IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 8, 2020, accepted July 27, 2020, date of publication July 31, 2020, date of current version August 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3013446

Combinational Randomized Response Mechanism for Unbalanced Multivariate Nominal Attributes

XUEJIE FENG[1,2], CHIPING ZHANG[1], JING LI[1,3], AND LINLIN DAI[1,4]

[1]Department of Mathematics, Harbin Institute of Technology, Harbin 150001, China
[2]School of International Business, Qingdao Huanghai University, Qingdao 266427, China
[3]School of Data Science, Qingdao Huanghai University, Qingdao 266427, China
[4]School of Finance and Economics, Qingdao Huanghai University, Qingdao 266427, China

Corresponding author: Chiping Zhang (zcp@hit.edu.cn)

This work was supported by the Construction team project of the introduction and cultivation of young innovative talents in Colleges and universities of Shandong Province of China, 2019 (Project name: Big data and business intelligence social service innovation team).

ABSTRACT At present, many enterprises provide users with better services by collecting their sensitive information. However, these enterprises will inevitably cause the leakage of users' information, thereby infringing on users' privacy. Local differential privacy resolves this problem by only aggregating randomized values from each user, providing plausible deniability. However, different users might have diverse privacy requirements for different attributes. Moreover, the dimensions of these attributes may be unbalanced. Traditional local differential privacy algorithms usually assign the same privacy budget to all attributes, resulting in undesired frequency estimation. To obtain highly accurate of the results while satisfying local differential privacy, the aggregator needs to implement a reasonable privacy budget allocation scheme. Motivated by this, this paper proposed a novel local differential privacy scheme. The proposed method combines the advantages of BRR and MRR to address the problem of high and low privacy requirements. It employs the Lagrange multiplier algorithm to transform the privacy budget allocation problem between unbalanced attributes into a problem of calculating minima from unconditionally constrained convex functions. The solution to the resulting nonlinear equation is used as the final privacy budget allocation scheme. Simulation experiments show that the novel local differential privacy scheme proposed by this paper can significantly reduce the estimation error under the premise of satisfying the local differential privacy.

INDEX TERMS Data privacy, estimation error, nonlinear equations, optimization, probability.

I. INTRODUCTION

With the rapid development of artificial intelligence technology, information from crowd-sourcing system has brought great convenience to people's production and life. Mobile payments, map navigation, hospital consultation and other convenient services all come from the analysis of people's data. Particularly, with high-dimensional heterogeneous data (data with unbalanced multivariate nominal attributes), there are many hidden rules and much hidden information behind the data that can be mined to provide better services for individuals or groups. While ensuring the rapid development of information technology, the protection of the privacy of personal data has become a top priority for governments and enterprises. In April 2016, the EU passed the General Data Protection Regulation, GDPR [1], which stipulates that the protection of personal data crosses national boundaries, and at the same time clarifies the right of users to knowwhat personal information is being collected and to be forgotten.

The emphasis on privacy issues has promoted research on privacy protection technologies, for which the degree of privacy protection and the utility of data are the most important metrics. In line with this developmental trend, differential privacy technology [2], [3] has been proposed. As a privacy protection model, it strictly defines the strength of privacy protection, that is, the addition or deletion of any record will not affect the final query result. Compared with k-anonymity [4], l-diversity [5], t-compactness [6] and other methods that require special attack assumptions and background knowledge, differential privacy has become a research hot spot in the current academic community due to its unique advantages. However, there are still two major challenges

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

in the application of differential privacy to high-dimensional unbalanced multivariate nominal attribute data.

(1) **Nonlocal privacy protection.** Differential privacy [7], as one of the currently effective privacy protection mechanisms, randomizes query output by adding noise to sensitive data to achieve privacy protection. Traditional differential privacy technologies [8]–[10] aggregate raw data into a data centre, and then release relevant statistical information that meets the requirements of differential privacy. We call such methods centralized differential privacy technology. The protection of sensitive information by centralized differential privacy is based on the idea of trusted third-party data collectors. However, even if third-party data collectors claim that they will not steal and disclose confidential user information, privacy is still not guaranteed. In practise, it is difficult to find a truly reliable third-party data collection platform, which greatly limits the use of centralized differential privacy technology. Users prefer that data security is ensured on their side, enabling them to process and protect their own confidential information separately (*i.e.*, local differential privacy [11]–[14]). (2) **High-dimensional disaster.** In crowd-sourced systems, high-dimensional heterogeneous data are ubiquitous. With the increases in data dimensionality and the dimensional differences between different attributes, many existing local differential privacy mechanisms such as RAPPOR [15] and [16], [17], if straightforwardly applied to multiple attributes with unbalanced dimensions, will become extremely unavailable. Their fatal drawbacks are the use of non-optimized privacy budget allocation schemes and their high computational complexities, which lead to great data utility loss and high latency. Attributes with different dimensions require the allocation of different privacy budgets. Determining the best allocation scheme is the key to improving data utility. Furthermore, considering that the level of privacy concern required by users for different data is inconsistent, it is also important to find the optimal privacy mechanism under high and low privacy regimes.

In response to the above challenges, many existing methods have proven their effectiveness from different perspectives. One group of methods ensures that users' privacy is not leaked by providing users with a local privacy guarantee. As a result, local differential privacy [11]–[14] technologies have emerged as effective methods based on the inheritance of centralized differential privacy technology to quantify the definition of privacy attacks. It delegates the right to randomize data to the users. Currently, local differential technologies are used by many companies to provide users with more convenient and high-quality services. For example, Apple has applied the technology to their iOS 10 operating system to protect users' device data, and the US Census Bureau uses differential privacy for demographics [18] data. However these methods are extremely complicated in terms of communication, and data availability can drop sharply when processing high-dimensional unbalanced multivariate nominal attribute data. Another group of methods privately release high-dimensional data [19]–[21]. These methods mainly use specific algorithm to reduce the dimensionality of the data and then release it privately. These methods not only have high computational complexity but also have low data utility due to unreasonable privacy budget allocation schemes.

In addressing the above issues, this paper proposes a novel local differential privacy scheme: combinational randomized response, or CRR. CRR combines the advantages of BRR [12], [15] and MRR [22], [23] in different privacy regimes and dimensions. We first divide the data into two parts according to the size of the dimensions. Then, we apply BRR to the part of the data with higher dimensions. Similarly, MRR is applied to the part of the data with lower dimensions. The reasons are detailed at the end of Section III. To determine whether the attributes use BRR or MRR, we design an h-index method to divide the attributes. When using BRR and MRR to perturb the data, we adopt the optimal privacy budget allocation scheme proposed in this paper to find the final allocation scheme. In calculating the optimal privacy budget allocation scheme, we use the square error (SE) as the metric to evaluate the estimation, which refers to the difference between the unbiased estimation and the real histogram. The optimal privacy budget allocation scheme is as follows. First, we use the Lagrange multiplier (LM) algorithm to transform the privacy budget allocation problem into the problem of calculating minima from unconditionally constrained convex functions. Next, we use the first derivative to transform the minimum value problem into the problem of finding the roots of the univariate cubic equation. Then, we employ the Cardano Formula (CF) method to derive the roots of the univariate cubic equation. The resulting roots comprise the privacy budget allocation scheme. In the end, we apply the optimal privacy budget allocation scheme to the CRR to perturb the data. CRR is very flexible; it can be adjusted to the corresponding Optimal Binary Randomized Response (OBRR) and Optimal Multivariate Randomized Response (OMRR) according to the different requirements of the privacy levels of the data, which will be introduced in detail later. To verify the effectiveness of our method, we compare CRR with BRR [12], [15], OBRR, MRR [22], [23], and OMRR. The results show that our method greatly improves the utility of the data while ensuring local differential privacy.

II. RELATED WORK

Multivariate frequency statistics can be applied when each user sends multiple variable values. After the user sends the data to the data collector, the data collector obtains the candidate value list according to the statistics. They count the frequency of each candidate value and publish it. Different from the single-valued frequency statistics problem, the multivariate situation needs to consider the division of privacy budget. An unreasonable privacy budget allocation scheme can lead to a significant reduction in the utility of the sanitized data.

A. LOCAL PRIVACY GUARANTEE

To address the shortcomings of differential privacy which cannot guarantee users' local privacy, the concept of local differential privacy was proposed to provide users with local privacy guarantees in a crowd-sourcing system [24]. In simple terms, single-valued frequency statistics can be reused for each variable in a multivariate situation, but this will bring new problems. Without loss of generality, we assume that the dataset $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_l\}$, and each attribute $\mathbf{a}_i$ has a specific 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, that is, $|\mathbf{a}_i| = k_i, i = 1, 2, \cdots, l$, where $|\mathbf{a}_i|$ refers to the dimension level for i-th attribute. $d = k_1 + \cdots + k_l$ denotes the total number of candidate values. Each user u_{i_u} possesses a fixed l number of variables $u_{i_u} = \{v_{i_u1}, v_{i_u2}, \cdots, v_{i_ul}\}$, $i_u = 1, \cdots, n$, where n is the number of users. In this situation, the privacy budget needs to be divided into l parts. When the number of variables l is large, the data utility decreases dramatically. For example, if we employ the S-Hist [25] algorithm to do the same for each variable, after splitting the privacy budget, each variable is assigned a privacy budget of $\frac{\epsilon}{l}$, which will directly lead to asymptotic error boundaries, and the results of the variance will increase l-times. Our method comprehensively considers the size of the dimensions and allocates a reasonable privacy budget for the attributes of different dimensions to minimize the asymptotic error. For the RAPPOR [15] method, the asymptotic error boundary increases from $O(\frac{k}{\epsilon\sqrt{n}})$ to $O(\frac{lk}{\epsilon\sqrt{n}})$. The communication cost will be asymptotic to $\prod_{i=1}^{l} k_i$, which requires exponential storage space in terms of d. However, our method only needs $d = \sum_{i=1}^{l} k_i$. Additionally, different differential privacy mechanisms are suitable for different dimension attributes. The BRR is suitable for high-dimensional attributes with a low budget, and the MRR is suitable for low-dimensional attributes with a high budget. If a univariate local differential privacy is used repeatedly, it cannot take advantages of its strengths with regard to dimensions for which it is not suitable. Therefore, directly repeating the single-valued frequency publishing method l times as the frequency publishing method in the multivariate case is not feasible in terms of data availability and transmission cost. In addition, there are many improved local differential privacy algorithms suitable for single-valued frequency statistics, such as O-RAPPOR [26], PCE [27], k-RR [28] and k-Subset [29]. All multivariate frequency statistics have a sharp drop in data utility due to the splitting of the privacy budget.

B. HIGH DIMENSION

For the high-dimensional case, an effective way to solve the problem of multivariate nominal attributes is to group related records into clusters and then allocate the privacy budget for each low-dimensional cluster. PriView [30] constructs k_p marginal distributions of low-dimensional attribute sets and then estimates the joint distribution of the high-dimensional sets. However, this method only works based on the assumption that all attributes are independent of each other and that attribute pairs are processed equally. Actually, this assumption is not in line with the fact that the attributes in crowd-sourcing systems are associated with each other. PDP-PCAO [31] improves the principal component analysis (PCA) algorithm by employing attribute importance and reduces the dimensionality of the data with the improved PCA, reducing time and space costs. This method considers the existence of multi-sensitive attributes in high-dimensional data, while the traditional methods of allocating privacy budgets cannot satisfy the requirements of local privacy protection. PPDP-PCAO introduces a sensitivity preference, combines with the optimal matching theory, and designs a sensitive attribute hierarchical protection strategy. There are also some other dimension-reducing differential privacy mechanisms [19], [32]–[34]. However, to determine the association between attributes and achieve cluster distribution, these methods need to access the original dataset twice. The two visits are calculated independently, which can lead to a consistent privacy guarantee. The privacy budgets allocated by these methods are completely unrelated. These methods do not specify how to allocate privacy budgets reasonably to achieve adequate privacy guarantees and maximize utility, however. Moreover, although unbalanced data with multivariate nominal attributes can be reduced into several low-dimensional clusters, the sparsity caused by combinations in each cluster will persist and may result in lower utility. Rather than the absolutely centralized settings, Su *et al*. [35] proposed a distributed multiparty setting to publish a new dataset from multiple data curators. However, their multiparty computing does not guarantee local personal privacy within the data server. Instead, they can only protect privacy between data servers.

To solve the problem of privacy leakage caused by centralized differential privacy, the RAPPOR-unknown [36] employs an expectation maximization (EM) method to estimate the joint probability distribution of multiple variables. Its purpose is to perform queries on contingency tables. RAPPOR-unknown is an improvement of the orginal RAPPOR method. Its asymptotic error boundary is $O(\frac{d}{\epsilon\sqrt{n}})$, but its communication cost is higher, $O(d) + O(r)$, where ϵ refers to the privacy budget, n is the number of users, d denotes the total number of candidate values and r is the number of the substring. Obviously, RAPPOR-unknown is not suitable for situations with many substrings. Otherwise, not only is the communication cost high but also the data utility is reduced. Li *et al*. [37] proposed a dichotomy of the privacy budget by publishing differential privacy histograms in groups, which obviously cannot maximize statistical accuracy. Wang *et al*. [38] added additional processing to the output to improve the accuracy of the published data. The purpose of this additional processing is to restore the consistency of the count specified in the structure. However, this method cannot solve the inherent error caused by high dimensionality. There are also some methods, such

as [39]–[41], which release a database through a matrix mechanism to minimize query noise. However, the optimization costs of those methods are very high, and the assumption that the query distribution is known in advance is not reasonable.

To solve the shortcomings of the above methods, which cannot meet pricacy locality requirements or handle high-dimensional data, some effective methods have been proposed. Ren *et al*. [20], [21] introduced a multivalued frequency statistics method that combines the RAPPOR method and the probability map model. They first transform each attribute value into a random bit string using a Bloom filter [42] and then send it to the cen- tral server. Subsequently, similar to the high-dimensional data publishing method based on centralized differential privacy in [19], the data collector performs frequency statistics on the collected data to construct a Markov network and uses the correlation between attributes to obtain the maximal clique. Next, the joint probability distribution of the attributes is expressed in the form of a maximal cluster to achieve data dimension reduction. Finally, the data set that is resynthesized through the joint probability distribution is released. Analogously, Ju *et al*. [43] proposed employing a Bayesian network to recognize the dimensional correlation of high-dimensional data. Then, they divided the high-dimensional data attribute set into multiple relatively independent low-dimensional attribute sets and sequentially synthesized the new dataset. However, the disadvantages of these methods are a high computational delay and an unreasonable privacy budget allocation. A comparison of several existing multivalued frequency statistics methods under local differential privacy is shown in Table 1, where r is the number of the substring in RAPPOR-unknown, s is the encoding length of string in S-Hist, d, l, n, ϵ is defined as previously.

To overcome the problems between those schemes, we propose a novel privacy budget allocation scheme to publish unbalanced multivariate nominal attribute data while guaranteeing local privacy. As far as we know, no literature has been published on this issue. In this paper, we combine the advantages of BRR and MRR in different privacy regimes. Then, we turn the privacy budget allocation problem into one that involves solving the univariate cubic equation. The experimental results show that our method can greatly improve the low query accuracy caused by the defects in privacy budget allocation.

III. PRELIMINARIES

A. LOCAL DIFFERENTIAL PRIVACY

Local differential privacy [8] is a rigorous privacy notion in the local setting, which provides a stronger privacy guarantee than centralized differential privacy. The formal definition of local differential privacy is as follows:

Definition 1: Given n users, with each user corresponding 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 $t^* \subseteq Range(\mathcal{F})$,

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

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

For local differential privacy technology, the privacy process is transferred from the data collector to a single client, so that no trusted third party intervention is required. It also eliminates privacy attacks that may be caused by untrusted third-party data collectors.

B. MULTIVARIATE RANDOMIZED RESPONSE

We consider the multivariate input domain $\chi = \{X_1, X_2, \cdots, X_m\}$, which is the category data with m categories. Similar to binary category data, a real category x is released with a finite probability q, and a false category is uniformly and randomly selected from $\chi - \{x\}$ with probability $1.0 - q$. The definition of MRR [22] is as follows:

Definition 2: For category data $x = X_i \in \chi$, where $\chi = \{X_1, X_2, \cdots, X_m\}$, suppose that the output of the multivariate random response mechanism is z; then, z is equal to X_i with probability $q(0.0 \leq q \leq 1.0)$, and to $X_j(X_j \in \chi, X_j \neq X_i)$ with probability $(1 - q)/(m - 1)$.

The following theorem 1 guarantees that the multivariate randomized response mechanism meets the local differential privacy protection. The theorem also provides the randomized response parameters required to achieve the corresponding level of privacy protection.

Theorem 1: The local differential privacy protection level satisfied by the multivariate randomized response mechanism is

$$\epsilon = \log\left(\max\left\{\frac{q(m-1)}{1-q}, \frac{1-q}{q(m-1)}\right\}\right) \tag{2}$$

For the statistical analysis task of local differential privacy, the service provider usually needs to determine the frequency distribution of each category in the user group. Since the multivariate randomized response mechanism disturbs the real category data x, the frequency distribution of z observed by the service provider is different from the frequency distribution of x. Assuming the number of users is n and the true frequency of category X_i is F_i, the following relationship exists for the frequency F_i' of X_i observed from z:

$$E[F_i'] = F_i q + (n - F_i)\frac{1-q}{m-1} \tag{3}$$

Therefore, an unbiased estimate of F_i can be obtained from F_i':

$$F_i = E\left[\frac{F_i' - n(1-q)/(m-1)}{q - (1-q)/(m-1)}\right] \tag{4}$$

In the above multivariate randomized response mechanism for multivariate category data, the input domain is the same as the output domain. When the privacy protection level $\epsilon = \log\left(\max\left\{\frac{q(m-1)}{1-q}, \frac{1-q}{q(m-1)}\right\}\right)$ is fixed, as the attribute candidate value dimension m becomes larger, q will decrease. The probability of outputting the real category is reduced, thus increasing the final single category frequency estimation or distribution estimation error.

TABLE 1. Multi-valued frequency statistics method under local differential privacy.

Method	Advantage	Disadvantage	Communication cost	Error boundary	Computational overhead	Candidate value is known	Privacy budget allocation
RAPPOR-unknown	no need to know the candidate values	the parameters of the Bloom filter	$O(d)+O(r)$	$O(\frac{d}{\epsilon\sqrt{n}})$	high	no	divided by the number of substrings
Harmony-frequency	high data utility	reduced accuracy due to sampling	$O(s)$	$O(\frac{\sqrt{d\log\frac{d}{l}}}{\epsilon\sqrt{n}})$	middle	yes	all ϵ assigned to sampling variables
LDPMiner	high data utility	only applicable to heavy hitter	$O(d)+O(s)$	$O(\frac{d}{\epsilon\sqrt{n}})$	high	yes	ϵ is divided into two phases
LoPub	weakening the impact of high latitudes on accuracy	high communication cost and computational overhead	$O(d)$	$O(\frac{d}{\epsilon\sqrt{n}})$	high	yes	split ϵ by attribute

C. BINARY RANDOMIZED RESPONSE

In the aforementioned multivariate randomized response mechanism for multivariate category data, the output domain is the same as the input domain. Given a privacy protection level ϵ, if the number of categories is large, the value of q will decrease as the number of categories l increases, so that the final frequency estimation or distribution of a single category will have a larger error.

To solve the effect of the number of categories on the error of frequency estimation of a single category, Duchi *et al*. [12] proposed a binary randomized response mechanism, BRR. This mechanism first represents the category data in the form of a d-length bitmap, and then a binary random response is performed independently for each bit in the bitmap. The formal definition of the BRR mechanism is as follows:

Definition 3: For category data $x = X_i \in \chi$, where $\chi = \{X_1, X_2, \cdots, X_m\}$, the bitmap of x is represented as $bx \in \{0, 1\}^m$. The output of the binary randomized response is $z \in \{0, 1\}^m$. For any given $j \in [1, m]$, the jth bit z^j of z is equal to bx^j with probability $q(0.0 \le q \le 1.0)$, and to $1 - bx^j$ with probability $(1 - q)$.

The following theorem 2 guarantees that the binary randomized response mechanism meets the local differential privacy protection requirements. The theorem also gives the randomized response parameters required to achieve the corresponding level of privacy protection.

Theorem 2: The local differential privacy protection level satisfied by the binary randomized response mechanism is:

$$\epsilon = 2\log\left(\max\left\{\frac{q}{1-q}, \frac{1-q}{q}\right\}\right) \tag{5}$$

Now, for the statistical analysis of local differential privacy based on a binary randomized response mechanism, the service provider usually needs to determine the frequency distribution of each category in the user group. Since the binary randomized response mechanism disturbs the real category data x, the frequency distribution of z observed by the service provider is different from the frequency distribution of x. Assuming the number of users is n and the true frequency of category X_i is F_i, the following relationship exists for the frequency F_i' of X_i observed from z:

$$E[F_i'] = F_i q + (n - F_i)(1 - q) \tag{6}$$

Therefore, an unbiased estimate of F_i can be obtained from F_i':

$$F_i = E\left[\frac{F_i' - n(1-q)}{2q-1}\right] \tag{7}$$

The BRR mechanism incurs $O(m)$ communication costs for each user, and the MRR incurs $O(1)$ communication costs in single attribute scenario. As far as the communication cost is concerned, MRR is superior to BRR. In work proposed by Kairouz *et al*. [22], BRR and MRR are called staircase mechanisms. BRR has been proven to be optimal in the high-privacy regime, and MRR has been proven to be optimal in the low-privacy regime [26]. In fact, BRR is suitable for high-dimensionality cases with low privacy budgets, while MRR is suitable for low-dimensionality cases with high privacy budgets. We assume that the privacy budget ϵ is fixed, and the data have only one attribute X. The dimensionality of the candidate value of attribute X is m. The number of users is n. Then, the mean square error of the frequency estimation of BRR and MRR is calculated as follows:

$$SE(BRR) = \frac{nm\exp(\frac{\epsilon}{2})}{(\exp(\frac{\epsilon}{2})-1)^2}$$
$$SE(MRR) = \frac{n(m-1)(2\exp(\epsilon)+m-2)}{(\exp(\epsilon)-1)^2}$$

We let $\exp(\frac{\epsilon}{2}) = x$, then

$$\begin{aligned} f &= SE(BRR) - SE(MRR) \\ &= n\frac{mx^3 + mx - m^2 + 3m + 2x^2 - 2}{(x-1)^2(x+1)^2} \end{aligned}$$

By calculating the first partial derivative, we can see that when m is fixed, a larger x means a larger SE(BRR). In contrast, when x is fixed, a larger m means a smaller SE(BRR). This also verifies the conclusion that BRR is suitable for high-dimensionality cases with low budgets, and MRR is suitable for low-dimensionality cases with high budgets. Since the dimensions are fixed, we apply BRR to the part of the data with higher dimensionality and apply MRR to the part of the data with lower dimensionality. The optimal privacy

budget allocation scheme is designed to take advantage of the strengths of BRR and MRR. Considering the advantages of BRR and MRR, in the next section, we will present a combinational randomized response mechanism, CRR, for multiple unbalanced categorical data.

IV. COMBINATIONAL RANDOMIZED RESPONSE

A. PROBLEM DESCRIPTION

This paper focuses on frequency estimation over unbalanced multivariate nominal attributes. The unbalanced multivariate nominal attributes indicate that the data have a set of attributes $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_l\}$, and each attribute $\mathbf{a}_i$ has a specific number of categories $\mathbf{a}_i = \{a_{i1}, a_{i2}, \cdots, a_{ik_i}\}$, where k_i is the number of candidate values for the i-th attribute, that is, $|\mathbf{a}_i| = k_i, i = 1, 2, \cdots, l$. Specifically, each user u_{i_u} possesses a set $\mathbf{v}_{i_u} = \{v_{i_u1}, v_{i_u2}, \cdots, v_{i_ul}\}$ of items, where $v_{i_ui} \in \mathbf{a}_i, i = 1, 2, \cdots, l$. The aggregator queries participants about the domain $\mathbf{V}$, in which each participant u_{i_u} holds a secret value $\mathbf{v}_{i_u}$. To ensure that the participant's privacy response will not be disclosed, each participant randomizes their secret response $\mathbf{v}_{i_u}$ independently using a privacy preserving randomizer $\mathcal{F}$ to obtain a sanitized version of the response $\mathbf{v}'_{i_u} \in Range(\mathcal{F})$ and then publishes $\mathbf{v}'_{i_u}$ to the aggregator. After receiving the sanitized data list $\{\mathbf{v}'_1, \mathbf{v}'_2, \cdots, \mathbf{v}'_n\}$, the aggregator attempts to decode an estimation over the domain $\mathbf{V}$. According to the estimated results from the sanitized data set, the aggregator tries to provide users with better network services. In the process of data release with local differential privacy, no one knows the secret information they release except for the participants themselves.

Let n be the total number of users, $d = k_1 + k_2 + \cdots + k_l$ be the total number of candidate values, and f_j be the frequency of the j-th item(denoted as v_j), the portion of users possessing item v_j, $1 \le j \le d$. Formally, we have

$$f_j = \frac{|\{u_i | v_j \in \mathbf{v}_i, 1 \le i \le n|}{n}$$

Unbalanced multivariate nominal data have different items for each attribute, which indicates that if the user applies the same privacy budget to all attributes, the estimation of the frequency may be unsatisfactory. To obtain greater accuracy of the results while satisfying local differential privacy, the aggregator needs to implement a reasonable privacy budget allocation scheme. Some of the notations employed in this paper are listed in Table 2.

TABLE 2. Notation.

A	multiple unbalanced categorical data sets
i	index of attributes
j	index of candidate values of the i-th attribute
i_u	index of users
s	encoding length of string in S-Hist method
k_{avg}	average dimensions of each attribute
t^*	an output within a range of values used in Definition 1
m	number of candidate values used in Definition 2 and 3
m_d	digit precision
k_p	dimensions of marginal distributions used in PriView method
t	average number of iterations used in our method
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 $\|\mathbf{a}_j\|$ is k_j
$\mathbf{v}_i$	private values possessed by the i-th user, its length $\|\mathbf{v}_i\|$ is l
v_{i_uj}	j-th value of $\mathbf{v}_{i_u}$
$\mathbf{h}_{i_u}$	private bit vector of i_u-th users, its length is d
H	true histogram, $\mathbf{H} = sum\{\mathbf{h}_1, \cdots, \mathbf{h}_n\}$
H'	sanitized histogram of $\mathbf{H}$, $\mathbf{H}' = sum\{\mathbf{h}'_1, \cdots, \mathbf{h}'_n\}$
H''	estimated histogram of $\mathbf{H}'$, $\mathbf{H}'' = sum\{\mathbf{h}''_1, \cdots, \mathbf{h}''_n\}$
ϵ_i	privacy budget of i-th attribute
ϵ	privacy budget
h	divided index
CF	Cardano Formula
NRM	Newton-Raphson Method
SE	Square Error
NSE	Normalized Square Error, $NSE = \frac{SE}{n}$
BRR	Binary Randomized Response
$OBRR$	Optimal Binary Randomized Response
MRR	Multivariate Randomized Response
$OMRR$	Optimal Multivariate Randomized Response
CRR	Combinational Randomized Response

B. COMBINATIONAL RANDOMIZED RESPONSE MECHANISM

The essence of our method is to solve the problem that the estimation error increases due to the imbalance of differential privacy budget allocation under the condition of high-dimensional heterogeneous data. In this paper, we propose a combinational randomized response mechanism CRR. CRR gives full play to the advantages of BRR and MRR in different dimensions to solve the problem of the excessive error caused by an uneven allocation of privacy budget. In this paper, we start with the mean square estimation error of CRR, so that the optimal solution satisfying the minimum mean square error of CRR is taken as our final privacy budget allocation scheme.

The basic idea of optimizing multiple unbalanced categorical histogram aggregation errors is explicit: more privacy budgets should be allocated to a large number of items than to a small number of attributes. Due to the different number of candidate attribute values, the distribution of the privacy budget is different, and we can reasonably combine BRR and MRR to take advantage of their strengths in different privacy regimes.

To separate the attributes into two groups, we first sort them in ascending order according to the number of candidate values of each attribute, that is, $k_1 \le k_2 \le \cdots \le k_l$. According to the number of candidate values in the attribute category, we choose a parameter h as the dividing index. We then take the first h attributes as the low-privacy regime domain $\mathbf{S}_l$, and the remaining attributes as the high-privacy regime domain $\mathbf{S}_h$. We then apply BRR to the high privacy regime $\mathbf{S}_h$ and apply MRR to a low privacy regime $\mathbf{S}_l$. In the next section, we will introduce the privacy budget allocation method; here, we assume that the optimal privacy budget sequence $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_l\}$ has been obtained. The mechanism of CRR is shown in Algorithm 1. In summary, the randomizer naturally embeds a multivariate randomized response into a binary randomized response on a bitmap. This randomizer is carried out by each participant, and its privacy guarantee is

Algorithm 1 Combinational Randomized Response

Input: ϵ-privacy budget; $\{k_1, k_2, \cdots, k_l\}$-number of items for each attribute; $\mathbf{v} \in \{0,1\}^{k_1+\cdots+k_l}$-a secret value that is represented as a bit map. $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_l\}$-optimal budget allocation; h-divided index.

Output: $\mathbf{v}' \in \{0,1\}^{k_1+\cdots+k_l}$-a sanitized bit map that satisfies local ϵ-differential privacy.

initialize $d = k_1 + k_2 + \cdots + k_l$; $\mathbf{v}' = \mathbf{0} \in 0^d$;$m = 0$
for i = 1 to h **do**
  if $i \neq 1$ **then**
    $m = m + k_{i-1}$
  end if
  for $j = 1$ to k_i **do**
    if $v_j == 1$ **then**
      $t = j$
    end if
  end for
  $p = random[0, 1]$
  if $p < \frac{\exp(\epsilon_i)}{\exp(\epsilon_i)+k_i-1}$ **then**
    $t' = t$
  else
    $t' = random([1, k_i] \backslash \{t\})$
  end if
  $v'_{t'+m} = 1$
end for
$m = k_1 + k_2 + \cdots + k_h$
for i = h+1 to l **do**
  for $j = 1$ to k_i **do**
    $p = random[0, 1]$
    if $p < \frac{\exp(\epsilon_i)}{\exp(\epsilon_i)+1}$ **then**
      $v'_{j+m} = v_{j+m}$
    else
      $v'_{j+m} = 1 - v_{j+m}$
    end if
  end for
  $m = m + k_{i-1}$
end for

declared in Theorem 3, the proof of which is clear and will not be shown here.

Theorem 3: The randomizer shown in Algorithm 1 satisfies the local ϵ-differential privacy constraints in Definition 1, where $\epsilon = \epsilon_1 + \cdots + \epsilon_h + 2\epsilon_{h+1} + \cdots + 2\epsilon_l$.

Now consider the statistical analysis of local differential privacy based on the CRR mechanism. Each user u_{i_u} publishes a length-l bit vector $\mathbf{v}'_{i_u} = \{v'_{i_u 1}, v'_{i_u 2}, \cdots, v'_{i_u l}\}$. The data collectors aggregate the collected $\mathbf{v}'_{i_u}$ into $\mathbf{h}'_{i_u}$, which is obtained by perturbing the original bit vector $\mathbf{h}_{i_u}$. The true histogram $\mathbf{H} = sum\{\mathbf{h}_1, \cdots, \mathbf{h}_n\}$. The sanitized histogram $\mathbf{H}' = sum\{\mathbf{h}'_1, \cdots, \mathbf{h}'_n\}$. Let $\mathbf{H}'' = \{H''_{11}, \cdots, H''_{1k_1}, H''_{21}, \cdots, H''_{2k_2}, \cdots, H''_{lk_l}\}$ denote the unbiased estimation of $\mathbf{H}$; for each attribute, we have $H''_{ij}p_i + (n - H''_{ij})(1 - p_i) = H'_{ij}, i = 1, \cdots, h, j = 1, \cdots, k_i$, where $p_i = \frac{\exp(\epsilon_i)}{\exp(\epsilon_i)+1}$; $H''_{ij}p_i + (n - H''_{ij})(1 - p_i) = H'_{ij}, j = 1, \cdots, k_i, i = h+1, \cdots, l$, where $p_i = \frac{\exp(\epsilon_i)}{\exp(\epsilon_i)+k_i-1}$. Therefore, we have:

$$\begin{cases} H''_{ij} = \dfrac{H'_{ij}(\exp(\epsilon_i) + 1) - n}{\exp(\epsilon_i) - 1}, & i = h+1, \cdots, h \\ H''_{ij} = \dfrac{H'_{ij}(\exp(\epsilon_i) + k_i - 1) - n}{\exp(\epsilon_i) - 1}, & i = h+1, \cdots, l \\ & j = 1, \cdots, k_i \end{cases} \tag{8}$$

All that remains is to develop a method to optimize the parameters h and $\{\epsilon_1, \cdots, \epsilon_l\}$, which we will introduce in the next section.

Algorithm 2 Divided Index Selection

Input: $\{k_1, k_2, \cdots, k_l\}$-number of items for each attribute;
Output: h-divided index

initialize $\{ad_1, ad_2, \cdots, ad_l\} = \mathbf{0}$
for $i = 1$ to l **do**
  for $j = 1$ to l **do**
    $ad_i = ad_i + abs(k_i - k_j)$
  end for
end for
for $i = 1$ to $l - 1$ **do**
  $diff_i = ad_{i+1} - ad_i$
end for
$[h, value] = max(diff_1, \cdots, diff_{l-1})$

C. OPTIMAL PARAMETER SELECTION

1) h-DIVIDED INDEX

How should we choose a suitable h for (8)? We first define the attribute dispersion (AD) as given in Definition 4. By definition, we can derive a discrete AD value for each attribute $\mathbf{a}_i$, denoted as $ad_i, i = 1, 2 \cdots, l$. Then, we calculate the AD difference between two adjacent attributes, denoted as $diff_i = ad_{i+1} - ad_i, i = 1, \cdots, l - 1$. Next, we choose the maximum $diff_{i'}$. Finally, we let $h = i'$ be the dividing index. The procedure is detailed in Algorithm 2.

Definition 4: Let $\mathbf{A} = \mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_l$, and the number of categories for each attribute $\mathbf{a}_i$ is equal to k_i. The dispersion of attribute $\mathbf{a}_i$ is defined as $AD(\mathbf{a}_i) = \sum_{j=1, j\neq i}^{l} |k_i - k_j|$

2) OPTIMAL ϵ-PRIVACY BUDGET ALLOCATION

In this section, we will present a method for choosing the optimal budget allocation scheme. The parameters k, l, ϵ are defined in the same way as before, and the h-divided index is calculated by employing the method proposed in Algorithm 2. According to (8), the square error, SE, from decoding CRR (Algorithm 1) is given as follows:

$$\begin{aligned} &\mathbf{SE}(\epsilon, h, l, d) \\ &= E[\sum_{i=1}^{l}\sum_{j=1}^{k_i}(H''_{ij} - H_{ij})^2] \\ &= \sum_{i=1}^{l}\sum_{j=1}^{k_i} E[(H''_{ij} - H_{ij})^2] = \sum_{i=1}^{l}\sum_{j=1}^{k_i} Var[H''_{ij}] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{h}\sum_{j=1}^{k_i}\left(\frac{\exp(\epsilon_i)+1}{\exp(\epsilon_i)-1}\right)^2 Var[H'_{ij}] \\
&\quad + \sum_{i=h+1}^{l}\sum_{j=1}^{k_i}\left(\frac{\exp(\epsilon_i)+k_i-1}{\exp(\epsilon_i)-1}\right)^2 Var[H'_{ij}] \\
&= \sum_{i=1}^{h}\frac{nk_i\cdot\exp(\epsilon_i)}{(\exp(\epsilon_i)-1)^2} \\
&\quad + \sum_{i=h+1}^{l}\sum_{j=1}^{k_i}\frac{H_{ij}\exp(\epsilon_i)(k_i-1)+(n-H_{ij})\exp(\epsilon_i)+k_i-2}{(\exp(\epsilon_i)-1)^2} \\
&= \sum_{i=1}^{h}\frac{n(k_i-1)(2\exp(\epsilon_i)+k_i-2)}{(\exp(\epsilon_i)-1)^2} \\
&\quad + \sum_{j=h+1}^{l}\frac{nk_j\exp(\epsilon_j)}{(\exp(\epsilon_j)-1)^2} \qquad (9)
\end{aligned}
$$

Let $\exp(\epsilon_i) = x_i$. To satisfy the privacy guarantee, our goal is to minimize the following equation $L(x)$:

$$
L(x) = \min_{x_1\cdots x_h x_{h+1}^2\cdots x_l^2=\exp(\epsilon)}\left[\sum_{i=1}^{h}\frac{n(k_i-1)(2x_i+k_i-2)}{(x_i-1)^2} + \sum_{j=h+1}^{l}\frac{nk_jx_j}{(x_j-1)^2}\right] \qquad (10)
$$

To prove that the equation (10) has a minimum, let's first look at two general theorems.

Theorem 4: If $f(x)$ is continuous in $[a, b]$ and has first and second derivatives in (a, b). Then if $f''(x) > 0$ in (a, b), $f(x)$ is concave in $[a, b]$.

Theorem 5: If $f(x, y)$ is a concave function with continuous partial derivatives in the open region D, $(x_0, y_0) \in D$ and $f'_x(x_0, y_0) = 0, f'_y(x_0, y_0) = 0$, then $f(x_0, y_0)$ must be the minimum value of $f(x, y)$ in D.

The above theorem in low dimensional space can be easily extended to high dimensional space. In high dimensional space, since the second-order partial derivative $\frac{\partial^2 L(x)}{\partial^2 x_i} > 0, i = 1, 2, \cdots, l$, $L(x)$ is a strictly concave function for the variable x_i. The solution to which the first derivative of equation (10) is zero is the minimum solution, that is $\exists\ x^* = (x_1^*, \cdots, x_l^*)$, $s.t.\ \frac{\partial L(x^*)}{\partial x_i^*} = 0, i = 1, 2, \cdots, l$, $x_1^*\cdots x_h^* {x_{h+1}^*}^2\cdots {x_l^*}^2 = \exp(\epsilon)$. x^* is the minimum solution.

Equation (10) is a conditional constrained optimization problem, which is difficult to solve directly. Therefore, we employ LM method to translate the conditional restrictions into unconditional constraints:

$$
\begin{aligned}
L(x,\lambda) = &\left[\sum_{i=1}^{h}\frac{n(k_i-1)(2x_i+k_i-2)}{(x_i-1)^2} + \sum_{j=h+1}^{l}\frac{nk_jx_j}{(x_j-1)^2}\right] \\
&+ \lambda(x_1\cdots x_h x_{h+1}^2\cdots x_l^2 - \exp(\epsilon)) \qquad (11)
\end{aligned}
$$

Its optimal solution is obtained by solving the following equations:

$$
\begin{cases}
\dfrac{\partial L(x,\lambda)}{\partial x_i} = \lambda\exp(\epsilon)(x_i-1)^3 \\
\qquad\qquad -2n(k_i-1)(x_i+k_i-1)x_i \\
\qquad\qquad = 0, i = 1, 2, \cdots, h \\
\dfrac{\partial L(x,\lambda)}{\partial x_i} = 2\lambda\exp(\epsilon)(x_i-1)^3 - k_i n(x_i+1)x_i \\
\qquad\qquad = 0, i = h+1, \cdots, l \\
\dfrac{\partial L(x,\lambda)}{\partial \lambda} = x_1\cdots x_h x_{h+1}^2\cdots x_l^2 - \exp(\epsilon) = 0
\end{cases} \qquad (12)
$$

Let us make a simple transformation to the equation, and we can obtain:

$$
\begin{cases}
(1)\ 2\lambda\exp(\epsilon)x_i^3 - (6\lambda\exp(\epsilon)+nk_i)x_i^2 + \\
\quad (6\lambda\exp(\epsilon) - nk_i)x_i - 2\lambda\exp(\epsilon) = 0 \\
\quad i = 1, 2, \cdots, h \\
(2)\ \lambda\exp(\epsilon)x_i^3 - (3\lambda\exp(\epsilon)+2nk_i-2n)x_i^2 + \\
\quad (3\lambda\exp(\epsilon) - 2nk_i^2 - 4nk_i + 2n)x_i - \lambda\exp(\epsilon) = 0 \\
\quad i = h+1, h+2, \cdots, l \\
(3)\ x_1\cdots x_h x_{h+1}^2\cdots x_l^2 = \exp(\epsilon)
\end{cases} \qquad (13)
$$

The above equation relates to the problem of solving the univariate cubic equation. There are a variety of methods to solve the univariate cubic equation; here we employ the CF method. We let

$$
\begin{aligned}
(1)\ a_1^i &= 2\lambda\exp(\epsilon), \quad b_1^i = -(3a_1^i + nk_i), \\
c_1^i &= 3a_1^i - nk_i, \quad d_1^i = -a_1^i \\
i &= 1, 2, \cdots, h \\
(2)\ a_2^i &= \lambda\exp(\epsilon), \quad b_2^i = -(3a_2^i + 2nk_i - 2n), \\
c_2^i &= 3a_2^i - 2nk_i^2 - 4nk_i + 2n, \quad d_2^i = -a_2^i \\
i &= h+1, h+2, \cdots, l
\end{aligned}
$$

The univariate cubic equation in (13) can be changed into:

$$
\begin{cases}
a_1^i x_i^3 + b_1^i x_i^2 + c_1^i x_i + d_1^i = 0 \\
i = 1, 2, \cdots, h \\
a_2^i x_i^3 + b_2^i x_i^2 + c_2^i x_i + d_2^i = 0 \\
i = h+1, h+2, \cdots, l \\
x_1\cdots x_h x_{h+1}^2\cdots x_l^2 = \exp(\epsilon)
\end{cases} \qquad (14)
$$

To find the root of the equation, we let $x_i = y_i - \frac{b_k^i}{3a_k^i}, k = 1, 2; i = 1, \cdots, h$. The first two equations in (14) can be changed into:

$$
y_i^3 + \left(\frac{c_k^i}{a_k^i} - \frac{{b_k^i}^2}{3{a_k^i}^2}\right)y_i + \left(\frac{d_k^i}{a_k^i} + \frac{2{b_k^i}^3}{27{a_k^i}^3} - \frac{b_k^i c_k^i}{3{a_k^i}^2}\right) = 0. \qquad (15)
$$

We let $p = \frac{c_k^i}{a_k^i} - \frac{{b_k^i}^2}{3{a_k^i}^2}$, $q = \frac{d_k^i}{a_k^i} + \frac{2{b_k^i}^3}{27{a_k^i}^3} - \frac{b_k^i c_k^i}{3{a_k^i}^2}$, so (15) can be expressed as

$$
y_i^3 + py_i + q = 0. \qquad (16)
$$

By using the CF method, we can obtain the root of (16) as follows:

$$\begin{aligned} y_{i1} &= \sqrt[3]{-\frac{q}{2}+\sqrt{(\frac{q}{2})^2+(\frac{p}{3})^3}}+\sqrt[3]{-\frac{q}{2}-\sqrt{(\frac{q}{2})^2+(\frac{p}{3})^3}} \\ y_{i2} &= \omega\sqrt[3]{-\frac{q}{2}+\sqrt{(\frac{q}{2})^2+(\frac{p}{3})^3}}+\omega^2\sqrt[3]{-\frac{q}{2}-\sqrt{(\frac{q}{2})^2+(\frac{p}{3})^3}} \\ y_{i3} &= \omega^2\sqrt[3]{-\frac{q}{2}+\sqrt{(\frac{q}{2})^2+(\frac{p}{3})^3}}+\omega\sqrt[3]{-\frac{q}{2}-\sqrt{(\frac{q}{2})^2+(\frac{p}{3})^3}} \end{aligned} \tag{17}$$

where $\omega = \frac{-1+\sqrt{3}i}{2}$. By solving l equations, we can obtain the roots $x_{ij} = y_{ij} - \frac{b}{3a}, j = 1, 2, 3$ of (14) represented by λ. We only take x_{i1} as our final real root. The finally obtained l solutions $x_1^*, x_2^*, \cdots, x_l^*$ are applied to equation $f(\lambda) = x_1^* x_2^* \cdots x_l^* - exp(\frac{\epsilon}{2}) = 0$. We can also obtain a higher order equation of λ. We employ the existing Newton-Raphson Method (NRM) to solve equations of high degree with one unknown. The NSM first chooses two initial values λ_0, λ_1. At each iteration, let λ_k, λ_{k-1} be the initial value of the next iteration, which is given as:

$$\lambda_{k+1} = \lambda_k - \frac{f(\lambda_k)}{f(\lambda_k')} \tag{18}$$

The NSM will produce an infinite sequence $\{\lambda_1, \lambda_2, \cdots\}$, and this sequence converges to the true root of the function $f(\lambda)$. After obtaining the asymptotic answer λ^*, we can obtain the value of $\{x_1, x_2, \cdots, x_l\}$. The privacy budget ϵ_i can be obtained by $\epsilon_i = \log x_i$, $i = 1, \cdots, l$ for each attribute. To analyse the optimal answer $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_l\}$, we arrive at the following conclusions:

Theorem 6: For multiple unbalanced categorical data, the optimal privacy budget value ϵ_i of CRR is positively correlated with the number of items k_i.

D. ERROR BOUNDS, COMPUTATIONAL COMPLEXITIES AND COMMUNICATION COSTS

1) ERROR BOUNDS

In this subsection, we present several methods to get the upper error boundaries under special conditions, for example $h = 0$, $h = l$ or $k_1 = k_2 = \cdots = k_l$. These error boundaries are all derived from equation (10). If we set $h = 0$, CRR will degenerate into OBRR. Then, the equation (10) will be changed into:

$$L(x) = \min_{x_1^2 \cdots x_l^2 = \exp(\epsilon)} \sum_{i=1}^{l} \frac{nk_i x_i}{(x_i - 1)^2} \tag{19}$$

OBRR is optimal in the high-privacy regime when dealing with multivariate unbalanced nominal attributes. If we assume the solution of equation (19) is $\{x_1, \cdots, x_l\}$. We can calculate the privacy budget $\epsilon_i = \log x_i$ allocated for each attribute, where $i = 1, \cdots, l$ refers to the index of attributes. The mean square error of OBRR will be changed into $\sum_{i=1}^{l} \frac{nk_i \exp(\epsilon_i)}{(\exp(\epsilon_i)-1)^2}$. Then we have:

$$\mathbf{SE}(CRR) \le \sum_{i=1}^{l} \frac{nk_i \exp(\epsilon_i)}{(\exp(\epsilon_i) - 1)^2}$$

If we set $h = l$, CRR would be degenerate into OMRR. Then, the equation (10) will be changed into:

$$L(x) = \min_{x_1 \cdots x_l = \exp(\epsilon)} \sum_{i=1}^{l} \frac{n(k_i - 1)(2x_i + k_i - 2)}{(x_i - 1)^2} \tag{20}$$

OMRR is optimal in the high-privacy regime when dealing with multivariate unbalanced nominal attributes. If we assume the solution of equation (20) is $\{x_1', \cdots, x_l'\}$. We can calculate the privacy budget $\epsilon_i' = \log x_i'$ allocated for each attribute, where $i = 1, \cdots, l$ refers to the index of attributes. The mean square error of OMRR will be changed into $\sum_{i=1}^{l} \frac{n(k_i-1)(2\exp(\epsilon_i')+k_i-2)}{(\exp(\epsilon_i')-1)^2}$. Then we have:

$$\mathbf{SE}(CRR) \le \sum_{i=1}^{l} \frac{n(k_i - 1)(2\exp(\epsilon_i') + k_i - 2)}{(\exp(\epsilon_i') - 1)^2}$$

Thus, we have:

$$\mathbf{SE}(CRR) \le \min\{\sum_{i=1}^{l} \frac{n(k_i - 1)(2\exp(\epsilon_i') + k_i - 2)}{(\exp(\epsilon_i') - 1)^2}, \sum_{i=1}^{l} \frac{nk_i \exp(\epsilon_i)}{(\exp(\epsilon_i) - 1)^2}\}$$

Actually, for multiple unbalanced categorical data, the optimal privacy budget value ϵ_i of OBRR or OMRR is positively correlated with the number of items k_i. Specially, if $k_1 = k_2 = \cdots = k_l$, the allocation scheme $\epsilon_1 = \cdots = \epsilon_l$ is optimal. To meet the local differential privacy guarantee, if we set $h = 0$, the allocation scheme $\epsilon_1 = \cdots = \epsilon_l = \frac{\epsilon}{2l}$ for OBRR is optimal, and its mean square error equals to $\frac{dn\exp(\frac{\epsilon}{2l})}{(\exp(\frac{\epsilon}{2l})-1)^2}$. When the dimensions of attributes are different, we have

$$SE(CRR) < \frac{dn\exp(\frac{\epsilon}{2l})}{(\exp(\frac{\epsilon}{2l}) - 1)^2}$$

If we set $h = l$, the allocation scheme $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_l = \frac{\epsilon}{l}$ for OMRR is optimal, and its mean square error equals to $\frac{n(d-l)(2\exp(\frac{\epsilon}{l})+\frac{d}{l}-2)}{(\exp(\frac{\epsilon}{l})-2)^2}$. When the dimensions of attributes are different, we have

$$SE(CRR) < \frac{n(d-l)(2\exp(\frac{\epsilon}{l}) + \frac{d}{l} - 2)}{(\exp(\frac{\epsilon}{l}) - 2)^2}$$

Thus, we have

$$\mathbf{SE}(CRR) \le \min\{\frac{dn\exp(\frac{\epsilon}{2l})}{(\exp(\frac{\epsilon}{2l})-1)^2}, \frac{n(d-l)(2\exp(\frac{\epsilon}{l})+\frac{d}{l}-2)}{(\exp(\frac{\epsilon}{l})-1)^2}\}$$

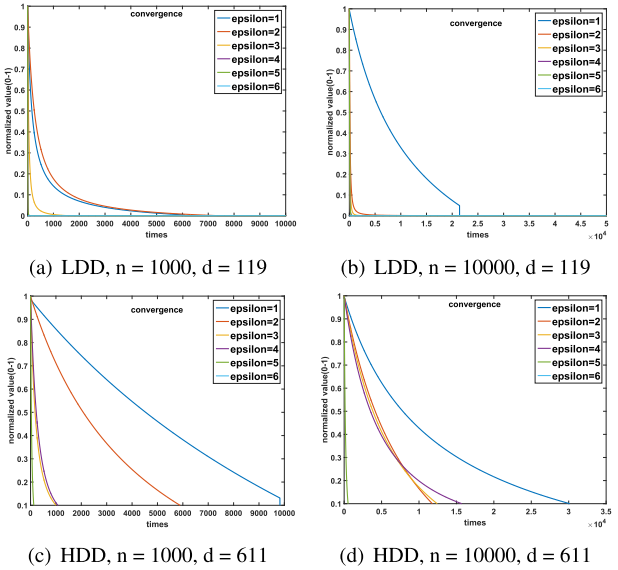
Therefore the estimated histogram in the CRR mechanism is no less favorable than the estimated histogram by the state-of-art BRR [12], [15] or MRR [23]. CRR is also superior to OBRR and OMRR which are presented in this paper.

2) COMPUTATIONAL COMPLEXITIES

In this part, we discuss the issue of time complexity from two perspectives, aggregators and participants, and compare them with existing methods. For participants, CRR does not cause additional computing overhead. For each participant, the CRR mechanism proposed in Algorithm 1 has a computational complexity of $O(d)$, where d is the domain of the participants' secret values and the domain of histogram buckets. The time complexity of CRR on the client side is the same as that of other local differential privacy technologies using Bloom filter technology, such as LoPub [20], RAPPOR [15], RAPPOR-unknown [36] and DLDP [44].

For the aggregator, searching the optimal divided index h requires $O(l^2)$ computational complexity where l is the number of attributes, and finding the optimal budget allocation scheme $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_l\}$ with m_d-digit precision, provided that a good initial approximation is known, requires $O(\log(m_d)F(m_d))$ complexity, where $F(m_d)$ is the cost of calculating $\frac{f(x)}{f'(x)}$ with m_d-digit precision. The biggest problem using the NRM method lies in the selection of the initial iteration values. If the initial value is far from the true solution, it is difficult for the NRM method to converge. To improve the shortcomings of the overreliance of the NRM on the initial value, we add the selection of the best initial value to the iteration process. The iteration is divided into two processes. We first calculate whether $|f(\lambda_k) - f(\lambda)|$ falls within a reasonable interval $[a, b]$ on the basis of the given initial value λ_0. If it does not match, then we add a fixed step size $\lambda_{k+1} = \lambda_k + \delta$ and recalculate until a suitable initial value λ_0' is found. Based on the best initial value λ_0', the NRM method is used to improve the iteration accuracy. The global threshold is set to $\xi = 0.01$. When the iteration error $f(\lambda^*) - f(\lambda) \leq \xi$, the iteration is terminated. To show the relationship between the overall number of iterations and the number of iteration errors, we performed experiments on two data sets. The data sets are detailed in section V. The results are shown in Figure 1. To facilitate the comparison, the error is normalized to [0, 1]. It can be seen from the figure that the number of iterations is on the order of $1e4$. Therefore, the time complexity of calculating the optimal budget allocation scheme is approximately $O(t \log(m_d)F(m_d))$, where t is the average number of iterations. Estimating the histogram from the observed sanitized data costs $O(nd + n)$ time, where n is the number of participants.

The total time complexities of the CRR mechanism are $O(l^2 + nd + d + t\log(m_d)F(m_d))$. For the aggregator, RAPPOR-unknown needs to learn the correlations between dimensions via an EM-based learning algorithm. However, the EM algorithm will have an exponentially higher complexity. When the dimension is high, the time complexities will be far larger than those from the CRR mechanism. Lopub also uses the EM algorithm to estimate the joint probability distribution. Their total time complexities are $O(ndk_{avg}^l + tnk_{avg}^{2l})$, where $k_{avg} = \frac{d}{l}$ denotes the average dimensions of each attribute, and t denotes the number of iterations. When the number of attributes is large, the time complexity of Lopub is very high.



(a) LDD, n = 1000, d = 119 (b) LDD, n = 10000, d = 119

(c) HDD, n = 1000, d = 611 (d) HDD, n = 10000, d = 611

FIGURE 1. The relationship between the number of iterations and convergence.

3) COMMUNICATION COSTS

In the face of high-dimensional data, compared with the existing local differential privacy mechanisms, our method will not cause extra delay for participants. For the aggregator, our method has a relatively small delay, and as the number of dimensions increases, the effectiveness of our method will become more prominent. In conclusion, the CRR mechanism is highly efficient for high-dimensional categorical data aggregation.

The communication cost of CRR is

$$C = \sum_{i=1}^{l} k_i = d$$

If we assume that the domain of each attribute is publicly known by both users and the server, and the dividing index h is known, then the minimal communication cost of CRR is

$$C_{min} = h + \frac{\ln(1/p)}{(\ln 2)^2} \sum_{i=h+1}^{l} |k_i|$$

The reason is that only randomly flipped bit strings (not original data record) are sent.

For comparison, under the same condition, when RAPPOR is directly applied to l-dimensional data, all $\mathbf{a_1} \times \cdots \mathbf{a_l}$ candidate values will be regarded as 1-dimensional data; then, the cost is

$$C_{RAPPOR} = \frac{\ln(1/p)}{(\ln 2)^2} \prod_{i=1}^{l} k_i$$

where $\prod_{i=1}^{l} k_i$ is due to the size of the candidate set $\mathbf{a_1} \times \cdots \mathbf{a_l}$. This is one of the reasons why we cannot directly apply the

TABLE 3. Communication costs of different local differential privacy mechanisms.

CRR	RAPPOR-unknown	Lopub	LDPminer	RAPPOR
$O(d)$	$O(d+r)$	$O(d)$	$O(d+s)$	$\prod_{i=1}^{l} k_i$

univariate local differential privacy to the high-dimensional data. Additionally, we compare CRR with existing local differential privacy mechanisms, and the results are shown in Table 3. r represents the number of substrings in the RAPPOR-unknown [36] method, s is the encoding length of the string in the S-Hist [25] method and d is defined as previously. It can be seen from the table that our method has a smaller communication cost than other mechanisms.

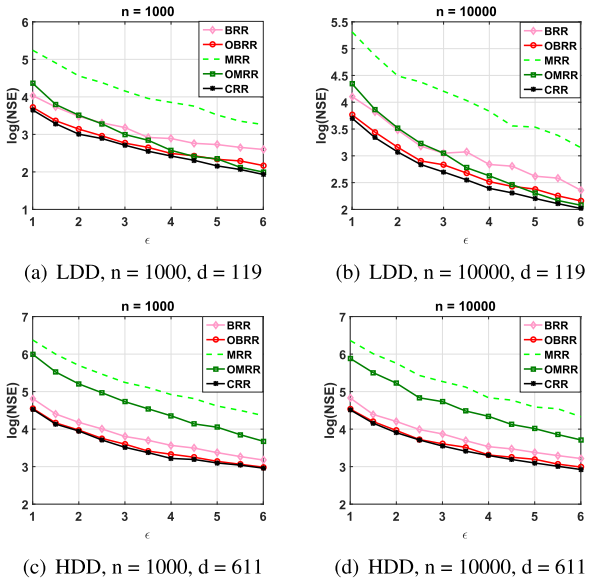
V. SIMULATION

A. DATA SET

We assume that each participant's secret data value is drawn from histogram H, which is uniform randomly generated during each aggregation. The dimensions of the data set is $[n, d]$. The selection of the data set guarantees the following criteria. First, each participant can only vote for l tickets, that is, the sum of each row of the data set matrix is l. Second, the total number of tickets for all participants is $l * n$. Without loss of generality, we assume there are 5 attributes, and each attribute has a different number of candidate values. We selected two data sets in total. The number of attribute categories is randomly selected to demonstrate the optimal effect of budget allocation for unbalanced data. Without loss of generality, we let $\{k_{11}, k_{12}, \cdots, k_{15}\} = \{5, 6, 150, 200, 250\}$ and $\{k_{21}, k_{22}, \cdots, k_{25}\} = \{2, 4, 6, 7, 100\}$. For simplicity, we denote the data set $\{k_{11}, k_{12}, \cdots, k_{15}\}$ as a higher degree of dispersion (HDD) and the data set $\{k_{21}, k_{22}, \cdots, k_{25}\}$ as a lower degree of dispersion (LDD). The number of participants in the two datasets was 1000 and 10000. The privacy budget ranged from 1.0 to 6.0, and we employed normalized square error ($NSE = \frac{SE}{n}$) as the metric to measure the performance of the mechanisms, where SE is the square error.

B. COMBINATIONAL RANDOMIZED RESPONSE MECHANISM

Considering the two extreme cases, $h = 0$ and $h = 1$ in (9), when $h = 0$, CRR allocates all privacy budgets to BRR. At this time, MRR does not work. This is equivalent to finding the optimal privacy budget allocation scheme in BRR. At this time, the differential privacy mechanism is called OBRR. Similarly, when $h = l$, we call it OMRR. OBRR and OMRR have greatly improved data availability than BRR and MRR, respectively. CRR combines the advantages of OBRR and OMRR. By properly adjusting the parameters, CRR is optimal regardless of the attribute category or the privacy budget. We compared the local differential privacy mechanism CRR proposed in this paper with BRR, OBRR, MRR, and OMRR for the different data sets HDD (n = 1000, 10000) and LDD (n = 1000, 10000). In particular,



(a) LDD, n = 1000, d = 119 (b) LDD, n = 10000, d = 119

(c) HDD, n = 1000, d = 611 (d) HDD, n = 10000, d = 611

FIGURE 2. The relationship between the estimated histogram error measured by log(NSE) and privacy budget ϵ.

we need to calculate the dividing index using the DIS method proposed in Algorithm 2. The calculated dividing index is 4 for LDD and 2 for HDD. We employ 5 different mechanisms to randomize the secret data sets HDD and LDD, and then use the decoder to estimate the frequencies of each item. The detailed experimental results are presented in Fig. 2.

Fig. 2(a) and Fig. 2(b) denote the NSE of encoding LDD when the number of participants is 1000 and 10000, respectively. Fig. 2(c) and Fig. represent the NSE of encoding HDD when the number of participants is 1000 and 10000, respectively. The light green dotted lines denote the estimated square error of MRR, and the bottom green lines represent the estimated square error of OMRR. The pink lines and the red lines denote the estimated square errors of BRR and OBRR, respectively. The black lines denote the combinational mechanism CRR. As seen from the figure, the optimal privacy budget allocation scheme proposed by this paper has had a beneficial effect. In Fig. 2(a) and Fig. 2(b) respectively, the OBRR mechanism can reduce the estimated square error by 41.6% and 40.2% compared to the BRR and the OMRR reduces the estimated square error by 72.8% and 72.0% compared to the MRR. In Fig. 2(c) and Fig. 2(d), respectively, the OBRR mechanism can reduce the estimated square error by 33.2% and 36.4% compared to the BRR and the OMRR reduces the estimated square error by 73.0% and 73.7% compared to the MRR. Because the number of items in HDD is 611, the situation where MRR is not suitable-resulting in the effect of OMRR-performs worse than the OBRR. CRR has the best performances over the two different secret data sets, the results of which are consistent with our theoretical analysis. In summary, CRR reduces the mean square estimation error by approximately 55% compared with differential privacy mechanism using other budget allocation schemes. It can be concluded from the experimental results that the

magnitude of the error reduction is independent of the number of participants n but is related to the number of attribute values $(k_1, k_2, \cdots, k_l)$. In fact, the larger the dimension difference between attributes, the more effective the privacy budget allocation scheme proposed in this paper will be.

VI. DISCUSSION

VII. CONCLUSION

Aiming to solve the privacy budget allocation problem for data with unbalanced multivariate nominal attributes, traditional local differential privacy algorithms usually assign the same privacy budget to all attributes with a different number of values, resulting in an undesired frequency estimation. To solve the problem, we propose an optimal privacy budget allocation scheme with high-dimensional heterogeneous data based on the Lagrange multiplier algorithm, Cardano Formula and Newton-Raphson methods. In addition, to meet the local privacy guarantee and the different needs of data for different privacy concern levels, we use the optimal privacy budget allocation scheme obtained by the above processes to improve BRR and MRR, which are then called OBRR and OMRR respectively, and propose a novel combinational randomized response mechanism, CRR. CRR combines the advantages of BRR and MRR to address the problem of high and low attribute dimensionality. The simulation results demonstrate that the proposed mechanism can achieve considerable improvement by reducing the estimated square error by 55% compared to that of the BRR and MRR on average.

REFERENCES

[1] E. Union. *General Data Protection Regulation*. Accessed: Feb. 2, 2020. [Online]. Available: http://en.wikipedia.org/wiki/General_Data_Protection_Regulation

[2] T. Zhu, G. Li, W. Zhou, and P. S. Yu, "Differentially private data publishing and analysis: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1619–1638, Aug. 2017.

[3] X. Yang, T. Wang, X. Ren, and W. Yu, "Survey on improving data utility in differentially private sequential data publishing," *IEEE Trans. Big Data*, early access, Jun. 15, 2017, doi: 10.1109/TBDATA.2017.2715334.

[4] S. P. Wang, "Generalizing data to provide anonymity when disclosing information," in *Proc. PODS*, vol. 98, Oct. 1998, p. 188.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k -anonymity," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, p. 3, Mar. 2007, doi: 10.1145/1217299.1217302.

[6] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 1–7.

[7] 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.

[8] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata, Lang. Program.*, 2006, pp. 1–6.

[9] C. Dwork and J. Lei, "Differential privacy and robust statistics," in *Proc. 41st Annu. ACM Symp. Symp. Theory Comput.*, 2009, pp. 371–380.

[10] A. Smith, "Privacy-preserving statistical estimation with optimal convergence rates," in *Proc. 43rd Annu. ACM Symp. Theory Comput.*, 2011, pp. 813–822.

[11] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" in *Proc. 49th Annu. IEEE Symp. Found. Comput. Sci.*, Oct. 2008, pp. 793–826.

[12] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, Oct. 2013, pp. 429–438.

[13] Q. Ye, H. Hu, X. Meng, and H. Zheng, "PrivKV: Key-value data collection with local differential privacy," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 317–337.

[14] Z. Zhang, T. Wang, N. Li, S. He, and J. Chen, "CALM: Consistent adaptive local marginal for marginal release under local differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Jan. 2018, pp. 212–229.

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

[16] M. M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, "Enhancing privacy in participatory sensing applications with multidimensional data," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun.*, Mar. 2012, pp. 142–152.

[17] J. Sun, R. Zhang, J. Zhang, and Y. Zhang, "PriStream: Privacy-preserving distributed stream monitoring of thresholded PERCENTILE statistics," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–6.

[18] S. Ruggles, C. Fitch, D. Magnuson, and J. Schroeder, "Differential privacy and census data: Implications for social and economic research," *Aea Papers Process.*, vol. 109, pp. 403–408, May 2019.

[19] R. Chen, Q. Xiao, Y. Zhang, and J. Xu, "Differentially private high-dimensional data publication via sampling-based inference," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 129–138.

[20] 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, Sep. 2018.

[21] X. Ren, C.-M. Yu, W. Yu, S. Yang, X. Yang, and J. McCann, "High-dimensional crowdsourced data distribution estimation with local privacy," in *Proc. IEEE Int. Conf. Comput. Inf. Technol. (CIT)*, Dec. 2016, pp. 226–233.

[22] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2879–2887.

[23] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2007, pp. 94–103.

[24] X. Fang, Q. Zeng, and G. Yang, "Local differential privacy for human-centered computing," *EURASIP J. Wireless Commun. Netw.*, vol. 2020, p. 1, Dec. 2020.

[25] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in *Proc. Forty-Seventh Annu. ACM Symp. Theory Comput.*, 2015, pp. 127–135.

[26] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, vol. 48, Jun. 2016, pp. 2436–2444.

[27] 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.

[28] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Stat. Assoc.*, vol. 60, no. 309, pp. 63–69, Mar. 1965.

[29] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under local differential privacy," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 5662–5672, Sep. 2017.

[30] W. Qardaji, W. Yang, and N. Li, "PriView: Practical differentially private release of marginal contingency tables," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 1435–1446, doi: 10.1145/2588555.2588575.

[31] W. LI, X. ZHANG, X. LI, G. CAO, and Q. ZHANG, "PPDP-PCAO: An efficient high-dimensional data releasing method with differential privacy protection," *IEEE Access*, vol. 7, pp. 176429–176437, 2019.

[32] W.-Y. Day and N. Li, "Differentially private publishing of high-dimensional data using sensitivity control," in *Proc. 10th ACM Symp. Inf., Comput. Commun. Secur.*, 2015, pp. 451–462.

[33] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "PrivBayes: Private data release via Bayesian networks," *ACM Trans. Database Syst.*, vol. 42, no. 4, pp. 1–41, 2014.

[34] X. Zhang, L. Chen, K. Jin, and X. Meng, "Private high-dimensional data publication with junction tree," *J. Comput. Res. Develop.*, vol. 55, pp. 2794–2809, 2018.

[35] S. Su, P. Tang, X. Cheng, R. Chen, and Z. Wu, "Differentially private multi-party high-dimensional data publishing," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 205–216.
[36] G. Fanti, V. Pihur, and Ú. Erlingsson, "Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries," *Proc. Privacy Enhancing Technol.*, vol. 2016, no. 3, pp. 41–61, Jul. 2016.
[37] H. Li, J. Cui, X. Lin, and J. Ma, "Improving the utility in differential private histogram publishing: Theoretical study and practice," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2016, pp. 1100–1109.
[38] N. Wang, Y. Gu, J. Xu, F. Li, and G. Yu, "Differentially private high-dimensional data publication via grouping and truncating techniques," *Frontiers Comput. Sci.*, vol. 13, no. 2, pp. 382–395, Apr. 2019.
[39] X. Cheng, P. Tang, S. Su, R. Chen, Z. Wu, and B. Zhu, "Multi-party high-dimensional data publishing under differential privacy," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1557–1571, Aug. 2020.
[40] G. Cormode, T. Kulkarni, and D. Srivastava, "Marginal release under local differential privacy," in *Proc. Int. Conf. Manage. Data*, 2018, pp. 131–146.
[41] 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.
[42] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, Jul. 1970.
[43] G. W. Chunhua Ju, Q. Gu, and S. Zhang, "Local differential privacy protection of high-dimensional perceptual data by the refined Bayes network," in *Sensors*, vol. 12, no. 7, pp. 226–233, 2020.
[44] F. Peng, S. Tang, B. Zhao, and Y. Liu, "A privacy-preserving data aggregation of mobile crowdsensing based on local differential privacy," in *Proc. ACM Turing Celebration Conf. China*, 2019, pp. 1–5.

XUEJIE FENG received the M.S. degree in business management from the University of Portsmouth, Portsmouth, U.K., in 2013. She is currently pursuing the Ph.D. degree with the Department of Mathematics, Harbin Institute of Technology, Harbin, China, in 2017.

Since 2016, she was a Lecturer with the School of International Business, Qingdao Huanghai University, Qingdao, China. 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, in 2019, China.

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

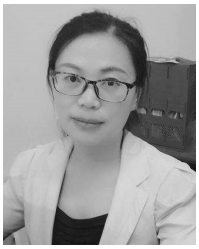
Since 2006, he was a Professor with the Department of Mathematics, Harbin Institute of Technology, Harbin. He is currently an Assistant Dean and the Office Director of the Department of Mathematics. He has presided over and undertaken key projects of the National Natural Science Foundation of China, 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, the Teaching Achievement Award of Mathematical Modeling Teaching Research and Quality Education Practice Teaching Achievement Award, in 2003, and so on.

JING LI received the M.S. degree in operational research and cybernetics from Qufu Normal University, Qufu, China, in 2008. She is currently pursuing the Ph.D. degree with the School of Mathematics, Harbin Institute of Technology.

Since 2009, she was an Associate Professor with Qingdao Huanghai University, Qingdao, China. She has presided over and undertaken many projects such as the Shandong Province Higher Educational Science and Technology Program. Her research interests include operation research optimization, data fusion, and intelligent algorithm.

Prof. Li's awards and honors include the Shandong Provincial Teaching Achievement Award, in 2018, the Famous Teacher of Qingdao, in 2016, and so on.



LINLIN DAI received the B.S. degree in mathematics from Liaocheng University, Liaocheng, China, in 2003, and the M.S. degree in mathematics education from Guizhou Normal University, Guiyang, China, in 2006. She is currently pursuing the Ph.D. degree with the Harbin Institute of Technology, Harbin, China.

Since 2014, she was a Professor with the School of Finance and Economics, Qingdao Huanghai University, Qingdao, China. She is currently an Assistant Dean with the School of Bigdata. She has presided over and undertaken humanities and social sciences projects of universities and Twelfth Five-Year Plan project in Shandong, and a number of provincial and ministerial key projects. Her research interests include the statistical analysis, data processing, and distributed computing.

Prof. Dai's awards and honors include the Teaching Achievement Award of Mathematical Modeling Teaching Research and Quality Education Practice Teaching Achievement Award, in 2011, the Teaching Master of Qingdao, in 2018, and so on.