Feng and Zhang *Hum. Cent. Comput. Inf. Sci.* (2020) 10:25
https://doi.org/10.1186/s13673-020-00233-x

Human-centric Computing and Information Sciences

RESEARCH Open Access

Local differential privacy for unbalanced multivariate nominal attributes

Xuejie Feng[1,2] and Chiping Zhang[1*]

*Correspondence: zcp@hit.edu.cn
[1] Department of Mathematics, Harbin Institute of Technology, Harbin 150001, China
Full list of author information is available at the end of the article

Abstract

Data with unbalanced multivariate nominal attributes collected from a large number of users provide a wealth of knowledge for our society. However, it also poses an unprecedented privacy threat to participants. Local differential privacy, a variant of differential privacy, is proposed to eliminate the privacy concern by aggregating only randomized values from each user, with the provision of plausible deniability. However, traditional local differential privacy algorithms usually assign the same privacy budget to attributes with different dimensions, leading to large data utility loss and high communication costs. To obtain highly accurate results while satisfying local differential privacy, the aggregator needs a reasonable privacy budget allocation scheme. In this paper, the Lagrange multiplier (LM) algorithm was used to transform the privacy budget allocation problem into a problem of calculating the minimum value from unconditionally constrained convex functions. The solution to the nonlinear equation obtained by the Cardano formula (CF) and Newton-Raphson (NS) methods was used as the optimal privacy budget allocation scheme. Then, we improved two popular local differential privacy mechanisms by taking advantage of the proposed privacy budget allocation techniques. Extension simulations on two different data sets with multivariate nominal attributes demonstrated that the scheme proposed in this paper can significantly reduce the estimation error under the premise of satisfying local differential privacy.

Keywords: Differential privacy, Multivariate, Information, Lagrange multiplier, Cardano formula

Introduction

At present, with the development of various integrated sensors and crowdsensing systems, crowdsourced information from all aspects can be collected and analyzed among various data attributes to better produce rich knowledge about a group, thus benefiting everyone in the crowdsourcing system [1]. 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. For example, in the process of providing cloud services, user gender, age, and habits when using operating systems and browsers should be deeply explored to provide different special services to different user groups. In hospital staff, people's historical medical records and genetic information [2] can be followed closely to better diagnose and monitor patient health status.

However, in practical crowdsourcing systems, high-dimensional heterogeneous data cannot be utilized effectively. There are two main reasons for this situation. (1) Non-local privacy guarantee. Differential privacy [3, 4], as one of the currently effective privacy protection mechanisms, randomizes the query output by adding noise to sensitive data to achieve the purpose of privacy protection. Many existing works [5–8] focus on centralized data sets under the assumption of trusted third-party data collectors. These concentrate raw data into a data center and then publish relevant statistical information that satisfies differential privacy. However, even if third-party data collectors claim that they will not steal and disclose confidential user information, the privacy of users is still not guaranteed. It is difficult to find a truly trusted third-party data collection platform in practical applications, which significantly limits the use of centralized differential privacy technologies. As users, they prefer to ensure data security on the user side, enabling themselves to process and protect their confidential information separately (i.e., local differential privacy [9, 10]). (2) High-dimensional disaster. In crowdsourcing systems, high-dimensional heterogeneous data are ubiquitous. With the increases in data dimensions and the dimensional difference between different attributes, many existing local differential privacy mechanisms such as RAPPOR [11] and [12, 13], if straightforwardly applied to multiple attributes with unbalanced dimensions, will become extremely unavailable. Their fatal drawbacks are their non-optimized privacy budget allocation schemes and high computational complexities, which lead to large data utility loss and high latency. Different dimensions of attributes need to allocate different privacy budgets. How to find the best allocation scheme is the key to improving data utility.

In addition to privacy vulnerability and data utility, collecting a large amount of data from distributed user groups means that the efficiency of data processing is low, especially in the application of the Internet of Things (IoT). Thus, it is important to provide an efficient privacy-preserving method with high-dimensional heterogeneous data. Furthermore, considering that the privacy concern level 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 addressing the above issues, many existing methods have proved their effectiveness from different perspectives. One is to ensure that user privacy is not leaked when users are provided a local privacy guarantee, such as [11–13]. However, these methods become extremely complicated in communication, and the data availability drops sharply when processing high-dimensional heterogeneous data. The other is to privately release high-dimensional data [14–16]. These methods mainly use specific methods 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 their unreasonable privacy budget allocation schemes.

In this paper, we aim at designing an efficient and effective privacy budget allocation scheme for high-dimensional heterogeneous data under the local privacy guarantee. Our main contributions are as follows:

- We propose an optimal privacy budget allocation scheme with high-dimensional heterogeneous data. In this scheme, we use the Lagrange multiplier (LM) algorithm to transform the privacy budget allocation problem into a problem of calculating the minimum value from unconditionally constrained convex functions. Then, the Cardano formula (CF) and Newton-Raphson (NS) methods are employed to iteratively calculate the optimal solution.
- To meet the local privacy guarantee and the different needs of different data for the privacy concern levels, we use the optimal privacy budget allocation scheme obtained by the above procedure to improve the BRR and MRR and call it the OBRR and OMRR, respectively, which are optimal in the high and low privacy regimes with high-dimensional heterogeneous data, respectively.
- Finally, we conduct simulation experiments to show that the two improved algorithms, OBRR and OMRR, can significantly reduce the estimation error under the premise of satisfying local differential privacy, with lower time and communication complexities.

Related work

This paper focuses on the frequency statistics problem of high-dimensional heterogeneous data with local differential privacy, which refers to the situation where each user sends multiple variable values voted from candidate attributes. The candidate attributes always have different dimensions. Without loss of generality, we assume that the candidate attributes $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l\}$, where each attribute $\mathbf{a}_i$ has a specific dimension k_i, and we assume $d = k_1 + k_2 + \cdots + k_l$. Each user needs to translate a fixed value of l variables. Unlike single-valued frequency statistics, in multivariate scenarios, we need to consider not only the locality of users' privacy but also the segmentation of privacy budgets. An unreasonable privacy budget allocation scheme will result in a sharp drop in sanitized data utility.

Local privacy guarantee

Despite the privacy protection reaction against difference and inference attacks from aggregate queries, individuals' data may also suffer from privacy leakage before aggregation. Given the privacy flaws of differential privacy, the notion of local privacy has been proposed to provide the local privacy guarantee to distributed users [9, 10]. Recently, local privacy has aimed to learn particular aggregation features from distributed users with some public knowledge. Groat et al. [12] proposed the technique of negative surveys, which is based on randomized response techniques, to identify the true distributions from noisy participant data. Similarly, Bassily et al. [17] proposed the S-Hist algorithm. To reduce the transmission cost, they use random response technology to perturb the original data and then randomly select one of the bits and send it to the data collector. However, when the dimension is high, the sparsity of the data will lead to much utility loss. At the same time, their high computational complexities will lead to high latency.

Many single-valued frequency statistical mechanisms that satisfy local differential privacy have been proposed. Erlingsson et al. [11] proposed RAPPOR to estimate the frequencies of different strings in a candidate set. Their subsequent research RAPPOR-unknown [18] proposed learning the correlations between dimensions via an EM-based learning algorithm.

Intuitively, single-valued frequency statistics can be used repeatedly on each variable in high-dimensional cases. However, when the dimension is high, the data utility decreases dramatically, and the computational complexities increase exponentially. For the RAPPOR method, the length of the Bloom filters over the multi-attributes domain becomes:

$$m_{RAPPOR} \propto |k_1 \times k_2 \times k_l| = \prod_{i=1}^{l} k_i.$$

Their asymptotic error boundary rises from $O(\frac{k}{\epsilon\sqrt{n}})$ to $O(\frac{dk}{\epsilon\sqrt{n}})$. Moreover, the EM algorithm has an exponentially higher complexity. Therefore, if the single-valued frequency publishing method is used as the frequency publishing method in the high-dimensional case, the data utility and communication cost cannot be optimized. In addition, there are many improved local differential privacy algorithms suitable for single-valued frequency statistics, such as O-RAPPOR [19], PCE [20], k-RR [21], and k-Subset [22]. When addressing high-dimensional frequency statistics, they all have irreparable deficiencies in terms of data utility, communication costs or computational complexity.

High dimension

Currently, for the issue of high-dimensional data publishing, there are many methods that have had their effectiveness proved from different perspectives. For example, Cai et al. [23] studied the trade-off between statistical accuracy and privacy in average estimation and linear regression with high-dimensional data, mainly by improving the setting strategies of parameters such as the minimum-maximum lower bound and iterative threshold to ensure the statistical accuracy under the premise of satisfying differential privacy. However, this approach does not satisfy the locality of users' privacy, and they did not discuss how to allocate the privacy budget effectively. Li et al. [24] put forward the dichotomy of the privacy budget by using the method of publishing differential privacy histograms in groups. When it comes to high-dimensional heterogeneous data, there is no theoretical basis for their division. Since the dimensions of attributes are different, allocating the same privacy budget inevitably leads to a decline in data utility. Similarity, the method in [25] improves the accuracy of published data by adding additional processing to the output to restore the consistency of the count specified in the structure. However, this method cannot solve the problem of data utility decline caused by the sparsity of high-dimensional data before aggregation. There are also some methods such as [26, 27] that use the matrix mechanism to publish the database to minimize the query noise. However, the optimization cost of this method is very high, and the assumption that the query distribution is known in advance is not reasonable.

Another solution to mitigate the high dimension issue is to group the correlated records into clusters and then allocate the privacy budget to each low-dimensional cluster. However, in the existing schemes [14, 28, 29], the original data set is explicitly accessed twice to understand the correlation between properties and to generate the distribution of the cluster. The biggest problem with these methods is that the two accesses are computed separately and that there is no consistent privacy guarantee. That is, two different privacy budgets are allocated separately, but it is not clear how to allocate the privacy budget to achieve a sufficient privacy guarantee and utility maximization.

Moreover, although the unbalanced data with the multivariate nominal attribute can be reduced into several low-dimensional clusters, the sparsity caused by the combinations in each cluster still exists and may result in lower utility. In contrast to the totally centralized setting in [14], Su et al. [30] proposed a distributed multiparty setting to publish a new data set from multiple data curators. However, their multiparty computation can protect only the privacy between data servers. Instead, there is no guarantee of local personal privacy in a data server. In addition, Zhang et al. [31] proposed a self-adaptive regression-based multivariate data compression scheme. They used a correlation matrix to compress different data streams from the same node to reduce communication costs. However, this method does not solve how to effectively compress a high-dimensional data stream when there is only one.

To solve the shortcomings of the above methods, which cannot meet the privacy locality nor handle high-dimensional data, some effective methods have been proposed. For example, Ren et al. proposed LoPub [15, 16], which combines the RAPPOR and probability graph model. They first transform each attribute value into a random bit string using a Bloom filter [32] and then send it to the central server. Subsequently, similar to the high-dimensional data publishing method based on centralized differential privacy in [14], the data collector determines the frequency statistics of the collected data and then constructs a Markov network. The joint probability distribution of attributes is expressed as a maximal clique to reduce the dimensions of the data. Finally, a data set is resynthesized by a joint probability distribution for data release. However, the biggest disadvantage of this method is that it does not consider the allocation of the privacy budget before the high-dimensional heterogeneous data are aggregated to the server. Moreover, if each attribute is mutually independent, they propose using the EM to estimate the multivariate distribution, which will increase the computational complexities exponentially.

To overcome the shortcomings of low data utility, nonlocal privacy and high computational complexities within those schemes, we propose a novel privacy budget allocation scheme to publish unbalanced multivariate nominal attribute data while guaranteeing local privacy. At present, many similar optimization theories and methods have been proposed [33–35], but different objective functions lead to different solutions. In this paper, we turn the privacy budget allocation problem into a problem of solving the univariate cubic equation. The experimental results show that our method can greatly improve the low query accuracy caused by the defect of privacy budget allocation.

System model

The demonstrative aggregation model is depicted in Fig. 1, where several users and a central aggregator are interconnected, constituting a crowdsourcing system. At first, the aggregator releases or publishes unbalanced multivariate aggregation query $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l\}$ to each participant, along with the global parameters, including the optimal privacy budget allocation scheme $\epsilon = \{\epsilon_1, \ldots, \epsilon_l\}$ for each attribute $\mathbf{a}_i$, and other specific mechanism parameters, such as the sign of the high or low privacy regime. The different regimes require different privacy mechanisms i.e., OBRR or OMRR). In our mechanism, both secret data $\mathbf{v}$ and sanitized data $\mathbf{v}'$ are expressed as bit maps; specifically, if a participant's secret value equals the j-th element V_j in data domain V, then the secret data $v_i \in \{0, 1\}^{|V|}$ is a bit map of length $|V|$, with the j-th bit set to 1 and other bits

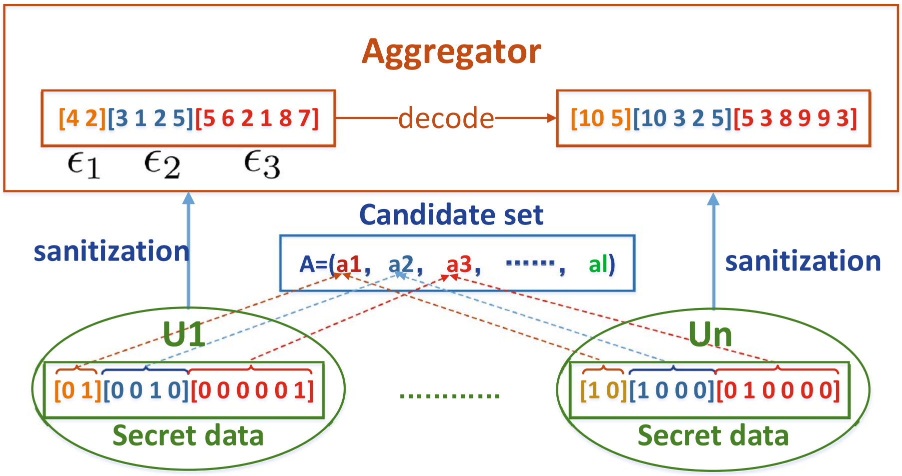


Fig. 1 Aggregation model for differential private histogram estimation with multivariate unbalanced nominal attributes

set to 0. After receiving the sanitized data list $\{\mathbf{v}'_1, \mathbf{v}'_2, \ldots, \mathbf{v}'_n\}$, 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 releasing with local differential privacy, no one knows the secret information they release except for the participants themselves.

Problem statement

Given a collection of data records with l attributes from different users, the dimensions of different attributes are different. Our goal is to help the aggregator design a reasonable privacy budget allocation scheme to improve the utility of releasing data under different privacy regimes. Formally, the unbalanced multivariate nominal attributes $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l\}$, and each attribute $\mathbf{a}_i$ has a specific number of categories $\mathbf{a}_i = \{a_{i1}, a_{i2}, \ldots, a_{ik_i}\}$, where k_i is the number of categories for the i-th attribute, that is, $|\mathbf{a}_i| = k_i, i = 1, 2, \ldots, l$. We assume that if $i \neq j$, we have $k_i \neq k_j$. Specially, each user u_i possesses l-length attributes $\mathbf{v}_i = \{v_{i1}, v_{i2}, \ldots, v_{il}\}$. Let n be the total number of users and $d = k_1 + k_2 + \cdots + k_l$ be the length of the bit maps. $\mathbf{v}_i$ is first translated into bit maps: $\mathbf{h}_i = \{h_{11}, \ldots, h_{1k_i}, h_{2k_1} \ldots, h_{2k_2}, \ldots, h_{lk_l}\}$, where $h_{ij} \in \mathbf{a}_i, i = 1, 2, \ldots, l$. Then, the sanitized data $\mathbf{h}'_i$ is sent to the aggregator. The frequency of the true histogram is denoted as $\mathbf{H} = \sum\{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$. The estimated frequency of the sanitized histogram can be expressed as $\mathbf{H}'' = \sum\{\mathbf{h}''_1, \ldots, \mathbf{h}''_n\}$.

With the above notations, our problem can be formulated as follows: Given the fixed privacy budget (located in the high or low regime), our goal is to find the optimal privacy budget allocation scheme $\{\epsilon_1, \ldots, \epsilon_l\}$ to minimize the error of estimated histogram $\mathbf{H}''$ of the true histogram $\mathbf{H}$. This can be expressed as follows:

$$SE(\mathbf{H}, \mathbf{H}'') = \min_{\epsilon = \epsilon_1 + \cdots + \epsilon_l} E\left[\sum_{i=1}^{l} \sum_{j=1}^{k_i} (H_{ij}'' - H_{ij})^2\right]$$

Moreover, different privacy regimes require different privacy mechanisms, which require a flexible privacy budget allocation scheme that can be easily applied to different local privacy mechanisms. Some notations employed in this paper are listed in Table 1.

Preliminaries

Local differential privacy

The protection model under local differential privacy (LDP) fully considers the possibility of data collectors stealing or revealing user privacy during data collection. In this model, each user first randomizes the data and then sends the sanitized data to data collectors; data collectors collect statistics on the collected data to obtain valid analysis results. Local differential privacy [5] is a rigorous privacy notion in the local setting, which provides a stronger privacy guarantee than does centralized differential privacy. The formal definition of local differential privacy is as follows:

Definition 1 Given n users, where 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 all $M \subseteq Range(\mathcal{F})$:

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

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

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

Table 1 Notations

A	Multiple unbalanced categorical data sets
l	Number of attributes
n	Number of participants
k_i	Number of items of the i-th attribute
d	Total number of items, $d = \sum_i k_i$
$\mathbf{a}_j$	j-th attributes of A, the length $\lvert\mathbf{a}_j\rvert$ of which is k_j
$\mathbf{v}_i$	Private values possessed by the i-th user, the length $\lvert\mathbf{v}_i\rvert$ of which is l
v_{ij}	j-th value of $\mathbf{v}_i$
$\mathbf{h}_i$	Private bit vector of i-th users, the length of which is d
H	True histogram, $\mathbf{H} = \sum\{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$
H'	Sanitized histogram of $\mathbf{H}$, $\mathbf{H}' = \sum\{\mathbf{h}'_1, \ldots, \mathbf{h}'_n\}$
H''	Estimated histogram of $\mathbf{H}'$, $\mathbf{H}'' = \sum\{\mathbf{h}''_1, \ldots, \mathbf{h}''_n\}$
ϵ_i	Privacy budget of the i-th attribute
CF	Cardano formula
NS	Newton-Raphson method
SE	Square error
NSE	Normalized square error, $NSE = \frac{SE}{n}$

Binary randomized response

The binary randomized response (BRR) [11] is a technique that requires each user to send a sanitized bit to the aggregator, where the perturbation is based on a randomized response (RR). Each participant is asked to flip a biased coin with probability p in secret and tell the truth if it comes up heads but tell a lie otherwise (if the coin comes up tails). To solve the perturbation problem of multiple unbalanced categorical data, the binary random response first initializes a length-d binary vector $\mathbf{h} = (\underbrace{0, 0, \ldots, 0}_{d})$ of zeros, next maps the input $\mathbf{v}_i = \{v_{i1}, v_{i2}, \ldots, v_{il}\}$ of a user u_i to a position in $\mathbf{h}$ and then sets the rest of the positions to 0, *i.e.*, $\mathbf{h} = (\underbrace{0, \ldots, 1, \ldots, 0}_{k_1}, \ldots, \underbrace{0, \ldots, 1, \ldots, 0}_{k_l})$. For each bit in $\mathbf{h}$, the output $\mathbf{h}'$ is given by:

$$\mathbf{BRR}(h_i'|h_i) = \begin{cases} p, & \textit{if } h_i' = h_i \\ 1-p, & \textit{if } h_i' \neq h_i \end{cases} \tag{2}$$

Yet, how to determine the value of p to make the sanitized data released by each user satisfy the need for differential privacy is the key problem. To do so, we analyze the sensitivity of releasing a length-d bit vector to each user. Since each user possesses exactly l items, there are l ones in $\mathbf{H}$. Therefore, two such bit vectors can differ by at most $2l$ bits, meaning that the sensitivity is $2l$. To meet the requirements of differential privacy, the probability p follows the method applied by RAPPOR [11]:

$$p = \frac{\exp\left(\frac{\epsilon}{2l}\right)}{\exp\left(\frac{\epsilon}{2l}\right) + 1} \tag{3}$$

The BRR allocates the same privacy budget $\frac{\epsilon}{2l}$ for each attribute, regardless of whether the number of categories of attributes is the same. If the number of categories between attributes varies widely, for example, the user's browsing site and the user's gender, the same privacy budget will likely bring a large estimated deviation.

Multivariate randomized response

The multivariate randomized response (MRR) mechanism [36] is a locally differentiable private mechanism whose noisy output alphabet $\mathcal{Y}$ is the original input domain $\mathcal{X}$. Specially, each user possesses a set $\mathbf{v}_i = \{v_{i1}, \ldots, v_{il}\}$ of an item; after being perturbed by the MRR, the sanitized output turns into $\mathbf{v}_i' = \{v_{i1}', \ldots, v_{il}'\}$, where $v_{ij}', v_{ij} \in \mathbf{a}_j, j = 1, 2, \ldots, l$. Then, the user u_i publishes the sanitized set $\mathbf{v}_i'$ of items to the aggregator. The conditional probabilities are given by:

$$\mathbf{MRR}(v_{ij}'|v_{ij}) = \begin{cases} \frac{\exp(\epsilon_m)}{\exp(\epsilon_m)+k_j-1}, & \textit{if } v_{ij}' = v_{ij} \\ \frac{1}{\exp(\epsilon_m)+k_j-1}, & \textit{if } v_{ij}' \neq v_{ij} \end{cases} \tag{4}$$

To satisfy the requirements of the differential privacy, we analyze the sensitivity of releasing a length-l vector to each user in a manner similar to that mentioned above. Two such vectors can differ by at most l positions, meaning that the sensitivity is l. Thus, when ϵ_m satisfies $\epsilon_m = \frac{\epsilon}{l}$, the MRR mechanism satisfies the differential privacy requirements.

The MRR allocates the same privacy budget $\frac{\epsilon}{l}$ for each attribute. The same unreasonable budget allocation problem will also appear in the MRR mechanism.

The BRR mechanism incurs $O(d)$ communication costs for each user, and the MRR incurs $O(l)$ communication costs. The number of attributes l is usually far smaller than the total number of items d, that is, $l \ll d$. As far as the communication cost is concerned, the MRR is superior to the BRR. In the work proposed by Kairouz et al. [36], the BRR and MRR are called staircase mechanisms. The BRR has been proved to be optimal in the high privacy regime, and the MRR has been proved to be optimal in the low privacy regime [19]. However, their unreasonable privacy budget allocation schemes are fatal problems. In the next section, we present evidence showing how to obtain the optimal allocation schemes over multiple unbalanced categorical data. Then, we apply the optimal budget allocation scheme to the BRR and MRR, resulting in the optimal mechanisms in the high and low regimes, respectively.

Optimal privacy budget allocation

Optimal budget allocation for the BRR

The main goal of the aggregator is to estimate the frequency of the items without disclosing the privacy of the users. Therefore, we adopt the square error (SE) as the metric to evaluate the estimation. Without loss of generality, we assume there are l attributes $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l$ and that the number of items for each attribute $\mathbf{a}_i$ is k_i, that is, $|\mathbf{a}_i| = k_i$. The total number of items $d = k_1 + k_2 + \cdots + k_l$. We allocate budgets $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_l\}$ to the set of attributes $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l\}$, respectively, and $\epsilon_1 + \epsilon_2 + \cdots + \epsilon_l = \frac{\epsilon}{2}$. Each user u_i publishes a length-d bit vector $\mathbf{h}'_i$ obtained by perturbing the original bit vector $\mathbf{h}_i$. The true histogram $\mathbf{H} = \sum\{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$. The sanitized histogram $\mathbf{H}' = \sum\{\mathbf{h}'_1, \ldots, \mathbf{h}'_n\}$. Let $\mathbf{H}'' = \{H''_{11}, \ldots, H''_{1k_1}, H''_{21}, \ldots, H''_{2k_2}, \ldots, H''_{lk_l}\}$ denote the unbiased estimation of $\mathbf{H}$; for each attribute, we have $H''_{ij}p_i + (n - H''_{ij})(1 - p_i) = H'_{ij}, i = 1, \ldots, l, j = 1, \ldots, k_i$, where $p_i = \frac{\exp(\epsilon_i)}{\exp(\epsilon_i)+1}$. Thus, we have:

$$H''_{ij} = \frac{H'_{ij}(\exp(\epsilon_i) + 1) - n}{\exp(\epsilon_i) - 1}$$

The SE is given as follows:

$$\begin{aligned}
\mathbf{SE}(\epsilon, l, d) &= E\left[\sum_{i=1}^{l}\sum_{j=1}^{k_i}(H''_{ij} - H_{ij})^2\right] = \sum_{i=1}^{l}\sum_{j=1}^{k_i} E[(H''_{ij} - H_{ij})^2] \\
&= \sum_{i=1}^{l}\sum_{j=1}^{k_i} Var[H''_{ij}] = \sum_{i=1}^{l}\sum_{j=1}^{k_i}\left(\frac{\exp(\epsilon_i) + 1}{\exp(\epsilon_i) - 1}\right)^2 Var[H'_{ij}] \\
&= \sum_{i=1}^{l}\frac{nk_i \cdot \exp(\epsilon_i)}{(\exp(\epsilon_i) - 1)^2}
\end{aligned}$$

where H'_{ij} is the Bernoulli probability distribution, with the variance of H'_{ij} being equal to $np_i(1 - p_i)$. Our goal is given as:

$$L(\epsilon) = \min_{\epsilon_1+\cdots+\epsilon_l=\frac{\epsilon}{2}} \sum_{i=1}^{l}\frac{nk_i \cdot \exp(\epsilon_i)}{(\exp(\epsilon_i) - 1)^2}$$

To solve the optimization problem under restricted conditions, we employ the LM method to translate the conditional restrictions into unconditional constraints:

$$L(\epsilon,\lambda)=\sum_{i=1}^{l}\frac{nk_i\cdot\exp(\epsilon_i)}{(\exp(\epsilon_i)-1)^2}+\lambda\Big(\epsilon_1+\epsilon_2+\cdots+\epsilon_l-\frac{\epsilon}{2}\Big) \tag{5}$$

where $\epsilon_i\geq 0, i=1,\ldots,l$. The task now is to obtain the minimum value of $L(\epsilon,\lambda)$. Since the second-order partial derivative $\frac{\partial^2 L(\epsilon,\lambda)}{\partial^2\epsilon_i}=\frac{nk_i\exp(3\epsilon_i)+4nk_i\exp(2\epsilon_i)+nk_i\exp(\epsilon_i)}{(\exp(\epsilon_i)-1)^4}>0, i=1,2,\ldots,l,$ $L(\epsilon,\lambda)$ is strictly a convex function for the variable ϵ_i, there must exist a minimum solution for $L(\epsilon,\lambda)$. For simplicity, let $x_i=\exp(\epsilon_i)$; then, the equation $L(\epsilon,\lambda)$ becomes:

$$L(x,\lambda)=\sum_{i=1}^{l}\frac{nk_ix_i}{(x_i-1)^2}+\lambda\Big(x_1x_2\ldots x_l-\exp\left(\frac{\epsilon}{2}\right)\Big) \tag{6}$$

where $x_i>1, i=1,2,\ldots,l$. Its optimal solution is obtained by solving the following equations:

$$\begin{cases}\frac{\partial L(x,\lambda)}{\partial x_i}=-\frac{nk_i(x_i+1)}{(x_i-1)^3}+\lambda\frac{x_1\ldots x_l}{x_i}=0\\ \frac{\partial L(x,\lambda)}{\partial\lambda}=x_1,x_2\ldots x_l=\exp\left(\frac{\varepsilon}{2}\right)=0\end{cases} \tag{7}$$

where $i=1,2,\ldots,l$. We carry out a simple transformation of the equation to obtain:

$$\begin{cases}\lambda\exp\left(\frac{\varepsilon}{2}\right)(x_i-1)^3-nk_i(x_i+1)x_i=0\\ x_1,x_2\ldots x_l=\exp\left(\frac{\varepsilon}{2}\right)\end{cases} \tag{8}$$

where $i=1,2,\ldots,l$. The above equation is related to the problem of solving the univariate cubic equation. There are a variety of methods for solving the univariate cubic equation. Here, we employ the Cardano formula (CF) to solve this problem. The univariate cubic equation in Eq. (8) can be changed to:

$$\lambda\exp\left(\frac{\epsilon}{2}\right)x_i^3-\Big(3\lambda\exp\left(\frac{\epsilon}{2}\right)+nk_i\Big)x_i^2+\Big(3\lambda\exp\left(\frac{\epsilon}{2}\right)-nk_i\Big)x_i-\lambda\exp\left(\frac{\epsilon}{2}\right)=0 \tag{9}$$

we let

$$a=\lambda\exp\left(\frac{\epsilon}{2}\right), b=-(3a+nk_i), c=3a-nk_i, d=-a$$

Then, Eq. (9) can be expressed as:

$$ax_i^3+bx_i^2+cx_i+d=0 \tag{10}$$

To find the root of the equation, we let $x_i=y_i-\frac{b}{3a}$. Eq. (10) can then be changed to:

$$y_i^3+\left(\frac{c}{a}-\frac{b^2}{3a^2}\right)y_i+\left(\frac{d}{a}+\frac{2b^3}{27a^3}-\frac{bc}{3a^2}\right)=0 \tag{11}$$

We let $p=\frac{c}{a}-\frac{b^2}{3a^2}$ and $q=\frac{d}{a}+\frac{2b^3}{27a^3}-\frac{bc}{3a^2}$; thus, Eq. (11) can be expressed as:

$$y_i^3 + py_i + q = 0 \tag{12}$$

By using the CF method, we can obtain the root of Eq. (12) as follows:

$$\begin{aligned}
y_{i1} &= \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} \\
y_{i2} &= \omega\sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \omega^2\sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} \\
y_{i3} &= \omega^2\sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \omega\sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}
\end{aligned} \tag{13}$$

where $\omega = \frac{-1+\sqrt{3}i}{2}$. Thus, the roots of Eq. (10) are obtained by solving $x_{ij} = y_{ij} - \frac{b}{3a}, j = 1, 2, 3$. We take only x_{i1} as our final real root.

The finally obtained l solutions $x_1, x_2, \ldots, x_l$ are applied to equation $f(\lambda) = x_1x_2 \ldots x_l - \exp(\frac{\epsilon}{2}) = 0$. We can thus obtain a higher-order equation for λ. We employ the existed Newton-Raphson (NS) method to solve the problem of high degree with one unknown. The NS method first chooses an initial approximate value λ_0. At each iteration, let λ_k be the initial value of the next iteration, which is given as:

$$\lambda_{k+1} = \lambda_k - \frac{f(\lambda_k)}{f'(\lambda_k)} \tag{14}$$

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

Theorem 1 For multiple unbalanced categorical data, the optimal privacy budget value ϵ_i of the BRR is positively correlated with the number of items k_i. Specially, if $k_1 = k_2 = \cdots = k_l$, the allocation scheme $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_l = \frac{\epsilon}{2l}$ is optimal.

Theorem 2 *For any given number of items $\{k_1, k_2, \ldots, k_l\}$, there exists only one optimal budget allocation scheme $\epsilon^* = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_l\}$, s.t. $\epsilon_1 + \epsilon_2 + \cdots + \epsilon_l = \frac{\epsilon}{2}$, and its upper bound is $\frac{dn\exp(\frac{\epsilon}{2l})}{(\exp(\frac{\epsilon}{2l})-1)^2}$.*

When we apply the optimal privacy budget allocation scheme to the BRR, we obtain the OBRR mechanism in a high privacy regime, which greatly improves the original mechanism. The encoder algorithm of the OBRR is shown in Algorithm 1.

Algorithm 1 Optimal Binary Randomized Response

```
Input: $\epsilon$-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
Output: $\mathbf{v}' \in \{0,1\}^{k_1+\cdots+k_l}$-a sanitized bit map that satisfies local $\epsilon$-differential privacy.
initialize $d = k_1 + k_2 + \cdots + k_l$; $\mathbf{v}' = \mathbf{0} \in 0^d$;$m = 0$
for $i = 1$ to $l$ do
    if $i \neq 1$ then
        $m = m + k_{i-1}$
    end if
    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
end for
```

Optimal budget allocation for the MRR

In this section, we also employ the SE as a metric to evaluate the estimation. We assume that the parameters used in this section are the same as in the previous definition. The true histogram $\mathbf{H} = \sum\{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$. The sanitized histogram $\mathbf{H}' = \sum\{\mathbf{h}'_1, \ldots, \mathbf{h}'_n\}$, and $domain(\mathbf{H}) = domain(\mathbf{H}')$. Let $\mathbf{H}'' = \{H''_{11}, \ldots, H''_{1k_1}, H''_{21}, \ldots, H''_{2k_2}, \ldots, H''_{lk_l}\}$ denotes the unbiased estimation of $\mathbf{H}$; for each attribute, we have $H''_{ij}p_i + (n - H''_{ij})(1 - p_i) = H'_{ij}$, $j = 1, \ldots, k_i, i = 1, \ldots, l$, where $p_i = \frac{\exp(\epsilon_i)}{\exp(\epsilon_i)+k_i-1}$. We obtain:

$$H''_{ij} = \frac{H'_{ij}(\exp(\epsilon_i) + k_i - 1) - n}{\exp(\epsilon_i) - 1}$$

The SE is given as follows:

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

where H_{ij} represents the j-th item of the i-th attribute. One can use prior knowledge on $\mathbf{H}$ as a substitution; here, we assume only that it is a uniform histogram such that $H_{ij} = \frac{n}{k_i}$. Thus, our goal is to minimize the following equation:

$$L(\epsilon) = \min_{\epsilon_1 + \cdots + \epsilon_l = \epsilon}\sum_{i=1}^{l}\frac{n(k_i - 1)\cdot(2\exp(\epsilon_i) + k_i - 2)}{(\exp(\epsilon_i) - 1)^2}$$

We also employ the LM to translate the conditional restrictions into unconditional constraints and let $x_i = \exp(\epsilon_i)$; thus, we obtain:

$$L(x,\lambda) = \sum_{i=1}^{l} \frac{n(k_i-1)\cdot(2x_i+k_i-2)}{(x_i-1)^2} + \lambda(x_1x_2\ldots x_l - \exp(\epsilon)) \tag{15}$$

where $x_i > 1, i = 1, 2, \ldots, l$. Since the second-order partial derivative $\frac{\partial^2 L(x,\lambda)}{\partial^2 x_i} = \frac{n(k_i-1)(4x_i+6k_i-4)}{(x_i-1)^4} > 0, i = 1, 2, \ldots, l$ and $L(x,\lambda)$ is strictly a convex function for the variable x_i, there must exist a minimum solution for $L(x,\lambda)$. Its optimal solution is obtained by solving the following equations:

$$\begin{cases} \dfrac{\partial L(x,\lambda)}{\partial x_i} = \lambda \exp(\epsilon)(x_i-1)^3 - 2n(k_i-1)(x_i+k_i-1)x_i = 0 \\ \dfrac{\partial L(x,\lambda)}{\partial \lambda} = x_1x_2\ldots x_l - \exp(\epsilon) = 0 \end{cases} \tag{16}$$

where $i = 1, 2, \ldots, l$. We use the same CF and NS methods introduced in the last section to solve the roots of the above equation. Finally, we obtain the optimal allocation scheme $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_l\}$, and $\epsilon_1 + \epsilon_2 + \cdots + \epsilon_l = \epsilon$. To further analyze the properties of the optimal budget, we can draw the following conclusion:

Theorem 3 *For multiple unbalanced categorical data, the optimal privacy budget value* ϵ_i *of the MRR is positively correlated with the number of items* k_i*. Specifically, if* $k_1 = k_2 = \cdots = k_l$*, the allocation scheme* $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_l = \frac{\epsilon}{l}$ *is optimal.*

Theorem 4 *For any given number of items* $\{k_1, k_2, \ldots, k_l\}$*, there exists only one optimal budget allocation scheme* $\epsilon^* = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_l\}$ *s.t.* $\epsilon_1 + \epsilon_2 + \cdots + \epsilon_l = \epsilon$*, and its upper bound is* $\frac{n(d-l)(2\exp(\frac{\epsilon}{l})+\frac{d}{l}-2)}{(\exp(\frac{\epsilon}{l})-1)^2}$.

When we apply the optimal privacy budget allocation scheme to the MRR, we can obtain the OMRR mechanism in a high privacy regime, which improves the original mechanism significantly. The encoder algorithm of the OMRR is shown in Algorithm 2.

Algorithm 2 Optimal Multivariate Randomized Response

```
Input: $\epsilon$-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
Output: $\mathbf{v}' \in \{0,1\}^{k_1+\cdots+k_l}$-a sanitized bit map that satisfies local $\epsilon$-differential privacy.
initialize $d = k_1 + k_2 + \cdots + k_l$; $\mathbf{v}' = \mathbf{0} \in 0^d$;$m = 0$
for $i = 1$ to $l$ 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
```

Theoretical analysis

Convergence

When using the NS method to calculate the roots of the equation $f(\lambda) = x_1 x_2 \ldots x_l - \exp(\frac{\epsilon}{2}) = 0$, the biggest problem lies in the selection of the initial iteration values. If the initial value is far from the true solution, it is difficult for the NS method to converge. To improve the shortcomings of the over-reliance of the NS 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 NS 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 perform experiments on two data sets. The data sets are detailed in "Simulation" section. The comparison results are shown in Fig. 2. To facilitate the comparison, the error is normalized to [0, 1].

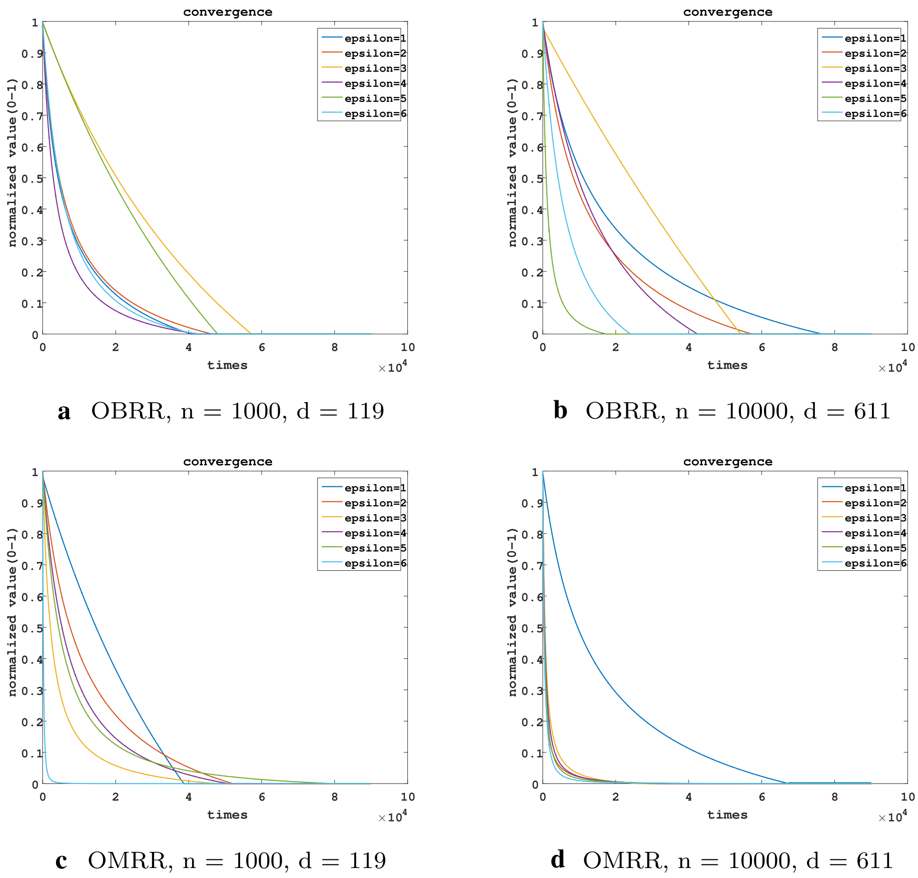
It can be seen from the experiment that the number of iterations has a great relationship with the selection of the initial value λ_0 and with the selection of the step size δ. After the improvement, all optimization equations can stably converge to the real root λ. Based on the obtained approximate solution, the optimal privacy allocation schemes can be obtained, which are illustrated in Table 2.

Analysis

To prove the effectiveness of the optimal methods, we carry out an experimental analysis of the theoretical error. Without loss of generality, we assume that there are 5 attributes and that each attribute has different categories; specifically, we let $\{k_1, k_2, \ldots, k_5\} = \{5, 6, 150, 200, 250\}$. These numbers are randomly chosen, but it does exist in reality, for example, the number of sexes is 2, while the number of websites visited by users may be in the thousands. We assume that there are 1000 participants. We conduct experiments on the BRR, MRR, OBRR, and OMRR. The experimental results are shown in Table 3. Here, we use $\log_{10}$(NSE) as the reference point. In this experiment, the OBRR has the best performance, and the MRR has the worst performance. The reason for this result is the excessive number of items. The conditional probability of the BRR satisfies $p_b = \frac{\exp(\epsilon_b)}{\exp(\epsilon_b)+1}$, but the probability of the MRR meets $p_m = \frac{\exp(\epsilon_m)}{\exp(\epsilon_m)+k-1}$; thus, we find that if k is large, the probability p_m becomes small, which will incur a bad performance. Compared to the BRR and MRR, the OBRR method can reduce the estimated square error by 40%, and the OMRR method can reduce the estimated square error by approximately 73%.

Error bounds and computational complexities

The OBRR is optimal in the high privacy regime when addressing multivariate unbalanced nominal attributes. The OMRR is optimal in the low privacy regime when addressing multivariate unbalanced nominal attributes. Thus, the estimated histograms



a OBRR, n = 1000, d = 119

b OBRR, n = 10000, d = 611

c OMRR, n = 1000, d = 119

d OMRR, n = 10000, d = 611

Fig. 2 The relationship between the number of iterations and convergence

in these mechanisms are no less favorable than the histogram estimated by the BRR [11, 37] and MRR [38]. Thus, we have:

$$\mathbf{SE}(OBRR) \leq \frac{dn \exp\left(\frac{\epsilon}{2l}\right)}{\left(\exp\left(\frac{\epsilon}{2l}\right)-1\right)^2}, \quad \mathbf{SE}(OMRR) \leq \frac{n(d-l)\left(2\exp\left(\frac{\epsilon}{l}\right)+\frac{d}{l}-2\right)}{\left(\exp\left(\frac{\epsilon}{l}\right)-1\right)^2}$$

For each participant, both the OBRR mechanism proposed in Algorithm 1 and the OMRR proposed in Algorithm 2 have a computational complexity of $O(d)$, where d is the length of the bit maps. For the aggregator, finding the optimal budget allocation scheme $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_l\}$ requires approximately $O(\log(l)F(l))$ computational complexity, where $F(l)$ is the cost of calculating $\frac{f(x)}{f'(x)}$ with l-digit precision, and estimating the histogram from the observed sanitized data requires $O(nd+n)$ time, where n is the number of participants. The OBRR and OMRR mechanisms have only linear time complexities concerning d or n, except when optimizing the budget allocation scheme. The optimal privacy allocation scheme can be calculated offline, that is, it can be calculated in advance before aggregating users' sanitized information. In short, the OBRR and OMRR mechanisms have only linear complexities with respect to the domain size $|D|$ or number of participants n for both participants and the aggregator. Hence, the OBRR and

Table 2 The optimal privacy budget allocation scheme

ϵ	2	4	6	7	100
(a) OBRR, n = 1000, d = 119					
1	0.0568	0.0716	0.0820	0.0863	0.2094
2	0.1127	0.1420	0.1626	0.1711	0.4152
3	0.1687	0.2126	0.2433	0.2562	0.6214
4	0.2248	0.2832	0.3242	0.3413	0.8277
5	0.2810	0.3541	0.4053	0.4267	1.0338
6	0.3374	0.4251	0.4866	0.5122	1.2393
(b) OBRR, n = 10,000, d = 611					
1	0.0412	0.0438	0.1281	0.1410	0.1519
2	0.0818	0.0869	0.2541	0.2797	0.3013
3	0.1224	0.1301	0.3803	0.4186	0.4509
4	0.1631	0.1733	0.5067	0.5576	0.6007
5	0.2038	0.2166	0.6331	0.6968	0.7505
6	0.2446	0.2599	0.7597	0.8360	0.9003
(c) OMRR, n = 1000, d = 119					
1	0.0436	0.0787	0.1063	0.1186	0.6564
2	0.0955	0.1711	0.2295	0.2553	1.2499
3	0.1573	0.2791	0.3715	0.4120	1.7805
4	0.2293	0.4023	0.5307	0.5862	2.2518
5	0.3109	0.5390	0.7040	0.7743	2.6719
6	0.4018	0.6872	0.8882	0.9725	3.0503
(d) OMRR, n = 1000, d = 611					
1	0.0266	0.0304	0.2644	0.3173	0.3649
2	0.0562	0.0643	0.5317	0.6309	0.7182
3	0.0899	0.1026	0.8037	0.9424	1.0618
4	0.1284	0.1464	1.0793	1.2507	1.3953
5	0.1726	0.1967	1.3571	1.5548	1.7188
6	0.2235	0.2543	1.6355	1.8541	2.0326

Table 3 The relationship between $\log_{10}$(NSE) and the privacy budget ϵ

ϵ	Mechanism										
	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6
BRR	4.7857	4.4330	4.1825	3.9879	3.8285	3.6935	3.5761	3.4723	3.3791	3.2944	3.2168
OBRR	4.5683	4.2144	3.9325	3.7488	3.5955	3.4639	3.3484	3.2454	3.1523	3.0672	2.9889
MRR	6.4056	6.0087	5.7135	5.4736	5.2686	5.0874	4.9235	4.7727	4.6320	4.4995	4.3737
OMRR	5.9710	5.5472	5.2254	4.9578	4.7310	4.5274	4.3408	4.1675	4.0048	3.8507	3.7041

OMRR mechanisms are highly efficient for multiple unbalanced categorical data aggregation.

Simulation

Optimal binary randomized response mechanism

In this section, we conduct an experiment to compare the performances of the BRR and OBRR mechanisms. We assume that each participant's secret data value is drawn from histogram *H*, which is uniformly and randomly generated during each aggregation. The dimension of the data set is $[n, d]$. The selection of the data set guarantees the following criteria: each participant can vote for only l tickets, that is, the sum of each row of the data set matrix is l. The total number of tickets for all participants is $l * n$. The data set generation algorithm is given in Algorithm 3. All of the experiments mentioned in this paper are run on a notebook with Windows 8.1, $i7 - 4710MQ$, a 2.50 GHz CPU and 8.0 GB of RAM. The coding platform is MATLAB R2015b. Without loss of generality, we assume that there are 5 attributes, and each attribute has a different number of categories. 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 $\{k_{11}, k_{12}, \ldots, k_{15}\} = \{5, 6, 150, 200, 250\}$ and $\{k_{21}, k_{22}, \ldots, k_{25}\} = \{2, 4, 6, 7, 100\}$. We conduct two experiments, one with 1000 participants and the other with 10000 participants. The privacy budget ranges from 1.0 to 6.0, and we employ the normalized square error ($NSE = \frac{SE}{n}$) as the metric to measure the performance of the mechanisms, where SE is the square error. The comparison results are shown in Fig. 3.

Algorithm 3 Simulation Data Set Generation Algorithm

```
Input: $n$-number of participants; $l$-number of attributes; $\{k_1, k_2, \cdots, k_l\}$-number of items for each
    attribute
Output: $D_{n*l}$-data set.
initialize $index = 0$; $m = 0$
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
```

The black lines denote the $\log_{10}$(NSE) of the BRR mechanism. The BRR ignores the number of categories of attributes and treats all attributes as equal, encoding each attribute with the same privacy budget $\frac{\epsilon}{2l}$. Our OBRR method takes into account the nature of all attributes and finds a more reasonable privacy budget allocation scheme, that is, it allocates more budget to attributes with more items, and then encodes each attribute using the method proposed in Algorithm 1. When $(k_1, k_2, \ldots, k_l) = (5, 6, 150, 200, 250)$, Fig. 3a, b represent the estimated errors for 1000 and 10,000 users, respectively. When $(k_1, k_2, \ldots, k_l) = (2, 4, 6, 7, 100)$, Fig. 3c, d represent the estimated errors for 1000 and 10000 users, respectively. Due to the randomness of perturbation, we perform three experiments for each case and take the average of the tests for the mapping. The error bars in the figures are calculated using the standard deviation.

As can be seen from the figure, the optimal privacy budget allocation scheme proposed by us plays an important role. According to Fig. 3a, b, the OBRR mechanism can reduce the estimated square error by 41.6% and 40.2% compared with the BRR, respectively. According to Fig. 3c, d, the OBRR mechanism can reduce the estimated square error by 33.2% and 36.4% compared with the BRR, respectively. It can be concluded

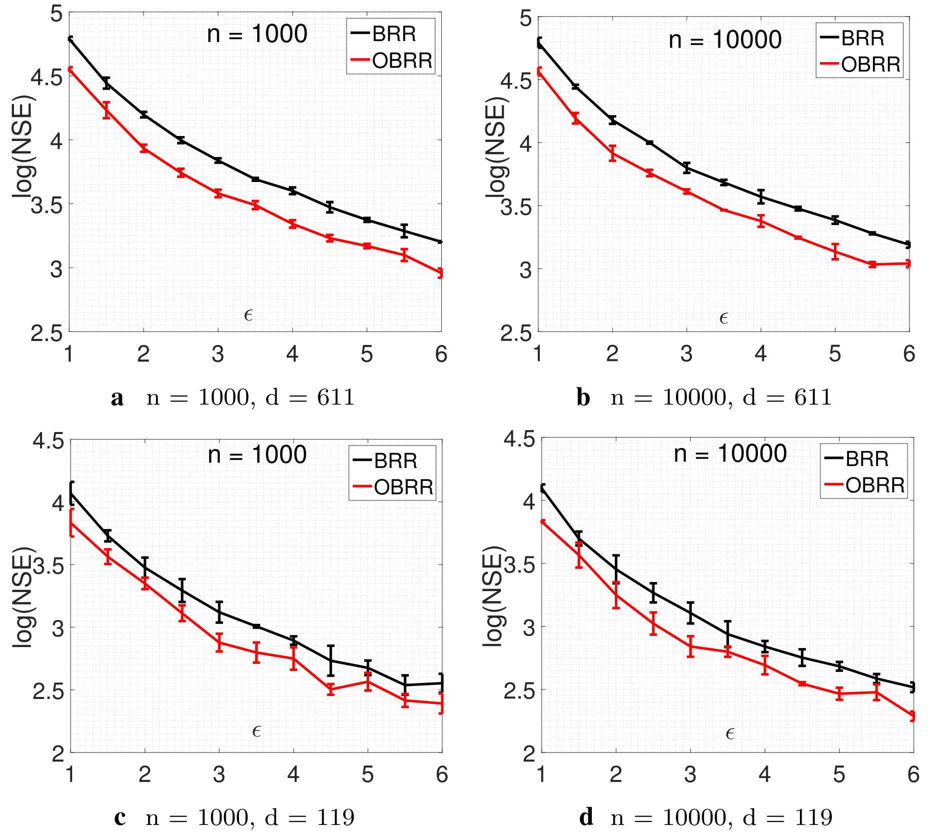


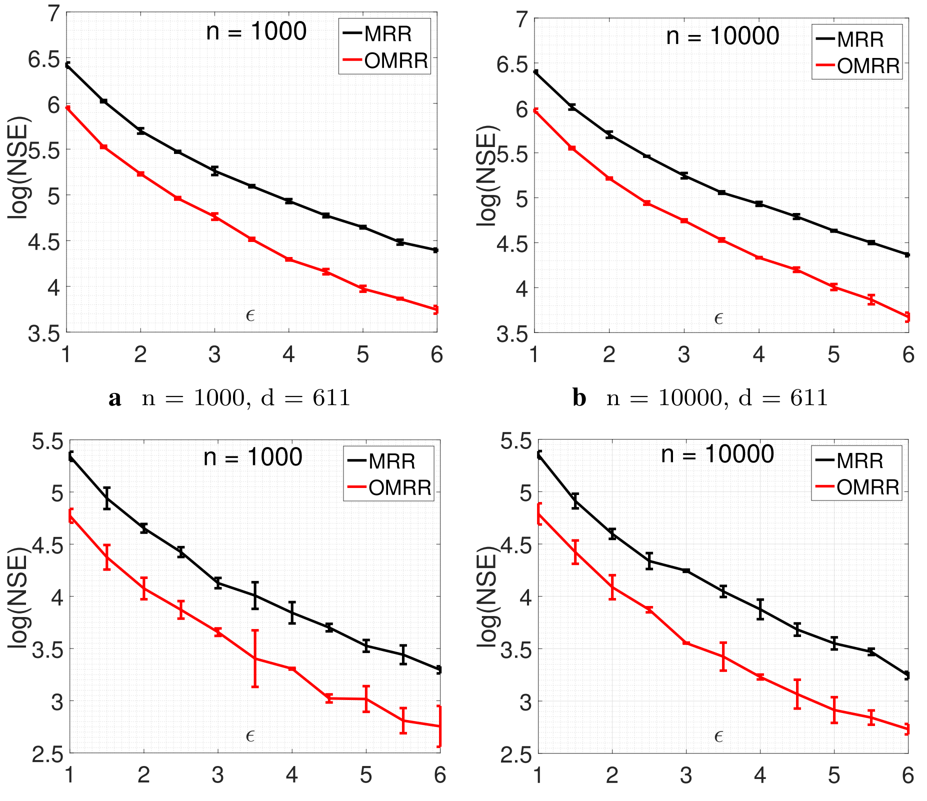
Fig. 3 The relationship between the estimated histogram error measured by $\log_{10}$ (NSE) and the privacy budget ϵ

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 values of the attributes $(k_1, k_2, \ldots, k_l)$. In fact, the larger the dimensional difference between attributes is, the better the privacy budget allocation scheme proposed in this paper will be.

Optimal multivariate randomized response mechanism

We first introduce the implementation principle of the MRR, which was introduced in "Preliminaries" section. The MRR treats all the attributes as equal and allocates the same privacy budget $\frac{\epsilon}{l}$ to each attribute. If the dimensions between attributes differ greatly, assignment of the same privacy budget is bound to result in inaccurate estimates of the results. Taking into account the drawbacks of the MRR, we allocate the privacy budget more reasonably, assigning more budget to attributes with more items.

To demonstrate the effectiveness of the OMRR, we experiment on the same data set created above. The compared results are presented in Fig. 4. The black lines denote the MRR, and the red lines indicate the OMRR. The number of participants assumed in Fig. 4a, c is 1000, while in Fig. 4b, d, it is 10,000. When the budget increases, the estimated error gradually declines. In Fig. 4a, b , the OMRR reduces the estimated square error by 72.8% and 72.0% compared with the MRR, respectively. In Fig. 4c, d, the OMRR



a n = 1000, d = 611

b n = 10000, d = 611

c n = 1000, d = 119

d n = 10000, d = 119

Fig. 4 The relationship between the estimated histogram error measured by $\log_{10}$(NSE) and the privacy budget ϵ

reduces the estimated square error by 73.0% and 73.7% compared with the MRR, respectively.

In fact, there is currently no research on privacy budget allocation schemes for unbalanced multivariate nominal attributes. The purpose of our comparison with the BRR and MRR in "Simulation" section is to prove the effectiveness of our method. Our approach is highly scalable. In the process of local differential privacy processing, as long as it involves the allocation of privacy budgets for categorically unbalanced data, it can be solved by our privacy budget allocation scheme.

Conclusion

Traditional local differential privacy techniques typically assign the same privacy budget to unbalanced multivariate nominal attributes, leading to large data utility loss and high communication costs. To solve this 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 different data for the privacy concern levels, we use the proposed optimal privacy budget allocation scheme to improve the BRR and MRR and call it the OBRR and OMRR, respectively. The OBRR and OMRR are optimal in the high and low privacy regimes with high-dimensional heterogeneous data, respectively. To prove the effectiveness of our improved local differential privacy mechanisms, we carry out simulation experiments on two different data sets with unbalanced multivariate nominal attributes. The simulation results demonstrate that the proposed mechanism can achieve a considerable improvement by reducing the estimated square error by 53.2% compared to the BRR and MRR on average.

Acknowledgements
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).

Author' contributions
XF designed and implemented the optimal allocation algorithm of the differential privacy budget and proved the effectiveness of the algorithm through experiments. She was a major contributor to the writing of the manuscript. CZ participated in the drafting of the article and made great contributions to the structure of the article. Both authors read and approved the final manuscript.

Funding
Not applicable.

Data availability statement
The data sets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Mathematics, Harbin Institute of Technology, Harbin 150001, China. [2] School of International Business, Qingdao Huanghai University, Qingdao 266427, China.

Received: 9 November 2019 Accepted: 8 May 2020
Published online: 15 June 2020

References

1. Li G, Wang J, Zheng Y, Franklin MJ (2016) Crowdsourced data management: a survey. IEEE Trans Knowl Data Eng 28(9):2296–2319
2. Aziz MMA, Sadat MN, Alhadidi D, Wang S, Jiang X, Brown CL, Mohammed N (2019) Privacy-preserving techniques of genomic data—a survey. Brief Bioinform 20(3):887–895
3. Zhu T, Li G, Zhou W, Philip SY (2017) Differentially private data publishing and analysis: a survey. IEEE Trans Knowl Data Eng 29(8):1619–1638
4. Yang X, Wang T, Ren X, Yu W (2017) Survey on improving data utility in differentially private sequential data publishing. IEEE Trans Big Data
5. Dwork C (2006) Differential privacy. In: International Colloquium on Automata, Languages, & Programming
6. Dwork C, Lei J (2009) Differential privacy and robust statistics. In: ACM symposium on theory of computing, pp 371–380
7. Smith A (2011) Privacy-preserving statistical estimation with optimal convergence rates. In: ACM symposium on theory of computing, pp 813–822
8. Gu K, Yang L, Yin B (2018) Location data record privacy protection based on differential privacy mechanism. ITC 47(4):639–654
9. Kasiviswanathan SP, Lee HK, Nissim K, Raskhodnikova S (2008) What can we learn privately? In: Proc IEEE 49th annual IEEE symp on foundations of computer science (FOCS), vol, 40, no 3, pp 793–826
10. Duchi JC, Jordan MI, Wainwright MJ (2013) Local privacy and statistical minimax rates. In: Annual IEEE symposium on foundations of computer science, pp 429–438
11. Erlingsson Ú, Korolova A, Pihur V (2014) Rappor: Randomized aggregatable privacy-preserving ordinal response. In: ACM Sigsac conference on computer and communications security, pp 1054–1067
12. Groat MM, Edwards B, Horey J, He W, Forrest S (2012) Enhancing privacy in participatory sensing applications with multidimensional data. In: 2012 IEEE international conference on pervasive computing and communications, IEEE, New York, pp 144–152
13. Sun J, Zhang R, Zhang J, Zhang Y (2016) Pristream: privacy-preserving distributed stream monitoring of thresholded percentile statistics. In: IEEE INFOCOM 2016-the 35th annual IEEE international conference on computer communications, IEEE, New York, pp 1–9
14. Chen R, Xiao Q, Zhang Y, Xu J (2015) Differentially private high-dimensional data publication via sampling-based inference. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 129–138

15. Ren X, Yu CM, Yu W, Yang S, Yang X, Mccann JA, Yu PS (2016) Lopub: high-dimensional crowdsourced data publication with local differential privacy. IEEE Trans Inf Forensics Secur PP(99):1–1
16. Ren X, Yu CM, Yu W, Yang S, Yang X, Mccann J (2017) High-dimensional crowdsourced data distribution estimation with local privacy. In: IEEE international conference on computer and information technology, pp 226–233
17. Bassily R, Smith A (2015) Local, private, efficient protocols for succinct histograms. In: Proceedings of the forty-seventh annual ACM symposium on theory of computing, pp 127–135
18. Fanti G, Pihur V, Erlingsson Ú (2015) Building a rappor with the unknown: privacy-preserving learning of associations and data dictionaries. Proc Privacy Enhanc Technol 2016(3):41–61
19. Kairouz P, Bonawitz K, Ramage D (2016) Discrete distribution estimation under local privacy. arXiv preprint arXiv:160207387
20. Chen R, Li H, Qin AK, Kasiviswanathan SP, Jin H (2016) Private spatial data aggregation in the local setting. In: IEEE international conference on data engineering, pp 289–300
21. Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc 60(309):63–69
22. Ye M, Barg A (2017) Optimal schemes for discrete distribution estimation under local differential privacy. IEEE Trans Inf Theory PP(99):1–1
23. Cai TT, Wang Y, Zhang L (2019) The cost of privacy: optimal rates of convergence for parameter estimation with differential privacy. Statistics PP(99)
24. Li H, Cui J, Lin X (2017) Improving the utility in differential private histogram publishing: theoretical study and practice. In: Proceedings of IEEE international conference on Big Data, pp 1100–1109
25. Wang N, Gu Y, Xu J, Li F, Yu G (2019) Differentially private high-dimensional data publication via grouping and truncating techniques. Front Comput Sci 13(2)
26. Cheng X, Tang P, Su S, Chen R, Wu Z, Zhu B (2019) Multi-party high-dimensional data publishing under differential privacy. IEEE Tran Knowl Data Eng 1–1
27. Kulkarni T, Cormode G, Srivastava D (2018) Marginal release under local differential privacy. In: Proceedings of the 2018 international conference on management of data, SIGMOD conference 2018, pp 131–146
28. Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X (2017) Privbayes: private data release via bayesian networks. ACM Trans Database Syst 42(4):1–41
29. Day WY, Li N (2015) differentially private publishing of high-dimensional data using sensitivity control. In: The 10th ACM symposium on information, computer and communications security, pp 451–462
30. Su S, Tang P, Cheng X, Chen R, Wu Z (2016) Differentially private multi-party high-dimensional data publishing. In: 2016 IEEE 32nd international conference on data engineering (ICDE), IEEE, New York, pp 205–216
31. Zhang J, Yang K, Xiang L, Luo Y, Xiong B, Tang Q (2013) A self-adaptive regression-based multivariate data compression scheme with error bound in wireless sensor networks. Int J Distrib Sensor Netw 9(3):68–96
32. Bloom BH (1970) Space/time trade-offs in hash coding with allowable errors. Ipsj Mag 12(7):422–426
33. Peng J, Li S, Zhu C, Liu W, Lin K (2015) A joint subcarrier selection and power allocation scheme using variational inequality in ofdm-based cognitive relay networks. Wirel Commun Mob Comput 16(8):977–991
34. Jian HR, Zhi JZ, Yi MX, Ji JY (2013) Topology optimization of finite similar periodic continuum structures based on a density exponent model. Comput Model Engineering Sci 90(3):211–231
35. Wang D, Huang L, Tang L (2017) Dissipativity and synchronization of generalized bam neural networks with multivariate discontinuous activations. IEEE Trans Neural Netw Learn Syst 29(8):3815–3827
36. Kairouz P, Oh S, Viswanath P (2014) Extremal mechanisms for local differential privacy. In: International conference on neural information processing systems, pp 2879–2887
37. Duchi JC, Jordan MI, Wainwright MJ (2013) Local privacy and statistical minimax rates. In: 2013 IEEE 54th annual symposium on foundations of computer science, IEEE, New York, pp 429–438
38. Mcsherry F, Talwar K (2007) Mechanism design via differential privacy. In: IEEE symposium on foundations of computer science, 2007. FOCS '07, IEEE, New York, pp 94–103

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.