruichun@ouc.edu.cn; Peng Tang, tangpeng@sdu.edu.cn; Ge Yu, yuge@mail.neu.edu.cn

1 Introduction

Huge amounts of private and sensitive personal data in the cyber world are being collected via terminal devices involved in crowd sensing systems. The valuable data enables us to model the product logistics, analyze the preference of subscribers and improve the productivity of the healthcare industry workers. Such applications are very crucial and worth billion dollars in the business. However, directly collecting the true data from users will pose serious risks on personal privacy. Therefore, developing effective approaches to collect data with privacy guarantee becomes an urgent need in the crowd sensing system [**?**, **?**, **?**]. A promising methodology for collecting sensitive data without violating users' privacy is *local differential privacy* [**?**, **?**, **?**, **?**], which has come to be the *de facto* standard for individual privacy protection. In the LDP setting, random noise is locally injected into the data values at each individual user and then only the perturbed version is transmitted to the curator. Thus, users do not need to rely on the trustworthiness of the curator. This desirable feature of LDP has led to wide deployment in popular software systems such as Apple iOS [**?**], macOS [**?**], Microsoft Windows Insiders [**?**] and Google Chrome browser [**?**].

In this paper, we tackle the problem of estimating the frequency distribution on the crowdsourced multidimensional categorical data under LDP. For example, given two attributes *Sex* and *Race*, the curator aims to collect data from users participating in crowd sensing to derive the frequency distribution on *Sex* with value domain {*Male*, *Female*} and *Race* with

{*White*, *Latino*, *African*, *Native Americans*, *Asian*, *others*}. A straightforward solution is to use the composition property of differential privacy and apply existing LDP protocols tailored for single-dimension to collect data for each dimension. However, that incurs much more noise since dimensionality can be large. Other pioneering works [**?**, **?**, **?**] evenly divide users into disjoint groups among dimensions, so as to collect a single dimension data from each group. This design reduces the risk of posing privacy and hence noise added. In fact, the noise scale on each frequency is proportional to $\frac{1}{\sqrt{x}}$, where x is the group size. As shown in Fig. **??**(a), by adding the same scale of noise into the frequencies on *Sex* and *Race*, the frequency distribution estimation is ({55%, 45%} and {53%, 13%, 18%, 10%, 1%, 6%}, where the red line indicates the true distribution {51%, 49%} and {57%, 18%, 13%, 6%, 5%, 1%}. Clearly, the noisy estimation on *Sex* can reflect the ground truth value more accurately, compared with *Race*. The reason is the ground truth distribution on *Sex* is relatively uniform which is more resilient to the noise compared with the heavily skewed distribution on *Race*. In order to improve the overall result utility on all dimensions, we would like to add less noise to the distribution with smaller frequencies, and vice versa. Fig. **??**(b) shows such an example, i.e., {57%, 43%} on *Sex* and {54%, 21%, 10%, 8%, 3%, 3%} on *Race*). Note that following the traditional metric *mean absolute error* given in Def. **??** in Section **??**, the two examples shown in Fig. **??** have the same performance with 0.0425, although Fig. **??**(b) is actually better. The false positively high error in Fig. **??**(b) is because big noise added into dimensions with high frequencies strikes a compromise with small noise added into dimensions with low frequencies. Further, the intrinsical reason of such a compromise is that the metric uses the absolute difference between noisy and ground truth values. We thereby adopt a more reasonable metric *relative error* measuring the relative difference between two values to clearly distinguish the two scenarios, which is given in Def. **??** in Section **??**. Accordingly, in this paper, the goal of LDP mechanism design, is to minimize the *relative error* of frequency distribution estimation computed from the collected noisy data, while satisfying LDP.

Although the frequency estimation problem under LDP [**?**, **?**, **?**, **?**] has attracted a lot of attention in recent years, currently, to our knowledge, the existing works are all devoted to optimizing the absolute error, rather than relative error. Unlike the work for the former, the one targeting at the latter should take the true frequency distribution into consider and design true frequency distribution-sensitive data collection protocol so that the skewed distribution, which involves s-

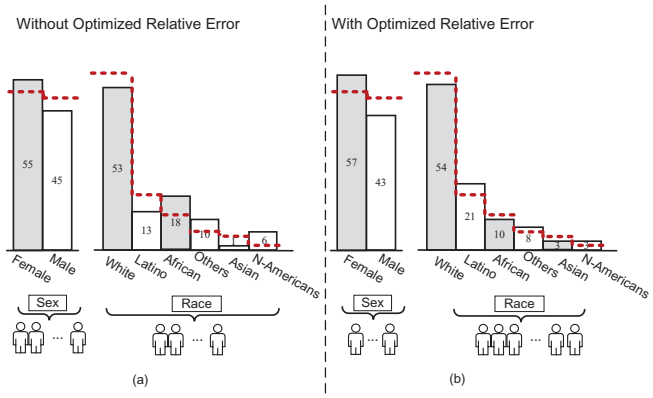


Fig. 1 An example for illustrating the motivation

maller frequency in high probability, is with less noise. However, it is challenging to fulfill such a requirement, since the true frequency distribution is private information and not available. An intuitive method to crack this nut is that we can leave the relative error behind and use the existing LDP methods to derive a rough version of estimation of the true frequency distribution based on a part of users. Following that, the rough version is regarded as a priori knowledge to guide the data collection strategy made with the relative error optimization for the remaining users. Such a method usually allocates a small part of users to derive the priori estimation in order to leave more optimization space in the relative error optimization phase. However, since the result accuracy in LDP is proportional to the number of users contributing data for collection, a small number of users make the priori estimation deviate the true one severely, and hence, optimization in the following phase does not make sense.

Since collecting data from more users contributes to improving the result accuracy under the LDP setting, the relative error optimization for frequency estimation on multidimensional data can be solved by assigning more users to dimensions with skewed frequency distribution, and fewer users to dimensions with uniform distribution. To make the user assignment adaptive to the true distribution, this paper proposes *iterUA*, a novel iterative LDP framework to construct a user allocation strategy between users and dimensions. Specifically, *iterUA* firstly costs a part of users to learn a rough version of frequency distribution without relative error optimization under LDP, in which users are allocated among dimensions. And then it refines the rough version iteratively by using a batch of users. In the process of iteration, based on the previous rough version, *iterUA* makes the user allocation strategy with the goal of minimizing relative error and collects data under the strategy to calibrate the noise in the rough version.

In this way, the priori estimation is refined steadily, which contributes to deriving the optimal allocation strategy for the next iteration. Undoubtedly, the iterative refining operation can significantly alleviate the negative impacts brought by the inaccurate priori estimation in the first several iterations. Besides, considering the output of each iteration is a combined version of the noisy frequency estimation generated in the current iteration and the ones from all previous iterations, we optimize the user allocation strategy used *iterUA* with the goal of minimizing the relative error of the combined version.

In summary, we make the following contributions in this paper:

- We construct the relationship between user allocation for dimensions and relative error of frequency distribution estimation under LDP, based on which we design the user allocation strategy.
- We propose *iterUA*, the first LDP framework for the relative error optimization of frequency distribution estimation on multidimensional data, which iteratively constructs user allocation strategy between users and dimensions. To guarantee LDP while reduce the relative error, *iterUA* uses the frequency estimation generated from the previous iteration steps to guide the user allocation in the current step.
- Based on the iterative feature of *iterUA*, we optimize the allocation strategy by considering the relative error for the estimation accumulated from all performed iterations, rather than the one only from the current iteration, which further boosts the performance of *iterUA*.
- We conduct extensive experiments to evaluate the performance of different approaches using real datasets. The results validate the effectiveness of our proposals.

In the following, Section **??** reviews related work. Section **??** provides the necessary background on LDP and problem definition. Section **??** describes our user allocation strategy with relative error optimization. Section **??** elaborates the LDP-compliant iterative framework for the frequency distribution estimation with reduced relative error. Section **??** contains an extensive set of experiments. Finally, Section **??** concludes this paper.

2 Related Work

Differential privacy [**?**, **?**], as a rigorous privacy protection model, has attracted a lot of attention, since it can provide theoretical privacy guarantee against adversaries with arbitrary background information. Many representative efforts under differential privacy have been devoted to publishing the frequency distribution on the multidimensional data which is one fundamental component involved in some complex analysis tasks. The existing works can be classified into two categories, one is under centralized differential privacy (CDP) and the other is under local differential privacy (LDP). We elaborate them and then distinguish our work from them as follows.

Frequency distribution publication under CDP. Different from the setting in LDP, CDP assumes that a data curator collects unfettered data from all users and aims to publish a noisy version of frequency distribution that preserves privacy. A straightforward way [**?**, **?**] to guarantee ϵ-CDP is to add the noise following Laplace distribution with parameters $u = 0$ and $b = \frac{d}{\epsilon}$ to the frequency distribution, where d is the dimensionality and ϵ is the privacy budget controlling the strength of privacy protection. Since the noise scale added to each frequency is independent of the true frequency, the above method focuses on absolute error, rather than relative error. Considering the latter is a more reasonable measurement in real applications, Xiao et al. [**?**] propose *iReduct* to optimize the relative error. In particular, they firstly publish a rough version of frequency distribution with a small part of ϵ, then iteratively refine the rough version by picking up some dimensions with higher relative error and re-publishing the frequency distribution on these dimensions with some budgets. Although *iReduct* provides the estimation with reduced relative error, it cannot be employed to solve the problem under LDP. The main reason is that it is not an effective way in the LDP setting to split privacy budget into multiple parts, each of which is used to publish a part of information from one user. So the iterative framework of *iReduct* involving privacy budget split can not work well in the LDP setting. Besides, *iReduct* has another limitation that the same privacy budget has to be paid to refine the frequency estimation for each picking up dimension. However, it is undoubtedly that even if the dimensions are all picked up as the ones with high relative error, privacy budget needs to be allocated based the value of relative error of each dimension. To solve the above problem, this paper proposes a framework with relative error optimization to iteratively construct user allocation strategy for each dimension, in which one user just publishes her value one time with all privacy budget and the allocation strategy is adaptive to the change of relative error. In addition, a number of recent works study the synthesis of high-dimensional datasets with CDP. They first-

ly choose some low-dimensional marginals either based on one optimization problem [**?**] or based on Bayes network [**?**] and Markov Random Fields [**?**], then generate the synthetic datasets based on the chosen marginals. With the synthetic datasets, we can deal with any analysis task without posing privacy, including the frequency distribution estimation. However, none of these works takes the relative error optimization into consider, leading to inferior utility.

Frequency distribution publication under LDP. Many representative efforts have been devoted to designing LDP-based protocols for frequency estimation. OUE [**?**], OLH [**?**] and Random Matrix Projection [**?**], which rely on techniques like hashing and Hadamard transform for good utility, are proposed to collect information from users to derive the frequency distribution estimation on one dimension. They are often used as the basis components of some complex LDP algorithms. Besides, for the multidimensional problem under LDP, LoPub [**?**] is proposed to synthesize an LDP-compliant multidimensional dataset. To randomize the transformed user information, a bit string, LoPub splits the privacy budget into several parts, each of which is used to randomize one bit. However, the following works [**?**, **?**, **?**, **?**] validate that the privacy budget split is not an effective way for multidimensional data publication. They adopt PAD [**?**, **?**, **?**] to tackle this problem, which splits users into d disjoint groups and uses a group of users to compute the frequency distribution on one dimension. PrivTrie [**?**] and PEM [**?**] also adopt the idea that one user is just involved in information collection one time. In addition, Zhang et al. [**?**] propose the technique Calm to publish any k-way marginal under LDP, which privately publishes m l-way marginals as a synopsis for k-way marginals. Although the above methods can be used to publish the frequency distribution estimation on multidimensional data, they all focus on the absolute error optimization. Besides, there are also some works [**?**, **?**] for optimizing the result accuracy measured by absolute error for the range queries on the multidimensional data, which are orthogonal to our work.

3 Preliminaries

Section 3.1 provides preliminaries on local differential privacy (LDP), as well as the frequently used LDP protocol for collecting data. And Section 3.2 describes the problem that this paper focuses on.

3.1 Local Differential Privacy

In the problem setting, an untrusted curator collects data from a number of (say, n) individual users, each of which possesses a data record with sensitive information, and then computes statistics based on the collected data. To protect private information from being posed, each user needs to perturb her own data with the guarantee of LDP model before sending. The LDP model is defined in the following.

Definition 1. *(ϵ-Local Differential Privacy). A randomized algorithm f satisfies ϵ-local differential privacy, if and only if for any output $\tilde{y}$ of f and any two input tuples y and y', we have*

$$\Pr[f(y) = \tilde{y}] \leq e^{\epsilon} \cdot \Pr[f(y') = \tilde{y}] \quad (1)$$

In the above definition, ϵ is called the *privacy budget*, which controls the strength of privacy protection. A smaller ϵ leads to stricter privacy protection, and vice versa.

An important property of LDP is the *parallel composition* rule, which can be used to analyze the privacy guarantee of the complex task consisting of multiple LDP queries. The following lemma shows the rule.

Lemma 1. *(Parallel Composition [?]). Given a randomized algorithm $\mathcal{A}$ consisting of sub-procedures $\{\mathcal{A}_1, ..., \mathcal{A}_i, ..., \mathcal{A}_m\}$, if every sub-procedure $\mathcal{A}_i$ applies one ϵ-LDP mechanism to the disjoint users, then $\mathcal{A}$ satisfies ϵ-LDP.*

Besides, any post-processing operation on the output of a publication algorithm under ϵ-LDP does not destroy privacy further, i.e., the publication algorithm with the post-processing operation satisfies ϵ-LDP.

In the LDP model, each user needs to use LDP protocols to perturb her data under the constraint described in Eqn. **??**. Due to the high accuracy of the perturbed result and succinct perturbation procedure, OUE [**?**], as the representative LDP protocol, has been adopted frequently as the perturbation component. We introduce it in the following.

LDP Protocol-OUE. Suppose that there are n individual users and each user has one value from domain $D = \{1, 2, ..., d\}$. And the curator aims to estimate the frequency distribution on D. In particular, on the user side, each user u_i encodes her data v_i as one vector $\mathbf{y}_i$ with size d, in which just the $(v_i - 1)$th bit is set as 1 and others are set as 0. Then she perturbs the bit in $\mathbf{y}_i$ one by one to get a noisy version $\mathbf{y}_i^*$ of $\mathbf{y}_i$. Specifically, for the $(j + 1)$th bit, if $\mathbf{y}_i[j]$ is equal to 1, u_i perturbs $\mathbf{y}_i[j]$ as 1 with probability $p = 0.5$ and 0 with probability $1 - p$. If $\mathbf{y}_i[j]$ is 0, u_i perturbs $\mathbf{y}_i[j]$ as 1 with

probability $q = \frac{1}{e^\epsilon+1}$ and 0 with probability $1-q$. Once the curator receives all perturbed data from users, she computes the final statistical result $\frac{\sum_{i=1}^{n} \mathbf{y}_i^*[j]-n\cdot q}{n(p-q)}$ as the frequency estimation of value $j+1$ ($0 \le j \le d-1$), whose absolute error scale is $\frac{\sqrt{q(1-q)}}{\sqrt{n}(p-q)}$.

Algorithm 1 Optimized Unary Encoding

Input: $y \in \{0,1\}$, privacy budget ϵ.
Output: $\tilde{y} \in \{0,1\}$.

```
p1=0.5, p2 = 1/(e^ε+1);
if y = 1 then
    p = p1;
else
    p = p2;
end if
Sample a Bernoulli variable ỹ_i that equals 1 with p probability;
return ỹ_i.
```

3.2 Problem Definition

This paper focuses on the classical problem of collecting multidimensional categorical data involving sensitive information to derive frequency distribution estimation. Specifically, there are n individual users. Each user u_i's private data is represented by a tuple t_i, which contains d categorical attributes $A_1, ..., A_d$. And t_{ij} denotes the value of A_j. Let C_j indicate the domain of $A_j (1 \le j \le d)$, i.e., $t_{ij} \in C_j$ and C_{jm} denote the mth value in C_j. Without loss of generality, we assume attribute A_j with $|C_j|$ distinct values has a discrete domain $\{1, 2, ..., |C_j|\}$. Then based on the users' data, we can use $F_{jm} = \frac{\sum_{i=1}^{n} \mathbf{1}(t_{ij}=m)}{n}$ to compute the frequency that value m on attribute A_j appears in the users' data, where $\mathbf{1}()$ is an indicator function. In this way, the frequency distribution on the multidimensional data can be derived by computing the frequency for any value $m \in C_j$ on any attribute $A_j (1 \le j \le d)$.

Since the data on the user side is sensitive in our problem setting, we design the LDP mechanisms to collect the data from users. The goal of our design is to maximize the utility of frequency distribution estimation computed from the collected noisy data, while satisfying LDP. We measure the utility of derived estimation by *relative error*, which is defined as below. A smaller relative error means a better result utility.

Definition 2. *(Relative Error). Let F be the true frequency distribution and F^* be the noisy frequency distribution estimation, where F_{jm} (F^*_{jm}) be the true (noisy) frequency of value m on attribute A_j. If F^* is regarded as the estimation of F, the relative error with sanity bound δ is*

$$\frac{1}{d}\sum_{j=1}^{d}\frac{1}{|C_i|}\sum_{m=1}^{|C_i|}\frac{|F^*_{jm}-F_{jm}|}{\max(F_{jm},\delta)}. \quad (2)$$

where δ is a user-specified constant and can mitigate the effect of extremely small frequency on overall relative error. For the noisy distribution given in Fig. **??**(a), we can compute its relative error with $\delta = 0$ as follows: $\frac{|51-55|/51+|49-45|/49}{2\times2} + \frac{|57-53|/57}{2\times6} + \frac{|18-13|/18+|13-18|/13+|6-10|/6+|5-1|/5+|1-6|/1}{2\times6} = 0.64$. The relative error for the distribution in Fig. **??**(b) is 0.33, which can be computed in a similar way.

For comparison with *relative error*, we also show the definition of *absolute error* as follows.

Definition 3. *(Absolute Error). Let F be the true frequency distribution and F^* be the noisy frequency distribution estimation, where F_{jm} (F^*_{jm}) be the true (noisy) frequency of value m on attribute A_j. If F^* is regarded as the estimation of F, the absolute error is*

$$\frac{1}{d}\sum_{j=1}^{d}\frac{1}{|C_i|}\sum_{m=1}^{|C_i|}|F^*_{jm}-F_{jm}|. \quad (3)$$

For the noisy distribution given in Fig. **??**(a), we can compute its absolute error as follows: $\frac{|51-55|+|49-45|}{2\times2\times100} + \frac{|57-53|}{2\times6\times100} + \frac{|18-13|+|13-18|}{2\times6\times100} + \frac{|6-10|+|5-1|+|1-6|}{2\times6\times100} = 0.0425$. The absolute error for the distribution in Fig. **??**(b) is also 0.0425, which can be computed in a similar way.

4 User Allocation Strategy

In the LDP setting, to improve the accuracy of frequency estimation on multidimensional data, a widespread adopted way is to let a user involve the frequency computation for only one dimension or attribute, which incurs that the absolute error in each frequency is inversely proportional to x, where x is the number of users fed up to one dimension. As a result of that, allocating more users into the dimensions with small frequencies contributes to reducing the relative error. However, the available users for all dimensions are finite. So it is important to make user allocation strategy which splits the available users into d disjoint groups and each is fed up to one dimension, with the goal of minimizing the overall relative error.

Since the relative error is sensitive to the true value, our basic idea for relative error optimization is depending on the priori frequency distribution of true one to make the user allocation strategy, so that the dimensions with extremely small true

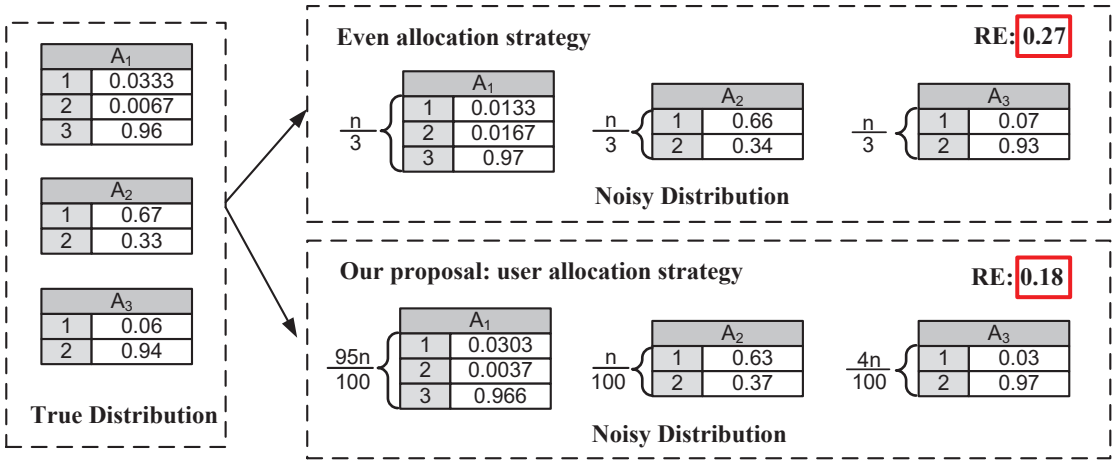


Fig. 2 An example for illustrating the improvement on result utility brought by UAS.

frequencies have more users, and hence less noise is added. In the following, to elaborate the key idea of our proposal *User Allocation Strategy* (UAS), we take the true frequency distribution as the priori knowledge temporarily, which incurs privacy leakage risk. Later, Section **??** will show one way to learn the priori distribution under LDP.

Now we describe the user allocation problem for relative error optimization in UAS as an optimization problem in the following. Given d attributes $\{A_1, A_2, ..., A_d\}$ and true frequency distribution F on all the attributes, we aim to split n users into d disjoint groups and the users in group g_i ($1 \le i \le d$) are involved in computing the frequency estimation F_i^* on A_i, so that the relative error of F^* is minimized. Suppose the LDP protocol OUE is adopted to collect the data from users, then the absolute error scale in frequency for the value in A_i is $\sqrt{\frac{q(1-q)}{|g_i|(p-q)^2}}$, where p and q are the parameters of OUE. Then the expected overall relative error on all attributes can be written as :

$$E(g_1, g_2, ..., g_d) = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{|A_i|}\sum_{j=1}^{|A_i|}\frac{\sqrt{q(1-q)}}{\sqrt{|g_i|(p-q)^2}\max(\delta, F_{ij})}.$$

Based on the above equation, we have

$$E(g_1, g_2, ..., g_d) \propto \sum_{i=1}^{d}\frac{1}{|A_i|}\sum_{j=1}^{|A_i|}\frac{1}{\sqrt{|g_i|}\max(\delta, F_{ij})}.$$

We aim to derive the number of users in each group g_1, g_2, ..., g_d, such that

$$\sum_{i=1}^{d}\frac{1}{|A_i|}\sum_{j=1}^{|A_i|}\frac{1}{\sqrt{|g_i|}\max(\delta, F_{ij})}$$

is minimized subject to the constraint that $\sum_{i=1}^{d}|g_i| = n$.

We can use the Lagrange multiplier method to derive the values of $|g_1|$, $|g_2|$, ..., and $|g_d|$, where $|g_i|$ is equal to

$$|g_i| = n \cdot \frac{\left(1/|A_i|\sum_{j=1}^{|A_i|}\frac{1}{\max(\delta, F_{ij})}\right)^{2/3}}{\sum_{i=1}^{d}\left(1/|A_i|\sum_{j=1}^{|A_i|}\frac{1}{\max(\delta, F_{ij})}\right)^{2/3}} \quad (4)$$

Then we can choose $|g_i|$ users from the user set randomly, which are fed up to the attribute A_i. Fig. **??** presents an example of our proposal UAS on three attributes. It is observed that attribute A_1 is with a skewed frequency distribution, in which only 3.33% of users possess value "1". Based on the true distribution on all dimensions including A_1, A_2 and A_3, we invoke UAS to derive the groups with $|g_1| = \frac{95n}{100}$, $|g_2| = \frac{n}{100}$ and $|g_3| = \frac{4n}{100}$. Obviously, more users are allocated to Attribute A_1, which leads to reduced relative error. With this example, compared with the even allocation strategy that allocating the same number of users to participate the frequency computation of each dimension, UAS can reduce the relative error from 0.27 to 0.18.

Although UAS performs well on boosting the accuracy on the relative error, it is true data-sensitive method due to the allocation depending on the true frequency, which incurs privacy leakage. In the next section, we will explore the technique to derive the LDP-compliant distribution estimation with high accuracy, which is involved in making the user allocation strategy.

5 Iterative Framework for Relative Error Optimization

This section proposes an LDP-compliant iterative framework *iterUA* to construct the user allocation strategy between dimensions and users with the objective of optimizing the relative error of derived frequency distribution estimation. In the following, Section 5.1 presents the simple yet effective framework, *iterUA*, which refines the noisy estimation of frequency distribution and allocates the users to dimensions iteratively. Section 5.2 further optimizes UAS by considering the feature of *iterUA*.

5.1 The Framework of IterUA

IterUA includes two phases: one is to use αn users to derive a rough estimation F^* of true frequency distribution, and the other one is to combine the UAS module to refine F^* by collecting information from the remaining $(1-\alpha)\cdot n$ users. In the following, we elaborate the two phases in detail.

In the first phase, the key idea is to generate a noisy estimation F^* of F as the input of the UAS module so that the allocation strategy from UAS is not sensitive to the true frequency. In particular, as shown in Algorithm **??**, αn users are split into d disjoint parts, in which each part is associated with one dimension and has the same number of users. Then the frequency distribution estimation F_i^* on the ith dimension can be derived by invoking the basic LDP protocol OUE to collect information from the associated group of users.

In the second phase, we use the UAS module to allocate the users to each dimension dynamically and iteratively by optimizing the relative error of F^*. In particular, considering that the first phase aims to derive a priori estimation of F, rather than optimize the relative error, we should set the parameter α with a small number so that more users can be left to the second phase for reducing the relative error. However, that leads to large error in F^*, since the absolute error of F^* is inversely proportional to the number of users contributing information. As a result, with the rough estimation F^* from the first phase, the UAS module usually generates a user allocation strategy mis-calibrating the relative error. It is a fact that, on the one hand, we expect to use more users in the first phase to make F^* as accurate as possible; on the other hand, we aim to leave more users to optimize the relative error in the second phase.

Algorithm 2 Iterative User Allocation Algorithm

Input: the number of users n, the ratio of users used in the first and second phases α, privacy budget ϵ, the number of iterations τ, sanity bound δ.
Output: noisy frequency distribution estimation F^*.

1: Split n users into two disjoint parts, U_a and U_b, where $|U_a|$ is αn;
/* Phase 1 */
2: Split the users in U_a into d disjoint groups with the same size;
3: **for** each Attribute $A_i \in \{A_1, A_2, ...A_d\}$ **do**
4: Apply OUE to collect information from the users in the ith group to derive the frequency distribution estimation F_i^* on A_i using privacy budget ϵ;
5: **end for**
/* Phase 2 */
6: $\omega_1^* = \frac{dq(1-q)}{|U_a|(p-q)^2}$, $\omega_2^* = \frac{dq(1-q)}{|U_a|(p-q)^2}$, ..., $\omega_d^* = \frac{dq(1-q)}{|U_a|(p-q)^2}$;
7: $\omega^\Delta = <0, 0, ..., 0>$, $b = \frac{|U_b|}{\tau}$;
8: **for** $t \in [1, \tau]$ **do**
9: Extract a batch of users U_Δ with size b from U_b;
10: $g_1, g_2, ..., g_d$ = UAS(F^*, δ, U_Δ);
11: **for** Attribute $A_i \in \{A_1, A_2, ...A_d\}$ **do**
12: Apply OUE to collect information from the users in g_i to derive the frequency distribution estimation F_i^Δ on A_i using privacy budget ϵ;
13: $\omega_i^\Delta = \frac{q(1-q)}{|g_i|(p-q)^2}$;
14: $F_i^* = \frac{\omega_i^\Delta}{\omega_i^* + \omega_i^\Delta} F_i^* + \frac{\omega_i^*}{\omega_i^* + \omega_i^\Delta} F_i^\Delta$;
15: $\omega_i^* = \frac{\omega_i^* \cdot \omega_i^\Delta}{\omega_i^* + \omega_i^\Delta}$;
16: **end for**
17: Remove the users in U_Δ from U_b;
18: **end for**
19: **return** F^*.

To strive a balance between the above two aspects, *iterUA* refines F^* with a relative error-optimized frequency distribution estimation F^Δ from a batch of users iteratively and progressively. In particular, as shown in Algorithm **??**, in each iteration of *iterUA*, a batch of users U_Δ and the available frequency distribution estimation F^* are fed up to the UAS module, which outputs a user allocation strategy for U_Δ with relative error optimization. Following that, we apply OUE to collect the information from the different groups of users to derive the frequency distribution estimation F^Δ on different dimensions. Then F^Δ is used to refine F^*, by using the *weight average technique* [**?**] which is an effective way to improve result accuracy by combining multiple versions of estimations. Specifically, let ω_i^* (ω_i^Δ) denote the error variance of F_i^* (F_i^Δ). We use $\frac{\omega_i^\Delta}{\omega_i^* + \omega_i^\Delta} F_i^* + \frac{\omega_i^*}{\omega_i^* + \omega_i^\Delta} F_i^\Delta$ as the more accurate estimation of F_i to update F_i^*. Then the error variance in the updated F_i^* is $\frac{\omega_i^* \omega_i^\Delta}{\omega_i^* + \omega_i^\Delta}$. Up to now, one step of iteration is

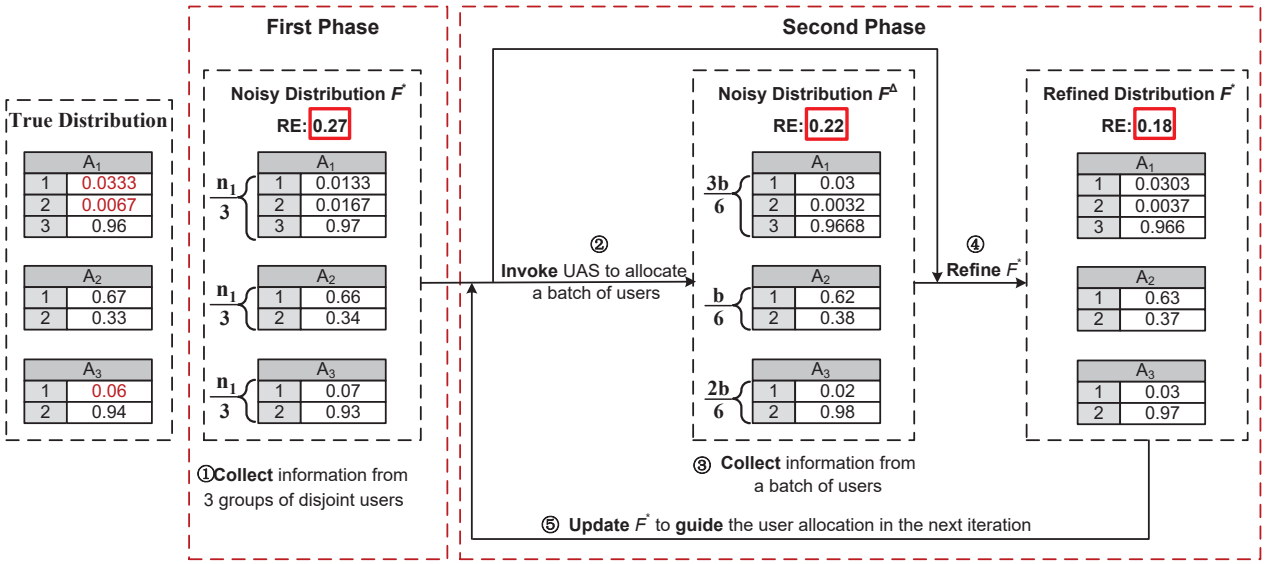


Fig. 3 An example for illustrating the process of *iterUA*.

finished with outputting a more accurate F^* than the version in the beginning of this step, which can lead to better optimization on the relative error in the UAS module. With the processing of iterations, since the number of users participating in frequency computation increases, the accuracy of F^* should be improved gradually. Further, each batch of users are allocated reasonably based on F^*, leading to reduced relative error.

Fig. **??** shows a simple example of the iterative framework for the frequency distribution estimation on attributes $\{A_1, A_2, A_3\}$, where n_1 is the number of users involved in the first phase. Firstly, *iterUA* allocates $\frac{n_1}{3}$ users to each dimension and each user sends the OUE-perturbed data on the corresponding attribute to the curator. Then the latter derives the rough estimation of frequency distribution F^*. It is observed that the distributions on Attributes A_1 and A_3 are more skewed with smaller frequencies than that on A_2. So UAS allocates b users appropriately according to the noisy frequency distribution, i.e., $\frac{3b}{6}$ users for A_1, $\frac{b}{6}$ users for A_2, and $\frac{2b}{6}$ users for A_3, where b is the size of one batch. Obviously, the attributes with skewed distributions are allocated more users to. That is significantly important to reduce the relative error. After the allocation is completed, each user still invokes OUE to perturb her data on the corresponding attribute and the curator derives an estimation F^Δ. F^Δ is adopted to refine F^* in a weight average way. Following that, a more accurate estimation F^* is regarded as the input of the UAS module in the next iteration, which contributes to achieving a better allocation strategy for the relative error optimization.

Lemma 2. *The iterative framework* iterUA *satisfies ϵ-local differential privacy.*

Proof. IterUA splits users into two parts, each of which is fed up to the first phase and second phase respectively. Based on the property of differential privacy, *parallel composition*, *iterUA* satisfies ϵ-LDP if and only if both of the two phases satisfy ϵ-LDP. Now we show the proofs of LDP guarantee for the two phases. In the first phase, each user invokes OUE to publish her value on only a dimension under ϵ-LDP. The simple application of OUE makes the first phase satisfy ϵ-LDP. On the other hand, the second phase is an iterative process, in which each iteration invokes the UAS module to split a batch of users U_Δ into d disjoint groups and derives a frequency distribution estimation F^Δ based on the users in the d disjoint groups to update F^*. The UAS module just takes an ϵ-LDP compliant noisy frequency distribution estimation F^* and the size of U_Δ as input, and both are not sensitive information. So the UAS module does not consume extra privacy budget. To derive F^Δ, each user just invokes OUE to publish her data on one dimension under ϵ-LDP. As a result of that, the process of deriving F^Δ satisfies ϵ-LDP. Besides, the weight average technique does not consume privacy budget, since it deals with two ϵ-LDP compliant estimations. In conclusion, the iteration satisfies ϵ-LDP. In addition, as different iterations in the second phase collect information from disjoint users to compute F_Δs, the second phase satisfies ϵ-LDP due to the *parallel composition* property. Hence, this lemma is proved. □

5.2 IterUA with Optimized User Allocation Strategy

Recall that *iterUA* is an iterative framework, in which each iteration invokes UAS to derive the allocation strategy for a batch of users U_Δ to minimize the relative error of the frequency distribution estimation F^Δ computed from U_Δ. However, the goal of *iterUA* is to output an estimation F^* with reduced relative error as the final result, which is achieved by a weight average technique based on F^Δ and a previous version of F^*. So in *iterUA*, it is more reasonable to construct a user allocation strategy to minimize the relative error of F^*, rather than F^Δ. Since F^* depends on not only the values of the users in the current iteration but also the information from the previous iterations, it leaves some space of the optimization for the UAS module.

We propose *optimized user allocation strategy* (OUAS) by combining the feature of *iter*UA, which constructs a user allocation strategy to minimize the relative error of the output F^* from each iteration . Now we can describe the user allocation problem in the tth iteration of *iterUA* as follows. Given d attributes $\{A_1, A_2, ..., A_d\}$ and a rough frequency distribution estimation F^* on all attributes, OUAS aims to split the users in U_Δ into d disjoint groups and the users in group g_{ti} ($1 \le i \le d$) are involved in computing the frequency estimation F_i^Δ on A_i, so that the relative error in the updated version F^* is minimized, where F^* is equal to the weight average of F^Δ and the rough estimation F^* (Line 14 in Alg. **??**). Since the error variance in the frequency of each value in the updated F_i^* is $\frac{\omega_i^\Delta \omega_i^*}{\omega_i^\Delta + \omega_i^*}$, the expected overall relative error on all attributes after the tth iteration can be written as:

$$E(g_{t1}, g_{t2}, ..., g_{td}) = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{|A_i|}\sum_{j=1}^{|A_i|}\frac{\sqrt{\omega_i^\Delta \omega_i^*}}{\sqrt{\omega_i^\Delta + \omega_i^*}\max(\delta, F_{ij}^*)}. \tag{5}$$

Since the LDP protocol OUE is adopted to collect data from users, the error variance ω_i^Δ of each frequency in F_i^Δ is $\frac{q(1-q)}{|g_{ti}|(p-q)^2}$. As for the error variance ω_i^* in the rough version of F_i^*, it is equal to $\frac{q(1-q)}{(p-q)^2(\sum_{k=1}^{t-1}|g_{ki}|+\alpha n/d)}$, where $|g_{ki}|$ denotes the number of users involved in computing the frequency distribution on A_i in the kth iteration. By taking the two variances into Eqn. **??**, we have

$$\begin{aligned} &E(g_{t1}, g_{t2}, ..., g_{td}) \\ &= \frac{1}{d}\sum_{i=1}^{d}\frac{1}{|A_i|}\sum_{j=1}^{|A_i|}\frac{\sqrt{q(1-q)}}{\sqrt{(\frac{\alpha n}{d}+\sum_{k=1}^{t}|g_{ki}|)(p-q)^2}\max(\delta, F_{ij}^*)} \\ &\propto \sum_{i=1}^{d}\frac{1}{|A_i|}\sum_{j=1}^{|A_i|}\frac{1}{\sqrt{(\sum_{k=1}^{t-1}|g_{ki}|+\frac{\alpha n}{d})+g_{ti}}\max(\delta, F_{ij}^*)} \end{aligned} \tag{6}$$

In Eqn. **??**, $\sum_{k=1}^{t-1}|g_{ki}|$ is a constant, since the user allocation strategies of the previous $t-1$ iterations have already been available when at the tth iteration. We can derive the number of users in each group g_{t1}, g_{t2}, ..., g_{td}, such that

$$\sum_{i=1}^{d}\frac{1}{|A_i|}\sum_{j=1}^{|A_i|}\frac{1}{\sqrt{(\sum_{k=1}^{t-1}|g_{ki}|+\alpha n/d)+|g_{ti}|}\max(\delta, F_{ij}^*)}$$

is minimized subject to the constraint that

$$\sum_{i=1}^{d}|g_{ti}| = |U_\Delta|.$$

Similar to UAS, we can use the Lagrange multiplier method to derive the values of $|g_{t1}|$, $|g_{t2}|$, ..., and $|g_{td}|$, where $|g_{ti}|$ is equal to

$$\frac{(\alpha n + |U_\Delta|t)\left(1/|A_i|\sum_{j=1}^{|A_i|}\frac{1}{\max(\delta, F_{ij}^*)}\right)^{2/3}}{\sum_{k=1}^{d}\left(1/|A_i|\sum_{j=1}^{|A_k|}\frac{1}{\max(\delta, F_{kj}^*)}\right)^{2/3}} - \frac{\alpha n}{d} - \sum_{k=1}^{t-1}|g_{ki}| \tag{7}$$

Then we can choose $|g_{ti}|$ users from the user set U_Δ randomly as the users in g_{ti}, which are fed up to the attribute A_i.

The OUAS module can be embedded in *iterUA* with slight modifications on Algorithm **??**. In particular, we need to use d variables to record the number of users having been involved in computing frequency distributions on d attributes respectively. The value $\alpha n + |U_\Delta|t$ in Eqn. **??** is equal to U_Δ plus the sum of these variables. And the value $\frac{\alpha n}{d} + \sum_{k=1}^{t-1} g_{ki}$ for Attribute A_i is equal to the variable for A_i. Then with these variables, we can derive the user allocation strategy with minimizing the relative error in F^*.

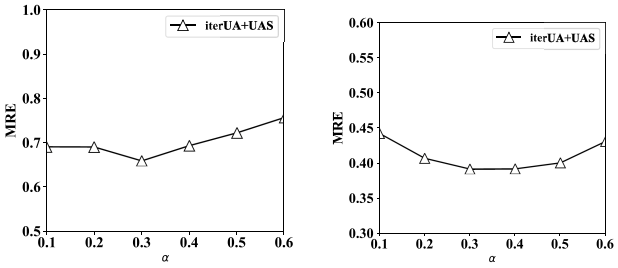
Note that although OUAS seems like so complex, it is privacy free. This is because the user allocation strategies in the first $t-1$ iterations and the rough frequency distribution estimation F^*, which are used to compute $|g_{ti}|$ in Eqn. **??**, are not sensitive information.

6 Experiments

We have implemented the proposed methods and evaluated them using two pubic datasets extracted from the *Integrated Public Use Microdata Series* [**?**], termed as BR and MX, which contain census records from Brazil and Mexico, respectively. BR contains about 4M tuples and 10 categorical attributes; MX contains about 4M records and 14 categorical attributes. We compare our proposals *iterUA* involving UAS (*iterUA*+UAS) and *iterUA* involving OUAS (*iterUA*+OUAS) with two methods on the measurement MRE ($\delta = 2 \times 10^{-4} n$): (i) AU [**?**, **?**]: Since no existing solution is devoted to optimizing relative error for frequency estimation under LDP, we take the state-of-the-art technique UA for optimizing absolute error as the baseline. It splits users into d disjoint parts, each of which is involved in computing the frequency estimation on a dimension. And the recently proposed works [**?**, **?**] for range queries also adopt the user split technique to collect multiple values; (ii) TTP: it is designed to show the upper bound of the optimization on relative error which we can achieve. In particular, we use αn users to compute F^* in the non-private setting, and then invoke UAS to derive the user allocation strategy for the remaining $(1-\alpha)n$ users with F^* as input. Finally, we estimate the frequency distribution based the derived allocation strategy as the result under LDP. Note that TTP does not satisfy LDP due to collecting true data from αn users directly. In all experiments, we report average results over 20 runs.

6.1 Impacts of α and τ

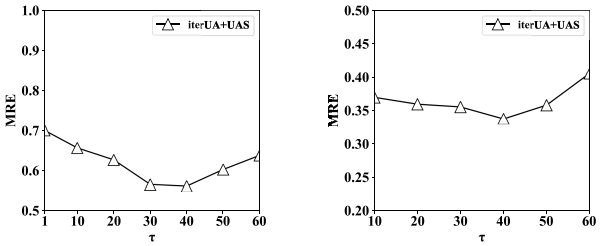
Recall that *iterUA* has one internal parameter: the ratio α of users involved in the first phase and second phase. To evaluate the impact of α, we examine the performance of *iterUA* in the frequency distribution estimation with varied α but fixed iteration counter ($\tau = 1$) and privacy budget ($\epsilon = 1$). Fig. 4 illustrates the results. Observe that the MRE of the frequency estimation tends to be higher when α is very small or very large. This is consistent with our analysis in Section 5.1 that (i) small α leads to very noisy frequency distribution estimation F^*, which makes the UAS module output a user allocation strategy mis-calibrating the relative error, and (ii)large α makes the number of users involved in optimizing relative error decrease in the second phase, leading to smaller optimization space and inferior results. Based on Fig. 4, we infer that an appropriate value for α should be in the range of $[0.2, 0.4]$. Without loss of generality, for all subsequent experiments, we set $\alpha = 0.3$.



(a) BR (b) MX

Fig. 4 Impact of α

The other important parameter in *iterUA* is the number of iterations τ in the second phase, which decides the batch size in the iteration process. Fig. 5 evaluates its impact by varying τ while privacy budget ϵ is fixed as 1. With the increase of τ, the MRE displays the tendency declining at the beginning and rising up later. That is because a smaller number τ leaves fewer opportunities to remedy the negative impact on relative error. Such impact is caused by an unreasonable user allocation generated from the mis-calibrated F^*. On the other hand, when τ becomes large, the number of users involved in each iteration is reduced, resulting in inaccurate or even meaningless frequency distribution estimation F^{Δ}. In Fig. 5, it is observed that an appropriate value for τ should be in the range of $[30, 40]$. For all subsequent experiments, we set $\tau = 40$ when $\epsilon = 1$. According to Ref. [**?**], the number of iterations under different privacy budgets should be proportional to ϵ^2. In this way, by determining the number of iterations when $\epsilon = 1$, we can easily derive the reasonable value of τ for other privacy budget.



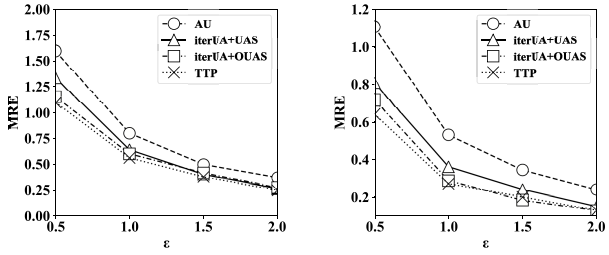
(a) BR (b) MX

Fig. 5 Impact of τ

6.2 Results on 1-way marginals

We compares proposals *iterUA*+UAS and *iterUA*+OUAS with the AU and TTP approaches on 1-way marginals over

the BR and MX datasets. The 1-way marginals publication is consistent with the frequency distribution estimation on multidimensional data this paper focuses on. To distinguish the publication task in the next section, we use the term 1-way marginals in the experiment part. Fig. 6 plots the MRE results on 1-way marginals as a function of the privacy budget ϵ. Overall, the proposed solutions consistently and significantly outperform the existing method AU. Notably, when ϵ is 1, the gap between *iterUA* and AU is about 0.2. The good performance of *iterUA* is mainly due to the iterative user allocation with the relative error optimization. In addition, the MSE of *iterUA*+OUAS is noticeably better than that of *iterUA*+UAS in all cases and close to the upper bound of the relative error optimization which is described by the line of TTP. This is because *iterUA*+OUAS adopts a more reasonable user allocation strategy with relative error optimization, which considers not only the information from the users in the current iteration, but also that from the users in the previous iterations. Fig. 7 plots the MRE under $\epsilon = 1$ as a function of δ, the sanity bound of relative error. It is observed that a larger δ contributes to reducing the relative error since a larger denominator is adopted in Eqn. **??** when the frequency is small. Besides, Fig. 7 shows a similar phenomenon to that in Fig. 6, i.e., *iterUA* including *iterUA*+UAS and *iterUA*+OUAS considerably outperforms AU and the performance of *iterUA*+OUAS is better than that of *iterUA*+UAS.

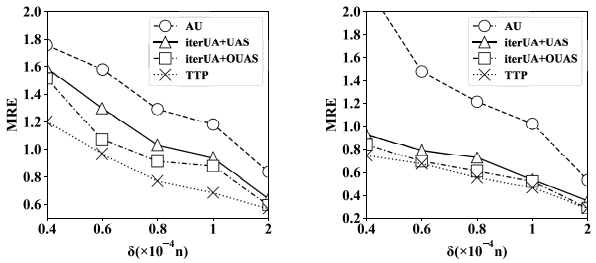


(a) BR (b) MX

Fig. 6 1-way marginals on two datasets by varying ϵ

6.3 Results on 2-way Marginals

In the last set of experiments, we evaluate different methods for 2-way marginals, which include the joint frequency distributions on any two attributes or dimensions. To derive the distribution estimation on any two attributes, we regard each combination of two attributes as one new dimension. In particular, if there exist d attributes in the dataset, C_d^2 new dimensions are generated. And given a new dimension from

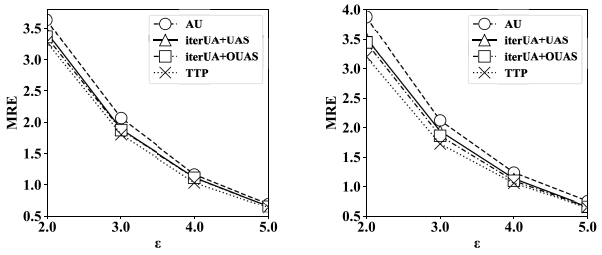


(a) BR (b) MX

Fig. 7 1-way marginals on two datasets by varying δ

attributes A_i and A_j, its domain is $C_i \times C_j$, where "×" denotes the operation of Cartesian product. Then the tuple from each user is transformed into one new tuple with C_d^2 values. In this way, we can invoke our methods or compared methods to deal with the transformed data to derive the 2-way marginals.

Fig. 8 shows the MRE of each method on the two real datasets. The results on 2-way marginals lead to similar conclusions with that on 1-way marginals. We then omit the explanation for brevity. Note that the MRE for 2-way marginals is considerably higher than that for 1-way marginals. The main reason is that the transformed tuples for computing 2-marginals have more dimensions, which makes fewer users be allocated to each dimension, and hence, incurs larger amount of noise in the results. Besides, the frequency of each value in the new dimensions is much smaller than that in the original dimensions, which also has a negative impact on the result accuracy measured by relative error. Another observation is that the improvement of result accuracy with *iterUA* on the MX dataset is much larger than that on the BR dataset. This is because higher dimensionality on MX provides more chances to optimize relative error for *iterUA*.



(a) BR (b) MX

Fig. 8 2-way marginals on two datasets by varying ϵ

7 Conclusion

In this paper, an iterative framework *iterUA* is designed for publishing frequency distribution with reduced relative error on multidimensional data under LDP. In each iteration step, the optimized user allocation strategy OUAS is invoked to reduce the relative error in the derived results, which takes the information not only from the current iteration step but also from the previous steps into consideration. The combination of *iterUA* and OUAS dramatically improves the result accuracy measured by relative error, as verified over real datasets. In the furture, we are going to investigate the problem of the relative error optimization for mean estimation under LDP.

References

1. Haiming Jin, Lu Su, Houping Xiao, and Klara Nahrstedt. Incentive mechanism for privacy-aware data aggregation in mobile crowd sensing systems. *IEEE/ACM Trans. Netw.*, 26(5):2019–2032, 2018.
2. Leye Wang, Daqing Zhang, Dingqi Yang, Brian Y. Lim, Xiao Han, and Xiaojuan Ma. Sparse mobile crowdsensing with differential and distortion location privacy. *IEEE Trans. Inf. Forensics Secur.*, 15:2735–2749, 2020.
3. Yaliang Li, Houping Xiao, Zhan Qin, Chenglin Miao, Lu Su, Jing Gao, Kui Ren, and Bolin Ding. Towards differentially private truth discovery for crowd sensing systems. In *40th IEEE International Conference on Distributed Computing Systems, ICDCS 2020, Singapore, November 29 - December 1, 2020*, pages 1156–1166. IEEE, 2020.
4. Shaowei Wang, Yuqiu Qian, Jiachun Du, Wei Yang, Liusheng Huang, and Hongli Xu. Set-valued data publication with local privacy: Tight error bounds and efficient mechanisms. *Proc. VLDB Endow.*, 13(8):1234–1247, 2020.
5. Apple's 'differential privacy' is about collecting your data – but not *your* data. `https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/`, 2016.
6. Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint 1709.02753*, 2017.
7. Rappor (randomized aggregatable privacy preserving ordinal responses). `https://www.chromium.org/developers/design-documents/rappor`.
8. Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *NIPS*, pages 3571–3580, 2017.
9. Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multidimensional data with local differential privacy. In *ICDE*, pages 638–649, 2019.
10. Ruixuan Liu, Yang Cao, Hong Chen, Ruoyang Guo, and Masatoshi Yoshikawa. FLAME: differentially private federated learning in the shuffle model. In *AAAI*, pages 8688–8696, 2021.
11. Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *USENIX Security*, pages 729–745, 2017.
12. Tianhao Wang, Ninghui Li, and Somesh Jha. Locally differentially private heavy hitter identification. *IEEE Trans. Dependable Secur. Comput.*, 18(2):982–993, 2021.
13. Graham Cormode, Samuel Maddock, and Carsten Maple. Frequency estimation under local differential privacy. *Proc. VLDB Endow.*, 14(11):2046–2058, 2021.
14. Tianhao Wang, Milan Lopuhaä-Zwakenberg, Zitao Li, Boris Skoric, and Ninghui Li. Locally differentially private frequency estimation with consistency. In *NDSS*, 2020.
15. Cynthia Dwork. Differential privacy. In *ICALP*, pages 1–12, 2006.
16. Cynthia Dwork. Differential privacy: A survey of results. In *TAMC*, pages 1–19, 2008.
17. Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. pages 1423–1434, 2014.
18. Wahbeh H. Qardaji, Weining Yang, and Ninghui Li. Priview: practical differentially private release of marginal contingency tables. In *SIGMOD*, pages 1435–1446, 2014.
19. Xiaokui Xiao, Gabriel Bender, Michael Hay, and Johannes Gehrke. ireduct: differential privacy with reduced relative errors. In *SIGMOD*, pages 229–240, 2011.
20. Zhikun Zhang, Tianhao Wang, Ninghui Li, Jean Honorio, Michael Backes, Shibo He, Jiming Chen, and Yang Zhang. Privsyn: Differentially private data synthesis. In *USENIX Security Symposium*, pages 929–946, 2021.
21. Kuntai Cai, Xiaoyu Lei, Jianxin Wei, and Xiaokui Xiao. Data synthesis via differentially private markov random field. *Proc. VLDB Endow.*, 14(11):2190–2202, 2021.
22. Raef Bassily and Adam D. Smith. Local, private, efficient protocols for succinct histograms. In *STOC*, pages 127–135, 2015.
23. Xuebin Ren, Chia-Mu Yu, Weiren Yu, Shusen Yang, Xinyu Yang, Julie A. McCann, and Philip S. Yu. Lopub: High-dimensional crowdsourced data publication with local differential privacy. *IEEE Trans. Inf. Forensics Secur.*, 13(9):2151–2166, 2018.
24. Ning Wang, Xiaokui Xiao, Yin Yang, Ta Duy Hoang, Hyejin Shin, Junbum Shin, and Ge Yu. Privtrie: Effective frequent term discovery under local differential privacy. In *ICDE*, pages 821–832, 2018.
25. Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. Marginal release under local differential privacy. In Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein, editors, *SIGMOD*, pages 131–146, 2018.
26. Zhikun Zhang, Tianhao Wang, Ninghui Li, Shibo He, and Jiming Chen. CALM: consistent adaptive local marginal for marginal release under local differential privacy. In *CCS*, pages 212–229, 2018.
27. Jianyu Yang, Tianhao Wang, Ninghui Li, Xiang Cheng, and Sen Su. Answering multi-dimensional range queries under local differential

privacy. *Proc. VLDB Endow.*, 14(3):378–390, 2020.

28. Linkang Du, Zhikun Zhang, Shaojie Bai, Changchang Liu, Shouling Ji, Peng Cheng, and Jiming Chen. AHEAD: adaptive hierarchical decomposition for range query under local differential privacy. In *CCS*, pages 1266–1288, 2021.
29. Frank McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *SIGMOD*, pages 19–30, 2009.
30. Ninghui Li, Wahbeh Qardaji, Dong Su, and Jianneng Cao. Privbasis: Frequent itemset mining with differential privacy. *PVLDB*, 5(11):1340–1351, 2012.
31. Integrated public use microdata series international. `https://international.ipums.org`.