Overcoming Challenges in PLDP: A New Paradigm for Protecting Privacy Budgets and Enhancing Data Utility

Nengfa Wu
School of computer science and technology, Huazhong University of Science and Technology
Wuhan, China
nengfawu8@gmail.com

Hong Zhu*
School of computer science and technology, Huazhong University of Science and Technology
Wuhan, China
zhuhong@hust.edu.cn

Zhiqiang Zhang
School of computer science and technology, Huazhong University of Science and Technology
Wuhan, China
chijiang.cheung@gmail.com

Meiyi Xie
School of computer science and technology, Huazhong University of Science and Technology
Wuhan, China
xiemeiyi@hust.edu.cn

ABSTRACT

With the growing awareness of security, privacy protection has become a core concern in modern life. Personalized Local Differential Privacy (PLDP) meets users' privacy protection needs and enhances data usability. However, PLDP faces a challenge: differences in users' privacy budgets could be exploited by malicious third parties to infer private information. Existing PLDP-based frequency estimation schemes have not fully addressed this issue, or have shortcomings. To tackle this, we propose a new framework: PLDP-Frequency Estimation with Budget Security Framework(PLDP-FEBSF). Users can choose a privacy budget to perturb their data according to their needs, with an additional budget to protect the actual privacy budget. On the server side, we developed an approximate unbiased frequency distribution estimation method with protected privacy budgets. We also analyzed the usability and potential shortcomings of this method and proposed an expectation-maximization-based approach to improve estimation accuracy. Evaluations on real datasets show that PLDP-FEBSF improves data utility by 87% compared to schemes protecting privacy budgets and achieves comparable results to those that do not.

PVLDB Reference Format:
Nengfa Wu, Hong Zhu, Zhiqiang Zhang, and Meiyi Xie. Overcoming Challenges in PLDP: A New Paradigm for Protecting Privacy Budgets and Enhancing Data Utility. PVLDB, 14(1): XXX-XXX, 2024.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/Couchy-wu/PLDP-FEBSF.

1 INTRODUCTION

With the advent of the intelligent era, various services based on data analysis have become an indispensable part of daily life. Service providers have collected a large amount of data. However, recent data disclosure incidents on major platforms such as Facebook [1] and Apple's iCloud [18] have raised widespread concerns about privacy and security. Consequently, users have become more cautious about sharing their personal information. Therefore, it has become an important issue to effectively de-sensitize user data and minimize its impact on statistics while preserving privacy.

Local differential privacy(LDP) [14] is a mainstream privacy protection method. Compared with centralized differential privacy [11], LDP does not rely on a trusted third-party server, and allows users to perturb the original data locally before sending it to the central server for analysis, thus effectively protecting user privacy. As a primary task in statistics, frequency distribution estimation has been widely used in data analysis and machine learning [5, 19]. Researchers have recently proposed many LDP-based frequency distribution estimation schemes [4, 9, 33] and applied them to data collection tasks in various enterprises [14, 30]. However, existing LDP-based frequency estimation schemes [41, 44] usually assume that the privacy protection requirements are uniform among users, ignoring the diversity of users' privacy needs. To address this problem, researchers have proposed personalized Local differential privacy (PLDP) [7], which allows users to apply the data with different perturbations according to their privacy protection requirements. Several frequency estimation methods based on PLDP have been proposed [31, 34], and the representative schemes include RCF [31] and APLDP [34].

For the RCF scheme [31], users send their accurate privacy budget and perturbed data to the server. The server groups the perturbed data according to the privacy budget and then aggregates each group's estimation results by adopting minimum mean square error principle to obtain the final estimation value. Similarly, the APLDP scheme [34] also uses the grouping method to estimate the frequency distribution of the original data.Researchers consider that different LDP protocols perform best in different privacy budget ranges, allowing each group of users to choose the best-fit LDP protocol based on their accurate privacy budget, thus enhances

[*] Corresponding author

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

the accuracy of the final estimation. Unfortunately, in PLDP, the difference in privacy budget between users may be exploited by a malicious third party to infer users' privacy information [24]. Therefore, it is crucial to protect users' actual privacy budget. Currently, only some PLDP-based frequency distribution estimation schemes [35, 37] pay attention to the protection of user privacy budget, and some limitations are existed in these schemes. For example, in the PBP scheme [35], users can hide their accurate privacy budget from the data collector. However, to accurately estimate the frequency distribution of the original data, the authors assume that the distribution of users across individual privacy budgets is public. However, in pratice, obtaining this information may be challenging. In the PPDA scheme [37], users do not need to send their privacy budget to the data collector. However, this approach is only suitable for the local perturbation algorithm using onehot encoding [13]. Moreover, to ensure the accuracy of the estimation results, a substantial number of extra zero bits must be appended to the encoding vector. This approach not only increases additional communication overhead but may also indirectly leak privacy budgets. Since frequency distribution estimation in PLDP depends on the information of privacy budgets chosen by users, accurately estimating the original data's frequency distribution becomes challenging after protecting users' privacy budgets.

In order to better estimate the frequency distribution of original data while protecting user privacy budget, we propose a new frequency distribution estimation framework: PLDP-Frequency Estimation and Budget Security Framework (PLDP-FEBSF). In the proposed framework, users select an appropriate privacy budget according to their privacy requirements to perturb their private data locally. In order to protect the user's privacy budget and ensure the generality of the scheme, different from the existing schemes, we introduce an additional privacy budget to perturb the user's accurate privacy budget locally. To accurately estimate the frequency distribution of the original data, the framework develops a approximate unbiased estimation method on the server side. Furthermore,the potential deficiencies of this method is identified by performing a usability analysis of the approximately unbiased estimation results. In order to further improve the accuracy of estimation results, we transform the frequency distribution estimation problem under PLDP into a multi-distribution mixture problem and propose an optimization algorithm based on expectation maximization [45].

Finally, to verify the effectiveness of the PLDP-FEBSF method, we use two evaluation indicators to conduct a comparative analysis on three real datasets. The experimental results show that, compared with the frequency distribution estimation scheme that protects the user privacy budget, the proposed method improves data availability by 87%. Moreover, the proposed method is comparable in performance to the scheme that does not protect the user privacy budget.

In summary, the main contributions of this paper are as follows:

- We introduce the PLDP-Frequency Estimation with Budget Security Framework which simultaneously protects users'true privacy budgets and allows personalized data perturbation, ensuring user privacy.
- We propose an approximate unbiased estimation method under the scenario of preserving users' privacy budgets for perturbed data, and conduct variance analysis.
- We transform the frequency distribution problem under personalized local differential privacy into a multi-distribution mixture problem and propose an EM optimization algorithm to improve the estimation results' accuracy further.
- We conduct comparative experiments with existing schemes on three real data sets. The experimental results show that, compared with the frequency distribution estimation scheme with user privacy budget protection, our scheme improves the data availability index by 87%and is close to the scheme's performance without user privacy budget protection.

Outline. Recent advances and research results are reviewed in Section 2. Section 3 introduces Local Difference Privacy (LDP), Personalized Local Difference Privacy (PLDP) and frequency distribution estimation protocols. Section 4 clarifies the problem definition and outlines the PLDP-FEBS framework. Section 5, "Strategic Approaches to Data Privacy and Utility Analysis," details the client-side steps to protect data privacy and budget and explains the methodology for server-side approximate unbiased frequency distribution estimation from perturbed data. Section 6, "Usability Optimization," discusses how to improve the accuracy of the estimation results. Section 7 focuses on experimental evaluation and analysis. Finally, Section 8 reviews the main work and future directions.

2 RELATED WORK

LDP [14, 40] is one of the current mainstream privacy protection methods. Unlike centralized differential privacy [6, 11], it does not rely on a trusted third-party server, but allows users to locally perturb the private data before sending it to the data collector for statistical analysis.As a basic task in statistics, frequency distribution estimation has been widely used in machine learning and data analysis [5, 20, 25]. In recent years, researchers have proposed many frequency estimation schemes based on LDP [26, 41, 44], and some of them have been widely used in user data collection tasks by companies such as Google [14] and Apple [8]. However, LDP usually assumes that the privacy protection requirements of all users are uniform, ignoring the differences among users. Addressing these differences is crucial to improve data usability and meet the diverse privacy protection needs of users.

Alaggan et al. [3] and Niu et al. [32]proposed the concept of personalized differential privacy, taking into account different privacy preferences and adapting the Laplace mechanism accordingly. However, this approach does not work for certain query functions such as median or min/Max. To address this limitation, Jorgensen et al. [21] proposed two new approaches based on nonuniform sampling and exponential mechanisms. Li et al. [23] also proposed a partition-based personalized differential privacy protection method, which partitions the user data according to privacy preferences before applying the differential privacy protection mechanism. Although the partitioning procedure is effective, it incurs a large computational overhead. Recent research has also explored personalized privacy requirements in different scenarios, such as social recommendation, etc [2, 42, 43]. However, it is important to note that most of these approaches rely on a trusted third party for implementation.

Chen et al. [7] proposed a privacy-preserving mechanism called Personalized Local differential privacy (PLDP). The current research on frequency distribution estimation of discrete data based on PLDP mainly focuses on two aspects: 1. Value-oriented PLDP. 2. User-oriented PLDP. Value-oriented PLDP considers the sensitivity difference of data attribute values rather than the sensitivity difference of users. For example, the sensitivity of answering "yes" to the question of whether a person is HIV positive is significantly higher than that of answering "no". On this basis, Murakami et al. [28] proposed ULDP(Utility-optimized LDP), which divided the data into sensitive and non-sensitive parts and only protected the sensitive part to improve the estimation accuracy of frequency distribution. Gu et al. [16] further divided the sensitivity level and proposed an ID-LDP (Input Discrimination-LDP) scheme.

User-oriented PLDP focuses on the different sensitivity of different users to the same information. For example, women are more sensitive to weight information than men. Nie et al. [31] proposed a frequency estimation framework RCF (Recycle and Combination framework) based on user grouping. It estimates the distribution of the perturbed data according to the user's privacy budget. Then, the estimation results of each group are aggregated according to the variance minimization principle to obtain the final estimation result. The APLDP scheme [34] also employs a grouping method to estimate the frequency distribution of the original data. They consider that different LDP protocols have different ranges of optimal privacy budgets and thus allow each group of users to choose the most appropriate LDP protocol based on their accurate privacy budget, thereby improving the accuracy of the final estimation. However, these methods often ignore the protection of users' accurate privacy budgets because illegal third parties may use the difference between users' privacy budgets to speculate about users' privacy information. Presently, the research on frequency distribution estimation based on PLDP has paid limited attention to protecting the user privacy budget. Takagi et al. [35] proposed PBP (Parameter Blending Privacy), where users can hide their accurate privacy budget from the data collector. However, to ensure the accuracy of the estimation results, this method assumes that the distribution of users over each privacy budget is public, which is challenging in practice. Wang et al. [37] proposed PPDA (Personalized Private Data Aggregation) scheme, which is only applicable to the LDP protocol using one-hot encoding. In order to ensure the accuracy of the estimation result, a substantial number of extra zero bits need to be added to the encoding vector. This approach not only increases additional communication overhead but may also indirectly leak privacy budgets.

3 PRELIMINARIES

3.1 Local Differential Privacy

Local differential privacy (LDP) [14] is an extension of central differential privacy [11] used in scenarios where a trusted third-party server is not present. The concept of LDP can be defined as follows.

Definition 1. *For any privacy protection algorithm $\mathcal{M}(\cdot)$ that satisfies ϵ-LDP, where $\epsilon > 0$, and for any inputs $v_1, v_2 \in D$ with output $y \in \mathcal{M}(\cdot)$, the following inequality holds:*

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

The parameter ϵ is the privacy budget. The larger the ϵ is, the less the protection strength becomes. This definition formally ensures that at most e^{ϵ} output distinguishability in probabilities when the input data differs; that is, an adversary cannot distinguish the user's raw data just by seeing the perturbed versions.

3.2 Personal Local Differential Privacy

Personalized Local Differential Privacy(PLDP) [7]can be viewed as a generalized definition of LDP, allowing different users to select varying privacy budgets based on their individual privacy needs. Hence, we provide the following definition for PLDP:

Definition 2. *Let $\mathcal{D} = \{x_i\}_{i=1}^{k}$ represent a discrete and finite set of data, and let $\epsilon = \{\epsilon_i\}_{i=1}^{t}$ denote a discrete and finite set of privacy bugets. For each user u, with data x, a privacy budget ϵ_u is chosen from ϵ. The user employs a random perturbation method $\mathcal{M}$ that satisfies PLDP. The algorithm satisfies PLDP if and only if the following inequality is met:*

$$\Pr(\mathcal{M}(x) = Z_u) \leq e^{\epsilon_u} \Pr(\mathcal{M}(x') = Z_u),$$

Here, $x, x' \in \mathcal{D}$, and $Z_u \in \mathcal{M}(\cdot)$.

In contrast to traditional LDP, where all users adhere to a uniform global privacy budget, PLDP enables individual users to locally determine a privacy budget that aligns with their specific privacy requirements for perturbing their private data. Additionally, many properties of differential privacy are still applicable in the context of PLDP, such as Composition Theorem [27] and Post-processing Theorem [12].

3.3 Pure Local Differential Privacy Protocol

Wang [38] introduced a practical concept called "Pure Local Differential Privacy" for frequency estimation schemes on discrete data. The formally definition of the Pure local differential privacy(Pure LDP) protocol is following:

Definition 3. *By* PE(.) *and Support(.) a given LDP protocol is pure if and only if there are two probability values, denoted as p^* and q^*, such that $p^* > q^*$. Additionally, following condition holds for all entries v_1.*

$$\Pr[\mathrm{PE}(v_1) \in \{y | v_1 \in \mathrm{Support}(y)\}] = p^*,$$
$$\forall v_1 \neq v_2 \Pr[\mathrm{PE}(v_2) \in \{y | v_1 \in \mathrm{Support}(y)\}] = q^*, \tag{1}$$

The term "PE(.)" represents the combined process of encoding and perturbation, specifically denoted as "PE(.) = Perturb(Encode(.))," ensuring the satisfaction of ϵ-LDP. suppprt(y) represents the set of items in the input space, and the outputs are all equal to y.

For any frequency estimation protocol that satisfies the definition of Pure LDP, data collectors can use the following Eq. (2) to obtain an unbiased estimate of the original frequency for the corresponding term x_i:

$$\hat{p}_i = \frac{\sum_u \mathbb{I}_{support(t_u)}(x_i) - Nq^*}{N(p^* - q^*)}, \tag{2}$$

where N represents the total number of users, $\hat{p}_i$ is the estimated frequency of item x_i, t_u is the perturbation output of user u, and the function $\mathbb{I}_{Support(t_u)}(x_i)$ outputs 1 if item i is supported, and 0 otherwise. Intuitively if t_u supports x_i then the count is plus one. Finally, we normalize the counts using probabilities p^* and q^*. Different encoding and perturbation functions are used for different Pure LDP algorithms, so they have distinct Support functions and probabilities p^* and q^*. Notably, most of the current mainstream frequency estimation schemes based on local differential privacy comply with the definition of Pure LDP. Examples of such schemes include Basic-RAPPOR [38],k-RR [22], and OUE [38].

4 OVERVIEW

4.1 Problem Definition

Firstly, to meet the personalized privacy protection needs of users, we categorize user privacy budgets into t levels, denoted as $\boldsymbol{\varepsilon} = \{\epsilon_i\}_{i=1}^{t}$. Here, $\epsilon_i \in \boldsymbol{\varepsilon}$ and $i \in [1, t]$ represents the i-th privacy budget. The larger the value of i, the higher the value of ϵ_i, resulting in a lower level of privacy protection (with the 1st level offering the strongest and the t-th the weakest protection). Without loss of generality, the difference between the two sequential privacy budget levels is $\triangle_\epsilon = \epsilon_{i+1} - \epsilon_i > 0$. The data domain $\mathcal{D} = \{x_i\}_{i=1}^{k}$ is a discrete and finite set of data items, and the Data x_i represents the value of attribute $\mathcal{D}$ for the i-th data item. For example, if $\mathcal{D}$ represents disease information, then x_i would denote the name of the i-th disease.

Additionally, since malicious third parties may exploit a user's accurate privacy budget to infer private information, it is essential to protect the privacy budget. In existing studies [35, 37], researchers usually choose to hide users' privacy budgets from data collectors. Although this hiding strategy is effective in terms of security, it significantly diminishes the generality of the method and data usability. Therefore, unlike existing studies, we decided to introduce an additional privacy budget ϵ_p to perturb the user's accurate privacy budget locally, thereby achieving the goal of protecting the user's privacy budget. Therefore, based on the above analyses, we can define the problem as follows.

There are N users, each having a data item $X_u \in \mathcal{D}$. Each user u selects a privacy budget ϵ_u^τ from the privacy budget set $\boldsymbol{\varepsilon}$ (indicating that user u has chosen the τ-th privacy budget from the set $\boldsymbol{\varepsilon}$). We assume the choice of privacy budget ϵ_u^τ by user u is dependent on the representation of the data domain $\mathcal{D}$ (e.g., health information, hobbies, etc.) and not on its specific values. After selecting the privacy budget ϵ_u^τ, user, u uses it to locally perturb the original data X_u, generating the perturbed data Z_u. Simultaneously, the user locally perturbs the privacy budget ϵ_u^τ using the privacy budget ϵ_p, generating the perturbed privacy budget ϵ_u^p. Finally, the user send $\langle Z_u, \epsilon_u^p \rangle$ to data collector. The data collector estimates the raw data distribution based on the user-provided data $\{\langle Z_u, \epsilon_u^p \rangle\}_{u=1}^{N}$.

Our objective is to achieve as accurate as possible estimation of the original data frequency distribution while effectively meeting users' personalized privacy protection needs and rigorously safeguarding their true privacy budgets. Furthermore, our method assumes user integrity, that is, Users are expected to strictly adhere to the prescribed protocol to protect and transmit their perturbed data, ensuring that there is no intentional data manipulation or concealment.

4.2 Framework Description

To address the abovementioned problems, we propose a new framework named PLDP-Frequency Estimation with Budget Security Framework(PLDP-FEBSF), shown in Figure 1. From Figure 1, it is evident that our framework can be broadly divided into two major components named 'Perturbations and Estimation' and 'Utility Optimization'. Within these, the 'Perturbations and Estimation' component further splits into the client and server side. On the client, each local user chooses an appropriate privacy budget ϵ_u^τ based on their privacy requirements. The LDP method is then used to perturb the real data X_u, generating the perturbed data Z_u. To prevent inference of user's privacy information through the real privacy budget ϵ_u^τ, a uniform privacy budget ϵ_p is applied to perturb ϵ_u^τ, resulting in ϵ_u^p after perturbation. The perturbed privacy budget ϵ_u^p and data Z_u are then sent to the server. On the server side, after collecting all perturbation data $\{\langle Z_u, \epsilon_u^p \rangle\}_{u=1}^{N}$ from users, the approximate unbiased estimation module obtains the estimated distribution $\hat{\rho}$ of users over each privacy budget from $\{\epsilon_u^p\}_{u=1}^{N}$. Using $\hat{\rho}$ and the perturbed privacy data $\{Z_u\}_{u=1}^{N}$, the original data distribution is estimated,denote as $\tilde{P}$. In the 'Utility Optimization' component, utilizing the Frequency optimization module improves the usability of the data, resulting in the final frequency optimized result P_{opt}. Next, we will detailedly disuss the 'Perturbations and Estimation' and 'Utility Optimization' in Sections 5 and 6, respectively.

5 STRATEGIC APPROACHES TO DATA PRIVACY AND UTILITY ANALYSIS

This section discuss the 'Perturbations and Estimation' component, structured into three key part: 'Personalized Data Perturbation and Localized Budget Protection', 'Approximate Unbiased Estimation with Protected Privacy Budgets', and 'Utility Analysis'. The 'Personalized Data Perturbation and Localized Budget Protection' subsection discusses methods for individualized protection of private data and safeguarding users' privacy budgets. 'Approximate Unbiased Estimation with Protected Privacy Budgets' focuses on detailing the process of estimating the raw data frequency distribution. Lastly, 'Utility Analysis' offers a theoretical error analysis of the estimation results, assessing their accuracy and utility.

5.1 Personalized Data Perturbation and Localized Budget Protection

In Pure LDP [38], frequency estimation methods are typically decomposed into three steps:$\mathcal{A} = \{\text{Encode}, \text{Perturb}, \text{Aggregate}\}$. During the **Encode** step, a value v is taken as the input, and an encoded value x is produced as the output. In the **Perturb** step, the encoded data x is perturbed to generate a perturbed value y, where the perturbation probabilities p^* and q^* are calculated according to a given privacy budget ϵ as Eq. (1). The **Aggregate** step collects all users' perturbed data and obtains the aggregated information. In the abovementioned problem, we aim to satisfy personalized privacy protection requirements while preserving users' actual privacy budgets. Therefore, we extend the two user-local steps $\{\textbf{Encode}, \textbf{Perturb}\}$ into the following five steps:

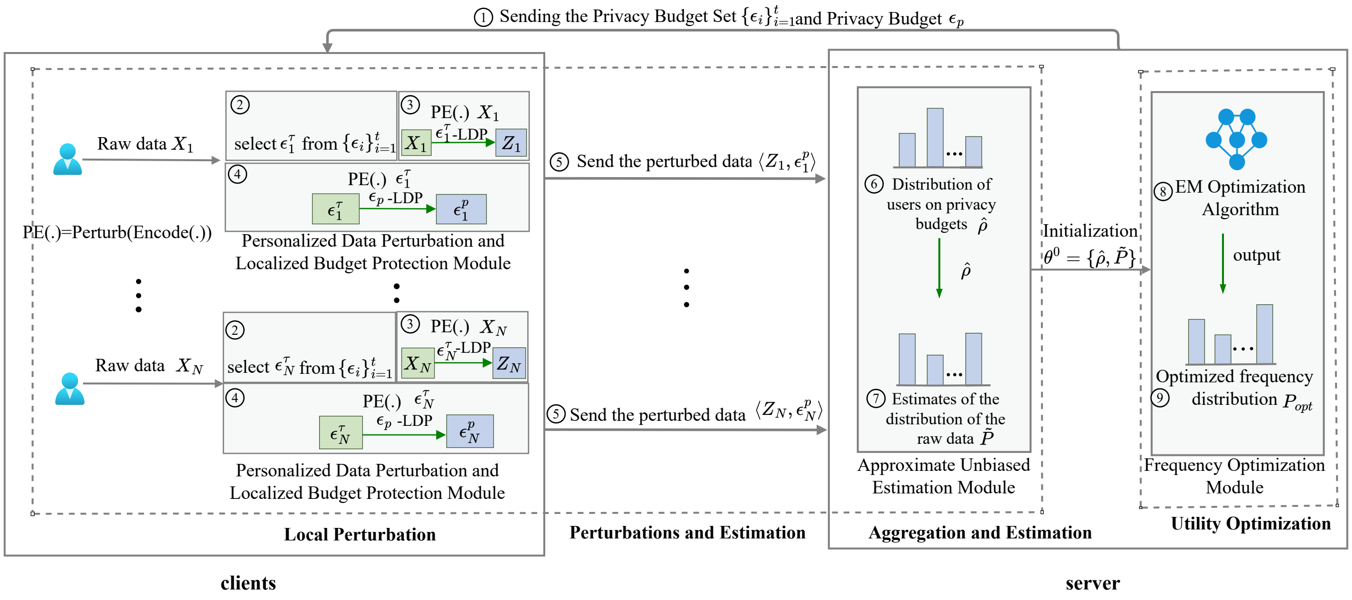

① Sending the Privacy Budget Set $\{\epsilon_i\}_{i=1}^t$and Privacy Budget ϵ_p
Raw data X_1
② select ϵ_1^τ from $\{\epsilon_i\}_{i=1}^t$
③ PE(.) X_1
ϵ_1^τ-LDP
X_1
Z_1
④ PE(.) ϵ_1^τ
ϵ_p -LDP
ϵ_1^τ
ϵ_1^p
PE(.)=Perturb(Encode(.))
Personalized Data Perturbation and Localized Budget Protection Module
Raw data X_N
② select ϵ_N^τ from $\{\epsilon_i\}_{i=1}^t$
③ PE(.) X_N
ϵ_N^τ-LDP
X_N
Z_N
④ PE(.) ϵ_N^τ
ϵ_p -LDP
ϵ_N^τ
ϵ_N^p
Personalized Data Perturbation and Localized Budget Protection Module
Local Perturbation
⑤ Send the perturbed data $\langle Z_1, \epsilon_1^p \rangle$
⑤ Send the perturbed data $\langle Z_N, \epsilon_N^p \rangle$
Perturbations and Estimation
⑥ Distribution of users on privacy budgets $\hat{\rho}$
$\hat{\rho}$
⑦ Estimates of the distribution of the raw data $\tilde{P}$
Approximate Unbiased Estimation Module
Aggregation and Estimation
Initialization
$\theta^0 = \{\hat{\rho}, \tilde{P}\}$
⑧ EM Optimization Algorithm
output
Optimized frequency
⑨ distribution P_{opt}
Frequency Optimization Module
Utility Optimization
clients
server

Figure 1: PLDP Frequency Estimation with Budget Security Framework(PLDP-FEBSF)

(1) **Select privacy budget:** Users choose a suitable privacy budget, denoted as ϵ_u^τ, from the given privacy budgets set $\boldsymbol{\varepsilon} = \{\epsilon_i\}_{i=1}^t$. This choice depends on their desired level of privacy protection.
(2) **Encode data:** Take the original data X_u as the input and then output the encoded data Y_u. This step is independent of the privacy budget selection and is solely related to the specific LDP protocol. For instance, Basic-RAPPOR uses onehot encoding.
(3) **Perturb data:** Based on the privacy budget ϵ_u^τ selected in the first step, calculate the perturbation probabilities p^* and q^* according to Eq. (1), denoted as $p_{\epsilon_u^\tau}$ and $q_{\epsilon_u^\tau}$, respectively. Then, perturb the encoded data Y_u to generate the perturbed data Z_u. This perturbation process satisfies ϵ_u^τ-LDP.
(4) **Encode privacy budget:** Similar to step (2), take ϵ_u^τ as input and output the encoded result ϵ_u^c. Users can choose a different LDP protocol to protect the privacy budget ϵ_u^τ compared to the one used for perturbing the data.
(5) **Perturb privacy budget:** Similar to step (3), based on the privacy budget ϵ_p, calculate the perturbation probabilities p^* and q^* according to Eq. (1), denoted as p_{ϵ_p} and q_{ϵ_p}. Then, perturb the encoded privacy budget ϵ_u^c from step 4 to generate the perturbed privacy budget ϵ_u^p. This process satisfies ϵ_p-LDP.

After completing the five steps on the user side, the perturbed privacy budget ϵ_u^p and the perturbed data Z_u are sent to the server side. It is essential to highlight that the choice of LDP methods for data perturbation and user privacy budget perturbation can differ in our framework.

In order to help readers better understand the overall operation process of the framework on the user side, we show the specific execution process through the example "**Example 1**".

Example 1: Consider this scenario: suppose the privacy budget set is $\{\epsilon_1, \epsilon_2, \epsilon_3\}$, and the privacy budget for protecting users' privacy budgets is ϵ_p. Two users, user1 and user2, have the same data item d_1 from the data set $\{d_1, d_2, d_3\}$, but they have different privacy protection needs.We assume data is perturbed with OUE [38], and the LDP protocol for protecting users' privacy budgets is k-RR [22]. They each locally execute the following five steps. **Select Privacy Budget:** user1 chooses ϵ_1, user2 chooses ϵ_2. **Encode Data :** Both user1 and user2 encode their data as [1,0,0]. **Perturb Data :** user1 calculates the perturbation probabilities as $p_{\epsilon_1^1} = 1/2, q_{\epsilon_1^1} = 1/(1+e^{\epsilon_1})$; user2 calculates the perturbation probabilities as $p_{\epsilon_2^2} = 1/2, q_{\epsilon_2^2} = 1/(1+e^{\epsilon_2})$. Suppose the perturbed data are [1,0,1] and [1,1,0]. **Encode Privacy Budget :** user1 encodes the privacy budget ϵ_1 as ϵ_1, and user2 encodes ϵ_2 as ϵ_2. **Perturb Privacy Budget:** Both user1 and user2 calculate the perturbation probabilities as $p_{\epsilon_p} = e^{\epsilon_p}/(e^{\epsilon_p}+2), q_{\epsilon_p} = 1/(e^{\epsilon_p}+2)$. Suppose the final perturbed privacy budget are ϵ_2 and ϵ_3. Finally, user1 and user2 send the perturbed data to the server as $\langle[1,0,1], \epsilon_2\rangle$ and $\langle[1,1,0], \epsilon_3\rangle$, respectively.

5.2 Approximate Unbiased Estimation with Protected Privacy Budgets

When the server side receives the perturbed data sent from the client side, in traditional LDP scenarios, if the local perturbation satisfies Pure LDP, the server can use Eq. (2) to achieve an unbiased estimation of the original data frequency distribution. However, in the context of PLDP, each user can choose different privacy budgets according to their privacy needs to perturb their private data. Therefore, unlike traditional LDP, when the local perturbation satisfies Pure LDP, we can use the following Eq. (3) to estimate the original data distribution,

$$\hat{p}_i = \frac{\sum_u \mathbb{I}_{support(Z_u)}(x_i) - Nq^\star}{N(p^\star - q^\star)}, \tag{3}$$

here, N represents the total number of users, and $\hat{p}_i$ is the estimated frequency of the data item x_i. Z_u is the perturbed encoding data output of user u. The function $\mathbb{I}_{Support(Z_u)}(x_i)$ outputs 1 if and only if Z_u supports the item x_i, otherwise, it outputs 0. $p^\star$ and $q^\star$ are defined as: $p^\star = \sum_{i=1}^t \rho_i p_{\epsilon_i}, \quad q^\star = \sum_{i=1}^t \rho_i q_{\epsilon_i}$. Consistent with the meanings of p^* and q^* as expressed in Eq. (1), p_{ϵ_i} represents the probability of an input value x_1 being perturbed to its own support set under the privacy budget ϵ_i, and q_{ϵ_i} represents the probability of an input value $x_2, x_2 \neq x_1$, being perturbed to the support set of x_1 under the privacy budget ϵ_i. ρ denotes the frequency of the user distribution over the privacy budget ϵ_i.

Theorem 5.1. *In PLDP, $\hat{p}_i$ is an unbiased estimate of the original frequency p_i of the data item x_i when the local perturbation satisfies the Pure LDP protocol.*

Proof. For a detailed proof, please refer to the appendix A.2. □

According to Theorem 5.1, in order to accurately estimate the frequency distribution P of the original data, we also need to understand the frequency distribution $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^t$ of the user privacy budget set $\boldsymbol{\varepsilon} = \{\epsilon_i\}_{i=1}^t$. Therefore, on the server side, the original **Aggregate** step in literature [38] is refined into two separate steps: {**Aggregate privacy budgets**, **Aggregate data**}. Among them, **Aggregate privacy budgets** unbiasedly estimate users' frequency distribution $\boldsymbol{\rho}$ on the privacy budgets set $\boldsymbol{\varepsilon}$ on perturbed privacy budgets. Denote it as $\hat{\boldsymbol{\rho}}$. **Aggregate data**, on the other hand, focuses on estimation of the frequency distribution P of the original data.

In the **Aggregate privacy budgets** step, because the user uses the local perturbation protocol that satisfies the Pure LDP protocol to perturb his actual privacy budget, combined with the properties of Pure LDP protocol [38], The server-side can provide an unbiased estimate of the distribution $\boldsymbol{\rho}$ of the user over the privacy budget set $\boldsymbol{\varepsilon}$ using Eq. (4),

$$\hat{\rho}_i = \frac{\sum_{u=1}^N \mathbb{I}_{support(\epsilon_u^p)}(\epsilon_i) - Nq_{\epsilon_p}}{N(p_{\epsilon_p} - q_{\epsilon_p})}. \tag{4}$$

where N denotes the total number of users, $\hat{\rho}_i$ is the estimated frequency of privacy budget ϵ_i, and ϵ_u^p is the perturbed privacy budget sent by user u. The indicator function $\mathbb{I}_{Support(\epsilon_u^p)}(\epsilon_i)$ outputs 1 if and only if ϵ_u^p supports ϵ_i, and 0 otherwise. The probability value p_{ϵ_p} represents the probability that the input value ϵ_1 is perturbed to its own support set. The probability value q_{ϵ_p} represents the probability that the input value ϵ_2 ($\epsilon_2 \neq \epsilon_1$) is perturbed to the support set of ϵ_1.

In the **Aggregate data** step, based on Eq. (3), and an unbiased estimate $\hat{\boldsymbol{\rho}}$ of the distribution $\boldsymbol{\rho}$ over the privacy budgets $\boldsymbol{\varepsilon}$ obtained in the **Aggregate privacy budgets** step, and the perturbed data set $\{Z_u\}_{u=1}^N$ sent by the users, the approximation of the unbiased estimation of the original data frequency distribution P can be obtained through Eq. (5),

$$\tilde{p}_i = \frac{\sum_u \mathbb{I}_{support(Z_u)}(x_i) - N\hat{q}^\star}{N(\hat{p}^\star - \hat{q}^\star)}.$$

Where N represents the total number of users and $\tilde{P}_i$ represents the estimated value of the frequency of the data item x_i. Z_u denotes perturbed data sent by the user, and the indicator function $\mathbb{I}_{Support(Z_u)}(x_i)$ takes the value 1 if and only if Z_u supports x_i, and outputs 0 otherwise. The expressions for the parameters $\hat{p}^\star$ and $\hat{q}^\star$ are

$$\hat{p}^\star = \sum_{i=1}^t \hat{\rho}_i p_{\epsilon_i}, \quad \hat{q}^\star = \sum_{i=1}^t \hat{\rho}_i q_{\epsilon_i}.$$

Here, the probability value p_{ϵ_i} represents the probability of perturbing the input value x_1 to its support set with a privacy budget of ϵ_i. The probability q_{ϵ_i} represents the probability of perturbing the input value $x_2 (x_1 \neq x_2)$ to the support set of x_1 given a privacy budget of ϵ_i. $\hat{\rho}_i$ denotes an unbiased estimate of the distribution frequency ρ_i of the privacy budget ϵ_i.

Theorem 5.2. *$\tilde{p}_i$ is an approximately unbiased estimate of the original frequency p_i of the data item x_i.*

Proof. For a detailed proof, please refer to the appendix A.3. □

5.3 Utility Analysis

Variance analysis can help us better understand the accuracy of frequency distribution estimation and the degree of deviation in the real result.A lower variance indicates that the estimates are concentrated and have high precision, whereas a higher variance suggests that the estimations are more deviation. Additionally, variance analysis is helpful in identifying the sources of estimation error. By further analyzing these errors, we can improve the estimation methods and thereby enhance the overall accuracy of the estimation results. Due to the complexity of the expression of the estimation results, direct variance analysis is impractical. Therefore, we use approximate variance to analyze the accuracy and usability of the estimation results. We have formalized the process of approximate variance analysis into Theorem 5.3. Below is the detailed content of Theorem 5.3.

Theorem 5.3. *Let the approximate variance of the frequency estimation value $\tilde{p}_i$ be given by*

$$\mathrm{Var}^*(\tilde{p}_i) = V_1 + V_2 + V_3 + V_4,$$

where the parameters V_1, V_2, V_3, V_4 are respectively defined as follows:

$$A = \left[(1 - p_i)\sum_{j=1}^t \rho_j q_{\epsilon_j} + p_i \sum_{j=1}^t \rho_j p_{\epsilon_j}\right],$$

$$B_i = \frac{N\left[(q_{\epsilon_p} + \rho_i(p_{\epsilon_p} - q_{\epsilon_p})(1 - q_{\epsilon_p} - \rho_i(p_{\epsilon_p} - q_{\epsilon_p})\right]}{N^2(p_{\epsilon_p} - q_{\epsilon_p})^2},$$

$$V_1 = \frac{N(A(1 - A))}{(\sum_{i=1}^t \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^2},$$

$$V_2 = \frac{(NA)^2 \sum_{i=1}^t B_i(p_{\epsilon_i} - q_{\epsilon_i})^2}{(\sum_{i=1}^t \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^4},$$

$$V_3 = \frac{\sum_{i=1}^t B_i q_{\epsilon_i}^2}{(\sum_{i=1}^t \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^2},$$

$$V_4 = \frac{(\sum_{i=1}^{t} \rho_i q_{\epsilon_i})^2 (\sum_{i=1}^{t} B_i (p_{\epsilon_i} - q_{\epsilon_i})^2)}{(\sum_{i=1}^{t} \rho_i (p_{\epsilon_i} - q_{\epsilon_i}))^4},$$

The parameters p_{ϵ_p} and q_{ϵ_p} respectively represent the probabilities of perturbing the privacy budget, and their values are related to the privacy budget ϵ_p used for protecting the user's privacy budget.

Proof. For a detailed proof, please refer to the appendix A.4. □

By Theorem 5.3, the upper bound of the approximate variance for the estimate $\tilde{p}_i$ is given by $O(\mathrm{Var}^*(\tilde{p}_i)) = O(\max(V_1, V_2, V_3, V_4))$. Further analysis of the expressions V_1, V_2, V_3, V_4 reveals that V_1 is independent of the values of p_{ϵ_p} and q_{ϵ_p}. Since p_{ϵ_p} and q_{ϵ_p} are related to the privacy budget ϵ_p, V_1 is independent of the size of the privacy budget ϵ_p and is only related to the strength of the user's perturbation of the data. Therefore, when $V_1 = \max(V_1, V_2, V_3, V_4)$, the upper bound for the variance of the estimation result is determined by the strength of the data perturbation imposed by the user.

In addition, it can be observed that the expression V_2, V_3, V_4 vary inversely with $(p_{\epsilon_p} - q_{\epsilon_p})$, and the privacy budget ϵ_p tends to 0. $p_{\epsilon_p} - q_{\epsilon_p}$ also tends to 0. In this case, the value of V_2, V_3, V_4 tends to infinity, $O(\mathrm{Var}^*(\tilde{p}_i))$ also tends to infinity and $V_1 \neq \max(V_1, V_2, V_3, V_4)$. The user's perturbation for his privacy budget becomes the main source of error.

6 UTILITY OPTIMIZATION

In Section 5.3, we performed an approximate Variance analysis on the estimated results $\tilde{p}_i$. By analyzing its theoretical expression in detail, we find that the upper bound of the approximate variance is affected by the value of the privacy budget ϵ_p. Specifically, when $\epsilon_p \longrightarrow 0$, the approximate variance of the model goes to infinity. In this case, protecting the privacy budget becomes the main source of error in estimation results. This finding is contrary to our original intention of improving the accuracy of estimation results under the premise of protecting the user privacy. Therefore, the further work of this paper will explore how to improve the accuracy of the estimation results and the data utility when the privacy budget ϵ_p is small.

Further analysis of the perturbation process of the local user u for its privacy data X_u and privacy budget ϵ_u^τ in Section 5.1 shows that these are two independent processes. Significantly, the perturbation of the private data X_u is not affected by the privacy budget ϵ_p. In the following, our aim is to mine the original data frequency distribution P from the perturbed data, particularly when the privacy budget ϵ_p is small, in order to enhance the accuracy of estimation.

6.1 Modeling Optimization Objective

It is well known that the original data distribution P is processed by a random perturbation algorithm satisfying ϵ-LDP to generate the perturbed distribution P_ϵ.The original distribution P is perturbed with$t$ distinct privacy budgets in $\boldsymbol{\varepsilon}$, resulting in the corresponding t perturbed distributions $\{P_{\epsilon_i}\}_{i=1}^t$. Therefore, we can transform the original problem into a problem of multi-distribution mixture generation.

The perturbed data set $\{Z_u\}_{u=1}^N$ received on the server side can be seen as generated by this mixture of t distributions. When the server receives the perturbed data Z_u sent by the client, data Z_u can be generated by any of the distributions in the set $\{P_{\epsilon_i}\}_{i=1}^t$. Without loss of generality, we assume that the distribution of users over the privacy budget set $\boldsymbol{\varepsilon}$ is $\boldsymbol{\rho}$. Therefore, the probability that data item Z_u comes from distribution P_{ϵ_i} is ρ_i. The observation probability of the data Z_u is given by

$$\begin{aligned}\Pr(Z_u) &= \sum_{i=1}^{t} \Pr(P_{\epsilon_i}) \Pr(Z_u | P_{\epsilon_i}) \\ &= \sum_{i=1}^{t} \rho_i \Pr(Z_u | P_{\epsilon_i}),\end{aligned}$$

where $\Pr(Z_u|P_{\epsilon_i})$ denotes the probability of observing Z_u under the condition that the distribution is P_{ϵ_i}. Assuming that the frequency distribution of the original data is $P = \{p_1, \ldots, p_k\}$, according to LDP protocol, it can be shown that $\Pr(Z_u|P_{\epsilon_i})$ can be expressed as:

$$\Pr(Z_u | P_{\epsilon_i}) = \sum_{j=1}^{k} p_j \Pr(Z_u | x_j, \epsilon_i),$$

where $\Pr(Z_u|x_j, \epsilon_i)$ denotes the probability of perturbing from x_j to Z_u, given the data domain $\mathcal{D} = \{x_i\}_{i=1}^k$ and the privacy budget ϵ_i. Simultaneously, according to Section 5.1, from the **Perturb data** step, we have

$$\Pr(Z_u | x_j, \epsilon_i) = \mathbb{I}_{support(Z_u)}(x_j) p_{\epsilon_i} + (1 - \mathbb{I}_{support(Z_u)}(x_j)) q_{\epsilon_i}. \quad (5)$$

From above analysis, the final maximum likelihood [29] function Eq. (6) can be obtained when the perturbed data $Z = \{Z_u\}_{u=1}^N$ from all users,

$$\begin{aligned}\mathcal{L} &= \prod_{u=1}^{N} \Pr(Z_u) \\ &= \prod_{u=1}^{N} \left[\sum_{i=1}^{t} \sum_{j=1}^{k} \rho_i p_j \Pr(Z_u | x_j, \epsilon_i) \right]. \end{aligned} \quad (6)$$

Now, our objective is to solve for the optimal parameter $\{p_1, p_2, \ldots, p_k, \rho_1, \ldots, \rho_t\}$, such that the objective function $\mathcal{L}$ attains its maximum value. However, the complexity of this objective function makes it impractical to find an analytical solution for the parameters. Therefore, we employ the Expectation-Maximization (EM) [45] algorithm for parameter estimation.

To solve the objective function $\mathcal{L}$ using the EM algorithm, the most important task is to construct an iterative solution equation of the following form:

$$\underset{\theta}{arg\,max} \quad Q(\theta, \theta^{(s)}),$$

$$\text{s.t.} \quad \begin{cases} p_i \in [0,1], & i = 1, \ldots, k, \\ \rho_j \in [0,1], & j = 1, \ldots, t, \\ \sum_{i=1}^{k} p_i = 1, \\ \sum_{j=1}^{t} \rho_j = 1. \end{cases} \quad (7)$$

In the Eq. (7), the parameter θ denotes the set of parameters to be solved for, $\{p_1, p_2, \ldots, p_k, \rho_1, \ldots, \rho_t\}$. Since the parameters $\{p_1, p_2, \ldots, p_k\}$ denote the distribution of the original data, while the parameter $\{\rho_1, \ldots, \rho_t\}$ denotes the distribution of users in the privacy budget set $\boldsymbol{\varepsilon}$, and thus they need to satisfy the constraints in Eq. (7).

Now our goal is to transform the likelihood function $\mathcal{L}$ into an iterative function $Q(\theta, \theta^{(s)})$. Since the parameter set is denoted by

θ, the Eq. (6) can be re-expressed as:

$$\mathcal{L}(\theta) = \prod_{u=1}^{N} \Pr(Z_u|\theta) = \Pr(Z|\theta), \tag{8}$$

the parameter Z denotes the set of perturbed data, defined as $Z = \{Z_u\}_{u=1}^{N}$, where Z_u denotes a single piece of data. Since the problem we study is a multi-distribution mixture generation problem, the objective function $\Pr(Z|\theta)$ is re-expressed as:

$$\underset{\theta}{arg\,max}\Pr(Z|\theta) = \underset{\theta}{arg\,max}\; \mathbb{E}_{\Pr(P_\epsilon|\theta)}[\Pr(Z, P_\epsilon|\theta)], \tag{9}$$

where, the expectation operation essentially integrates P_ϵ in the joint distribution $\Pr(Z, P_\epsilon|\theta)$ to yield the marginal distribution $\Pr(Z|\theta)$. Thus, the two are equivalent. Since we have no a priori knowledge of the probability of the distribution P_{ϵ_i}, the posterior probability is used instead. Furthermore, replace $\Pr(P_\epsilon|\theta)$ in Eq. (9) with $\Pr(P_\epsilon|\theta^{(s)})$, Where $\theta^{(s)}$ denotes the parameter solution obtained in the previous iteration.Due to the monotonicity of the $\log(\cdot)$ function and its property of transforming multiplication into addition, we can simplify the complexity of the objective function and derive the following the objective function:

$$\begin{aligned}\underset{\theta}{arg\,max}\; Q(\theta, \theta^{(s)}) &= \sum_{i=1}^{t}\left[\log\prod_{u=1}^{N}\Pr(Z_u, P_{\epsilon_i}|\theta)\right]\prod_{u=1}^{N}\Pr(P_{\epsilon_i}|Z_u, \theta^{(s)}) \\ &= \sum_{i=1}^{t}\left[\sum_{u=1}^{N}\log\Pr(Z_u, P_{\epsilon_i}|\theta)\right]\prod_{u=1}^{N}\Pr(P_{\epsilon_i}|Z_u, \theta^{(s)}).\end{aligned} \tag{10}$$

After simplifying the previously mentioned equation (10), we derived the ultimate expression for the objective function $Q(\theta, \theta^{(s)})$(11) in Eq. (7)as follows,

$$\underset{\theta}{arg\,max}\; Q(\theta, \theta^{(s)}) = \sum_{u=1}^{N}\sum_{i=1}^{t}\log\Pr(Z_u, P_{\epsilon_i}|\theta)\,\Pr(P_{\epsilon_i}|Z_u, \theta^{(s)}). \tag{11}$$

6.2 Solving Optimization Objective

In this subsection, our goal is to solve the EM optimization Eq. (7) presented in Section 6.1. Since the EM algorithm finds the optimal solution through an iterative approach, in the initial stage, we use the parameter values $\{\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_k, \hat{\rho}_1, \ldots, \hat{\rho}_t\}$, which were estimated in Section 5.2, to initialize the parameters θ, denoted as $\theta^{(0)}$. Next, a series of iterative solutions $\{\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(s-1)}, \theta^{(s+1)}\}$, such that the objective function increases gradually, i.e., it satisfies $Q(\theta^{(s-1)}, \theta) \le Q(\theta^{(s)}, \theta)$. Eventually, the iteration is stopped when the iteration termination condition is reached.

In the **E-step**, we utilize the parameter values $\theta^{(s)}$ obtained in step s to compute the posterior probability of the distribution P_{ϵ_i},

$$\begin{aligned}\Pr(P_{\epsilon_i}|Z_u, \theta^{(s)}) &= \frac{\Pr(P_{\epsilon_i}, Z_u|\theta^{(s)})}{\Pr(Z_u|\theta^{(s)})} \\ &= \frac{\Pr(P_{\epsilon_i})\Pr(Z_u|P_{\epsilon_i}, \theta^{(s)})}{\sum_{j=1}^{t}\Pr(P_{\epsilon_j})\Pr(Z_u|P_{\epsilon_j}, \theta^{(s)})} \\ &= \frac{\rho_i^{(s)}\sum_{j=1}^{k}p_j^{(s)}\Pr(Z_u|x_j, \epsilon_i)}{\sum_{i=1}^{t}\rho_i^{(s)}\sum_{j=1}^{k}p_j^{(s)}\Pr(Z_u|x_j, \epsilon_i)}.\end{aligned} \tag{12}$$

In **M-step**, we use the a posteriori probability values computed in **E-step** to update the EM Eq. (11) and solve the parameter set $\theta^{(s+1)}$ by maximizing the objective function (7). For the solution of the objective function (7), since this is an optimization problem with constraints, we use the Lagrange multiplier method. Specifically, we treat the objective function as follows:

$$\begin{aligned}\underset{\theta}{\text{argmax}}\; L(\theta, \lambda_1, \lambda_2|\theta^s) = Q(\theta, \theta^{(s)}) + \lambda_1\left(\sum_{j=1}^{t}\rho_j - 1\right) \\ + \lambda_2\left(\sum_{i=1}^{k}p_i - 1\right).\end{aligned} \tag{13}$$

Meanwhile, for the joint conditional probability $\Pr(Z_u, P_{\epsilon_i}|\theta)$ in the Eq. (11), it can be expanded according to the chain rule [46] as:

$$\Pr(Z_u, P_{\epsilon_i}|\theta) = \Pr(P_{\epsilon_i}|\theta)\Pr(Z_u|P_{\epsilon_i}, \theta).$$

Also, since the conditional probability $\Pr(P_{\epsilon_i}|\theta)$ denotes the probability that the distribution is P_{ϵ_i} given the set of parameters θ, it follows that $\Pr(P_{\epsilon_i}|\theta) = \rho_i$. Thus, we can derive the final form of $L(\theta, \lambda_1, \lambda_2|\theta^s)$:

$$\begin{aligned}L(\theta, \lambda_1, \lambda_2|\theta^s) = &\sum_{u=1}^{N}\sum_{i=1}^{t}\log\left[\rho_i\sum_{j=1}^{k}p_j\Pr(Z_u|x_j, \epsilon_i)\right] \\ &\times\Pr(P_{\epsilon_i}|Z_u, \theta^s) \\ &+ \lambda_1\left(\sum_{j=1}^{t}\rho_j - 1\right) + \lambda_2\left(\sum_{i=1}^{k}p_i - 1\right).\end{aligned}$$

Furthermore, due to the convexity of the function $\log(\cdot)$, we can derive the following inequality:

$$\begin{aligned}L(\theta, \lambda_1, \lambda_2|\theta^s) \ge &\sum_{u=1}^{N}\sum_{i=1}^{t}\Bigg[\log(\rho_i)\Pr(P_{\epsilon_i}|Z_u, \theta^s) \\ &+ \left(\sum_{j=1}^{k}\log(p_j)\Pr(Z_u|x_j, \epsilon_i, \theta)\right)\Pr(P_{\epsilon_i}|Z_u, \theta^{(s)})\Bigg] \\ &+ \lambda_1\left(\sum_{j=1}^{t}\rho_j - 1\right) + \lambda_2\left(\sum_{i=1}^{t}p_i - 1\right) \\ = &\,\mathcal{L}(\theta, \lambda_1, \lambda_2|\theta^s).\end{aligned}$$

By taking the partial derivative of $\mathcal{L}(\theta, \lambda_1, \lambda_2|\theta^{(s)})$ with respect to the parameter $\rho_j \in \theta$ and setting it to zero, we can solve for:

$$\rho_i = \frac{1}{N}\sum_{u=1}^{N}\Pr(P_{\epsilon_i}|Z_u, \theta^{(s)}). \tag{14}$$

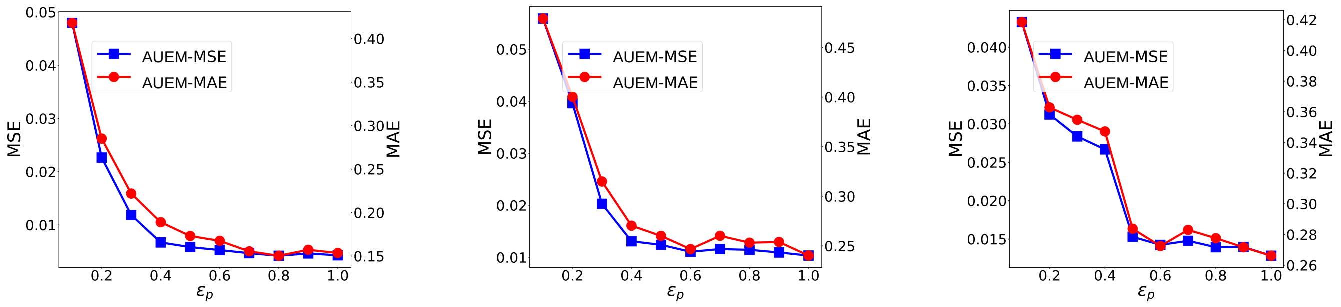
Similarly, for the parameter $p_i \in \theta$, we can obtain:

$$p_j = \frac{1}{N}\sum_{u=1}^{N}\sum_{i=1}^{t}\Pr(Z_u|x_j, \epsilon_i)\Pr(P_{\epsilon_i}|Z_u, \theta^{(s)}). \tag{15}$$

For the value of $\Pr(P_{\epsilon_i}|Z_u, \theta^{(s)}$, please refer to Eq. (12). After the aforementioned derivation and analysis, we can obtain the complete process of Expectation-Maximization Optimization Algorithm, with the detailed iterative steps outlined in Algorithm 1.

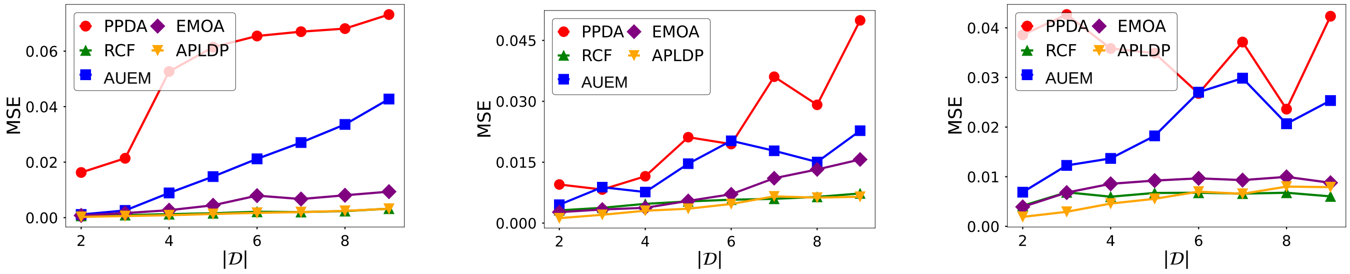
7 EXPERIMENTAL EVALUATION

To show the performance of our proposed method, we evaluated it on real datasets. We extensively compared existing PLDP-based frequency estimation schemes (e.g., PBP [35], RCF [31], APLDP [34], and PPDA [37]). For a detailed description of these comparison



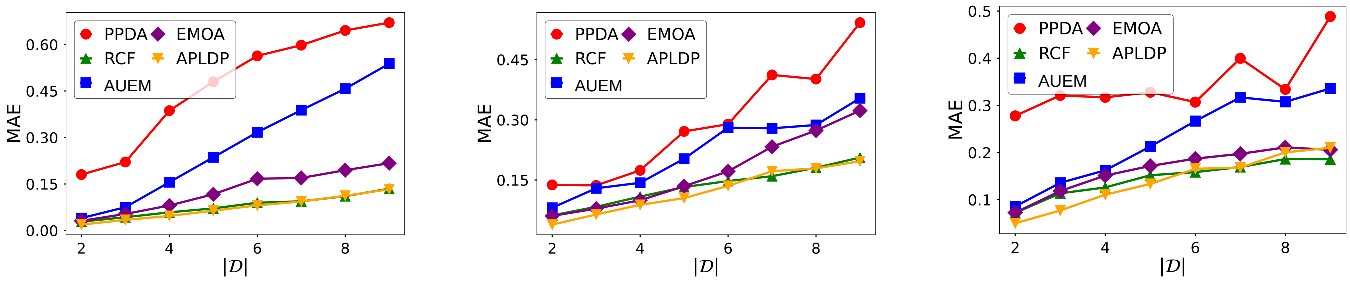
(a) Education **(b) Job** **(c) Country**

Figure 2: The performance of AUEM in terms of MSE and MAE for different values of the privacy budget ϵ_p.



(a) Education **(b) Job** **(c) Country**

Figure 3: Performance of PPDA, RCF, AUEM, EMOA, and APLDP in terms of MSE across different domain sizes $|\mathcal{D}|$.



(a) Education **(b) Job** **(c) Country**

Figure 4: Performance of PPDA, RCF, AUEM, EMOA, and APLDP in terms of MAE across different domain sizes $|\mathcal{D}|$.

schemes, please refer to the Appendix A.5. The detailed experimental results are presented below.

7.1 Experimental Environment and Parameter Settings

All experiments were conducted on a Windows personal computer equipped with a 2.9 GHz Intel Core i7 processor and 32GB of RAM. Firstly, to achieve precise control over the experimental parameters, we assumed that the intervals $\Delta\epsilon$ between adjacent privacy budgets in the given set ε were equal. Secondly, to evaluate the impact of privacy budget protection on the final experimental results, we set the initial value of the privacy budget ϵ_p to 0.1, and varied it from 0.1 to 1.0 in increments of 0.1, systematically investigating the effect of ϵ_p on the outcomes.

Given the diverse privacy protection needs of users in real-world scenarios, we selected two typical distributions for our experiments: Uniform(Uni) distribution and Power-Law(PL) distribution. Additionally, we discuss the impact of data domain size $|\mathcal{D}|$ on the experimental results.

For the choice of local perturbation algorithms, we followed the methodology of comparative schemes APLDP and PPDA. In APLDP, the authors used k-RR and Basic-RAPPOR to perturb privacy data, while in PPDA, the one-hot encoding local perturbation algorithm was required. To eliminate the influence of different perturbation algorithms on the experimental results, we uniformly

Algorithm 1: Expectation-Maximization Optimization Algorithm

```
Input   : Initial values $\theta^{(0)}$, Convergence threshold $maxgap$,
          User perturbed dataset $\{Z_u\}_{u=1}^N$, Maximum
          iteration steps $maxsteps$
Output  : Optimal frequency distribution estimation result
          $P_{opt} = \{p_i\}_{i=1}^k$
Initialize: $flag \leftarrow 0$, $steps \leftarrow 0$;
while $flag = 0$ and $steps < maxsteps$ do
    $steps \leftarrow steps + 1$;
    for $i = 1$ to $t$ do
        $\rho_i^{(steps)} \leftarrow \frac{1}{N} \sum_{u=1}^N \Pr(P_{\epsilon_i}|Z_u, \theta^{(steps-1)})$;
    for $j = 1$ to $k$ do
        $A(u, j) \leftarrow \Pr(Z_u|x_j, \epsilon_i, \theta) \cdot \Pr(P_{\epsilon_i}|Z_u, \theta^{(steps-1)})$;
        $p_j^{(steps)} \leftarrow \frac{1}{N} \sum_{u=1}^N \sum_{i=1}^t A(u, i)$;
    $gap \leftarrow \left|Q(\theta, \theta^{(steps)}) - Q(\theta, \theta^{(steps-1)})\right|$;
    if $gap < maxgap$ then
        $flag \leftarrow 1$;
return Optimal estimation result $P = \{p_i\}_{i=1}^k$;
```

adopted Basic-RAPPOR for perturbing privacy data in our experiments. Additionally, to demonstrate the generality of our proposed framework, we chose the k-RR algorithm to protect users' privacy budgets.

To provide a more intuitive demonstration of the actual performance of each method, we employed Mean Squared Error (MSE) [39] and Mean Absolute Error (MAE) [17] as evaluation metrics.

7.2 Datasets

In this study, we utilized three distinct datasets for the experiments. The first dataset, "Education," was extracted from the Education attribute of the adult dataset [15]. The second dataset, "Job," was derived from the Job attribute of the bank marketing dataset [47]. The third dataset ,"Country," [36] was obtained from the native-country attribute of the U.S. Census database. Detailed information about these three datasets is presented in Table 1.

Table 1: Detailed Information of the Datasets

Dataset	Values Numbers	Users Numbers
Education	16	32,561
Job	9	45,211
Country	42	48,842

7.3 Experimental Results

In the experimental section, we present the performance of various methods. To show the fairness, we compare our method with existing schemes under the same dataset and parameter settings to evaluate its effectiveness. In our experiments, we explore the impact of different parameter settings on the estimation results, including the effect of the privacy budget ϵ_p used to protect user privacy, the impact of the data domain size, and the distribution of users across different privacy budget levels. For clarity, we call the output of our approximate unbiased estimation method as "AUEM" and the output of the EM optimization algorithm as "EMOA".

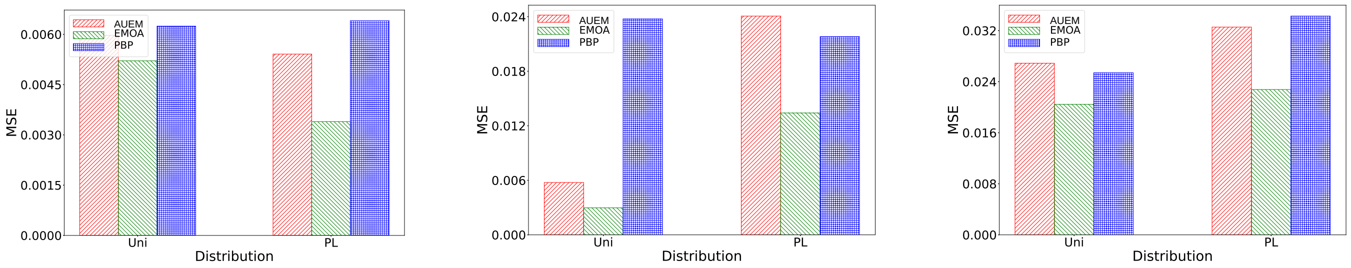
7.3.1 Impact of Privacy Budget ϵ_p. In our method, we protect user privacy budgets by introducing an additional privacy budget, ϵ_p. Furthermore, an approximate variance analysis of the approximate unbiased estimation result reveals that when the value of ϵ_p is small, the perturbation of users' privacy budgets becomes the primary source of estimation error. To verify this conclusion, in this experimental section, we set the privacy budget collection to $\boldsymbol{\varepsilon} = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and assume a uniform distribution of users across these privacy budgets. After these configurations, we set the value range of the privacy budget ϵ_p from 0.1 to 1.0, starting at 0.1 and increasing in intervals of 0.1, resulting in ten different values for our experiments. The experimental results are shown in Figures 2.

The experimental results presented in Figures 2 clearly indicate that, under both the MSE and MAE metrics, the estimation error of AUEM decreases as the privacy budget ϵ_p increases. Furthermore, an analysis of the slope of the experimental result curves reveals that when the privacy budget ϵ_p is less than 0.4, the curve is relatively steep. This suggests that, when ϵ_p is below 0.4, the impact on the accuracy of the AUEM's output is significantly heightened, making the perturbation of user privacy budgets become the primary source of estimation error. Conversely, when the privacy budget ϵ_p exceeds 0.4, the curve flattens, indicating that the impact of perturbing user privacy budgets on the accuracy of the AUEM's output is relatively minor.

These experimental findings further corroborate our usability analysis conclusion that when the privacy budget ϵ_p is small, the perturbation of users' privacy budgets becomes the primary source of estimation error in the AUEM's output, necessitating optimization.

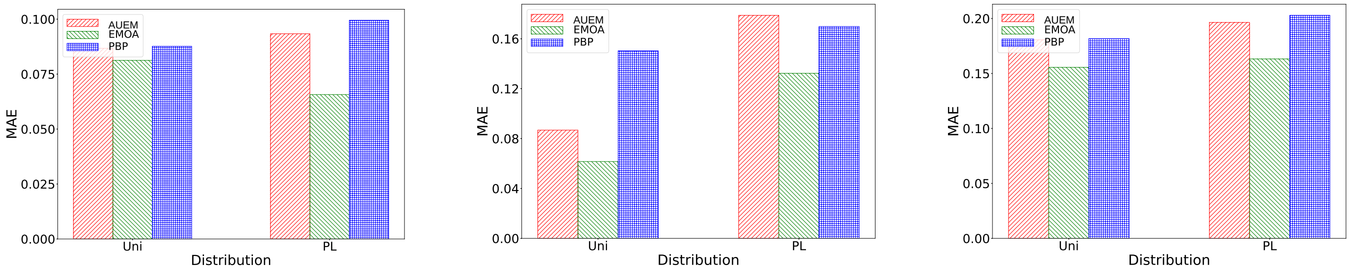
7.3.2 The Impact of Data Domain Size $|\mathcal{D}|$. Here, we primarily investigates the impact of the size of the data attribute domain, denoted as $|\mathcal{D}|$, on the experimental results. To comprehensively study the influence of $|\mathcal{D}|$ on the outcomes, the experimental parameters are set with a privacy budget set $\boldsymbol{\varepsilon} = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. According to the experimental results presented in Section 7.3.1, when the privacy budget ϵ_p is less than 0.4, it significantly impacts the AUEM's output. Therefore, to demonstrate the optimization effect of the Expectation Maximization algorithm proposed in the utility optimization on the AUEM's output, the privacy budget ϵ_p is set to 0.3. The distribution of users over the privacy budget set is configured as the commonly observed uniform distribution. The experimental results are shown in Figures 3 and 4.

First, as observed from Figures 3 and 4, with the increase in the size of the data attribute domain $|\mathcal{D}|$, the MSE and MAE values for the schemes AUEM, EMOA, RCF, PPDA, and APLDP also increase correspondingly. This indicates that as $|\mathcal{D}|$ increases, the estimation error for the original data frequency distribution P also increases.



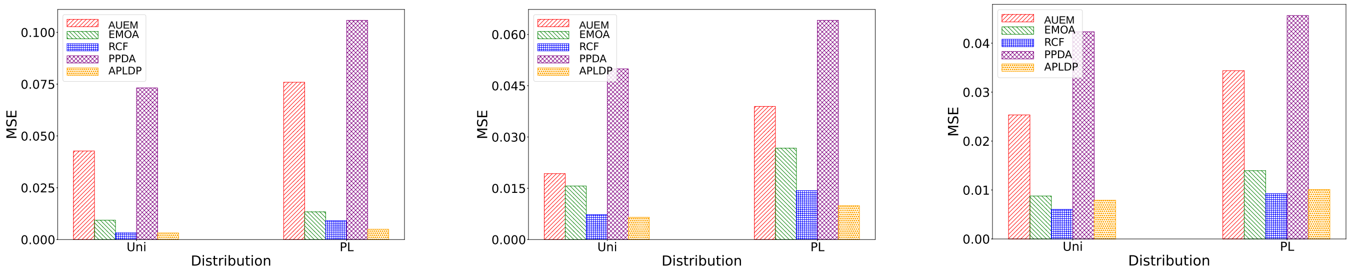
(a) Education **(b) Job** **(c) Country**

Figure 5: The performance of the PBP, AUEM, and EMOA schemes in terms of MSE under different user distributions.



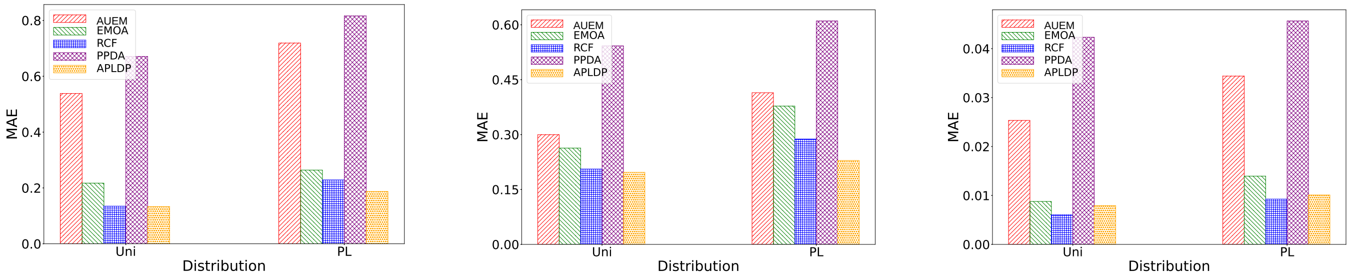
(a) Education **(b) Job** **(c) Country**

Figure 6: The performance of the PBP, AUEM, and EMOA schemes in terms of MAE under different user distributions.



(a) Education **(b) Job** **(c) Country**

Figure 7: Performance of AUEM, EMOA, RCF, PPDA, and APLDP Schemes under Different User Distributions in Terms of MSE.



(a) Education **(b) Job** **(c) Country**

Figure 8: Performance of AUEM, EMOA, RCF, PPDA, and APLDP Schemes under Different User Distributions in Terms of MAE.

Second, by comparing different schemes, it is evident that the proposed AUEM and EMOA schemes demonstrate significant advantages in terms of data usability and estimation accuracy on real-world datasets compared to the PPDA scheme. Furthermore, from the experimental results shown in Figures 3 and 4, it can be observed that the RCF and APLDP schemes outperform the proposed AUEM and EMOA schemes across the three real-world datasets. This is primarily because the RCF and APLDP schemes do not consider the protection of user privacy budgets, thereby reducing the estimation errors caused by the protection of user privacy budgets.

Finally, from the experimental results of AUEM and EMOA in Figures 3 and 4, it is evident that EMOA significantly optimizes AUEM when the privacy budget ϵ_p is set to a smaller value. Additionally, as seen in Figures 3 and 4, the output results after EMOA optimization are close to the experimental results of the comparison schemes RCF and APLDP, which is a sound proof that the proposed method achieves a better balance between security, accuracy of estimation results and data availability.

7.3.3 *Impact of user distribution $\boldsymbol{\rho}$ on the set of privacy budgets.*

We explore the impact of user distribution $\boldsymbol{\rho}$ on the privacy budget set $\boldsymbol{\varepsilon}$ in experimental results. In real-world scenarios, privacy needs vary among different user groups. For example, a smaller percentage of the healthy population has higher privacy requirements than the HIV population. Two representative distributions are selected for our experiments: Uniform (Uni) and Power Law (PL) distributions. Uniform distribution often represents activities like lotteries and dice rolling, while personal income follows a power law distribution, reflecting the Pareto principle. Additionally, users tend to choose stricter privacy standards than needed, leading to a preference for smaller privacy budgets. Thus, using a power law distribution simulates this behavior effectively.

The PBP comparison scheme only supports $|\boldsymbol{\varepsilon}|= 2$. For our experiments with the PBP scheme, we set $\boldsymbol{\varepsilon} = \{0.1, 0.2\}$ and $\epsilon_p = 0.1$. Since PBP requires a public user distribution over $\boldsymbol{\varepsilon}$, we randomly assigned this distribution in our experiments. Figures 5 and 6 show the PBP, AUEM, and EMOA methods' performance on MSE and MAE metrics under uniform and power law distributions.

Figures 5 and 6 demonstrate that AUEM and EMOA generally outperform PBP in MSE and MAE across three real-world datasets. This improvement is due to AUEM and EMOA's ability to estimate the user distribution $\boldsymbol{\rho}$ over the privacy budget set using perturbed budgets, enhancing estimation accuracy. Notably, EMOA achieves an 87% improvement in MSE over PBP in the uniform distribution scenario (Figure 5(b)).

Comparing the results between uniform and power law distributions, it is obvious that all methods perform better under uniform conditions. This is because users tend to choose smaller privacy budgets under power law, resulting in stronger privacy protection and higher experimental errors.

For comparisons involving PPDA, RCF, and APLDP, which support $|\boldsymbol{\varepsilon}|> 2$, we set $\boldsymbol{\varepsilon} = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and $\epsilon_p = 0.3$. AUEM, EMOA, PPDA, RCF, and APLDP were evaluated on three datasets: Education, Job, and Country, to examine the impact of user distributions $\boldsymbol{\rho}$ on the results. Figures 7 and 8 present the performance of these methods in terms of MSE and MAE under uniform and power law distributions.

Figures 7 and 8 indicate that AUEM and EMOA consistently outperform PPDA across all datasets, in both MSE and MAE metrics. The optimized EMOA scheme shows performance comparable to RCF and APLDP, which do not protect user privacy budgets, in most cases (Figures 7(a), (c), 8(a), and (c)). This validates the effectiveness of the proposed methods in terms of data utility and security.

Consistent with the PBP comparison experiments, the error is generally higher under power law distribution. Horizontally comparing the datasets shows significantly lower errors on the Country dataset compared to the Education and Job datasets, attributed to the larger data volume of the Country dataset, which reduces experimental error.

Based on the experimental results, it is evident that the AUEM and EMOA proposed in this paper outperform the PPDA and PBP on three real datasets. Additionally, in Section 7.3.1, we observe significant impact on the AUEM method's output when the privacy budget ϵ_p is small, with EMOA showing notable optimization benefits for AUEM. Furthermore, the error of our method is very close to APLDP and RCF. This is primarily because the APLDP and RCF do not protect user privacy budgets, reducing errors associated with safeguarding these budgets. The error in our method can only approach APLDP and RCF, and cannot be better than these methods.

8 CONCLUSION

In this paper, we focus on the problem of estimating the frequency distribution of discrete data based on PLDP. Unlike previous studies,this manuscript, in partiular, pays attention to the risk of leaking users' private information by user privacy budgets in PLDP scenarios. To address this problem, We propose a new frequency estimation framework called PLDP-FEBSF which is applicable to a wide range of LDP protocols. Within this framework, users protect their raw data and their private budgets to prevent inferring private information from malicious attackers. In addition, we introduce an EM algorithm to improve the utility and accuracy of the estimation results. Through comparative experiments with existing schemes on real datasets, we confirm the effectiveness and superiority of the frequency distribution frame work proposed in this paper.

In future, we aspire to broaden the applicability of our framework to diverse domains,such as mean estimation and frequency estimation of multi-dimensional data.

REFERENCES

[1] Chaabane Abdelberi, Gergely Ács, and Mohamed Ali Kâafar. 2012. You are what you like! Information leakage through users' Interests. In *19th Annual Network and Distributed System Security Symposium, NDSS 2012, San Diego, California, USA, February 5-8, 2012*. The Internet Society. https://www.ndss-symposium.org/ndss2012/you-are-what-you-information-leakage-through-users-interests

[2] Krishna Acharya, Franziska Boenisch, Rakshit Naidu, and Juba Ziani. 2024. Personalized Differential Privacy for Ridge Regression. *CoRR* abs/2401.17127 (2024). https://doi.org/10.48550/ARXIV.2401.17127 arXiv:2401.17127

[3] Mohammad Alaggan, Sébastien Gambs, and Anne-Marie Kermarrec. 2016. Heterogeneous Differential Privacy. *J. Priv. Confidentiality* 7, 2 (2016). https://doi.org/10.29012/JPC.V7I2.652

[4] Héber Hwang Arcolezi, Carlos Pinzón, Catuscia Palamidessi, and Sébastien Gambs. 2023. Frequency Estimation of Evolving Data Under Local Differential Privacy. In *Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28-31, 2023*, Julia Stoyanovich, Jens Teubner, Nikos Mamoulis, Evaggelia Pitoura, Jan Mühlig, Katja Hose, Sourav S. Bhowmick, and Matteo Lissandrini (Eds.). OpenProceedings.org, 512–525. https://doi.org/10.48786/EDBT.2023.44

[5] Dimitris Bertsimas and Vassilis Digalakis. 2023. Frequency Estimation in Data Streams: Learning the Optimal Hashing Scheme. *IEEE Trans. Knowl. Data Eng.* 35, 2 (2023), 1541–1553. https://doi.org/10.1109/TKDE.2021.3103819

[6] Luca Bonomi. 2013. Mining Frequent Patterns with Differential Privacy. *Proc. VLDB Endow.* 6, 12 (2013), 1422–1427. https://doi.org/10.14778/2536274.2536329

[7] Rui Chen, Haoran Li, A. Kai Qin, Shiva Prasad Kasiviswanathan, and Hongxia Jin. 2016. Private spatial data aggregation in the local setting. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016.* IEEE Computer Society, 289–300. https://doi.org/10.1109/ICDE.2016.7498248

[8] Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. 2018. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data.* 1655–1658.

[9] Graham Cormode, Samuel Maddock, and Carsten Maple. 2021. Frequency Estimation under Local Differential Privacy. *Proc. VLDB Endow.* 14, 11 (2021), 2046–2058. https://doi.org/10.14778/3476249.3476261

[10] Yuntao Du, Yujia Hu, Zhikun Zhang, Ziquan Fang, Lu Chen, Baihua Zheng, and Yunjun Gao. 2023. LDPTrace: Locally Differentially Private Trajectory Synthesis. *Proc. VLDB Endow.* 16, 8 (2023), 1897–1909. https://doi.org/10.14778/3594512.3594520

[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings (Lecture Notes in Computer Science)*, Shai Halevi and Tal Rabin (Eds.), Vol. 3876. Springer, 265–284. https://doi.org/10.1007/11681878_14

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

[13] Vamsidhar Enireddy, C. Karthikeyan, and D. Vijendra Babu. 2022. OneHotEncoding and LSTM-based deep learning models for protein secondary structure prediction. *Soft Comput.* 26, 8 (2022), 3825–3836. https://doi.org/10.1007/S00500-022-06783-9

[14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security.* 1054–1067.

[15] Sahil Girhepuje. 2023. Identifying and examining machine learning biases on Adult dataset. *CoRR* abs/2310.09373 (2023). https://doi.org/10.48550/ARXIV.2310.09373 arXiv:2310.09373

[16] Xiaolan Gu, Ming Li, Li Xiong, and Yang Cao. 2020. Providing Input-Discriminative Protection for Local Differential Privacy. In *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020.* IEEE, 505–516. https://doi.org/10.1109/ICDE48307.2020.00050

[17] Timothy O Hodson. 2022. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions* 2022 (2022), 1–10.

[18] Gail M Howser. 2022. *Known Personally Identifiable Information to Cloud Users.* Ph.D. Dissertation. Utica University.

[19] Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. 2019. Learning-Based Frequency Estimation Algorithms. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* OpenReview.net. https://openreview.net/forum?id=r1lohoCqY7

[20] Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. 2019. Learning-Based Frequency Estimation Algorithms. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* OpenReview.net. https://openreview.net/forum?id=r1lohoCqY7

[21] Zach Jorgensen, Ting Yu, and Graham Cormode. 2015. Conservative or liberal? Personalized differential privacy. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, Johannes Gehrke, Wolfgang Lehner, Kyuseok Shim, Sang Kyun Cha, and Guy M. Lohman (Eds.). IEEE Computer Society, 1023–1034. https://doi.org/10.1109/ICDE.2015.7113353

[22] Peter Kairouz, Kallista A. Bonawitz, and Daniel Ramage. 2016. Discrete Distribution Estimation under Local Privacy. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016 (JMLR Workshop and Conference Proceedings)*, Maria-Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. JMLR.org, 2436–2444. http://proceedings.mlr.press/v48/kairouz16.html

[23] Haoran Li, Li Xiong, Zhanglong Ji, and Xiaoqian Jiang. 2017. Partitioning-Based Mechanisms Under Personalized Differential Privacy. In *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part I (Lecture Notes in Computer Science)*, Jinho Kim, Kyuseok Shim, Longbing Cao, Jae-Gil Lee, Xuemin Lin, and Yang-Sae Moon (Eds.), Vol. 10234. 615–627. https://doi.org/10.1007/978-3-319-57454-7_48

[24] Junxu Liu, Jian Lou, Li Xiong, and Xiaofeng Meng. 2023. Personalized Differentially Private Federated Learning without Exposing Privacy Budgets. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 4140–4144. https://doi.org/10.1145/3583780.3615247

[25] Qidi Liu, Benjamin Gily, and Mable P Fok. 2021. Adaptive photonic microwave instantaneous frequency estimation using machine learning. *IEEE photonics technology letters* 33, 24 (2021), 1511–1514.

[26] Milan Lopuhaä-Zwakenberg, Zitao Li, Boris Skoric, and Ninghui Li. 2020. Improving Frequency Estimation under Local Differential Privacy. In *WPES'20: Proceedings of the 19th Workshop on Privacy in the Electronic Society, Virtual Event, USA, November 9, 2020*, Jay Ligatti, Xinming Ou, Wouter Lueks, and Paul Syverson (Eds.). ACM, 123–135. https://doi.org/10.1145/3411497.3420215

[27] Frank D McSherry. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data.* 19–30.

[28] Takao Murakami and Yusuke Kawamoto. 2019. Utility-Optimized Local Differential Privacy Mechanisms for Distribution Estimation. In *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, Nadia Heninger and Patrick Traynor (Eds.). USENIX Association, 1877–1894. https://www.usenix.org/conference/usenixsecurity19/presentation/murakami

[29] In Jae Myung. 2003. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology* 47, 1 (2003), 90–100.

[30] Thông T. Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. 2016. Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy. *CoRR* abs/1606.05053 (2016). arXiv:1606.05053 http://arxiv.org/abs/1606.05053

[31] Yiwen Nie, Wei Yang, Liusheng Huang, Xike Xie, Zhenhua Zhao, and Shaowei Wang. 2019. A Utility-Optimized Framework for Personalized Private Histogram Estimation. *IEEE Trans. Knowl. Data Eng.* 31, 4 (2019), 655–669. https://doi.org/10.1109/TKDE.2018.2841360

[32] Ben Niu, Yahong Chen, Boyang Wang, Jin Cao, and Fenghua Li. 2020. Utility-aware Exponential Mechanism for Personalized Differential Privacy. In *2020 IEEE Wireless Communications and Networking Conference, WCNC 2020, Seoul, Korea (South), May 25-28, 2020.* IEEE, 1–6. https://doi.org/10.1109/WCNC45663.2020.9120532

[33] Desong Qin and Zhenjiang Zhang. 2021. A Frequency Estimation Algorithm under Local Differential Privacy. In *15th International Conference on Ubiquitous Information Management and Communication, IMCOM 2021, Seoul, South Korea, January 4-6, 2021*, Sukhan Lee, Hyunseung Choo, and Roslan Ismail (Eds.). IEEE, 1–5. https://doi.org/10.1109/IMCOM51814.2021.9377325

[34] Haina Song, Hua Shen, Nan Zhao, Zhangqing He, Minghu Wu, Wei Xiong, and Mingwu Zhang. 2024. APLDP: Adaptive personalized local differential privacy data collection in mobile crowdsensing. *Comput. Secur.* 136 (2024), 103517. https://doi.org/10.1016/J.COSE.2023.103517

[35] Shun Takagi, Yang Cao, and Masatoshi Yoshikawa. 2020. POSTER: Data Collection via Local Differential Privacy with Secret Parameters. In *ASIA CCS '20: The 15th ACM Asia Conference on Computer and Communications Security, Taipei, Taiwan, October 5-9, 2020*, Hung-Min Sun, Shiuh-Pyng Shieh, Guofei Gu, and Giuseppe Ateniese (Eds.). ACM, 910–912. https://doi.org/10.1145/3320269.3405441

[36] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya G. Parameswaran, and Neoklis Polyzotis. 2015. SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics. *Proc. VLDB Endow.* 8, 13 (2015), 2182–2193. https://doi.org/10.14778/2831360.2831371

[37] Shaowei Wang, Liusheng Huang, Miaomiao Tian, Wei Yang, Hongli Xu, and Hansong Guo. 2015. Personalized privacy-preserving data aggregation for histogram estimation. In *2015 IEEE Global Communications Conference (GLOBECOM).* IEEE, 1–6.

[38] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally Differentially Private Protocols for Frequency Estimation. In *26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017*, Engin Kirda and Thomas Ristenpart (Eds.). USENIX Association, 729–745. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao

[39] Zhou Wang and Alan C. Bovik. 2009. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Process. Mag.* 26, 1 (2009), 98–117. https://doi.org/10.1109/MSP.2008.930649

[40] Xingxing Xiong, Shubo Liu, Dan Li, Zhaohui Cai, and Xiaoguang Niu. 2020. A comprehensive survey on local differential privacy. *Security and Communication Networks* 2020, 1 (2020), 8829523.

[41] Qiao Xue, Qingqing Ye, Haibo Hu, Youwen Zhu, and Jian Wang. 2023. DDRM: A Continual Frequency Estimation Mechanism With Local Differential Privacy. *IEEE Trans. Knowl. Data Eng.* 35, 7 (2023), 6784–6797. https://doi.org/10.1109/TKDE.2022.3177721

[42] Ruilin Yang, Hui Yang, Jiluan Fan, Changyu Dong, Yan Pang, Duncan S. Wong, and Shaowei Wang. 2023. Personalized Differential Privacy in the Shuffle Model. In *Artificial Intelligence Security and Privacy - First International Conference on Artificial Intelligence Security and Privacy, AIS&P 2023, Guangzhou, China, December 3-5, 2023, Proceedings, Part I (Lecture Notes in Computer Science)*, Jaideep

Vaidya, Moncef Gabbouj, and Jin Li (Eds.), Vol. 14509. Springer, 468–482. https://doi.org/10.1007/978-981-99-9785-5_33

[43] Mingyue Zhang, Junlong Zhou, Gongxuan Zhang, Lei Cui, Tian Gao, and Shui Yu. 2023. APDP: Attribute-Based Personalized Differential Privacy Data Publishing Scheme for Social Networks. *IEEE Trans. Netw. Sci. Eng.* 10, 2 (2023), 922–933. https://doi.org/10.1109/TNSE.2022.3224731

[44] Yue Zhang, Youwen Zhu, Yuqian Zhou, and Jiabin Yuan. 2024. Frequency Estimation Mechanisms Under 1013δ-Utility-Optimized Local Differential Privacy. *IEEE Trans. Emerg. Top. Comput.* 12, 1 (2024), 316–327. https://doi.org/10.1109/TETC.2023.3238839

[45] Mengbiao Zhao, Wei Feng, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. 2022. Mixed-Supervised Scene Text Detection With Expectation-Maximization Algorithm. *IEEE Trans. Image Process.* 31 (2022), 5513–5528. https://doi.org/10.1109/TIP.2022.3197987

[46] Huangjie Zheng and Mingyuan Zhou. 2021. Exploiting Chain Rule and Bayes' Theorem to Compare Probability Distributions. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 14993–15006. https://proceedings.neurips.cc/paper/2021/hash/7e0ff37942c2de60cbcbd27041196ce3-Abstract.html

[47] Imtiaz Masud Ziko, Eric Granger, Jing Yuan, and Ismail Ben Ayed. 2019. Clustering with Fairness Constraints: A Flexible and Scalable Approach. *CoRR* abs/1906.08207 (2019). arXiv:1906.08207 http://arxiv.org/abs/1906.08207

A APPENDICES

A.1 means and variance

Theorem A.1 (Mean and Variances of a Ratio [10]). *Given the unbiased frequency estimations of value x and value y (i.e., $\hat{g}(x)$ and $\hat{g}(y)$), Eq. (16) and Eq. (17) define the approximated mean and variance, respectively,*

$$\mathrm{E}^*\left[\frac{\hat{g}(x)}{\hat{g}(y)}\right] = \frac{f_x}{f_y}, \tag{16}$$

$$\mathrm{Var}^*\left[\frac{\hat{g}(x)}{\hat{g}(y)}\right] = \frac{(f_x)^2}{(f_y)^2}\left[\frac{\sigma_x^2}{(f_x)^2} - 2\frac{\mathrm{Cov}(x,y)}{f_x f_y} + \frac{\sigma_y^2}{(f_y)^2}\right], \tag{17}$$

where f_x and f_y represent the true frequencies of x and y respectively, and σ_x^2 and σ_y^2 denote the variances of the estimates $\hat{g}(x)$ and $\hat{g}(y)$.

Proof. For any function $f(X,Y)$, we can choose the expansion point to be $\theta = (\mu_x, \mu_y)$, and the first-order Taylor series approximation for $f(X,Y)$ is:

$$\begin{aligned}\mathrm{E}(f(X,Y)) &\approx \mathrm{E}(f(\theta)) + \mathrm{E}\left[f_x'(\theta)(X-\mu_x)\right] + \mathrm{E}\left[f_y'(\theta)(Y-\mu_y)\right]\\ &= \mathrm{E}[f(\theta)] + f_x'(\theta)\mathrm{E}\left[(X-\mu_x)\right] + f_y'(\theta)\mathrm{E}\left[(Y-\mu_y)\right]\\ &= f(\mu_x,\mu_y).\end{aligned}$$

We let $f(x,y) = x/y$, and we have $\mathrm{E}^*(f(X,Y)) = f(\mu_x,\mu_y) = \mu_x/\mu_y$. Therefore, we have the approximate mean of the ratio $\hat{g}(x)/\hat{g}(y)$ as f_x/f_y, where $\mathrm{E}(\hat{g}(x)) = f_x$ and $\mathrm{E}(\hat{g}(y)) = f_y$. In addition, we can express the variance of $f(X,Y)$ in the following form:

$$\begin{aligned}\mathrm{Var}\left[f(X,Y)\right] &= \mathrm{Var}\left\{\left[f(X,Y) - \mathrm{E}(f(X,Y))\right]^2\right\}\\ &\approx \mathrm{Var}\left\{\left[f(X,Y) - f(\theta)\right]^2\right\}.\end{aligned}$$

We use the first-order Taylor series expansion for $f(X,Y)$ at point θ:

$$\begin{aligned}\mathrm{Var}[f(X,Y)] &\approx \mathrm{Var}\{[f(\theta) + f_x'(\theta)(X-\theta_x) + f_y'(\theta)(Y-\theta_y) - f(\theta)]^2\}\\ &= f_x'(\theta)^2\mathrm{Var}(X) + 2f_x'(\theta)f_y'(\theta)\mathrm{Cov}(X,Y) + f_y'(\theta)^2\mathrm{Var}(Y).\end{aligned}$$

where $f(x,y) = x/y$, and the approximated variance is:

$$\mathrm{Var}^*\left[\frac{X}{Y}\right] = \frac{(\mu_x)^2}{(\mu_y)^2}\left[\frac{\sigma_x^2}{(\mu_x)^2} - 2\frac{\mathrm{Cov}(X,Y)}{\mu_x\mu_y} + \frac{\sigma_y^2}{(\mu_y)^2}\right].$$

□

A.2 The proof of Theorem 5.1

Proof. In Section 5.2, we presented the specific expression $\hat{p}_i$ for frequency estimation of data item x_i:

$$\hat{p}_i = \frac{\sum_{u=1}^{N}\mathbb{I}_{\mathrm{support}(Z_u)}(x_i) - Nq^\star}{N(p^\star - q^\star)},$$

Where, N represents the total number of users, and $\hat{p}_i$ is the estimated frequency of the data item x_i. Z_u is the perturbed encoding output of user u. The function $\mathbb{I}_{\mathrm{Support}(Z_u)}(x_i)$ outputs 1 if and only if Z_u supports the item x_i, otherwise, it outputs 0. $p^\star$ and $q^\star$ are defined as:

$$p^\star = \sum_{i=1}^{t}\rho_i p_{\epsilon_i}, \quad q^\star = \sum_{i=1}^{t}\rho_i q_{\epsilon_i},$$

p_{ϵ_i} represents the probability of an input value x_1 being perturbed to its own support set under the privacy budget ϵ_i, and q_{ϵ_i} represents the probability of an input value $x_2, x_2 \neq x_1$, being perturbed to the support set of x_1 under the privacy budget ϵ_i.

Since the frequency distribution of the original data is $\{p_i\}_{i=1}^{k}$ and the distribution of the users on the set of privacy budgets set $\boldsymbol{\varepsilon}$ is $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^{t}$, and also according to the LDP protocol in Section 5.1. We can get

$$\begin{aligned}\mathrm{E}(\hat{p}_i) &= \frac{\mathrm{E}\left[\sum_{u=1}^{N}\mathbb{I}_{\mathrm{support}(Z_u)}(x_i) - Nq^\star\right]}{\mathrm{E}\left[N(p^\star - q^\star)\right]}\\ &= \frac{\sum_{u=1}^{N}\mathrm{E}\left[\mathbb{I}_{\mathrm{support}(Z_u)}(x_i)\right] + N\mathrm{E}(q^\star)}{N\mathrm{E}(p^\star - q^\star)}\\ &= \frac{N\left[p_i\left(\sum_{j=1}^{t}\rho_j p_{\epsilon_j} - \sum_{j=1}^{t}\rho_j q_{\epsilon_j}\right)\right]}{N\left[\sum_{j=1}^{t}\rho_j p_{\epsilon_j} - \sum_{j=1}^{t}\rho_j q_{\epsilon_j}\right]}\\ &= p_i.\end{aligned}$$

So $\hat{p}_i$ is an unbiased estimator of the mean p_i for x_i. □

A.3 The proof of Theorem 5.2

Proof. In Section 5.2, we presented the specific expression $\tilde{P}_i$ for frequency estimation of data item x_i:

$$\tilde{p_i} = \frac{\sum_{u=1}^{N}\mathbb{I}_{\mathrm{support}(Z_u)}(x_i) - N\hat{q}^\star}{N(\hat{p}^\star - \hat{q}^\star)},$$

Here, N represents the total number of users, and $\hat{p}_i^*$ is the estimated frequency of the data item x_i. Z_u is the perturbed encoding output of user u. The function $\mathbb{I}_{\mathrm{Support}(Z_u)}(x_i)$ outputs 1 if and only if Z_u supports the item x_i, otherwise, it outputs 0. $\hat{p}^\star$ and $\hat{q}^\star$ are defined as:

$$\hat{p}^\star = \sum_{i=1}^{t}\hat{\rho}_i p_{\epsilon_i}, \quad \hat{q}^\star = \sum_{i=1}^{t}\hat{\rho}_i q_{\epsilon_i}.$$

p_{ϵ_i} represents the probability of an input value x_1 being perturbed to its own support set under the privacy budget ϵ_i, and q_{ϵ_i} represents the probability of an input value $x_2, x_2 \neq x_1$, being perturbed to the support set of x_1 under the privacy budget ϵ_i. $\hat{\rho}_i$ represents the unbiased estimation of the distribution frequency ρ_i of users on the privacy budget ϵ_i.

Also, by Theorem A.1, we can determine that the approximate mean of $\tilde{p}_i$ is as follows:

$$\begin{aligned} \mathrm{E}(\tilde{p}_i) &\approx \frac{\mathrm{E}\left[\sum_{u=1}^{N} \mathbb{I}_{\text{support}(Z_u)}(x_i) - N\hat{q}^{\star}\right]}{\mathrm{E}\left[N(\hat{p}^{\star} - \hat{q}^{\star})\right]} \\ &= \frac{\sum_{u=1}^{N} \mathrm{E}\left[\mathbb{I}_{\text{support}(Z_u)}(x_i)\right] + N\mathrm{E}(\hat{q}^{\star})}{N\mathrm{E}(\hat{p}^{\star} - \hat{q}^{\star})} \\ &= \frac{N\left[p_i\left(\sum_{j=1}^{t} \rho_j p_{\epsilon_j} - \sum_{j=1}^{t} \rho_j q_{\epsilon_j}\right)\right]}{N\left[\sum_{j=1}^{t} \rho_j p_{\epsilon_j} - \sum_{j=1}^{t} \rho_j q_{\epsilon_j}\right]} \\ &= p_i. \end{aligned}$$

So $\tilde{p}_i$ is an approximate unbiased estimator of the mean p_i for x_i. □

A.4 The proof of Theorem 5.3

Proof. First, based on Section 5.2 of the paper, we can determine the specific expression of $\tilde{p}_i$ as follows:

$$\tilde{p}_i = \frac{\sum_{u=1}^{N} \mathbb{I}_{\text{support}(Z_u)}(x_i) - N\hat{q}^{\star}}{N(\hat{p}^{\star} - \hat{q}^{\star})}.$$

Meanwhile, we define $X = \sum_u \mathbb{I}_{\text{support}(Z_u)}(x_i)$, $Y = \hat{q}^{\star}$, $Z = \hat{p}^{\star} - \hat{q}^{\star}$. Therefore, the expression of $\hat{p}_i^*$ can be simplified as follows:

$$\tilde{p}_i = \frac{X}{Z} + \frac{Y}{Z}.$$

Therefore, we can express the variance of $\hat{p}_i^*$ as:

$$\begin{aligned} \mathrm{Var}(\tilde{p}_i) &= \mathrm{Var}\left(\frac{X}{Z}\right) + \mathrm{Var}\left(\frac{Y}{Z}\right) - \mathrm{Cov}\left(\frac{X}{Z}, \frac{Y}{Z}\right) \\ &\leq \mathrm{Var}\left(\frac{X}{Z}\right) + \mathrm{Var}\left(\frac{Y}{Z}\right). \end{aligned}$$

Next, we provide the computation procedures for $\mathrm{Var}\left(\frac{X}{Z}\right)$ and $\mathrm{Var}\left(\frac{Y}{Z}\right)$ separately:

For $\mathrm{Var}\left(\frac{X}{Z}\right)$: According to Theorem A.1 and The random variables X and Z are independent of each other, we can conclude that:

$$\begin{aligned} \mathrm{Var}(\frac{X}{Z}) \approx& \mathrm{Var}^*(\frac{X}{Z}) \\ =& \frac{(\mathrm{E}(X))^2}{(\mathrm{E}(Z))^2}\left[\frac{\mathrm{Var(X)}}{(\mathrm{E}(X))^2} - 2\frac{\mathrm{Cov}(X, Z)}{\mathrm{E}(X)\mathrm{E}(Z)} + \frac{\mathrm{Var}(Z)}{(\mathrm{E}(Z))^2}\right] \\ =& \frac{(\mathrm{E}(X))^2}{(\mathrm{E}(Z))^2}\left[\frac{\mathrm{Var(X)}}{(\mathrm{E}(X))^2} + \frac{\mathrm{Var}(Z)}{(\mathrm{E}(Z))^2}\right]. \end{aligned}$$

Moreover, because we assume that the distribution frequency on the original data set $\mathcal{D}$ is $P = \{p_i\}_{i=1}^{k}$ The distribution frequency of users over each privacy budget is $\boldsymbol{\rho} = \{\rho\}_{i=1}^{t}$. Furthermore, combined with the perturbation rules outlined in Section 5.1, we can obtain the mean $\mathrm{E}(X)$ and variance $\mathrm{Var}(X)$ of $X = \sum_{u=1}^{N} \mathbb{I}_{\text{support}(Z_u)}(x_i)$ as follows.

$$\begin{aligned} A &= \left[(1 - p_i)\sum_{j=1}^{t} \rho_j q_{\epsilon_j} + p_i \sum_{j=1}^{t} \rho_j p_{\epsilon_j}\right], \\ \mathrm{E}(X) &= \sum_{u=1}^{N} \mathrm{E}(\mathbb{I}_{support(Z_u)}(x_i)) \\ &= N\left[\sum_{i=1}^{k} p_i \sum_{j=1}^{t} \rho_j q_{\epsilon_j} + p_i(\sum_{j=1}^{t} \rho_j p_{\epsilon_j} - \sum_{j=1}^{t} \rho_j q_{\epsilon_j})\right] \\ &= N\left[(1 - p_i)\sum_{j=1}^{t} \rho_j q_{\epsilon_j} + p_i \sum_{j=1}^{t} \rho_j p_{\epsilon_j}\right] \\ &= NA, \\ \mathrm{Var}(X) &= \sum_{u} \mathrm{Var}(\mathbb{I}_{support(Z_u)}(x_i)) \\ &= NA(1 - A). \end{aligned}$$

Similarly, we get the mean $\mathrm{E}(Z)$and variance $\mathrm{Var}(Z)$ of $Z = \hat{p}^{\star} - \hat{q}^{\star}$:

$$\begin{aligned} \mathrm{E}(Z) &= \mathrm{E}(\sum_{i=1}^{t} \frac{\sum_{u=1}^{N} \mathbb{I}_{support(\epsilon_u^p)}(\epsilon_i) - Nq_{\epsilon_p}}{N(p_{\epsilon_p} - q_{\epsilon_p})}(p_{\epsilon_i} - q_{\epsilon_i})) \\ &= \sum_{i=1}^{t} \frac{\sum_{u=1}^{N} \mathrm{E}(\mathbb{I}_{support(\epsilon_u^p)}(\epsilon_i)) - Nq_{\epsilon_p}}{N(p_{\epsilon_p} - q_{\epsilon_p})}(p_{\epsilon_i} - q_{\epsilon_i}) \\ &= \sum_{i=1}^{t} \rho_i(p_{\epsilon_i} - q_{\epsilon_i}), \\ B_i &= \frac{N\left[(q_{\epsilon_p} + \rho_i(p_{\epsilon_p} - q_{\epsilon_p}))(1 - q_{\epsilon_p} - \rho_i(p_{\epsilon_p} - q_{\epsilon_p}))\right]}{N^2(p_{\epsilon_p} - q_{\epsilon_p})^2}. \\ \mathrm{Var}(Z) &= \mathrm{Var}(\sum_{i=1}^{t} \frac{\sum_{u=1}^{N} \mathbb{I}_{support(\epsilon_u^p)}(\epsilon_i) - Nq_{\epsilon_p}}{N(p_{\epsilon_p} - q_{\epsilon_p})}(p_{\epsilon_i} - q_{\epsilon_i})) \\ &= \sum_{i=1}^{t} \frac{\sum_{u=1}^{N} \mathrm{Var}(\mathbb{I}_{support(\epsilon_u^p)}(\epsilon_i))}{N^2(p_{\epsilon_p} - q_{\epsilon_p})^2}(p_{\epsilon_i} - q_{\epsilon_i})^2 \\ &= \sum_{i=1}^{t} B_i(p_{\epsilon_i} - q_{\epsilon_i})^2. \end{aligned}$$

Thus we can obtain the final form of $\mathrm{Var}\left(\frac{X}{Z}\right)$:

$$\begin{aligned} \mathrm{Var}(\frac{X}{Z}) &\approx \mathrm{Var}^*(\frac{X}{Z}) \\ &= \frac{(\mathrm{E}(X))^2}{(\mathrm{E}(Z))^2}\left[\frac{\mathrm{Var(X)}}{(\mathrm{E}(X))^2} + \frac{\mathrm{Var}(Z)}{(\mathrm{E}(Z))^2}\right] \\ &= \frac{N(A(1 - A))}{(\sum_{i=1}^{t} \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^2} + \frac{(NA)^2 \sum_{i=1}^{t} B_i(p_{\epsilon_i} - q_{\epsilon_i})^2}{(\sum_{i=1}^{t} \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^4}. \end{aligned}$$

For $\mathrm{Var}\left(\frac{Y}{Z}\right)$: Similar to calculating the mean and variance of the random variable Z, we can obtain the mean $\mathrm{E}(Y)$ and variance $\mathrm{Var}(Y)$ of $Y = \hat{q}^{\star}$:

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}(\sum_{i=1}^{t} \frac{\sum_{u=1}^{N} \mathbb{I}_{support(\epsilon_u^p)}(\epsilon_i) - Nq_{\epsilon_p}}{N(p_{\epsilon_p} - q_{\epsilon_p})} q_{\epsilon_i}) \\ &= \sum_{i=1}^{t} \frac{\sum_{u=1}^{N} \mathrm{E}(\mathbb{I}_{support(\epsilon_u^p)}(\epsilon_i)) - Nq_{\epsilon_p}}{N(p_{\epsilon_p} - q_{\epsilon_p})} q_{\epsilon_i} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{t} \rho_i q_{\epsilon_i}, \\
\mathrm{Var}(Y) &= \mathrm{Var}(\sum_{i=1}^{t} \frac{\sum_{u=1}^{N} \mathbb{I}_{support(\epsilon_u^p)}(\epsilon_i) - N q_{\epsilon_p}}{N(p_{\epsilon_p} - q_{\epsilon_p})} q_{\epsilon_i}) \\
&= \sum_{i=1}^{t} \frac{\sum_{u=1}^{N} \mathrm{Var}(\mathbb{I}_{support(\epsilon_u^p)}(\epsilon_i))}{N^2(p_{\epsilon_p} - q_{\epsilon_p})^2} q_{\epsilon_i}^2 \\
&= \sum_{i=1}^{t} \frac{N\left[(q_{\epsilon_p} + \rho_i(p_{\epsilon_p} - q_{\epsilon_p})(1 - q_{\epsilon_p} - \rho_i(p_{\epsilon_p} - q_{\epsilon_p})\right]}{N^2(p_{\epsilon_p} - q_{\epsilon_p})^2} q_{\epsilon_i}^2.
\end{aligned}
$$

In the same way, according to theory 1, we can get the specific expression of $\mathrm{Var}\left(\frac{Y}{Z}\right)$:

$$
\begin{aligned}
\mathrm{Var}(\frac{Y}{Z}) &\approx \mathrm{Var}^*(\frac{Y}{Z}) \\
&\leq \frac{(\mathrm{E}(Y))^2}{(\mathrm{E}(Z))^2}\left[\frac{\mathrm{Var}(Y)}{(\mathrm{E}(Y))^2} + \frac{\mathrm{Var}(Z)}{(\mathrm{E}(Z))^2}\right] \\
&= \frac{\sum_{i=1}^{t} B_i q_{\epsilon_i}^2}{(\sum_{i=1}^{t} \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^2} \\
&\quad + \frac{(\sum_{i=1}^{t} \rho_i q_{\epsilon_i})^2(\sum_{i=1}^{t} B_i(p_{\epsilon_i} - q_{\epsilon_i})^2)}{(\sum_{i=1}^{t} \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^4}.
\end{aligned}
$$

In summary we can obtain that the approximate variance of $\tilde{p_i}$ is:

$$
\begin{aligned}
\mathrm{Var}(\tilde{p_i}) &\leq \mathrm{Var}\left(\frac{X}{Z}\right) + \mathrm{Var}\left(\frac{Y}{Z}\right) \\
&\approx \mathrm{Var}^*(\frac{Y}{Z}) + \mathrm{Var}^*(\frac{X}{Z}) \\
&\leq V_1 + V_2 + V_3 + V_4, \\
A &= \left[(1 - p_i)\sum_{j=1}^{t} \rho_j q_{\epsilon_j} + p_i \sum_{j=1}^{t} \rho_j p_{\epsilon_j}\right], \\
B_i &= \frac{N\left[(q_{\epsilon_p} + \rho_i(p_{\epsilon_p} - q_{\epsilon_p})(1 - q_{\epsilon_p} - \rho_i(p_{\epsilon_p} - q_{\epsilon_p})\right]}{N^2(p_{\epsilon_p} - q_{\epsilon_p})^2}, \\
V_1 &= \frac{N(A(1-A))}{(\sum_{i=1}^{t} \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^2}, \\
V_2 &= \frac{(NA)^2 \sum_{i=1}^{t} B_i(p_{\epsilon_i} - q_{\epsilon_i})^2}{(\sum_{i=1}^{t} \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^4}, \\
V_3 &= \frac{\sum_{i=1}^{t} B_i q_{\epsilon_i}^2}{(\sum_{i=1}^{t} \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^2}, \\
V_4 &= \frac{(\sum_{i=1}^{t} \rho_i q_{\epsilon_i})^2(\sum_{i=1}^{t} B_i(p_{\epsilon_i} - q_{\epsilon_i})^2)}{(\sum_{i=1}^{t} \rho_i(p_{\epsilon_i} - q_{\epsilon_i}))^4}.
\end{aligned}
$$

□

A.5 Overview of Comparative Approaches

To comprehensively validate the effectiveness of our method, we selected four representative schemes for comparative analysis. These schemes are based on personalized local differential privacy for estimating discrete data frequency distributions.

PBP(Parameter Blending Privacy) [35]:In this study, the authors propose a privacy definition called Parameter Blending Privacy (PBP). In this scheme, users conceal their privacy budgets from data collectors. However, to accurately estimate the frequency distribution of the original data, the authors assume that the distribution of users across different privacy budgets is publicly known. Additionally, the study only implements scenarios with two privacy budgets and two data attributes.

PPDA (Personalized Private Data Aggregation) [37]: In this scheme, the authors prioritize the importance of protecting user privacy budgets by concealing the users' true privacy budgets from data collectors. Users are required to perturb their data using a LDP method based on onehot encoding. Additionally, to ensure the accuracy of the estimation results, the scheme mandates appending extra zero bits to the encoded vectors.

RCF (Recycle and Combination Framework) [31]: In this scheme, the authors overlook the protection of user privacy budgets. Users send their perturbed data and privacy budgets to the data collectors. The data collectors group the perturbed data based on the users' privacy budgets to estimate the original data frequency distribution. Finally, the authors apply a weighted summation of the estimation results from each group, following the principle of minimizing the mean squared error, to obtain the final estimation.

APLDP(Adaptive Personal Local Differential Privacy) [34]:In this scheme, the authors also overlook the protection of user privacy budgets. Different local perturbation algorithms are optimal for different ranges of privacy budgets. The data collectors first group users according to their privacy budgets. Within each group, users select the optimal local perturbation algorithm based on their privacy budgets to perturb their private data. The data collectors then grouped the perturbed data and obtained estimates of the frequency distribution of the raw data for each group. Finally, the estimation results from each group are weighted and summed to produce the final estimation.