IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 June 2024, accepted 14 July 2024, date of publication 18 July 2024, date of current version 26 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3430863

RESEARCH ARTICLE

MPLDP: Multi-Level Personalized Local Differential Privacy Method

XUEJIE FENG[1] AND CHIPING ZHANG[2]
[1]School of International Business, Qingdao Huanghai University, Qingdao 266427, China
[2]Department of Mathematics, Harbin Institute of Technology, Harbin 150001, China

Corresponding author: Xuejie Feng (clairefxj@stu.hit.edu.cn)

This work was supported by the Second Batch of Doctoral Research Fund Project of Qingdao Huanghai University: Local Differential Privacy Optimization Method Based on Data Perturbation.

ABSTRACT Users have different sensitivities to different attributes for the same data set. Disregarding this can result in inadequate data confidentiality or reduced data availability. To address this, this paper proposes a multi-level personalized local differential privacy mechanism optimization method. In high-dimensional heterogeneous data scenario, this paper first adopts the optimal privacy budget allocation scheme to allocate the privacy budget of different attributes, and then categorizes the privacy levels into high, medium, and low. Users can freely select the privacy level for each attribute or choose the same level for all attributes. For data analysts, reorganizing data with different privacy levels to achieve histogram estimation is a challenging task. The paper introduces a histogram optimization estimation method based on two evaluation criteria. It proposes a combinatorial optimization method, OC, which minimizes mean square error, and a combinatorial optimization method, OP, based on perturbation theory, which minimizes maximum error. The paper comprehensively studies the balance between data availability and privacy protection based on these two rules.

INDEX TERMS Differential privacy, perturbation, nonlinear equations, optimization, personalized.

I. INTRODUCTION

Crowdsourcing technology has emerged as a powerful tool to understand users' needs, driven by the collection of vast amounts of private data. With the proliferation of smartphones and the Internet, data collectors can easily gather personal data from a variety of mobile apps, allowing them to create more apps that meet people's needs and are quietly changing the way we live. For example, real-time traffic conditions can be provided by collecting users' driving information; The user's mobile phone is used as a weather station, providing weather information for any given area; Businesses, marketers and urban planners can collect video from users' mobile devices and turn it into meaningful aggregate data. However, these apps pose a huge threat to users' privacy while collecting their data, which has become a major obstacle to the widespread acceptance of crowdsourcing [1]. Such as the Apple iCloud user photo leak and Yahoo data leak in 2014, people gradually realized the importance of user privacy. Without proper privacy protections, it is difficult for users to trust an organization. As a result, the sharing of personal data calls for increased caution.

The emergence of differential privacy technologies [2], [3] brings renewed optimism for users. This groundbreaking approach offers independent and verifiable privacy protection, impervious to an attacker's background knowledge and computing prowess. Unfortunately, users' trust is based on centralized third-party data managers. However, the frequent third-party privacy leakage in recent years makes people no longer trust any third-party data managers who claim to keep users secret.

The local differential privacy (ϵ-local differential privacy, referred to as LDP) originates from the differential privacy in the centralized database environment and has become the standard of data privacy protection in crowdsourcing environment. It protects the data privacy of each participant locally (e.g. on its own mobile device), without relying on

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

any other party (e.g. database administrator or aggregator, other participants). This is ideal for safeguarding data privacy in the crowdsourcing scenario outlined in this paper, as crowdsourcing aggregators, application servers, or other participants may potentially compromise privacy. In fact, people have done a lot of work on data aggregation of general categories with ϵ- LDP [4], [5], [6].

However, due to different professional backgrounds, people have different definitions of privacy and different levels of concern about the private data they care about. For example, for public figures, they pay more attention to their geographical location information, such as home address, vacation travel, etc. This is because some malicious people will use their exposed location information for personal interest tracking. On the contrary, for user groups like Hiking travel, they prefer people to know their geographical location in real time, so as to get more attention. Different historical backgrounds and different living environments make people have different definitions of the same type of privacy data. Therefore, it is worth studying to provide a personalized local differential privacy mechanism that meets different user groups. In certain circumstances, individuals may be open to trading some privacy protection for increased data availability or more precise services. For example, when sharing location data, if users are interested in accessing information about nearby services or deals, they may choose to lower their privacy level slightly to receive more accurate recommendations. In this instance, employing medium or low level LDP settings can effectively address this requirement. Therefore, in 2012, Li et al. [7] proposed a multi-level privacy data release method to provide a solution for users' diversified privacy needs. However, this method is centralized differential privacy, that is, it assumes that there is a trusted third-party data management organization to save user privacy data, which is not applicable to the local differential privacy model. Nie et al. proposed a multi-level personalized local differential privacy scheme [8] in 2018. This method is based on a single attribute scenario. However, under the condition of high-dimensional heterogeneous data collection, how to publish personalized local differential privacy scheme is still a problem worth studying.

Motivated by above problems, this paper proposes a novel multi-level personalized local differential privacy mechanism optimization method. Multi-level personalized local differential privacy means that different users can choose different levels of privacy protection for each attribute based on their privacy preferences and data sensitivity. These levels can be quantified by different privacy budgets ϵ, with a smaller ϵ value indicating a higher level of privacy protection. In order to achieve multiple levels of personalized local differential privacy, the aggregator first needs to define different privacy levels for users to choose from. To meet the needs of users, this paper divides the privacy level into *high*, *medium*, and *low* levels. You have the freedom to customize the privacy settings for individual attributes or set a universal privacy level for all attributes. This paper introduces innovative histogram optimization estimation methods tailored to the specific needs of data analysts, addressing two evaluation criteria and various privacy levels. It presents the combined optimization method OC, which minimizes the mean square error, as well as the combined optimization method OP, based on perturbation theory, which prioritizes minimizing the maximum error to cater to the diverse requirements of data analysts.

The main contributions of this paper are as follows:

- This paper introduces an innovative multi-level personalized local differential privacy mechanism tailored for high-dimensional heterogeneous attribute scenarios. By integrating an optimal privacy budget allocation strategy with a binary randomized response mechanism, this approach empowers users to customize privacy levels for different attributes.
- This paper introduces two novel histogram estimation optimization methods, OC and OP, based on distinct optimization strategies. Furthermore, it provides comprehensive theoretical support for each of these optimization strategies.
- We have successfully conducted simulation experiments to implement personalized perturbation of user privacy data across high-dimensional heterogeneous attributes. The experimental results of two histogram estimation optimizations validate the effectiveness of our proposed method.

II. RELATED WORK

A. HIGH DIMENSION

Nowadays, for the issue of high-dimensional data publishing, there are many ways have proved their effectiveness from different perspectives. For instance, Cai et al. [9] delved into the balance struck between statistical precision and privacy in the realm of average estimation and linear regression with high-dimensional datasets. Their approach primarily involved optimizing the parameter configurations, including the minimum-maximum lower bound and iterative threshold, to guarantee statistical accuracy while adhering to differential privacy principles. Nevertheless, this methodology neglects the localized aspect of user privacy, and the authors fail to elaborate on effective strategies for allocating the privacy budget.

Wang et al. [10] proposed LoCop and DR_LoCop, which guarantee local differential privacy using the randomized response technique while synthesizing and releasing high-dimensional crowdsourced data with high data utility. Specifically, LoCop leverages copula theory to synthesize data through univariate marginal distributions estimated by Lasso-based regression and models attribute dependencies as multivariate Gaussian copula. DR_LoCop further improves upon LoCop by utilizing C-vine copula to capture conditional dependencies and achieve dimension reduction. However, their proposed approach doesn't address how to precisely allocate the privacy budget to each attribute or each data point.

Ren et al. introduced LoPub [11], a novel approach that marries the principles of RAPPOR and the probability graph model. Their method initially converts each attribute's value into a randomized bit string utilizing a Bloom filter [12], followed by its transmission to a centralized server. Afterward, the data collector conducts frequency analysis on the received data and constructs a Markov network. The joint probability distribution of the attributes is then condensed into a maximal clique to minimize data dimensionality. The resulting joint probability distribution is then employed to regenerate a dataset for release. However, a significant limitation of this approach lies in its lack of foresight regarding privacy budget allocation prior to the aggregation of high-dimensional heterogeneous data at the server. Furthermore, when dealing with mutually independent attributes, they suggest employing the Expectation-Maximization (EM) algorithm to estimate the multivariate distribution, which can lead to a steep exponential rise in computational complexities.

Ren et al. [13] proposed LDP-IDS, a novel local differential privacy (LDP) paradigm for infinite streams. This work addresses the challenge of preserving end user privacy during streaming data collection, a crucial aspect of real-time data analytics in IoT and mobile-based systems. By adapting the budget division framework from centralized differential privacy (CDP) and developing a unified error analysis for LDP, the authors present two adaptive budget division-based LDP methods that enhance data utility by leveraging the non-deterministic sparsity in streams. Furthermore, a novel population division framework is introduced, which not only mitigates the high sensitivity of LDP noise to budget division but also significantly reduces communication requirements. Although two adaptive privacy budget allocation methods are proposed in this paper, they simply divide the privacy budget evenly into two parts according to the time window, and do not take into account the estimated loss caused by privacy budget division under the condition of heterogeneous privacy attributes.

B. MULTI-LEVEL DIFFERENTIAL PRIVACY

With the development of differential privacy technology, multi-level privacy methods are more and more popular. Because different users have different definitions of privacy, they may have different privacy attitudes towards the same type of data. Motivated by this, some researchers have proposed multi-level privacy methods. The existing multi-level privacy protection methods can be roughly divided into partition method [14] and randomization method [7], [15]. Xiao et al. proposed a multi-level privacy data publishing mechanism in 2010. The method they proposed solves the privacy leakage problem caused by user collusion under different privacy levels. Under the premise of centralized differential privacy, Jorgense et al. [16] proposed a personalized differential privacy scheme. To enable users to customize their privacy budgets, they proposed a sampling method for publishing privacy data, which automatically converts all existing differential privacy algorithms into the algorithms satisfying Personalized Differential Privacy (PDP). Min et al. [17] proposed a method for location privacy protection in 3D space, P3DLPPM, which uses a two-stage position perturbation mechanism. The mechanism maps its initial application in data publishing privacy protection to the location perturbation mechanism. It generates false locations with smaller perturbed distances while improving the balance between privacy and quality of service (QoS). In the method based on k-anonymity, Gedik and Liu [18] proposed a personalized data publishing method to share user's personal location, and Yuan et al. [19] proposed a method to realize personalized privacy protection in the structure of social network graph on the premise of assuming the attacker's background knowledge.

All the multi-level privacy methods mentioned above belong to the category of centralized differential privacy research. They realize data privacy protection of user's personalized privacy level on the basis of trusted third party. They are actually contradictory and do not intrinsically address the risk of user privacy breaches. Therefore, the personalized local differential privacy mechanism under the local differential privacy mechanism is studied in this paper. In the research of personalized privacy with local differential privacy, a small number of literatures are related to this research category. For example, Wang et al. [20] designed a feasible locally differential privacy protection scheme on the Bloom filter. The proposed method can ensure the confidentiality of users' privacy level and optimize the estimation accuracy through the selected randomization strategy. Chen et al. [6] proposed to control the intensity of privacy through the granularity of region division, and realized personalized local differential privacy protection based on the streaming histogram publication (SHP) method of sliding window division. The proposal of this scheme is mainly aimed at spatial data, which is not scalable for other data. In 2018, Nie et al. [8] proposed a post-processing optimization scheme for personalized local differential privacy mechanism. They processed the disturbed data of users with different privacy levels, assign privacy data no higher than that level according to the credit ratings of different data analysts, and finally output it through weighted optimization. Their approach is aimed at local privacy protection under univariate conditions, and there is no clear answer to how to handle high-dimensional data.

In general, personalized local differential privacy technology is still in its infancy and has a good application prospect, but the existing personalized centralized differential privacy mechanism can not guarantee the security of user privacy from the source. Besides, the existing personalized local differential privacy mechanism cannot clearly solve the data processing problem in the high-dimensional heterogeneous scenario. This paper will study the personalized local differential privacy mechanism in high-dimensional heterogeneous scenarios. In order to ensure the accuracy of the estimation results, this paper will optimize the result output according to the principle of minimizing the mean square estimation error and minimizing the maximum error.

III. MPLDP: MULTI-LEVEL PERSONALIZED LOCAL DIFFERENTIAL PRIVACY METHOD

A. LOCAL DIFFERENTIAL PRIVACY

Multi-level personalized local differential privacy [21] is essentially local differential privacy. So let's start with the definition of local differential privacy. Local differential privacy [22] is a rigorous privacy notion in the local setting, which provides a stronger privacy guarantee than the centralized differential privacy. The formal definition of local differential privacy is defined as follows:

Definition 1: Given n users, each user corresponds to a record, a randomized algorithm $\mathcal{F}$ satisfies ϵ-local differential privacy, if for any two records t and $t' \in D$, and for any output $m \subseteq Range(\mathcal{F})$,

$$Pr[\mathcal{F}(t) = m] \leq \exp(\epsilon) \cdot Pr[\mathcal{F}(t') = m] \tag{1}$$

where ϵ denotes the privacy budget, and D represents the domain of privacy data.

In this paper, we use OBRR [23] perturbation mechanism to realize the perturbation of user data. In fact, The perturbation mechanism in OBRR actually uses binary randomized response method (BRR) [24], so we'll introduce the loss function definition for the BRR mechanism next.

Definition 2: Suppose there is an attribute data $x = x_i \in \chi = \{x_1, \cdots, x_m\}$, and express x_i as a bitmap $bx \in \{0, 1\}^m$. $\forall$ $1 \leq j \leq m$, the j-th position of the bitmap bx_j outputs the true value with probability p and $1 - bx_j$ with probability $1 - p$.

Suppose there are l attributes, then each user possesses exactly l items (Only one value can be selected for each attribute). Two such bit vectors can differ by at most $2l$ bits, meaning that the sensitivity is $2l$. According to the definition of local differential privacy, the binary randomized response mechanism needs to be satisfied $\frac{p}{1-p} \leq e^{\frac{\epsilon}{2}}$ or $\frac{1-p}{p} \leq e^{\frac{\epsilon}{2}}$. We can get the following theorem.

Theorem 1: If the privacy budget ϵ in the binary randomized response mechanism meets:

$$\epsilon = 2\log(\max\{\frac{p}{1-p}, \frac{1-p}{p}\})$$

Then the binary randomized response mechanism satisfies the ϵ-local differential privacy guarantee.

Multi-level personalized local differential privacy (MPLDP) is an extension of local differential privacy (LDP) that takes into account the individual privacy needs and preferences of users. Under MPLDP, each user can choose a different level of privacy budget $\epsilon_{i\tau}$ for each attribute $\mathbf{a}_i$ according to their own preferences and privacy concerns, thus providing different levels of privacy protection for different attributes. "Multi-level" in MPLDP means that users can assign different levels of privacy to each attribute in a dataset. Similar to LDP, MPLDP ensures that the different noise or randomization is directly applied to the individual user's attributes before it leaves the user's device. So that no trusted third party intervention is required. This prevents the inference of original data even if the data is collected. The goal of MPLDP is to balance the privacy needs of users with the utility of the data. By allowing for personalized privacy budgets, MPLDP aims to provide sufficient privacy protection while still enabling useful data analysis and mining.

TABLE 1. Notation.

Notation	Description
A	multiple unbalanced categorical data sets
l	number of attributes
n	number of participants
k_i	number of items of i-th attribute
d	total number of items, $d = \sum_i k_i$
$\mathbf{a}_j$	j-th attributes of A, its length $\lvert\mathbf{a}_j\rvert$ is k_j
$\mathbf{v}_m$	private values possessed by the m-th user, its length $\lvert\mathbf{v}_m\rvert$ is l
v_{ij}	j-th value of $\mathbf{v}_i$
u_m	the m-th participant
$\mathbf{h}_m$	private bit vector of m-th users, its length is d
$\mathbf{h}'_m$	private bit vector of m-th users after disturbing
$[\mathbf{h}'_u]_{ij}$	the value of user u after disturbing the j-th candidate value of the i-th attribute
$\mathbf{H}$	true histogram, $\mathbf{H} = \sum\{\mathbf{h}_1, \cdots, \mathbf{h}_n\}$
$\mathbf{H}'$	sanitized histogram of $\mathbf{H}$, $\mathbf{H}' = \sum\{\mathbf{h}'_1, \cdots, \mathbf{h}'_n\}$
$\widetilde{\mathbf{H}}$	estimated histogram of $\mathbf{H}'$, $\widetilde{\mathbf{H}} = \sum\{\widetilde{\mathbf{h}_1}, \cdots, \widetilde{\mathbf{h}_n}\}$
$\widetilde{\mathbf{H}}^\tau$	represents the unbiased estimation result of the data whose privacy level group is τ after the disturbance, where $\tau = \{high, mid, low\}$
ϵ_i	privacy budget of i-th attribute
τ	privacy level
$\epsilon_{i\tau}$	privacy budget of i-th attribute with privacy level τ
$\mathbf{\Omega}$	The weight used to estimate the distribution, $\mathbf{\Omega} = [\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_l]^T$, $\mathbf{W}_i = [\frac{n\omega_i^{high}}{n^{high}}, \frac{n\omega_i^{mid}}{n^{mid}}, \frac{n\omega_i^{low}}{n^{low}}]$
MSE	Mean Square Error
BRR	Binary Randomized Response method [24].
OBRR	Optimal Binary Randomized Response [23]
OMRR	Optimal Multivariate Randomized Response [23]
OC	Optimal Combination Method of High Dimensional Data
OP	Optimization Method Based on Perturbation Theory
RSS	residual sum of squares

B. PROBLEM DESCRIPTION

This paper focuses on the problem of multi-level personalized local differential privacy in high-dimensional heterogeneous aggregate data. Given a set of l attribute set data from n different users, different attribute dimensions may be different. Each attribute will involve different privacy budget. This paper assumes that the optimal privacy budgets for all attributes have been calculated by OBRR [23]. Considering that each user has a personalized privacy level requirement for each attribute, the main purpose of this paper is twofold. First, it aims to help users design a multi-level personalized local differential privacy mechanism, so that users can freely choose the privacy level of different attributes; Second, it enables the data collector to create a more effective combination method for estimating the frequency statistics of each attribute candidate value under various privacy levels, subsequent to the user's independent selection of the privacy level for each attribute. Some notations employed in this paper are listed in Table 1.

Formally, suppose that there is high-dimensional heterogeneous set data $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_l\}$, and each attribute $\mathbf{a}_i$ has a number of candidate values, $\mathbf{a}_i = \{a_{i1}, a_{i2}, \cdots, a_{ik_i}\}$, where k_i is the number of candidate values for the i-th attribute, it can also be interpreted as a dimension, that is, $|\mathbf{a}_i| = k_i$, $i = 1, 2, \cdots, l$. The optimal privacy budget for each attribute is $\{\epsilon_1, \cdots, \epsilon_l\}$. Each user u_m holds a dataset of l attributes $\mathbf{v}_m = \{v_{m1}, v_{m2}, \cdots, v_{ml}\}$, where $v_{mj} \in \mathbf{a}_j$. Let n represents the total number of users, and $d = k_1 + k_2 + \cdots + k_l$ represents the total length of the bitmap. It is assumed that the privacy level of each attribute is divided into three levels: *high*, *medium* and *low*. The corresponding privacy budgets are ϵ_i^{high}, ϵ_i^{mid} and ϵ_i^{low}, where $i = 1, \cdots, l$ stands for attribute index. The allocation of different levels of privacy budget ϵ_i^{high}, ϵ_i^{mid} and ϵ_i^{low} is not within the scope of this paper, which assumes $\epsilon_i^{high} = \frac{\epsilon_i}{3}$, $\epsilon_i^{mid} = \frac{\epsilon_i}{2}$ and $\epsilon_i^{low} = \epsilon_i$.

Let $\mathbf{h}_m$ represents the private bit vector of the m-th user. The length of the $\mathbf{h}_m$ is d. Firstly, the private data $\mathbf{v}_m$ will be converted to a bitmap $\mathbf{h}_m$, and $\mathbf{h}_m = \{h_{11}, \cdots, h_{1k_1}, h_{21} \cdots, h_{2k_2}, \cdots, h_{lk_l}\}$, where $h_{ij} \in \mathbf{a}_i, i = 1, 2, \cdots, l, j = 1, 2, \cdots, k_i$. Assume that the privacy level selected by user u_m is $\tau_m = \{\tau_1, \cdots, \tau_l\}$, the user u_m then perturbs the data with the privacy levels corresponding to different attributes. Finally, the user u_m sends the perturbed data $\mathbf{h}'_m$ to the data collector. The true histogram frequency can be expressed as $\mathbf{H} = \sum \{\mathbf{h}_1, \cdots, \mathbf{h}_n\}$. The estimated frequency of the disturbed data can be expressed as $\widetilde{\mathbf{H}} = \sum \{\widetilde{\mathbf{h}}_1, \cdots, \widetilde{\mathbf{h}}_n\}$. In order to estimate the frequency of the interfered data, it is necessary to make a comprehensive estimation of the data of users with different privacy levels of the same attribute. Under different privacy levels, the estimation results are inconsistent, and the estimation errors are also very different.

The running process of the multi-level personalized local differential privacy histogram publishing method for aggregated data proposed in this paper is shown in Fig.1. The aggregator publishes the query request set data $\mathbf{A} = \{\mathbf{a}_1, \cdots, \mathbf{a}_l\}$ to each participant along with global parameters, including the optimal privacy budget allocation scheme $\epsilon = \{\epsilon_1, \cdots, \epsilon_l\}$ and privacy level $\tau = \{high, mid, low\}$. After the participant u_i selects the privacy level, the privacy data $\mathbf{v}_m$ held by u_m is converted into a bitmap $\mathbf{h}_m$ and perturbed to $\mathbf{h}'_m$. For example, the privacy levels selected by *user*1 for the three attributes are respectively *high*, *high* and *low*. Therefore, when the user disturbs the three attributes, ϵ_1^{high},ϵ_2^{high} and ϵ_3^{low} are used to independently perform differential privacy calculations for the attributes. After receiving the list of perturbed data, the aggregator attempts to decode the estimate on $\mathbf{H}'$. Based on the candidate estimation results of the perturb data set, the aggregator tries to provide better network services for users.

C. PERSONALIZED LOCAL DIFFERENTIAL PRIVACY SYSTEM MODEL

In this paper, the OBRR method proposed in [23] is used as the client data perturbation mechanism, and then the histogram of different attributes estimated according to different privacy levels of users. OBRR is an improved version of BRR [24] in high dimensional data conditions. BRR has been proved by literature [25] to be the optimal mechanism under the condition of high privacy level and low privacy budget. Secondly, according to the characteristics of histogram estimation under different privacy levels, different utility evaluation functions are designed, and the results are combined and optimized. Algorithm 1 shows the process of perturbing data on the client side. Algorithm 2 shows the process of estimating the histogram at the data collector side, where $[\mathbf{h}'_m]_{ij}$ represents the value of user m after disturbing the j-th candidate value of the i-th attribute.

Algorithm 1 Data Disturbance

Input: ϵ-Privacy budget; $\{k_1, k_2, \cdots, k_l\}$-Candidate values for each attribute; $\mathbf{v} \in \{0, 1\}^{k_1+\cdots+k_l}$-Private data expressed as bitmap; v_{ij} denotes the j-th value of the i-th attribute of $\mathbf{v}$; $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_l\}$-Optimal privacy budget allocation scheme; $\tau = \{\tau_1, \cdots, \tau_l\}$-User privacy level

Output: $\mathbf{h}' \in \{0, 1\}^{k_1+\cdots+k_l}$-Disturbed data satisfying ϵ-local differential privacy, where h'_{ij} is the j-th value of the i-th attribute of $\mathbf{h}'$

```
initialize $d = k_1 + k_2 + \cdots + k_l$, $\mathbf{h}' = \mathbf{0} \in 0^d$, $m = 0$
for $i = 1$ to $l$ do
    for $j = 1$ to $k_i$ do
        $p = random[0, 1]$
        if $p < \frac{\exp(\epsilon_{i\tau_i}/2)}{\exp(\epsilon_{i\tau_i}/2)+1}$ then
            $h'_{ij} = v_{ij}$
        else
            $h'_{ij} = 1 - v_{ij}$
        end if
    end for
end for
```

Algorithm 2 Histogram Estimation

Input: ϵ-Privacy Budget; $\{k_1, k_2, \cdots, k_l\}$-Candidate values for each attribute; $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_l\}$-Optimal privacy budget allocation scheme, $\tau = \{\tau_1, \cdots, \tau_l\}$-User privacy level; n_{τ_i}-Number of users with privacy level τ_i; $\mathbf{h}'$-Disturbed data; $\mathbf{h}'_m$ is the disturbed data of the m-th user; $[\mathbf{h}'_m]_{ij}$ is the j-th value of i-th attribute of $\mathbf{h}'_m$

Output: $\widetilde{\mathbf{H}}$-Histogram estimation

```
for $i = 1$ to $l$ do
    for $j = 1$ to $k_i$ do
        $\widetilde{H_{ij}} = \frac{\sum_{m=1}^{n_{\tau_i}} [\mathbf{h}'_m]_{ij}(\exp(\epsilon_{i\tau_i}/2)+1)-n_{\tau_i}}{\exp(\epsilon_{i\tau_i}/2)-1}$
    end for
end for
```

Firstly, OBRR is used to perturb user data. The total privacy budget ϵ not only affects the intensity of privacy protection, but also directly affects the availability of data. Smaller ϵ values, while providing greater privacy protection, may result in data distortion or reduced availability. Therefore, in local differential privacy, the relationship between privacy

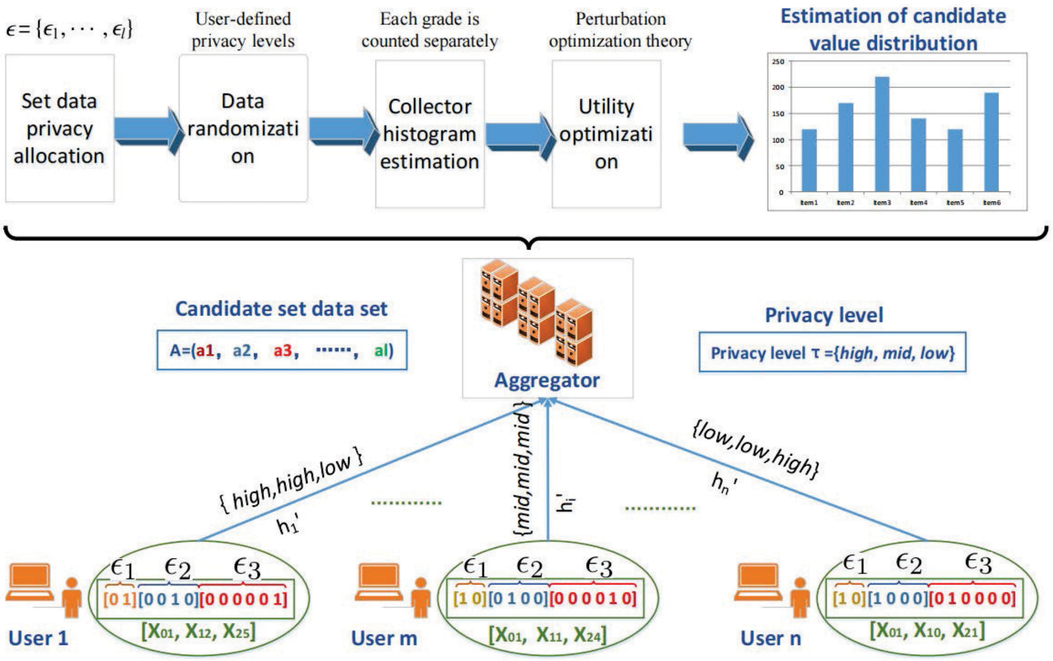


FIGURE 1. **Personalized local differential privacy framework for set data. For example, The first user in the figure holds values 2, 3, and 6 of the three attributes respectively. The corresponding positions of the bitmap are set to 1, and the remaining positions are 0, which noted as [X_{01}, X_{12}, X_{25}]. *User* 1 perturbs the bitmap they hold using the privacy budget ϵ_1^{high}, ϵ_2^{high} and ϵ_3^{low} for each attribute and send it to the aggregator. The aggregator combines and decodes the disturbed data and then publishes the results.**

protection and data availability needs to be carefully weighed to find the optimal ϵ value. Under multi-attribute conditions, the total privacy budget is still a global constraint even if each attribute is independent of each other. This means that when allocating privacy budgets to different attributes, it is necessary to ensure that the privacy protection needs of all attributes are met and that the privacy protection intensity of the entire process meets the predetermined requirements. In order to make the privacy budget allocated to each attribute maximize data availability in the process of unbiased estimation, we adopt the optimal privacy budget allocation scheme OBRR under the condition of multiple attributes. OBRR satisfies $\epsilon_{i\tau}$-local differential privacy guarantee in each attribute of high-dimensional set data. Assume $\epsilon = \epsilon_1 + \cdots + \epsilon_l$, then the local differential privacy guarantee is satisfied on l mutually independent disturbing attributes. When the data collector receives the disturbed data from the user, he/she first classifies the user data according to different privacy levels, and then uses the Algorithm 2 to unbiased estimate the data of each category. This involves a problem: how to combine the estimation results from different privacy levels but with the same attributes to ensure the correctness of the overall estimation results as much as possible? This is the focus of this paper's research, which is to find a solution suitable for the final estimated result by weighting the results of unbiased estimates of different privacy levels. In this paper, the utility function of minimizing square error and minimizing maximum error are used to evaluate the final results. The perturbation, estimation and optimization models proposed in this paper are shown in Fig.2.

IV. HISTOGRAM PUBLISHING OPTIMIZATION METHOD

Due to the different privacy level settings, the estimation accuracy within each level group is reduced. If the estimation results of different levels are directly added as the final estimation results, it will inevitably lead to a large difference between the estimated results and the actual results.

Excessive privacy levels create confusion and hinder user selections. For instance, a privacy level set at 10 makes it difficult for users to accurately evaluate their private data, rendering the rating meaningless. This paper proposes simplifying privacy levels into three categories: high, medium, and low. The corresponding privacy budgets are ϵ^{high}, ϵ^{mid}, ϵ^{low}. Assume that the total privacy budget assigned to the attribute i is ϵ_i, then let $\epsilon^{high} = \frac{\epsilon_i}{3}$, $\epsilon^{mid} = \frac{\epsilon_i}{2}$, $\epsilon^{low} = \epsilon_i$.

Let $\mathbf{H} = [\mathbf{H}_1, \mathbf{H_2}, \cdots, \mathbf{H_l}]$ represents the statistical results of the original privacy data, where $\mathbf{H}_i = [h_{i1}, \cdots, h_{ik_i}]$ represents the statistical result of the i-th attribute. $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{H}}^{high}, \widetilde{\mathbf{H}}^{mid}, \widetilde{\mathbf{H}}^{low}]^T$ represents the unbiased estimation result after user disturbance. $\widetilde{\mathbf{H}}^{\tau} = [\widetilde{h}^{\tau}_{11}, \cdots, \widetilde{h}^{\tau}_{1k_1}, \widetilde{h}^{\tau}_{21}, \cdots, \widetilde{h}^{\tau}_{2k_2}, \cdots, \widetilde{h}^{\tau}_{lk_l}]$ represents the unbiased estimation result of the data whose privacy level group is τ after the disturbance, where $\tau = \{high, mid, low\}$. There are l attributes in the user's aggregate data, and the number

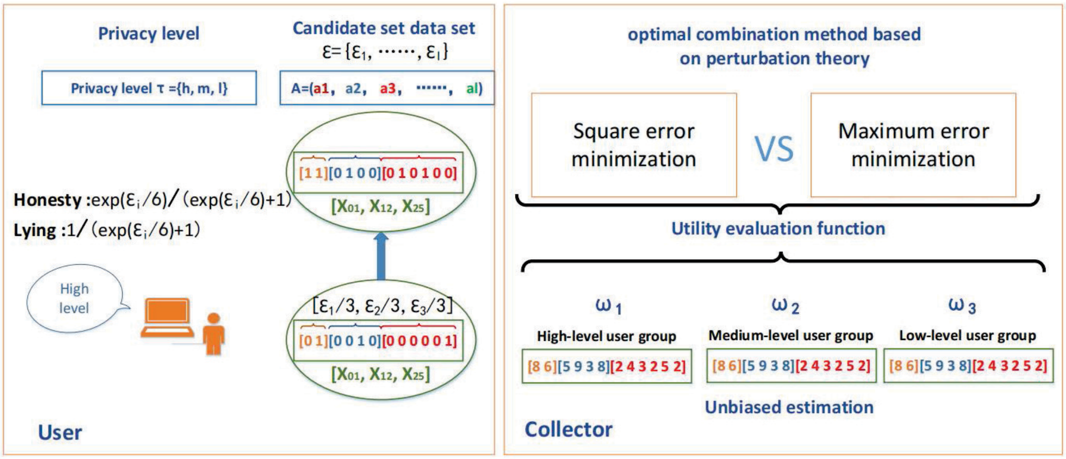


FIGURE 2. Collection Data Privacy Protection Publishing. If the user chooses a high privacy level $\tau = \{high, high, high\}$, then according to Algorithm III-C, the user has a probability of $\frac{\exp(\epsilon_i/6)}{\exp(\epsilon_i/6)+1}$ telling the truth and $\frac{1}{\exp(\epsilon_i/6)+1}$ telling lies for all the attributes. Assume that the bitmap of the privacy data held by the user is [[01][001][00001]], and the bitmap after the disturbance is [[11][0100][01000]]. The user sends the disturbed bitmap to the aggregator. The aggregator makes statistics on the disturbed user data according to the privacy level, and then estimates the real attributes of different privacy levels according to Algorithm 2.

of candidate values for the i attribute is k_i.

$$\widetilde{\mathbf{H}}_i = \begin{bmatrix} \tilde{h}_{i1}^{high} & \cdots & \tilde{h}_{ik_i}^{high} \\ \tilde{h}_{i1}^{mid} & \cdots & \tilde{h}_{ik_i}^{mid} \\ \tilde{h}_{i1}^{low} & \cdots & \tilde{h}_{ik_i}^{low} \end{bmatrix}.$$

Let $\Omega = [\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_l]^T$ represents the weight used to estimate the distribution, where $\mathbf{W}_i = [\frac{n\omega_i^{high}}{n^{high}}, \frac{n\omega_i^{mid}}{n^{mid}}, \frac{n\omega_i^{low}}{n^{low}}]$ denotes the weight of the combination, that is, each attribute is assigned three weights. So the combination result of the i-th attribute is

$$\widehat{\mathbf{H}}_i = \mathbf{W}_i\widetilde{\mathbf{H}}_i,\ \omega_i^{high} + \omega_i^{mid} + \omega_i^{low} = 1 \tag{2}$$

Then, on the i-th attribute, we can get the square sum residue between the unbiased estimate and the real result

$$\mathbb{E}(RSS(\mathbf{W}_i)) = \mathbb{E}[\mathbf{H}_i - \mathbf{W}_i\widetilde{\mathbf{H}}_i][\mathbf{H}_i - \mathbf{W}_i\widetilde{\mathbf{H}}_i]^T \tag{3}$$

Thus, the residual of the sum of squares between the unbiased estimate on all attributes and the true result can be expressed as

$$\mathbb{E}(RSS(\mathbf{W})) = \sum_{i=1}^{l} \mathbb{E}[\mathbf{H}_i - \mathbf{W}_i\widetilde{\mathbf{H}}_i][\mathbf{H}_i - \mathbf{W}_i\widetilde{\mathbf{H}}_i]^T \tag{4}$$

It can be seen from the Equation (4) that the weight $\mathbf{W}$ affects the accuracy of the evaluation results. Equation (4) is a kind of linear regression problem. If the evaluation standard is to minimize the mean square error, it can be solved by the least square method [26]; If the evaluation criterion is to minimize the maximum error, the perturbation method proposed in this paper can be used to solve the problem.

A. OPTIMAL COMBINATION METHOD OF HIGH DIMENSIONAL DATA

Next, we will illustrate the process of identifying the optimal weight based on various evaluation criteria. This paper addresses high-dimensional, heterogeneous data collections and accommodates users with the flexibility to choose from three levels of privacy. As can be seen from the Equation (4) that there are $3l$ parameters for weight Ω to be calculated. In fact, using the least square method to calculate the residual sum of square of $\widehat{\mathbf{H}}_i$ in Equation (4) is equivalent to calculating the minimum mean square estimation error of $\widehat{\mathbf{H}}_i$. Let's start with a lemma.

Lemma 1: $\widehat{\mathbf{H}}_i$ is the unbiased estimation result of real statistical data $\mathbf{H}_i$.

Proof: $\mathbb{E}(\widehat{\mathbf{H}}_i) = \sum_{\tau=1}^{3} \frac{n\omega_i^{\tau}}{n_\tau}\mathbb{E}(\widetilde{\mathbf{H}}_i^{\tau})$, Assume that there are n_τ users who choose privacy level τ, and $\widetilde{\mathbf{H}}_i^{\tau}$ is the unbiased estimation results on n_τ, it's easy to deduce $\frac{\mathbb{E}(\widetilde{\mathbf{H}}_i^{\tau})}{n\tau} = \frac{\mathbf{H_i}}{n}$. This is because the selection of user privacy level is independent of the specific private data held, so the estimate of the probability of different candidate values on the sample space n_τ is equivalent to the probability distribution on the entire sample space n. We can get

$$\mathbb{E}(\widehat{\mathbf{H}}_i) = \sum_{\tau=1}^{3} \frac{n\omega_i^{\tau}}{n_\tau}\frac{n_\tau\mathbf{H}_i}{n} = \sum_{\tau=1}^{3} \omega_i^{\tau}\mathbf{H}_i = \mathbf{H}_i \tag{5}$$

that is

$$\mathbb{E}(\widehat{\mathbf{H}}_i) = \mathbf{H}_i \tag{6}$$

Therefore, the lemma is proved. □

Based on unbiased estimation characteristics, the mean square error of $\widehat{\mathbf{H}}_i$ is calculated as follows

$$\begin{aligned} MSE(\widehat{\mathbf{H}}_i) &= \mathbb{E}[||\widehat{\mathbf{H}}_i - \mathbf{H}_i||_2^2] = \mathbb{E}[\sum_{j=1}^{k_i}(\widehat{\mathbf{H}}_{ij} - \mathbf{H}_{ij})^2] \\ &= \sum_{j=1}^{k_i}\mathbb{E}[\widehat{\mathbf{H}}_{ij}^2 - 2\widehat{\mathbf{H}}_{ij}\mathbf{H}_{ij} + \mathbf{H}_{ij}^2] \\ &= \sum_{j=1}^{k_i}\mathbb{E}[\widehat{\mathbf{H}}_{ij}^2] - \sum_{j=1}^{k_i}\mathbf{H}_{ij}^2 \end{aligned} \tag{7}$$

Since $\mathbb{E}(\widehat{\mathbf{H}}_{ij}) = \mathbf{H}_{ij}$, and $\mathbb{E}[\widehat{\mathbf{H}}_{ij}^2] = Var[\widehat{\mathbf{H}}_{ij}] + \mathbb{E}^2[\widehat{\mathbf{H}}_{ij}]$, Therefore, it can be obtained from the Equation (7).

$$\begin{aligned} MSE(\widehat{\mathbf{H}}_i) &= \sum_{j=1}^{k_i} Var[\widehat{\mathbf{H}}_{ij}] \\ &= \sum_{j=1}^{k_i}\sum_{\tau=1}^{3}(\frac{n\omega_i^\tau}{n_\tau})^2 Var[\widetilde{H}_{ij}^\tau] \\ &= \sum_{j=1}^{k_i}\sum_{\tau=1}^{3}(\frac{n\omega_i^\tau}{n_\tau})^2(\frac{\exp(\epsilon_{i\tau}/2)+1}{\exp(\epsilon_{i\tau}/2)-1})^2 Var[H'^{\tau}_{ij}] \\ &= \sum_{j=1}^{k_i}\sum_{\tau=1}^{3}(\frac{n\omega_i^\tau}{n_\tau})^2(\frac{\exp(\epsilon_{i\tau}/2)+1}{\exp(\epsilon_{i\tau}/2)-1})^2 n_\tau(\frac{\exp(\epsilon_{i\tau}/2)}{(\exp(\epsilon_{i\tau}/2)+1)^2}) \\ &= \sum_{\tau=1}^{3}\frac{k_i(n\omega_i^\tau)^2\exp(\epsilon_{i\tau}/2)}{n_\tau(\exp(\epsilon_{i\tau}/2)-1)^2} \end{aligned} \tag{8}$$

where H'^{τ}_{ij} denotes the statistical results of the data after the disturbance of the j-th candidate value of the i-th attribute. It has been proved that $\widehat{\mathbf{H}}_i$ is an unbiased estimate in Lemma 1, the sum of its squares and the expectation of residual $RSS(\mathbf{W})$ is equal to the mean square error of $\widehat{\mathbf{H}}_i$. Therefore, the optimal weight obtained by the least square method is consistent with the weight obtained by minimizing the mean square error. For simplicity, the optimal weight problem is directly solved by minimizing the mean square error. Therefore, in the l-dimensional set data, the weight value satisfying the user's personalized local differential privacy protection and utility maximization can be obtained by solving the following equation:

$$\begin{cases} \min & \sum_{\tau=1}^{3}\frac{k_i(n\omega_i^\tau)^2\exp(\epsilon_{i\tau}/2)}{n_\tau(\exp(\epsilon_{i\tau}/2)-1)^2} \\ & \sum_{\tau=1}^{3}\omega_i^\tau = 1 \end{cases} \tag{9}$$

where $i = 1, \cdots, l$, then

Theorem 2: when

$$\omega_i^\tau = \frac{D_\tau}{\sum_{\tau=1}^{3} D_\tau} \tag{10}$$

where $D_\tau = n_\tau\frac{(\exp(\epsilon_{i\tau}/2)-1)^2}{\exp(\epsilon_{i\tau}/2)}$,*equation* (9)*can get minimum value, and have*

$$MSE(\widehat{\mathbf{H}}_i) = \frac{k_i n^2}{\sum_{\tau=1}^{3} D_\tau} \tag{11}$$

Proof: Equation (9) is a minimum optimization problem under conditional constraints. By calculating the partial derivative of ω_i^τ, it can be concluded that the extreme point of the equation is the optimal parameter that minimizes the mean square error. Using the Lagrange method, the equation can be changed into

$$\ell(\mathbf{W}_i) = \sum_{\tau=1}^{3}\frac{k_i(n\omega_i^\tau)^2\exp(\epsilon_{i\tau}/2)}{n_\tau(\exp(\epsilon_{i\tau}/2)-1)^2} + C(1 - \sum_{\tau=1}^{3}\omega_i^\tau) \tag{12}$$

where C is a constant greater than zero. By calculating the partial derivative of ω_i^τ, we get

$$\frac{\partial\ell(\mathbf{W}_i)}{\partial\omega_i^\tau} = \frac{2k_i n^2\omega_i^\tau\exp(\epsilon_{i\tau}/2)}{n_\tau(\exp(\epsilon_{i\tau}/2)-1)^2} - C \tag{13}$$

Let $\frac{\partial\ell(\mathbf{W}_i)}{\partial\omega_i^\tau} = 0$, then have

$$\omega_i^\tau = \frac{Cn_\tau(\exp(\epsilon_{i\tau}/2)-1)^2}{2k_i n^2\exp(\epsilon_{i\tau}/2)} \tag{14}$$

where $\sum_{\tau=1}^{3}\omega_i^\tau = 1$. Bring in Equation (14) can get

$$\sum_{\tau=1}^{3}\frac{Cn_\tau(\exp(\epsilon_{i\tau}/2)-1)^2}{2k_i n^2\exp(\epsilon_{i\tau}/2)} = 1 \tag{15}$$

have

$$C = \frac{2k_i n^2}{\sum_{\tau=1}^{3} n_\tau\frac{(\exp(\epsilon_{i\tau}/2)-1)^2}{\exp(\epsilon_{i\tau}/2)}} \tag{16}$$

Substitute C back to the Equation (14), we have

$$\omega_i^\tau = \frac{n_\tau\frac{(\exp(\epsilon_{i\tau}/2)-1)^2}{\exp(\epsilon_{i\tau}/2)}}{\sum_{\tau=1}^{3} n_\tau\frac{(\exp(\epsilon_{i\tau}/2)-1)^2}{\exp(\epsilon_{i\tau}/2)}} = \frac{D_\tau}{\sum_{\tau=1}^{3} D_\tau} \tag{17}$$

where $D_\tau = n_\tau\frac{(\exp(\epsilon_{i\tau}/2)-1)^2}{\exp(\epsilon_{i\tau}/2)}$. then

$$MSE(\widehat{\mathbf{H}}_i) = \frac{k_i n^2}{\sum_{\tau=1}^{3} D_\tau} \tag{18}$$

So the theorem is proved. □

The value of ω_i^τ is the weight value calculated when the grade of the i-th attribute in the set data is τ, Similarly, the ownership values of 3-levels of l attributes can be calculated. The reasoning proof of Theorem 2 is based on the combination of weights obtained by minimizing the mean square error. The minimum mean square error serves as a valuable evaluation method in some cases, yet it's crucial to recognize that it may not always be the optimal approach. Minimizing the mean square error could result in disregarding the error values of certain individuals, leading to uneven error distribution and excessive error values for some. For seasonal commodities like watermelon and moon cakes, it's acceptable for estimation errors to fluctuate within a certain range due to minimal difference between purchase and sales. However, a substantial estimation error for a specific item can result in significant waste, contrary to the supermarket's desired outcome, even if the estimation of other items is highly accurate.

In order to solve this problem, this section studies the optimization method of frequency estimation based on perturbation theory, that is, the perturbation method is used to minimize the maximum estimated error on the premise of reducing the overall error as much as possible, and finally all the error values tend to be stable.

B. OPTIMIZATION METHOD BASED ON PERTURBATION THEORY

Since the mid-1980s, perturbation analysis has received enough attention in numerical linear algebra. Component perturbation theory and error analysis have been widely used in linear systems [27], [28], Matrix inversion [29], [30], Matrix decomposition [31], Least squares problem [32], [33], Eigenvalue and singular value solution [34], [35]. The problem we have encountered in this section is the approximate solution of the least squares. The norm perturbation theory was used to solve the perturbation problem of unitary least squares in [33], and the upper bound of error between the true solution and perturbation solution was given. Look at the following lemma:

Lemma 2: [33] Let $A \in \mathbb{R}^{m\times n}(m \geq n)$. A and $A+\Delta A$ are full rank. Let

$$\begin{aligned} &||Ax-b||_2 = \min, \, r = b - Ax, \\ &||(A+\Delta A)y - (b+\Delta b)||_2 = \min, \\ &||\Delta A||_2 \leq \epsilon ||A||_2, \, ||\Delta b||_2 \leq \epsilon ||b||_2. \end{aligned} \tag{19}$$

the premise is $\mathcal{K}_2(A)\epsilon < 1$, then have

$$\frac{||x-y||_2}{||x||_2} \leq \frac{\mathcal{K}_2(A)\epsilon}{1-\mathcal{K}_2(A)\epsilon}(1+\frac{||b||_2}{||A||_2||x||_2}+\mathcal{K}_2(A)\frac{||r||_2}{||A||_2||x||_2}) \tag{20}$$

where$\mathcal{K}_2(A) = ||A||_2||A^{+}||_2$, A^{+}represents the pseudo inverse of matrix A.

The above lemma gives the upper bound of the error between the solution x and y in the matrix equation. In fact, there is a similar relationship between the residuals of $||Ax - b||_2$ and $||(A+\Delta A)y - (b+\Delta b)||_2$. The basic idea of the perturbation method used in this section to solve the least squares problem is consistent with the above lemma, that is, adding perturbation to the matrix A to optimize the least squares solution. Different from the above lemma, the problem to be solved in this section is the equilibrium problem of a class of residual series, that is, to reduce the maximum error in the residual series. The following sections present problems and theorem proofs to verify the validity of the proposed method.

First, let's calculate the weight combination through the least square method. On the i-th attribute, there are

$$\widehat{\mathbf{H}}_i = \mathbf{W}_i\widetilde{\mathbf{H}}_i, \; \omega_i^{high} + \omega_i^{mid} + \omega_i^{low} = 1 \tag{21}$$

In order to express more simply and clearly, Let $\bar{\omega}_i^{\tau} = \frac{n\omega_i^{\tau}}{n_\tau}$, then we have $\mathbf{W}_i = [\bar{\omega}_i^{high}, \bar{\omega}_i^{mid}, \bar{\omega}_i^{low}]$. It has also been proved in Lemma 1 that $\widehat{\mathbf{H}}_i$ is an unbiased estimate of $\mathbf{H}_i$, so the following approximate equation is obtained

$$\mathbf{H}_i \approx \mathbf{W}_i\widetilde{\mathbf{H}}_i, \; \omega_i^{high} + \omega_i^{mid} + \omega_i^{low} = 1 \tag{22}$$

By simple transformation of Equation (21), we can get

$$\mathbf{H}_i\widetilde{\mathbf{H}}_i^T \approx \mathbf{W}_i\widetilde{\mathbf{H}}_i\widetilde{\mathbf{H}}_i^T \tag{23}$$

This leads to

$$\mathbf{W}_i = (\mathbf{H}_i\widetilde{\mathbf{H}}_i^T)(\widetilde{\mathbf{H}}_i\widetilde{\mathbf{H}}_i^T)^{-1} \tag{24}$$

The results obtained by the above least square method are consistent with the results obtained in the Theorem 2, which is the optimal solution obtained on the basis of minimizing the mean square error, that is, it guarantees the minimization of $\mathbb{E}[\sum_{j=1}^{k_i}(\widehat{\mathbf{H}}_{ij} - \mathbf{H}_{ij})^2]$. But there is no guarantee that the minimization of

$$d_k = \max_{1\leq j\leq k_i} |\widehat{\mathbf{H}}_{ij} - \mathbf{H}_{ij}| \tag{25}$$

The purpose of this program is to find the equilibrium point between the maximum single point error and the minimum mean square error. Look at a theorem.

Theorem 3: Assume that the solution of equation (24) obtained by least square method is $\mathbf{W}_i^{(0)}$, take $\mathbf{W}_i^{(0)}$ as initial value, The parameter obtained by using the perturbation method after 1 iteration is $\mathbf{W}_i^{(1)}$, If constant perturbation δ_1 satisfies

$$\begin{aligned} \mathrm{sign}(\delta_1) &= \mathrm{sign}(\mathbf{H}_{ik} - \widehat{\mathbf{H}}_{ik}^{(0)}) \\ |\delta_1| &\leq \frac{d_k^{(0)} - d_j^{(0)}}{|\widetilde{\mathbf{H}}_{ij}^T(\widetilde{\mathbf{H}}_i\widetilde{\mathbf{H}}_i^T)^{-1}\widetilde{\mathbf{H}}_{ij}|} \\ |\delta_1| &\leq \frac{d_k^{(0)}}{\widetilde{\mathbf{H}}_{ik}^T(\widetilde{\mathbf{H}}_i\widetilde{\mathbf{H}}_i^T)^{-1}\widetilde{\mathbf{H}}_{ik}} \end{aligned} \tag{26}$$

then we have

$$\max_{1\leq j\leq k_i} d_j^{(1)} \leq \max_{1\leq j\leq k_i} d_j^{(0)}$$

where

$$\begin{aligned} d_j^{(1)} &= |\widehat{\mathbf{H}}_{ij}^{(1)} - \mathbf{H}_{ij}| \\ d_j^{(0)} &= |\widehat{\mathbf{H}}_{ij}^{(0)} - \mathbf{H}_{ij}| \\ d_k^{(0)} &= \max_{1\leq j\leq k_i} d_j^{(0)} \\ \widehat{\mathbf{H}}_{ij}^{(0)} &= \bar{\omega}_i^{high(0)}\widetilde{h}_{ij}^{high} + \bar{\omega}_i^{mid(0)}\widetilde{h}_{ij}^{mid} + \bar{\omega}_i^{low(0)}\widetilde{h}_{ij}^{low} \\ \widehat{\mathbf{H}}_{ij}^{(1)} &= \bar{\omega}_i^{high(1)}\widetilde{h}_{ij}^{high} + \bar{\omega}_i^{mid(1)}\widetilde{h}_{ij}^{mid} + \bar{\omega}_i^{low(1)}\widetilde{h}_{ij}^{low} \end{aligned} \tag{27}$$

Proof: Let

$$d_k^{(0)} = \max_{1\leq j\leq k_i} d_j^{(0)} \tag{28}$$

That is, the maximum error is obtained at the k-th position. Define δ_1 as a non-negative constant perturbation. Let $\mathbf{H}_i^{(0)} = \mathbf{H}_i$, and let

$$\mathbf{H}_i^{(1)} = (h_{i1}, \cdots, h_{ik} + \delta_1, \cdots, h_{ik_i}) = \mathbf{H}_i^{(0)} + \theta_k^{(1)} \tag{29}$$

where $\theta_k^{(1)} = \overbrace{(0, 0, \ldots, \delta_1}^{k}, \ldots, 0)$. Let $\mathbf{H}_i^{(1)}$ replace $\mathbf{H}_i$ to solve the Equation (23) again:

$$\mathbf{H}_i^{(1)} \widetilde{\mathbf{H}}_i^T \approx \mathbf{W}_i^{(1)} \widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T \tag{30}$$

then obtain

$$\mathbf{W}_i^{(1)} = (\mathbf{H}_i^{(1)} \widetilde{\mathbf{H}}_i^T)(\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \tag{31}$$

where $\mathbf{W}_i^{(1)}$ is the perturbation parameter obtained after the first iteration, which can be expressed as $\mathbf{W}_i^{(1)} = \{\bar{\omega}_i^{high(1)}, \bar{\omega}_i^{mid(1)}, \bar{\omega}_i^{low(1)}\}$ Then $\widehat{\mathbf{H}}_{ij}^{(0)}$ and $\widehat{\mathbf{H}}_{ij}^{(1)}$ can be obtained in the qeuation (27). We have

$$\begin{aligned}
&\widehat{\mathbf{H}}_{ij}^{(1)} - \widehat{\mathbf{H}}_{ij}^{(0)} \\
&= \mathbf{W}_i^{(1)} \widetilde{\mathbf{H}}_{ij} - \mathbf{W}_i^{(0)} \widetilde{\mathbf{H}}_{ij} \\
&= (\mathbf{H}_i^{(1)} \widetilde{\mathbf{H}}_i^T)(\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ij} - (\mathbf{H}_i^{(0)} \widetilde{\mathbf{H}}_i^T)(\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ij} \\
&= \theta_k^{(1)} \widetilde{\mathbf{H}}_i^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ij} \\
&= \delta_1 \widetilde{\mathbf{H}}_{ij}^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ij}
\end{aligned} \tag{32}$$

where $\widetilde{\mathbf{H}}_{ij} = [\widetilde{h}_{ij}^{high}, \widetilde{h}_{ij}^{mid}, \widetilde{h}_{ij}^{low}]^T$. Then

$$\begin{aligned}
d_j^{(1)} &= |\widehat{\mathbf{H}}_{ij}^{(1)} - \mathbf{H}_{ij}| = |\mathbf{H}_{ij} - \widehat{\mathbf{H}}_{ij}^{(0)} - (\widehat{\mathbf{H}}_{ij}^{(1)} - \widehat{\mathbf{H}}_{ij}^{(0)})| \\
&= |\mathbf{H}_{ij} - \widehat{\mathbf{H}}_{ij}^{(0)} - \delta_1 \widetilde{\mathbf{H}}_{ij}^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ij}|
\end{aligned} \tag{33}$$

(1)when $j = 1, \cdots, l$ and $j \neq k$

$$d_j^{(1)} \leq d_j^{(0)} + |\delta_1| |\widetilde{\mathbf{H}}_{ij}^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ij}| \tag{34}$$

So, as long as δ_1 satisfied

$$|\delta_1| \leq \frac{d_k^{(0)} - d_j^{(0)}}{|\widetilde{\mathbf{H}}_{ij}^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ij}|} \tag{35}$$

then have

$$d_j^{(1)} = |\widehat{\mathbf{H}}_{ij}^{(1)} - \mathbf{H}_{ij}| \leq d_k^{(0)} = \max_j |\widehat{\mathbf{H}}_{ij}^{(0)} - \mathbf{H}_{ij}| \tag{36}$$

(2)when $j = k$

$$d_k^{(1)} = |\mathbf{H}_{ik} - \widehat{\mathbf{H}}_{ik}^{(0)} - \delta_1 \widetilde{\mathbf{H}}_{ik}^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ik}| \tag{37}$$

As long as δ_1 meets

$$sign(\delta_1) = sign(\mathbf{H}_{ik} - \widehat{\mathbf{H}}_{ik}^{(0)}) \tag{38}$$

and

$$|\delta_1 \widetilde{\mathbf{H}}_{ik}^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ik}| \leq d_k^{(0)} \tag{39}$$

or

$$|\delta_1| \leq \frac{d_k^{(0)}}{\widetilde{\mathbf{H}}_{ik}^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ik}} \tag{40}$$

have

$$d_k^{(1)} = |\mathbf{H}_{ik} - \widehat{\mathbf{H}}_{ik}^{(0)} - \delta_1 \widetilde{\mathbf{H}}_{ik}^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ik}| \leq d_k^{(0)} \tag{41}$$

The inequality $\widetilde{\mathbf{H}}_{ik}^T (\widetilde{\mathbf{H}}_i \widetilde{\mathbf{H}}_i^T)^{-1} \widetilde{\mathbf{H}}_{ik} \geq 0$ is used. The theorem is proved. □

The perturbation method can reduce the maximum error when iterating once, and it has the same effect when iterating n times. Let

$$d_{k_n}^{(n)} = \max_{1 \leq j \leq k_i} |\mathbf{H}_{ij} - \widehat{\mathbf{H}}_{ij}^{(n)}| \tag{42}$$

be the maximum estimation error after iteration n times. $d_{k_n}^{(n)}$ is a monotonically decreasing sequence. First, let's look at the following definitions

Definition 3: $C = \lim_n d_{k_n}^{(n)}$ is the error limit of frequency estimation of personalized local differential privacy model, actually $C > 0$.

From this definition, the following theorems can be obtained

Theorem 4: If $\mathbf{H}_{ij} > C > 0$, Then for sufficiently large iterations n, we have $\widehat{\mathbf{H}}_{ij}^{(n)} > 0$.

Proof: Assume $\mathbf{H}_{ij} - \widehat{\mathbf{H}}_{ij}^{(n)} \geq 0$ (If $\mathbf{H}_{ij} - \widehat{\mathbf{H}}_{ij}^{(n)} < 0$, we have $\widehat{\mathbf{H}}_{ij}^{(n)} > \mathbf{H}_{ij} \geq 0$, So the theorem is proved.) Since $\mathbf{H}_{ij} > C > 0$, then $\exists\, \epsilon_0 > 0$ make

$$\mathbf{H}_{ij} > C + \epsilon_0 > 0 \tag{43}$$

Consider the fact that $d_{k_n}^{(n)}$ is monotonically decreasing from condition

$$\lim_{n \to \infty} d_{k_n}^{(n)} = \lim_{n \to \infty} \max_{1 \leq j \leq k_i} (\mathbf{H}_{ij} - \widehat{\mathbf{H}}_{ij}^{(n)}) = C \tag{44}$$

Therefore, for any ϵ, as long as the number of iterations n is large enough, we have

$$\mathbf{H}_{ij} - \widehat{\mathbf{H}}_{ij}^{(n)} \leq d_{k_n}^{(n)} = \max_{1 \leq j \leq k_i} (\mathbf{H}_{ij} - \widehat{\mathbf{H}}_{ij}^{(n)}) < C + \epsilon \tag{45}$$

Without losing generality, let $\epsilon = \epsilon_0$, so we can get

$$\mathbf{H}_{ij} - C - \epsilon_0 < \widehat{\mathbf{H}}_{ij}^{(n)} \tag{46}$$

then

$$\widehat{\mathbf{H}}_{ij}^{(n)} > 0 \tag{47}$$

□

The result $\widehat{\mathbf{H}}_{ij}^{(n)} > 0$ of Theorem 4 is consistent with the non-negative weight combination parameter, that is, as long as $\mathbf{H}_{ij} > C > 0$ is met, it can be ensured that the weight parameters obtained are non-negative, which is actually required in this section. Because in all privacy levels, the estimated results are positively correlated. If the weight is negative, it is actually an incorrect frequency estimation. This theorem is particularly important in conditional heteroscedastic models. The specific perturbation algorithm is shown in Algorithm 3.

Algorithm 3 Perturbation Method

```
Input: $\mathbf{H}_i$-Real statistical results of users; $\widetilde{\mathbf{H}}_i$-Unbiased estimation result matrix of different levels; $iter$-Iterations
Output: $\mathbf{W}_\alpha$-Weight parameter vector
initialization $[row, column] = size(\mathbf{H}_i)$
for $i = 1$ to $iter$ do
    $[\mathbf{W}_\alpha, bint, r, rint, stats] = regress(\mathbf{H}_i, \widetilde{\mathbf{H}}_i)$
    $r = (\mathbf{H}_i - \widetilde{\mathbf{H}}_i * \mathbf{W}_\alpha)$
    $[d_k, k] = \max(abs(r))$
    for $t = 1$ to $row$ do
        if $t \neq k$ then
            $M(t) = \frac{|r(k)|-|r(t)|}{|\widetilde{\mathbf{H}}_{it}^T(\widetilde{\mathbf{H}}_i\widetilde{\mathbf{H}}_i^T)^{-1}\widetilde{\mathbf{H}}_{it}|}$
        else
            $M(k) = \frac{|r(k)|}{|\widetilde{\mathbf{H}}_{ik}^T(\widetilde{\mathbf{H}}_i\widetilde{\mathbf{H}}_i^T)^{-1}\widetilde{\mathbf{H}}_{ik}|}$
        end if
    end for
    $\delta = \min(M)$
    if $r(k) < 0$ then
        $\delta = -\delta$
    end if
    $\theta = (\overbrace{0, 0, \ldots, \delta}^{k}, \ldots, 0)$
    $\mathbf{H}_i = \mathbf{H}_i + \theta$
end for
```

V. ANALYSIS OF PERSONALIZED LOCAL DIFFERENTIAL PRIVACY MECHANISM

A. PERSONALIZED LOCAL DIFFERENTIAL PRIVACY MECHANISM PRIVACY GUARANTEE AND SECURITY ANALYSIS

The calculation steps involved in this paper are all based on the private data generated by the binary randomized response mechanism. The binary randomized response mechanism has been proved to meet the requirements of ϵ-local differential privacy. The optimization methods based on the criterion of mean square error minimization and maximum error minimization are both post-processing processes. Those based on post-processing also meet the local differential privacy [2]. The following theorem can be obtained.

Theorem 5: If the privacy data holder has l attributes in total, the privacy level set for attribute i is τ_i, the corresponding privacy budget is ϵ_{τ_i}, the output results of OC and OP optimization methods satisfy ϵ_{τ_i}-Local differential privacy guarantee on the i-th attribute, and satisfy $\sum_{i=1}^{l}\epsilon_{\tau_i}$-Local differential privacy on all attributes. Further, if the initial privacy budget of the i-th attribute is ϵ_i^*, then data records from different privacy levels on the i-th attribute meet the requirement of ϵ_i^*-local differential privacy, which satisfies ϵ-local differential privacy on all attributes.

Proof: Let's first prove that the ϵ_{τ_i}-local differential privacy is satisfied on the i-th attribute. Let the dimension of the i-th attribute be k_i, and for any two records t and t' in the data set, express them as bit bitmaps in the following forms:

$$B_t = (\overbrace{0, \ldots, 1}^{t^*}, \ldots, 0), B_{t'} = (\overbrace{0, \ldots, 1}^{t'^*}, \ldots, 0) \tag{48}$$

where the t^*-th position of the recorded t is 1, and the other positions are 0; The t'^*-th position of recording t' is 1, and the rest positions are 0. Binary randomized response equation is as follows:

$$\mathcal{F}(b_i'|b_i) = \begin{cases} p, & b_i = b_i' \\ 1-p, & b_i = 1 - b_i' \end{cases} \tag{49}$$

where $p = \frac{e^{\epsilon_{\tau_i}/2}}{e^{\epsilon_{\tau_i}/2}+1}$, and b_i is the value of the i-th position of B_t. Then for any identical output z, we just need to prove

$$e^{-\epsilon_{\tau_i}} \leq \frac{Pr[\mathcal{F}(t) = z]}{Pr[\mathcal{F}(t') = z]} \leq e^{\epsilon_{\tau_i}} \tag{50}$$

Because the global sensitivity of OBRR is 2, that is, for any same output, B_t and $B_{t'}$ have at most two different perturbation probabilities. So the max value of $\frac{Pr[\mathcal{F}(t)=z]}{Pr[\mathcal{F}(t')=z]}$ is $\max\{\frac{p^2}{(1-p)^2}, \frac{(1-p)^2}{p^2}\}$, that is $\max\{e^{\epsilon_{\tau_i}}, e^{-\epsilon_{\tau_i}}\}$, Thus the inequality is established.

For each attribute, the binary randomized response mechanism satisfies ϵ_{τ_i}-Local differential privacy, Let $\mathbf{x} = \{x_1, \cdots, x_l\}, \mathbf{x}' = \{x_1', \cdots, x_l'\}$ are any two records with l attributes in the dataset, $\mathbf{z} = \{z_1, \cdots, z_l\}$ is a bit sequence with length $k_1 + \cdots + k_l$, where z_i is a bit sequence with length k_i, then

$$\begin{aligned} &\frac{Pr[\mathcal{F}_{1,\cdots,l}(x_1, \cdots, x_l) = (z_1, \cdots, z_l)]}{Pr[\mathcal{F}_{1,\cdots,l}(x_1', \cdots, x_l') = (z_1, \cdots, z_l)]} \\ &= \prod_{j=1}^{l} \frac{Pr[\mathcal{F}_1(x_j) = z_j]}{Pr[\mathcal{F}_1(x_j') = z_j]} \\ &\leq \prod_{j=1}^{l} e^{\epsilon_{\tau_j}} \\ &= e^{\epsilon_{\tau_1}+\cdots+\epsilon_{\tau_l}} \\ &= \exp(\sum_{i=1}^{l}\epsilon_{\tau_i}) \end{aligned} \tag{51}$$

Similarly, $\frac{Pr[\mathcal{F}_{1,\cdots,l}(x_1,\cdots,x_l)=(z_1,\cdots,z_l)]}{Pr[\mathcal{F}_{1,\cdots,l}(x_1',\cdots,x_l')=(z_1,\cdots,z_l)]} \geq \exp(\sum_{i=1}^{l} -\epsilon_{\tau_i})$. Furthermore, for the attribute i, there are three independent data sets, which are subsets of records from high, medium and low privacy levels D_{hi}, D_{mi}, D_{li}. Suppose that the random disturbance mechanism on different subsets is $\mathcal{F}_{hi}$, $\mathcal{F}_{mi}$ and $\mathcal{F}_{li}$, And these three disturbance mechanisms meet ϵ_{h_i}, ϵ_{m_i}, ϵ_{l_i}-local differential privacy. Let $\mathcal{F}_i = \cup_\tau \mathcal{F}_{\tau i}$, by definition

$$Pr[\mathcal{F}_i(t) = z] \leq e^{\epsilon_{\tau_i}} Pr[\mathcal{F}_i(t') = z] \tag{52}$$

If

$$Pr[\mathcal{F}_i(t) = z] \leq e^{\epsilon_i} Pr[\mathcal{F}_i(t') = z] \tag{53}$$

if and only if $\epsilon_i \geq \epsilon_{\tau i}$. So ϵ_i can be expredssed as

$$\begin{aligned}\epsilon_i &= \min\{\epsilon_i | \wedge_{\tau=\{h,m,l\}} (\epsilon_i \geq \epsilon_{\tau i})\} \\ &= \min\{\epsilon_i | \epsilon_i \geq \max_{\tau=\{high,mid,low\}} \epsilon_{\tau i}\} \\ &= \max_{\tau=\{high,mid,low\}} \epsilon_{\tau i} \\ &\leq \epsilon_i^* \end{aligned} \tag{54}$$

Thus, it can be proved that the ϵ_i^*-local differential privacy guarantee is satisfied on the i-th attribute. Similarly, there are

$$\begin{aligned}&\frac{Pr[\mathcal{F}_{1,\cdots,l}(x_1,\cdots,x_l)=(z_1,\cdots,z_l)]}{Pr[\mathcal{F}_{1,\cdots,l}(x_1',\cdots,x_l')=(z_1,\cdots,z_l)]} \\ &= \prod_{j=1}^{l} \frac{Pr[\mathcal{F}_1(x_j)=z_j]}{Pr[\mathcal{F}_1(x_j')=z_j]} \\ &\leq \prod_{j=1}^{l} \exp(\max_{\tau=\{high,mid,low\}} \epsilon_{\tau j}) \\ &\leq \exp(\epsilon_1^* + \cdots + \epsilon_l^*) \\ &= e^{\epsilon}\end{aligned} \tag{55}$$

Thus, the theorem is proved. □

The threat of multi-level personalized local differential privacy mechanism can be summarized as follows. (1) Data reconstruction threat: If the perturbation mechanism is not strong enough or there are vulnerabilities, an attacker may be able to combine the disturbed data of multiple users to infer some sensitive information. (2) Collusion attack: When multiple users join forces, they may share information about the perturbing data or perturbing algorithms they receive, in an attempt to infer the original data of other users. (3) Threat of parameter disclosure: In the multi-level personalized differential privacy mechanism, the choice of parameters is crucial to protect privacy. If an attacker can obtain information about these parameters, they may use them to optimize their attack strategy.

In order to deal with these threats, the following measures are taken in our mechanism to overcome these problems: (1) Enhanced perturbation mechanism: We introduce noise intensity of different privacy levels in the multi-level personalized local differential privacy mechanism for users to choose to minimize the success rate of data reconstruction and collusion attacks. (2) Protect parameter security: Ensure the safe storage and transmission of privacy protection parameters (such as ϵ values) to prevent attackers from obtaining information about these parameters. Later, we will combine public key cryptography to ensure the safe transmission and storage of parameters, tamper-proof and so on. (3) Using theorem proof: In Theorem 5, we give that the mechanism proposed by me strictly meets the local differential privacy guarantee, and verify the effectiveness and security of the differential privacy mechanism through theorem proof to ensure its robustness under various attack scenarios.

B. ERROR BOUNDS AND COMPLEXITY OF PERSONALIZED LOCAL DIFFERENTIAL PRIVACY MECHANISM

It has been proved in Theorem 2 that when $\omega_i^\tau = \frac{D_\tau}{\sum_{\tau=1}^{3} D_\tau}$, the error can be taken to the minimum value $MSE(\widehat{\mathbf{H}}_i) = \frac{k_i n^2}{\sum_{\tau=1}^{3} D_\tau}$. This result is better than that when the ownership weight is 1, that is, $\mathbf{W}_i = \{1, 1, 1\}$. Therefore, in any case, the method proposed in this section has an error upper bound, which is

$$MSE(\widehat{\mathbf{H}}_i) \leq \sum_{\tau=1}^{3} \frac{k_i n_\tau \exp(\epsilon_{i\tau}/2)}{(\exp(\epsilon_{i\tau}/2)-1)^2} \tag{56}$$

As can be seen from the above formula, the relationship between privacy budget (i.e. privacy level) and data availability can be intuitively seen through the calculation error. Changes in privacy budgets will lead to changes in privacy levels, which will lead to changes in data availability. Specifically, the mean square estimate error is inversely proportional to the privacy budget. The larger the privacy budget, the lower the privacy level, the smaller the mean square estimation error, indicating the better the data availability; Conversely, smaller privacy budgets lead to lower data availability. For the user, the data disturbance mentioned in Algorithm 1 requires the computation of $O(d)$, and the communication complexity is also $O(d)$. In addition, no additional complexity is added. For data analysts, the optimization method OC based on the criterion of minimizing the mean square error needs to calculate additional weights of different privacy levels, which costs $O(1)$ computational complexity. The optimization criterion based on minimizing the maximum error requires additional calculation of weights with a complexity of $O(3k_i * d * m)$, where k_i is the dimension of the attribute, d is the binary bitmap length of all attributes, and m is the number of iterations. The number of iterations is often small. Actually, the number of iterations is directly related to the selected perturbation. In order to balance the minimum maximum error and the minimum mean square error, the number of iterations selected should not be too large. In addition, the calculation complexity of $O(n_\tau d + d)$ is required to estimate the frequency in different privacy levels separately, so the calculation complexity of $O(nd + 3d)$ is required to estimate the frequency in all levels, where n is the number of users. Therefore, the overall computational complexity for data analysts is $O((n + 3 + 3k_i m)d)$.

VI. SIMULATION EXPERIMENT OF PERSONALIZED LOCAL DIFFERENTIAL PRIVACY MECHANISM

Suppose that the privacy data value of each participant is extracted from histogram H, and H is generated uniformly and randomly in each aggregation process. The dimension of the dataset is $[n, d]$. The selection of data sets guarantees the following criteria. First, each participant can only vote l, that is, the sum of each row of the dataset binary matrix is l. Second, the total number of votes cast by all participants is $l * n$. Without losing generality, suppose there are 5 attributes, and each attribute has a different number of candidate values.

Without losing generality, select the dimension of each attribute as $\{k_1, k_2, k_3, k_4, k_5\} = \{5, 10, 15, 20, 25\}$. The data set generation algorithm is shown in Algorithm 4. The paper utilizes the optimal privacy budget allocation scheme from the OBRR method [23]. We conducted comparative experiments to demonstrate the correlation between privacy budget allocation and the number of users. Additionally, we tested the robustness of the proposed method with 1000 and 10000 users, respectively. Aggregators should ensure that the total amount of the privacy budget is limited and that the privacy budget allocated to each attribute and each user is manageable. This requires setting a reasonable initial privacy budget for the entire system according to the scenario requirements and the size of the user group. As the number of users increases or decreases or the sensitivity of data changes, the total privacy budget needs to be dynamically adjusted to ensure that it is always within reasonable bounds. In the experimental part, we assume that the range of the total privacy budget is $[1, 6]$, and the aggregator can decide which privacy budget to use as the upper limit according to the data availability under different privacy budget conditions. This section uses standardized square error ($NSE = \frac{SE}{n}$) as an indicator to measure the performance of the mechanism, where SE is the square error.

Algorithm 4 Dataset Generation Algorithm

Input: n- Number of participants; l- Total number of attributes; $\{k_1, k_2, \cdots, k_l\}$- Number of dimensions per attribute. $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_l\}$- Optimal privacy budget allocation scheme

Output: D_{n*l}-Data set.

```
initialization $index = 0$; $m = 0$
for $i = 1$ to $l$ do
    if $i \neq 1$ then
        $index = index + k_{i-1}$
    end if
    for $j = 1$ to $k_i$ do
        for $m = round(\frac{(j-1)n}{k_i} + 1)$ to $round(\frac{jn}{k_i})$ do
            $D[m, index + j] = 1$
        end for
    end for
end for
```

A. OPTIMAL COMBINATION METHOD OF HIGH DIMENSIONAL DATA, OC

Because the scenario targeted in this paper is an optimal combination method of high-dimensional data, the first step is to consider the allocation of the privacy budget. This section uses the optimal privacy budget allocation scheme proposed in [23] to allocate a reasonable privacy budget for each attribute. The allocation results are shown in Table 2. Based on the allocated privacy budget, the randomized response mechanism is used to disturb the user's private data, and then the frequency of the disturbed data is estimated. The experimental results are shown in Figure 3.

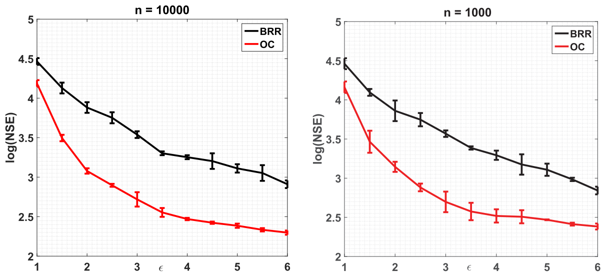
TABLE 2. The optimal privacy budget allocation scheme.

(a) n = 10000

k_i	ϵ=5	ϵ=10	ϵ=15	ϵ=20	ϵ=25
ine 1.0	0.0847	0.1067	0.1221	0.1344	0.1448
1.5	0.1078	0.1358	0.1555	0.1711	0.1844
2.0	0.1434	0.1807	0.2068	0.2276	0.2452
2.5	0.1789	0.2255	0.2581	0.2841	0.3060
3.0	0.2145	0.2703	0.3094	0.3405	0.3668
3.5	0.2502	0.3152	0.3608	0.3971	0.4277
4.0	0.2859	0.3602	0.4123	0.4538	0.4889
4.5	0.3215	0.4051	0.4637	0.5104	0.5498
5.0	0.3573	0.4501	0.5152	0.5670	0.6108
5.5	0.3930	0.4951	0.5667	0.6237	0.6719
6.0	0.4286	0.5400	0.6181	0.6803	0.7327

(b) n = 1000

k_i	ϵ=5	ϵ=10	ϵ=15	ϵ=20	ϵ=25
ine 1.0	0.0723	0.0911	0.1043	0.1148	0.1236
1.5	0.1078	0.1358	0.1555	0.1711	0.1843
2.0	0.1433	0.1806	0.2067	0.2275	0.2450
2.5	0.1789	0.2255	0.2581	0.2841	0.3060
3.0	0.2141	0.2697	0.3087	0.3398	0.3660
3.5	0.2502	0.3152	0.3608	0.3971	0.4278
4.0	0.2858	0.3600	0.4121	0.4536	0.4886
4.5	0.3214	0.4049	0.4635	0.5101	0.5495
5.0	0.3573	0.4501	0.5152	0.5671	0.6108
5.5	0.3930	0.4951	0.5668	0.6238	0.6719
6.0	0.4287	0.5401	0.6182	0.6804	0.7328



(a) n = 10000, d = 75 (b) n = 1000, d = 75

FIGURE 3. The relationship between the estimated histogram error measured by $\log_{10}$(NSE) and the privacy budget ϵ. The black line in the figure represents the total MSE (mean square error) of BRR method. The MSE of BRR obtained by directly adding the frequency estimation results of all different privacy levels and comparing them with the real estimation results, and the red line represents the total mean square error obtained by using the weighted combination method of OC.

The black line in the figure represents the total mean square error obtained by directly adding the frequency estimation results of all different privacy levels and comparing them with the real estimation results, and the red line represents the total mean square error obtained by using the weighted combination method of OC. Figure 3(a) and Figure 3(b) represent the error comparison results when the number of users is 10000 and 1000 respectively. In general, with the increase of privacy budget, the error gradually decreases, and OC method is superior to BRR method [24], with the overall error decreasing by about 60%. The estimated results in this section are the actual optimization results obtained based on

TABLE 3. The maximum estimation error.

(a) n = 10000

$log10$ iteration ϵ	1	2	3	4	5	6	7	8	9	10
ine 1.0	4.6184	4.6184	4.5966	4.5663	4.5645	4.5566	4.5541	4.5468	4.5449	4.5430
1.5	4.8277	4.8025	4.7486	4.7264	4.7173	4.7089	4.7014	4.6978	4.6955	4.6901
2.0	4.7433	4.7433	4.6838	4.6540	4.6287	4.6022	4.5892	4.5795	4.5794	4.5729
2.5	4.7625	4.7477	4.6901	4.6818	4.6719	4.6672	4.6639	4.6616	4.6611	4.6602
3.0	4.7216	4.7216	4.6690	4.6681	4.6503	4.6280	4.6227	4.6223	4.6086	4.6086
3.5	4.6723	4.6572	4.6489	4.6412	4.6402	4.6326	4.6310	4.6283	4.6272	4.6255
4.0	4.6482	4.6181	4.5774	4.5687	4.5630	4.5501	4.5470	4.5412	4.5396	4.5315
4.5	4.5431	4.5130	4.4869	4.4573	4.4541	4.4518	4.4504	4.4406	4.4393	4.4383
5.0	4.4968	4.4779	4.4569	4.4529	4.4436	4.4349	4.4197	4.4186	4.4180	4.4176
5.5	4.5261	4.5219	4.5024	4.4963	4.4902	4.4856	4.4722	4.4697	4.4683	4.4682
6.0	4.4780	4.4717	4.4534	4.4406	4.4194	4.4067	4.4067	4.4067	4.4067	4.3916

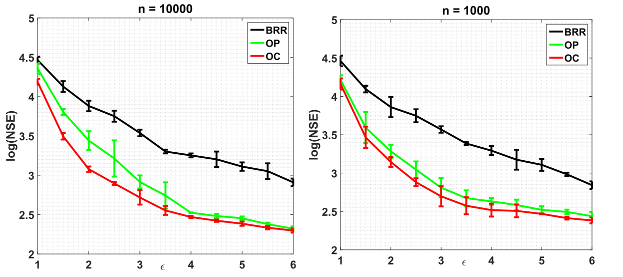
(b) n = 1000

$log10$ iteration ϵ	1	2	3	4	5	6	7	8	9	10
ine 1.0	3.0127	2.9568	2.9291	2.9103	2.9077	2.8937	2.8915	2.8561	2.8546	2.8388
1.5	3.0453	2.8930	2.8374	2.8250	2.8036	2.7930	2.7881	2.7873	2.7839	2.7822
2.0	3.0782	2.9541	2.8775	2.8734	2.8697	2.8665	2.8538	2.8515	2.8510	2.8505
2.5	2.9547	2.8861	2.8499	2.8467	2.8436	2.8417	2.8267	2.8256	2.8213	2.8170
3.0	2.9887	2.9318	2.8713	2.8553	2.8495	2.8321	2.8272	2.8101	2.8041	2.7940
3.5	3.0256	2.9793	2.9254	2.8887	2.8550	2.8288	2.8257	2.8213	2.8170	2.8170
4.0	2.9831	2.9086	2.8219	2.7927	2.7803	2.7740	2.7652	2.7621	2.7471	2.7392
4.5	2.9671	2.9003	2.8521	2.8298	2.8203	2.8013	2.7925	2.7816	2.7782	2.7696
5.0	2.9050	2.8929	2.8462	2.8136	2.7998	2.7975	2.7972	2.7954	2.7714	2.7695
5.5	3.0680	2.9977	2.9391	2.8997	2.8865	2.8685	2.8663	2.8606	2.8503	2.8444
6.0	2.9765	2.8200	2.7896	2.7537	2.7303	2.7028	2.6818	2.6767	2.6749	2.6739

the minimized mean square error. The figure clearly shows how the mean square estimation error of the proposed method changes based on the total privacy budget. A larger privacy budget weakens privacy protection, resulting in a smaller mean square estimation error and increased data availability. Next, we'll compare it with minimizing the mean square error by minimizing the maximum error.

B. OPTIMIZATION METHOD BASED ON PERTURBATION THEORY, OP

The experimental data used in this section are the same as those used in the OC method. It has been proved previously that the weight parameters obtained by the least square method are consistent with those obtained by minimizing the mean square error. Although these parameters can minimize the overall mean square error, they cannot minimize the maximum error. Therefore, the perturbation method is used to adjust the actual estimated value of each iteration to ensure that the overall mean square error does not increase too much and that the maximum error continues to decrease. The experimental results are shown in Table 3. There are 5 attributes in the experimental data. For the convenience of the display, the maximum error of the 5 attributes obtained by the least square method is added and its average value is taken. (a) and (b) in Table 3 represent the maximum error iterative decline results when the number of users is 10000 and 1000 respectively. The experimental process iterates a total of 20, but for convenience, only the first 10 results are shown.



(a) n = 10000, d = 75

(b) n = 1000, d = 75

FIGURE 4. The relationship between the estimated histogram error measured by $\log_{10}$(NSE) and the privacy budget ϵ. The black line in the figure represents the total MSE (mean square error) of BRR method. The MSE of BRR obtained by directly adding the frequency estimation results of all different privacy levels and comparing them with the real estimation results, and the red line represents the total MSE obtained by using the weighted combination method of OC. The green line represents the MSE of OP.

It is evident from the table that the maximum estimation error decreases as ϵ increases, and the maximum iteration error continues to decrease. The rate of error reduction is influenced by the selected perturbation δ, and in the experiment, we utilize the upper limit value that complies with perturbation theory. The table clearly demonstrates a rapid convergence rate, which is in line with our earlier observations. The primary objective of this section is to identify the optimal balance between the maximum error and the minimum mean square error, while proving the viability of our method. This approach can have a significant impact on future, more intricate machine learning tasks such as regression-based prediction.

Compared with the minimum mean square error, the mean square error of frequency estimation obtained by perturbation method will rise slightly, but it is still better than BRR method shown in the figure on the whole. The specific comparison results are shown in Figure 4. The OP method calculates its weight parameters after each iteration, and calculates its overall mean square error through the weight parameters. In a word, the weight combination optimization algorithm OC and the optimization algorithm OP based on perturbation theory proposed in this paper have obvious improvement compared with BRR.

VII. CONCLUSION

This paper introduces a multi-level personalized local differential privacy mechanism to address the varying privacy needs of individuals and protect their sensitive data. When estimating data frequency, data analysts must consider different user privacy levels. Current combination methods often lead to excessive overall estimation errors or maximum errors, which fail to adequately serve users. To address this, the paper proposes two weight combination optimization algorithms: one focuses on minimizing mean square error (OC), and the other on minimizing maximum error (OP). These algorithms calculate weight parameters for different privacy levels, allowing for overall frequency estimation through weighted combination. Experimental findings show that the OC algorithm can reduce overall mean square error by approximately 60% compared to BRR. While the OP algorithm results in a slightly higher overall mean square error than the OC method, it effectively minimizes the maximum estimation error for each attribute, mitigating overfitting or underfitting. Additionally, future plans include introducing a reward and punishment mechanism, such as privacy cost pricing compensation, to further clarify the privacy classification standard and analyze the balance between user privacy needs and data availability.

REFERENCES

[1] H. To, G. Ghinita, and C. Shahabi, "A framework for protecting worker location privacy in spatial crowdsourcing," *Proc. VLDB Endowment*, vol. 7, no. 10, pp. 919–930, Jun. 2014.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," *J. Privacy Confidentiality*, vol. 7, no. 3, pp. 17–51, May 2017.

[3] M. R. Hasan, R. Guest, and F. Deravi, "Presentation-level privacy protection techniques for automated face recognition—A survey," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–27, Dec. 2023.

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

[5] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren, "Heavy hitter estimation over set-valued data with local differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 192–203.

[6] R. Chen, H. Li, A. K. Qin, S. P. Kasiviswanathan, and H. Jin, "Private spatial data aggregation in the local setting," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 289–300.

[7] Y. Li, M. Chen, Q. Li, and W. Zhang, "Enabling multilevel trust in privacy preserving data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 9, pp. 1598–1612, Sep. 2012.

[8] Y. Nie, W. Yang, L. Huang, X. Xie, Z. Zhao, and S. Wang, "A utility-optimized framework for personalized private histogram estimation," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 655–669, Apr. 2019.

[9] T. T. Cai, Y. Wang, and L. Zhang, "The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy," *Ann. Statist.*, vol. 49, no. 5, pp. 2825–2850, Oct. 2021.

[10] T. Wang, X. Yang, X. Ren, W. Yu, and S. Yang, "Locally private high-dimensional crowdsourced data release based on copula functions," *IEEE Trans. Services Comput.*, vol. 15, no. 2, pp. 778–792, Mar. 2022.

[11] X. Ren, C. M. Yu, W. Yu, S. Yang, X. Yang, J. A. Mccann, and P. S. Yu, "LoPub: High-dimensional crowdsourced data publication with local differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2151–2166, Mar. 2018.

[12] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, Jul. 1970.

[13] X. Ren, L. Shi, W. Yu, S. Yang, C. Zhao, and Z. Xu, "LDP-IDS: Local differential privacy for infinite data streams," in *Proc. Int. Conf. Manage. Data*, Jun. 2022, pp. 1064–1077.

[14] X. Xiao and Y. Tao, "M-invariance: Towards privacy preserving re-publication of dynamic datasets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2007, pp. 689–700.

[15] X. Xiao, Y. Tao, and M. Chen, "Optimal random perturbation at multiple privacy levels," *Proc. VLDB Endowment*, vol. 2, no. 1, pp. 814–825, Aug. 2009.

[16] Z. Jorgensen, T. Yu, and G. Cormode, "Conservative or liberal? Personalized differential privacy," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 1023–1034.

[17] M. Min, H. Zhu, J. Ding, S. Li, L. Xiao, M. Pan, and Z. Han, "Personalized 3D location privacy protection with differential and distortion geo-perturbation," *IEEE Trans. Dependable Secure Comput.*, vol. 21, no. 4, pp. 3629–3643, Jul. 2024.

[18] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Proc. 25th IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Oct. 2005, pp. 620–629.

[19] M. Yuan, L. Chen, and P. S. Yu, "Personalized privacy protection in social networks," *Proc. VLDB Endowment*, vol. 4, no. 2, pp. 141–150, Nov. 2010.

[20] S. Wang, L. Huang, M. Tian, W. Yang, H. Xu, and H. Guo, "Personalized privacy-preserving data aggregation for histogram estimation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[21] Z. Shen, Z. Xia, and P. Yu, "PLDP: Personalized local differential privacy for multidimensional data aggregation," *Secur. Commun. Netw.*, vol. 2021, pp. 1–13, Jan. 2021.

[22] N. O. Attoh-Okine, "Differential privacy," in *Big Data and Differential Privacy: Analysis Strategies for Railway Track Engineering*. Berlin, Germany: Springer, 2017, pp. 241–247.

[23] X. Feng, C. Zhang, J. Li, and L. Dai, "Combinational randomized response mechanism for unbalanced multivariate nominal attributes," *IEEE Access*, vol. 8, pp. 143160–143172, 2020.

[24] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2014, pp. 1054–1067.

[25] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, Jun. 2016, pp. 2436–2444.

[26] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Philadelphia, PA, USA: SIAM, 1995.

[27] Z. V. Kovarik, "Compatibility of approximate solutions of inaccurate linear equations," *Linear Algebra Appl.*, vol. 15, no. 3, pp. 217–225, 1976.

[28] J. L. Rigal and J. Gaches, "On the compatibility of a given solution with the data of a linear system," *J. ACM*, vol. 14, no. 3, pp. 543–548, Jul. 1967.

[29] D. J. Higham, "Condition numbers and their condition numbers," *Linear Algebra Appl.*, vol. 214, pp. 193–213, Jan. 1995.

[30] J. Rohn, "New condition numbers for matrices and linear systems," *Computing*, vol. 41, nos. 1–2, pp. 167–169, Mar. 1989.

[31] G. W. Stewart, ''Perturbation bounds for the *QR* factorization of a matrix,'' *SIAM J. Numer. Anal.*, vol. 14, no. 3, pp. 509–518, Jun. 1977.

[32] G. H. Golub and J. H. Wilkinson, ''Note on the iterative refinement of least squares solution,'' *Numerische Math.*, vol. 9, no. 2, pp. 139–148, Dec. 1966.

[33] P.-Å. Wedin, ''Perturbation theory for pseudo-inverses,'' *BIT Numer. Math.*, vol. 13, no. 2, pp. 217–232, Jun. 1973.

[34] T. Ganai, ''Perturbation theory of structured matrix pencils with no spillover,'' *Appl. Math. Comput.*, vol. 458, Dec. 2023, Art. no. 128217.

[35] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. London, U.K.: Oxford Univ. Press, 1988.

XUEJIE FENG received the M.S. degree in business management from the University of Portsmouth, Portsmouth, U.K., in 2013, and the Ph.D. degree from the Department of Mathematics, Harbin Institute of Technology, Harbin, China, in 2022.

Since 2021, she has been an Associate Professor with the School of International Business, Qingdao Huanghai University, Qingdao, Shandong. Her research interests include perturbation method in computational economics and information perturbed method with local differential privacy and big data and business intelligence.

Dr. Feng is a member of Shandong Province Higher Education Youth Talent Induction Program Construction Team Project: Big Data and Business Intelligence Social Service Innovation Team, China, in 2019.

CHIPING ZHANG received the M.S. degree in thermal turbine and the Ph.D. degree in aircraft design from Harbin Institute of Technology, Harbin, China, in 1988 and 2006, respectively.

Since 2006, he was a Professor with the Department of Mathematics, Harbin Institute of Technology. He is currently the Assistant Dean and the Office Director of the Department of Mathematics. He has presided over and undertaken key projects of the National Natural Science Foundation of China, 863 key projects, 973 subprojects, a number of provincial and ministerial key projects, and a number of international cooperation projects. His research interests include the approximation theory, data fusion, and neural networks.

Prof. Zhang's awards and honors include the Mathematics Reform and Practical Teaching Achievement Award of Engineering University, in 1995, and the Teaching Achievement Award of Mathematical Modeling Teaching Research and Quality Education Practice Teaching Achievement Award, in 2003.