Associated Attribute-Aware Differentially Private Data Publishing via Microaggregation

GAOMING YANG, XINXIN YE, XIANJIN FANG, RONGSHI WU, AND LI WANG
School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China
Corresponding author: Xianjin Fang (xjfang@aust.edu.cn)
This work was supported in part by the National Natural Science Foundation of China under Grant 61572034, in part by the Major Science and Technology Projects in Anhui Province under Grant 18030901025, in part by the Anhui Province University Natural Science Fund under Grant KJ2019A0109, and in part by the Opening Project of Jiangsu Key Laboratory of Data Science and Smart Software under Grant 2019DS303.

ABSTRACT Releasing raw data sets with sensitive personal information will leak privacy. Therefore, various differential privacy methods have been proposed for efficient data sharing while preserving privacy. However, they focus on noise processing of all quasi-identifier attributes, which results in high time-space complexity and low data utility. In this paper, we propose a Differential Privacy Protection model considering the Correlations between Attributes, denoted DPPCA. DPPCA first computes the degree of correlations between the quasi-identifier attributes and the sensitive attributes, and determines the pair of attributes with maximal degree of correlation. Then based the attributes with the maximal degree of correlations, it uses microaggregation to partition the data set into clusters of size \( k \) \( (k \geq 2) \) according to three types of attributes, i.e., numerical, non-numerical, and hybrid attributes, such that there are \( l \) \( (l < k) \) values of sensitive attributes in a cluster. Finally, noise is added to each cluster separately such that each cluster satisfies \( \varepsilon \)-differential privacy. While keeping the same degree of preserving privacy, our experimental results demonstrate that DPPCA substantially reduces the amount of added noise to 11% for the Census data set and the Adult data set. Therefore, DPPCA greatly improve the data utility while reaching the same degree of differential privacy.

INDEX TERMS \( \varepsilon \)-Differential privacy, mutual information, microaggregation, privacy protection data publishing, data utility.

I. INTRODUCTION

Data-sharing mechanisms make cooperation and research among various organizations convenient but substantially increase the risk of privacy disclosure [1]–[3]. However, data owners (e.g., government agencies and hospitals) must outsource data to third parties to obtain better services. If the original data are released without being processed, the privacy of individuals or organizations will be easily disclosed. For example, the Centers for Disease Control and Prevention must collect cases from various medical institutions, and these data often contain an amount of sensitive information on patients. If the Centers for Disease Control and Prevention releases this original information, then the users’ private information will be disclosed. Therefore, privacy preserving data release has become desirable. However, conventional privacy protection models, such as \( k \)-anonymity and its extension [4], have various disadvantages. For example, these models always assume that the attacker has background knowledge; however, the background knowledge that is possessed by the attacker is difficult to predetermine. In addition, the anonymized privacy protection model cannot provide an effective and rigorous proof. Once the model parameters have been changed, it is impossible to quantitatively analyze the privacy protection level, which severely weakens the reliability of privacy protection. Hence, differential privacy was proposed.

Differential privacy is a research hotspot [5], [6], and it can overcome the disadvantages of conventional privacy protection without considering the background knowledge that is possessed by the attacker. However, differential privacy requires the addition of a substantial amount of noise to the query results, thereby resulting in a limited application scope. Typically, the shortcomings are manifested in two ways: the
noise that is added to the output is substantially increased, and the data utility is mainly a limited to the release of the query results [7]–[9]. It is challenging to prevent the disclosure of private data and to improve the utility of published data. To solve these problems, the noninteractive protection framework [10] converts or compresses the original data and subsequently adds noise to the query results to satisfy $\varepsilon$-differential privacy. Unlike conventional $\varepsilon$-differential privacy methods, the microaggregation methods [11], [12] aggregate the original records into clusters of size at least $k$ ($k \geq 2$) items and add noise to each cluster. For example, Soria-Comas et al. [12] proposed an insensitive microaggregation algorithm for increasing the within-cluster homogeneity. They considered the sensitivity of the input datasets to changes in only one record. Since much less noise is added to each cluster than to each record, the added noise will be smaller. To the best of our knowledge, the available methods focuses on all quasi-identifiers when performing microaggregation operations, and they assume that the data attributes are independent or fully correlated instead of considering partially correlated attributes in the input data set, which is likely to lead to a reduction in data utility while protecting privacy.

To balance data utility and protect sensitive information, we use microaggregation to realize differential privacy. Our method does not process all quasi-identifiers, and we process only the attributes that are most relevant to sensitive attributes, while other quasi-identifier attributes are anonymized. The objective is to reduce the loss of data utility. We calculate the mutual information between quasi-identifiers attributes and sensitive attributes. The greater the mutual information is, the higher the correlation. Hence, we identify the most dependent attribute pairs and microaggregate them. Then, we ensure that they satisfy differential privacy. The main contributions are as follow.

We proposed a new DPPCA model for differentially private data publishing, which were designed to reduce the influence of added noise. The model microaggregates an attribute pairs with the largest dependencies and makes the resulting cluster satisfy differential privacy. Since the amount of noise depends on the sensitivity of the data set. We add noise to a cluster much less than to each record in the dataset. To make DPPCA adapt to various types of data and improve its generalization performance, we have discussed numerical attributes, categorical attributes, and hybrid attributes. And we design suitable algorithms for each data type. In addition, we analyze each algorithm theoretically and calculate its time complexity and space complexity. We use the squared error sum (SSE) and the record linkage (RL) to evaluate the utility of the proposed model, that is, we select numerical attributes, categorical attributes, and hybrid attributes from the Census and Adult datasets for the evaluation of SSE and RL, respectively. In addition, we elaborated on the explanations for each experimental result.

This paper is organized as follows, some related works are introduced in Section 2, and the proposed microaggregation-based differential privacy protection model DPPCA are summarized in Section 3 and Section 4. Section 5 validates the effectiveness of the proposed model DPPCA. The paper ends with conclusion in Section 6.

II. RELATED WORK

The $\varepsilon$-differential privacy method does not impose any assumption regarding the attacker’s background knowledge. There are two scenarios for implementing differential privacy: interactive scenarios and noninteractive scenarios. For interactive scenarios, data analysis is limited because it enables only a limited number of queries to be answered; for noninteractive data publishing, all queries can be answered at once. According to the characteristics of interactive and non-interactive scenarios, noninteractive scenarios provide higher data usage flexibility [13]. The objectives are to approximate the data distribution by partitioning the data threshold and to calculate the number of records in each partitioned set. The accuracy of a histogram query depends mainly on the number of records that are included. If the number of records is small, the relative error is large, i.e., the accuracy of the histogram query cannot be guaranteed. To overcome this problem, Zhang et al. [14] considered the data utility of differential privacy and constructed a Bayesian network to the data prior to adding noise; in addition, they theoretically analyzed the privacy and provided a utility guarantee. Cormode et al. [15], [16] analyzed in detail the connection, advantages and disadvantages of using microaggregation to realize differential privacy. In short, the methods for improving the utility of differentially private data publishing differ in terms of their limitations.

Compared with conventional differential privacy methods, the microaggregation methods can improve the data utility and do not impose excessive restrictions on the query type. For example, Soria-Comas et al. [11] proposed the synergy between differential privacy and $k$-anonymity when publishing anonymous data. However, their method was not sufficient for considering the utility of numerical attributes. Martínez et al. [17] studied categorical attributes and proposed a semantic adaptive maximum distance to average record algorithm. However, this algorithm is not efficient since the size of each cluster cannot be determined in advance, which causes a large information loss during the data conversion of clusters. To reduce the information loss of microaggregation, Soria-Comas et al. [12] proposed the use of an insensitive microaggregation method to obtain an anonymized data set. They add random noise to every quasi-identifier attribute except sensitive attributes. Sánchez et al. [18] used the microaggregation of individual rankings to reduce the scale parameters of added noise in most cases. Masooma [19] proposed a stable microaggregation framework that incorporates microaggregation and differential privacy into the data dissemination process. In addition to this work, many researchers have proposed their own solutions in terms of the large-scale graphic [20] data release and privacy preserving big data publishing [21].
According to the above analysis, the available microaggregation-based differentially private data publishing models do not consider the association between quasi-identifier attributes and sensitive attributes. This paper proposes the realization of a differentially private data release via microaggregation by using the association attributes, namely, we investigate how to effectively reduce the sensitivity of the dataset and improve the signal-to-noise ratio while realizing differential privacy.

III. PRELIMINARIES

Prior to presenting our model formally, we briefly define the symbols that are used in this article. Suppose there are \( n \) records in the original data set \( D \) and the data attributes consist of the identifier (ID) attribute, quasi-identifiers (QI) attributes, and sensitive (SA) attributes. The identifier attribute is removed before the data are released; hence, we do not discuss it here. The QI attributes and SA attributes are denoted \( \{ A_i \}_{i=1, \ldots, m} \) and \( \{ S_j \}_{j=1, \ldots, s} \), respectively, where \( m \) and \( s \) are the numbers of QI attributes and SA attributes, respectively. For simplicity, we only discuss a single sensitive attribute, namely, \( \{ S_j \} \) with \( j = 1 \). For multisensitive attributes, the attribute can be generalized according to the processing method of the QI attribute, which is not described in detail here.

A. \( \varepsilon \)-DIFFERENTIAL PRIVACY

The \( \varepsilon \)-differential privacy model does not make any assumptions regarding the attacker’s background knowledge and has a rigorous mathematical basis. Under this model, the presence or absence of a single record has no effect on the dataset. Thus, this model has attracted the attention of many researchers in computer science and statistics. The definition of \( \varepsilon \)-differential privacy is as follows:

**Definition 1 (\( \varepsilon \)-Differential Privacy):** For any data sets \( D_1 \) and \( D_2 \) that differ by one record, a randomized algorithm \( A \) provides \( \varepsilon \)-differential privacy protection if and only if the following is satisfied:

\[
\Pr[A(D_1) = S] \leq \exp(\varepsilon) \times \Pr[A(D_2) = S] \tag{1}
\]

where \( S \) is the output threshold of algorithm \( A \). Definition 1 implies that even if the attacker knows most of the records in the original data set, she/he still cannot accurately determine whether a record is in \( D_1 \) or \( D_2 \). For realizing differential privacy, Dwork et al. [22] proposed the Laplace mechanism: for a query function \( f \), the differential privacy mechanism generates the actual result as middleware and adds Laplacian noise to the response of the query. The amount of added noise depends on the sensitivity, which is the maximum change in the query result after the removal of a record from the dataset or the addition of a record to the dataset, i.e., the added noise level is closely related to the global sensitivity.

**Definition 2 (\( L_1 \)-Sensitivity [23]):** For any two adjacent data sets \( D_1 \) and \( D_2 \), there is a function \( f: D \rightarrow \mathbb{R}^d \). The \( L_1 \)-sensitivity of the function \( f \) is defined as follows:

\[
\Delta f = \max_{D_1, D_2} ||f(D_1) - f(D_2)|| \tag{2}
\]

where \( R \) represents the mapped real space and \( d \) represents the query dimension of the function \( f \).

B. MICROAGGREGATION

To effectively reduce the amount of added noise when publishing differentially private data, we microaggregate the original data, namely, we divide the data set into several clusters, where each cluster contains at least \( k \) records. The data in each cluster are the most similar, and the data that differ in terms of their clusters are the most different. Microaggregation can be realized via two steps:

1) Partitioning: The original dataset is divided into clusters of size at least \( k \), and the records in each cluster are as similar as possible to each other.

2) Aggregation: After the data set has been divided into clusters, each original cluster will be replaced by the mean or median of the cluster, i.e., every record can be replaced by a representative record in this cluster.

C. MUTUAL INFORMATION

In probability theory or information theory, the mutual information is a measure of the mutual dependence between two random variables. The mutual information between two discrete random variables \( X \) and \( Y \) can be defined as follows:

\[
I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p(x) \cdot p(y)} \right) \tag{3}
\]

where \( p(x, y) \) is the joint probability distribution function of \( X \) and \( Y \), and \( p(x) \) and \( p(y) \) are the marginal probability distribution functions of \( X \) and \( Y \), respectively. For continuous random variables:

\[
I(X, Y) = \int_{x \in X} \int_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x) \cdot p(y)} \right) \tag{4}
\]

where \( p(x, y) \) is the joint probability density function of \( X \) and \( Y \), and \( p(x) \) and \( p(y) \) are the marginal probability density functions of \( X \) and \( Y \), respectively. The larger the mutual information value is, the higher the correlation. For the QI attribute \( Q_i \) and the sensitive attribute \( S \) in dataset \( D \), the mutual information is:

\[
I(Q_i, S) = \sum_{j=1}^{d_k} \sum_{k=1}^{d_i} \Pr(\varphi_i^j, \varphi_s^k) \log \left( \frac{\Pr(\varphi_i^j, \varphi_s^k)}{\Pr(\varphi_i^j) \Pr(\varphi_s^k)} \right) \tag{5}
\]

The range of category attributes is \( \varphi_x = \{ \varphi_x^1, \ldots, \varphi_x^{d_x}, \ldots, \varphi_x^{d_x} \} \), \( \Pr(\varphi_i^j) \) is the probability that each category attribute value is on the category attribute field, and \( \Pr(\varphi_i^j, \varphi_s^k) \) is the joint probability of the category attribute. \( d_i \) and \( d_s \) represent the number of attribute values of QI and SA, respectively.
IV. ATTRIBUTE-RELATED DIFFERENTIALLY PRIVATE DATA PUBLISHING ALGORITHM

To reduce the amount of added noise and to improve the utility of data publishing, we determine the correlation between the attribute pairs of the original dataset and add noise to each cluster after microaggregation.

A. DIFFERENTIALLY PRIVATE DATA PUBLISHING FOR NUMERICAL ATTRIBUTES

The dataset contains numerical attributes and categorical attributes, and each type of attribute is processed differently. We will discuss them separately below. This section mainly discusses the microaggregation of numerical attributes and the process of implementing differential privacy.

1) MICROAGGREGATION OF NUMERICAL ATTRIBUTES

There are many scenarios for microaggregation applications in which data is divided. For example, in [12], the microaggregation method is used to realize differential privacy that is based on k-anonymity. However, the microaggregation method is not sufficient for protecting the dataset directly. After comprehensively comparing these microaggregation methods, we decided to modify the MDAV [24] to realize differentially private data release. In Algorithm 1, we present an improved MDAV method for numerical attributes.

Since the sensitivity has a substantial impact on the results of microaggregation and the general MDAV does not require sensitivity between the attributes in the cluster, we use the nonsensitive microaggregation algorithm [12], for which deleting or adding a record does not substantially impact the entire cluster. For example, consider a dataset \( D \) that contains two attributes and 15 records. Figure 1 presents the original \( D \) and the insensitive MDAV clustering result after the modification of a record, where the cluster size is \( k = 5 \). and '×' represents the cluster core. According to Figure 1, changing only one record by applying insensitive microaggregation does not affect the entire clustering result. Such clustering has lower sensitivity and can also reduce the privacy budget of subsequent noise additions.

2) DIFFERENTIALLY PRIVATE PROTECTION OF NUMERICAL ATTRIBUTES

Suppose \( l(I,r) \) is a function that returns the attribute value of the \( r \)-th record that corresponds to \( D \). To realize differential privacy, we perform a nonsensitive microaggregation operation on the original dataset \( D \) to obtain a dataset \( D' \), and then we add Laplacian noise to \( D' \) to obtain a differentially private dataset \( D'' \).

**Theorem 1:** For any function of \( f : \rightarrow R^l \), if the output of algorithm \( A \) satisfies the following equation, then \( A \) satisfies \( \varepsilon \)-differential privacy.

\[
A(D) = f(D) + \langle Lap_1(\Delta f/\varepsilon), \ldots, Lap_d(\Delta f/\varepsilon) \rangle \quad (6)
\]

where \( Lap_i(\Delta f/\varepsilon)(1 \leq i \leq d) \) are independent Laplace variables.

**Proof:** From the Laplace distribution function, it follows that

\[
\frac{P[M(D) = S]}{P[M(D') = S]} = \frac{\frac{\varepsilon}{\Delta f} \exp(\frac{-f(D)}{\Delta f})}{\frac{\varepsilon}{\Delta f} \exp(\frac{-f(D')}{\Delta f})} = \frac{\varepsilon}{\Delta f} (|f(D)| - |f(D')|) \leq \varepsilon \quad (7)
\]
It is necessary to make \( |f(D) - f(D')| \leq \Delta f \), to satisfy \( |f(D) - f(D')| \leq \max_{D,D'} |f(D) - f(D')| \). This formula satisfies the absolute value inequality, so \( \text{Lap}(\Delta f / \epsilon) \) satisfies the definition of differential privacy. This concludes the proof.

For the attribute \( Q_{\max} \) that has the highest correlation with SA, we must calculate \( \Delta I_r(D) = [\text{Maximum attribute threshold} - \text{Minimum attribute threshold}] \). Here, the sensitivity is \( \Delta f = \frac{1}{n} \times \frac{q(x)}{k} \), \( n \) is the number of records in the data set, \( k \) is the size of the cluster, and a probability \( p(x) \in (0, 1) \) is randomly generated. From \( p(x) = \frac{1}{2\pi} \exp(-\frac{|x|}{2}) \), we can know that \( x = -\ln(2\pi x(x)) \) is the noise to be added. The amount of noise is proportional to \( \Delta f \) and inversely proportional to \( \epsilon \).

**B. CATEGORICAL ATTRIBUTE DIFFERENTIALLY PRIVATE DATA PUBLISHING METHOD**

For categorical attributes, we cannot add Laplacian noise to the numerical attributes; instead, we use an exponential mechanism [25]. The exponential mechanism must traverse the entire data domain, and we finally identify a dataset that satisfies differential privacy. The dataset responds to all queries regarding this concept category with the specified accuracy under the premise of privacy protection.

1) MICROAGGREGATION OF CATEGORICAL ATTRIBUTES

Considering the utility of anonymous data, we use the semantic distance to measure categorical attributes. Formally, given the distance function \( d \) of the data space, the centroid of a set \( x_1, x_2, \ldots, x_n \) can be defined as:

\[
\text{centroid}(x_1, x_2, \ldots, x_n) = \arg\min_c \left\{ \sum_{i=1}^{n} d(c, x_i) \right\}
\]

(8)

where \( x_1, x_2, \ldots, x_n \) are clusters that consist of \( n \) tuples and \( c \) is a centroid candidate set. In this paper, we use classification [26] to estimate the similarity between two category attribute values (such as \( a \) and \( b \)) [25]. The shortest path length between the connected elements is calculated via the taxonomic relationship.

\[
dis(a, b) = \log_2(1 + \frac{|T(a) \cup T(b)| - |T(a) \cap T(b)|}{|T(a) \cup T(b)|})
\]

(9)

Here, \( T(i) \) is a set of taxonomic subsets of element \( i \), including itself. The longer the path is, the larger the distance of the attribute-value pair. Let \( (a, b) = l_1, \ldots, l_k \) be the path that connects attribute values \( a \) and \( b \); [path\(a, b\)] = \( k \) is the length of this path. Considering all possible paths from \( a \) to \( b \), we use \( \text{dis}_{\text{sem}}(a, b) \) to represent the distance of the element pair \( (a, b) \), which can be defined as:

\[
\text{dis}_{\text{sem}}(a, b) = \min_{l} |\text{path}(a, b)|
\]

(10)

**Definition 3 (Weighted Distance of Categorical Attributes):** Consider a centroid candidate set \( c_j \) and single-variable attributes \( V = \{v_1, o_1\}, \ldots, \{v_n, o_n\} \), where \( v_i \) is an attribute value tag and \( o_i \) is the number of times it appears in the dataset. The weighted distance of the categorical attributes \( sd_o \) from all elements in \( V \) is defined as the sum of the distances between \( c_j \) and each \( v_i \in V \), weighted by the frequency \( o_i \) of \( v_i \) in the dataset, which is expressed as

\[
sd_o = \sum_{i=1}^{n} (o_i \times \text{dis}_{\text{sem}}(c_j, v_i)), v_i \in V
\]

(11)

An input dataset with univariate attributes can be expressed as \( V = \{v_1, o_1\}, \ldots, \{v_n, o_n\} \). The steps are as follows: (I) map the values in \( V \) to the concepts in taxonomic tree \( O \); (II) extract semantic-related concepts from \( O \); (III) analyze the taxonomic relationships; and (IV) retrieve new concepts (taxonomic ancestors), which become the centroid candidates for the values in \( V \). An example of solving for the distance between attribute values is presented as follows.

**Example 2:** Consider a univariate patient dataset in which each record corresponds to an individual. The following set of attribute values can be extracted: \( \{\{\text{colic}, 1\}, \{\text{lumbago}, 3\}, \{\text{appendicitis}, 1\}, \{\text{pain}, 1\}, \{\text{migraine}, 2\}, \{\text{gastritis}, 1\}\} \). These values are mapped to concepts that are found in the dictionary of WordNet, and the minimum hierarchical tree is extracted. Fig. 2 illustrates the hierarchical tree structure [27]. The process of constructing the centroid and microaggregating the categorical attributes is presented in Algorithm 2. We first extract a set of attribute values from \( V \), where \( V = \{v_1, o_1\}, \ldots, \{v_n, o_n\} \), in which \( v_i \) denotes the attribute tag value and \( o_i \) denotes the number of occurrences of \( v_i \), and then we map the attribute values to the WordNet model and extract the minimum hierarchical structure \( H_{WN} \).

2) DIFFERENTIAL PRIVACY PROTECTION OF CATEGORICAL ATTRIBUTES

For categorical attributes, we add exponential noise to realize differential private data release. The exponential mechanism is defined as follows.

**Theorem 2** [28]: Given a scoring function \( u : (O \times O) \rightarrow R \), if algorithm \( A \) satisfies the following equation, then \( A \) satisfies \( \epsilon \)-differential privacy:

\[
A(D, u) = \{r | \forall (r' \in O) \rightarrow \exp(-\frac{e^u(D, r)}{2\Delta u}) \}
\]

(12)

where \( \Delta u = \max_{D_1,D_2} |u(D_1, r) - u(D_2, r) | \) is the global sensitivity of the score function \( u(\cdot) \). According to this formula, the higher the score is, the higher the probability of being selected for output. A method for realizing differential
privacy is to select the centroid probabilistically that is proportional to the expected level \( \epsilon \) of differential privacy. Consider a function with a discrete output \( t \). The exponential mechanism selects a near-optimal output that is based on the input data \( D \) and the availability function \( q(D, t) \) while preserving privacy. Each output is associated with a selection probability \( \Pr(t) \), which increases with the quality criterion exponent, which is defined as follows:

\[
\Pr(t) \propto \exp\left(\frac{\epsilon q(D, t)}{2\Delta q}\right) \tag{13}
\]

Via this approach, the exponential mechanism will be more likely to choose the best output. The detailed process is presented in Algorithm 3.

Proof: The result is output with a probability, rather than directly adding noise to the numerical data, so it needs to be normalized and expressed in as a probability.

\[
\frac{P[M(x, q, T)]}{P[M(x', q, T)]} = \frac{\exp(\epsilon q(x, t)/2\Delta q)}{\sum_{t} \exp(\epsilon q(x, t)/2\Delta q)} = \frac{\exp(\epsilon q(x', t)/2\Delta q)}{\sum_{t} \exp(\epsilon q(x', t)/2\Delta q)}
\]

From the first term of the above formula

\[
\exp\left(\frac{\epsilon}{2\Delta q} (q(x, t) − q(x', t))\right) < \exp\left(\frac{\epsilon}{2}\right) \tag{14}
\]

The second term of formula the above formula

\[
\frac{\sum_{t} \exp(\epsilon q(x', t)/2\Delta q)}{\sum_{t} \exp(\epsilon q(x, t)/2\Delta q)}\]

Due to \( \sum_{t} \exp(\epsilon q(x, t)/2\Delta q)\leq \exp(\frac{\epsilon}{2}) \) from the above formulas, we can obtain

\[
\frac{P[M(x, q, T)]}{P[M(x', q, T)]} \leq \exp(\epsilon) \tag{15}
\]

Hence, Theorem 2 follows also in this case.

Algorithm 3 \( \epsilon \)-Differential Privacy of Categorical Attributes

**Input:** Original dataset \( D \) with \( n \) records, clusters \( c_i \), \( |c_i| \geq k \), \( i = 1, \ldots, n \)

**Output:** dataset \( D' \) with \( \epsilon \)-differential privacy

1. for each record of \( C_i \) do
2. Let \( A_j \) be the attribute of cluster \( C_i \) and \( |A_j| = t \) be the number of attribute values;
3. Count the frequency of occurrence of each \( A_j \);
4. Apply the quality standard function \( q(\ast, \ast) \) for each centroid candidate \( a_j' \) in \( \tau(A_j) \), where \( \tau(A_j) \) is the domain of \( A_j \);
5. Calculate the attribute value after noise has been added according to formula (13), and extract the value of \( S(A_i) \) randomly;
6. The calculated probability value of each attribute is denoted \( p(A_j') \) and satisfies argmin\( \sum_{j=1}^{t} p(A_j') > \Pr(a_j') \);
7. \( \Pr(a_j') \propto \exp\left(\frac{\epsilon x q(S(A_j), a_j')}{2\Delta q(q(A_j))}\right) \) satisfies the exponential mechanism;
8. end for
9. return \( D' \)

C. DIFFERENTIALLY PRIVATE DATA RELEASE FOR HYBRID ATTRIBUTES

In general, in addition to data sets with pure numeric attributes, data sets with hybrid class attributes are more common. Such data sets contain both numeric and non-numeric attributes. Before these data sets that include the hybrid attributes are released, noise must be added to satisfy \( \epsilon \)-differential privacy. For a dataset \( D \) with hybrid attributes [24], the record distance can be defined by combining the respective attribute values:

\[
d(x_1, x_2) = \sqrt{\left(\text{dist}(a_{11}, a_{21})\right)^2 + \ldots + \left(\text{dist}(a_{1m}, a_{2m})\right)^2}
\]

where \( a_{ij} \), \( i = 1, \ldots, m \), is the coordinate of \( x_j (j = 1, 2) \);

\( \text{dist}(a_{i1}, a_{i2}) \) is the distance between the values of the \( i \)-th
attribute $A_i$ in $x_1$ and $x_2$; and $\text{dist}(d'_i, d'_t)$ is the distance between the domain $\text{Dom}(A_i)$ boundaries that are used to eliminate the impact of the attribute size.

Similarly, we can use the weighted distance to microaggregate the attributes. During hybrid attribute clustering, we make a judgment regarding the attribute type and use Algorithm 1 or Algorithm 2 based on the judgment result. In summary, noises is added to the attribute types so that each attribute satisfies Definition 1. With the calculation method of the hybrid attributes, we can easily formulate a hybrid attribute differential privacy algorithm according to Algorithm 1 or Algorithm 2.

### D. TIME COMPLEXITY ANALYSIS

The time complexity of the algorithm reflects the magnitude of the increase in the program execution time with the increase in the input scale, which can be used to measure the advantages and disadvantages of the algorithm. For the above algorithm, the input scales are all $n$. For the algorithms of the numerical and categorical attributes, calculating the mutual information, the time complexities of Algorithm 2 and Algorithm 3 are $O(n^2)$.

## V. EXPERIMENTAL EVALUATION

### A. EXPERIMENTAL SETUP

To evaluate the performance of the model, we use the Census and Adult data sets. The Census data set contains 1080 records with only numeric attributes, and the Adult data set contains 30162 records. For the Census data set, we select the Fica, Fedtax, Intval and Pothval attributes as the QI attributes and Ptotval as the SA attribute. For the Adult data set, we select categorical attributes Education-Level and Native_Country as the QI attributes and Occupation as the SA attribute, and these attributes are denoted by Adult-Categorical. To evaluate the performance of our method on hybrid data sets, we select Occupation, Native_Country, Age, and Hours_per_week as the QI attributes and Occupation as the SA attribute from the Adult data sets, and these attributes are denoted Adult-Hybrid. Considering that attributes with a large value difference will have a greater impact on the results, we use $(0,1)$ normalization to normalize the attributes.
The experimental environment is an Intel Core 2.20 GHz CPU with 8 GB of RAM, the operating system is Windows 10 Ultimate, and the programming language is Java.

**B. EXPERIMENTAL METRICS**
We use the squared error sum (SSE) and the record linkage (RL) to measure the data utility and the leakage risk for the models. Here, SSE is defined as the squared sum of the attribute distances between the original data set $X$ and its anonymized set $X'$, which is expressed as follows:

$$SSE = \sum_{x_j \in X} \sum_{a_i \in x_j} (dist(a_i, (a_i'))^2$$

where $a_i'$ is the $i$th attribute value of the $j$th original record and $(a_i')'$ is the perturbed version. RL is the percentage of the original dataset that correctly matches the anonymous dataset $X'$, which is expressed as follows:

$$RL = 100 \times \frac{\sum_{x_j \in X} Pr(x_j')}{n}$$

where $n$ is the number of original records. RL measures the actual privacy from the natural perspective of privacy attacks.

The probability $Pr(x_j')$ is calculated as follows:

$$Pr(x_j') = \begin{cases} 
0, & x_j \notin G \\
\frac{1}{|G|}, & x_j \in G
\end{cases}$$

where $G$ is the set of original records with the minimum distance from $x_j'$. If the original record $x_j$ is in $G$, $Pr(x_j')$ is $1/|G|$; otherwise, $Pr(x_j')$ is 0.

**C. RESULTS AND ANALYSIS**
Before the experiment, we preprocessed and normalized the entire data set. To eliminate the effects of random errors, we execute the algorithm 10 times and average the results. DPPCA represents our model, and MDAV_QI_DP represents the MDAV [12] in which noise is added to satisfy $\epsilon$-differential privacy. We first apply Algorithm 1 and Algorithm 2, and add noise that satisfies $\epsilon$-differential privacy for numerical attributes and categorical attributes, respectively. Figure 3 presents SSE and RL for the numerical attributes in the Census data set as functions of parameters $k$ and $\epsilon$. Figure 3(a), (b) compares the DPPCA model with the general MDAV model. Figure 3(c) and (d) show the
performance of the DPPCA with conservational $\varepsilon$-differential privacy. According to Figure 3(a), the overall data utility ranking is as follows: MDAV > DPPCA > MDAV_QI_DP. However, the MDAV algorithm is only a clustering algorithm without noise, and its privacy protection is limited. For $k \approx \sqrt{n}$ or larger, SSE gradually decreases until it reaches a steady state, i.e., information loss is minimized. According to Figure 3(b), RL of the MDAV model fluctuates substantially, and the likelihood of identity leakage is high. As $k$ increases, the RL of DPPCA remains stable at approximately 0.1, i.e., the RL value of DPPCA is relatively small, and the privacy of anonymous output is improved. According to Figure 3(c), the SSE of DPPCA decreases regardless of the value of $\varepsilon$. When $\varepsilon$ is 10, the SSE is minimal, and it remains stable after $k > 33$. From Figure 3(d), the value of the overall RL is lower than 0.6%, although the RL of DPPCA fluctuates substantially with the increase in the $k$ value, thereby indicating that the risk of privacy leakage of DPPCA is low. In summary, DPPCA improves the data availability on the Census data set while demonstrating satisfactory data privacy. Even though the prior microaggregation step also causes a loss of information, the loss is small compared to the noise to be added, and it negligibly affects the threshold of the effect of the DPPCA display.

After evaluating the performance of our proposed model on numerical attributes, we evaluate the classification attributes. Figure 4(a) and (b) compare DPPCA with the general microaggregation algorithm MDAV. Figure 4(c) and (d) compare the DPPCA with the conservational $\varepsilon$-differential privacy. According to Figure 4(a), the MDAV fluctuates substantially since it does not apply to nonnumeric attributes. The DPPCA gradually decreases until it stabilizes. In Figure 4(a), regardless of the value of $\varepsilon$, the SSE of DPPCA is significantly lower than those of MDAV_QI_DP and MDAV; hence, the information loss of DPPCA is small, and it is dominant in the leakage of privacy risks. Figure 4(b) shows that the percentages of connection records of the MDAV are very high when $\varepsilon = 0.01$ and 0.1, and the RL of DPPCA is slightly larger. When $\varepsilon = 1$ and 10, the RL values of the two models are substantially reduced. According to Figure 4(c), the smaller the value of $\varepsilon$ is, the larger the SSE, i.e., the larger the information loss. According to Figure 4(d), the larger the value of $\varepsilon$ is, the larger the value of RL. In summary, the smaller the value of RL, the lower the risk of
identity leakage and the higher the privacy of anonymous output. From this experiment, the balance between data utility and privacy security can be satisfied when $\epsilon = 1$. DPPCA outperforms the conventional differential privacy method in terms of information loss and leakage risk, and the model design is more reasonable.

After evaluating the performance of our model on numerical and categorical attributes, we evaluate the performance of our model on hybrid attributes. Figure 5(a) and (b) compare DPPCA with MDAV. Figure 5(c) and (d) compare DPPCA with the conservational $\epsilon$-differential privacy. According to Figure 5(a), the SSE will decrease as $k$ increases. The DPPCA and MDAV_QI_DP methods have similar SSE curves. When $\epsilon = 1$ and $k > 174$, the SSE is relatively stable and relatively small. When $\epsilon = 10$ and $k > 66$, the SSE remains unchanged. For Figure 5(b), the RL for DPPCA is very close to zero. The privacy performance of the model is not easy to evaluate. Since the hybrid attributes differ in terms of their characteristics and their sensitivities to the added noise, the changes in the curves are not readily observed. In summary, we must consider the balance of data utility and privacy security. Some of the curves in the figure are from Figure 5(c) and (d). As the value of $k$ increases, the values of SSE and RL of DPPCA remain approximately constant and close to zero. For a hybrid attribute, for DPPCA or conservational $\epsilon$-differential privacy, the $k$ value and the $\epsilon$ value minimally affect the data; hence, the changes in their values are not large.

VI. CONCLUSION

We propose a model for publishing private data using microaggregated association attribute differences, referred to as DPPCA. This model fully considers the dependency relations of Quasi-Identifier attributes and sensitive attributes. The microaggregation operations, which can reduce the data sensitivity, were performed on these relevant attribute pairs to make each cluster contain more than $k(k > 2)$ items. By considering only the most relevant attribute pairs, the time and space complexity of the algorithm can be reduced. Then, we add noise to each cluster to realize $\epsilon$-differentially private data release. The experimental results demonstrate that the model can substantially improve data utility and reduce the amount of noise that is added to the dataset. However, many challenging problems remain to be solved. For example, in the processing of categorical attributes and hybrid attributes, the selection of a suitable value of parameter $\epsilon$ for multiple sensitive attributes requires further research.

REFERENCES

[1] J. W. Kim, D.-H. Kim, and B. Jang, “Application of local differential privacy to collection of indoor positioning data,” IEEE Access, vol. 6, pp. 4276–4286, Jan. 2018.
[2] L. Ni, C. Li, X. Wang, H. Jiang, and J. Yu, “DP-MCDBSCAN: Differential privacy preserving multi-core DBSCAN clustering for network user data,” IEEE Access, vol. 6, pp. 21053–21063, Apr. 2018.
[3] M. Yang, T. Zhu, B. Liu, Y. Xiang, and W. Zhou, “Machine learning differential privacy with multifunctional aggregation in a fog computing architecture,” IEEE Access, vol. 6, pp. 17119–17129, Mar. 2018.
[4] 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.
[5] J. Parra-Arnau, J. Domingo-Ferrer, and J. Soria-Comas, “Differentially private data publishing via cross-moment microaggregation,” Inf. Fusion, vol. 53, pp. 269–288, Jan. 2020.
[6] B. Cai, X. Wang, P. Li, and Z. Han, “A summary of data analysis based on differential privacy,” in Proc. IEEE 4th Int. Conf. Big Data Secur. Cloud (BigDataSecurity) Int. Conf. High Perform. Smart Comput. , (HPSC) IEEE Int. Conf. Intel. Data Secur. (IDS), Omaha, NE, USA, May 2018, pp. 79–82.
[7] S. Yu, “Big privacy: Challenges and opportunities of privacy study in the age of big data,” IEEE Access, vol. 4, pp. 2751–2763, 2016.
[8] M. Lecuyer, V. Attilidakis, R. Geambasu, D. Hsu, and S. Jana, “Certified robustness to adversarial examples with differential privacy,” in Proc. IEEE Symp. Secur. Privacy (SP), San Francisco, CA, USA, May 2019, pp. 656–672.
[9] N. Wang, X. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, and G. Yu, “Collecting and analyzing multidimensional data with local differential privacy,” in Proc. IEEE 35th Int. Conf. Data Eng. (ICDE), Macao SAR, China, Apr. 2019, pp. 638–649.
[10] A. H. M. Sarowar Sattar, J. Li, D. Jing, J. Liu, and M. Vincent, “A general framework for privacy preserving data publishing,” Knowl.-Based Syst., vol. 54, pp. 276–287, Dec. 2013.
[11] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, “Improving the utility of differentially private data releases via $k$-anonymity,” in Proc. 12th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun., Melbourne, VIC, Australia, Jul. 2013, pp. 372–379.
[12] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, “Enhancing data utility in differential privacy via microaggregation-based $k$-anonymity,” VLDB J., vol. 23, no. 5, pp. 771–794, Oct. 2014.
[13] Y. Nie, W. Yang, L. Huang, X. Xie, Z. Zhao, and S. Wang, “A utility-optimized framework for personalized private histogram estimation,” IEEE Trans. Knowl. Data Eng., vol. 31, no. 4, pp. 655–669, Apr. 2019.
[14] 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, Oct. 2017.
[15] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, “Amplification by shuffling: From local to central differential privacy via anonymization,” in Proc. 30th Ann. ACM-SIAM Symp. Discrete Algorithms, San Diego, CA, USA, Jan. 2019, pp. 2468–2479.
[16] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, “Empirical privacy and empirical utility of anonymized data,” in Proc. IEEE 29th Int. Conf. Data Eng. Workshops (ICDEW), Brisbane, QLD, Australia, Apr. 2013, pp. 77–82.
[17] S. Martínez, D. Sánchez, and A. Valls, “Semantic adaptive microaggregation of categorical microdata,” Comput. Secur., vol. 31, no. 5, pp. 653–672, Jul. 2012.
[18] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and D. Megias, “Individual differential privacy: A utility-preserving formulation of differential privacy guarantees,” IEEE Trans. Inf. Forensics Security, vol. 12, no. 6, pp. 1418–1429, Jun. 2017.
[19] M. Iftikhar, Q. Wang, and Y. Lin, “Publishing differentially private datasets via stable microaggregation,” in Proc. 22nd Int. Conf. Extending Database Technol., Lisbon, Portugal, 2019, pp. 662–665.
[20] X. Ding, C. Wang, K.-K.-R. Choo, and H. Jin, “A novel privacy preserving framework for large scale graph data publishing,” IEEE Trans. Knowl. Data Eng., early access, Jul. 30, 2019, doi: 10.1109/TKDE.2019.2931903.
[21] X. Ding, W. Yang, K.-K. Raymond Choo, X. Wang, and H. Jin, “Privacy preserving similarity joins using MapReduce,” Inf. Sci., vol. 493, pp. 20–33, Aug. 2019.
[22] C. Dwork, “Differential privacy and the US census,” in Proc. 38th ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst. (PODS), Portland, OR, USA, 2019, p. 1.
[23] J. Soria-Comas and J. Domingo-Ferrer, “Differentially private data publishing via optimal univariate microaggregation and record perturbation,” Knowl.-Based Syst., vol. 153, pp. 78–90, Aug. 2018.
[24] J. Domingo-Ferrer, O. Farràs, J. Ribes-González, and D. Sánchez, “Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges,” Comput. Commun., vol. 140–141, pp. 38–60, May 2019.
[25] M. Rodriguez-Garcia, M. Batet, and D. Sánchez, “Utility-preserving privacy protection of nominal data sets via semantic rank swapping,” Inf. Fusion, vol. 45, pp. 282–295, Jan. 2019.
[26] D. Sánchez, J. Domingo-Ferrer, S. Martínez, and J. Soria-Comas, “Utility-preserving differentially private data releases via individual ranking microaggregation,” Inf. Fusion, vol. 30, pp. 1–14, Jul. 2016.

[27] S. Martínez, A. Valls, and D. Sánchez, “Semantically-grounded construction of centroids for datasets with textual attributes,” Knowl.-Based Syst., vol. 35, pp. 160–172, May 2012.

[28] C. Xia, J. Hua, W. Tong, and S. Zhong, “Distributed K-means clustering guaranteeing local differential privacy,” Comput. Secur., vol. 90, pp. 101699–101710, Mar. 2020.

**GAOMING YANG** received the master’s degree in computer application from Guizhou University, in 2003, and the Ph.D. degree in computer application technology from Harbin Engineering University, in 2012. He is currently an Associate Professor and a M.S. Supervisor with the Anhui University of Science and Technology. He used to be a Visiting Scholar at the University of Arkansas for one year. His research interests include machine learning and privacy preserving.

**XINXIN YE** received the B.S. degree from the Anhui University of Science and Technology, in 2018, where she is currently pursuing the master’s degree. Her research interest includes privacy preserving.

**XIANJIN FANG** received the Ph.D. degree in computer application technology from Anhui University, in 2010. He is currently a Professor and a M.S. Supervisor with the Anhui University of Science and Technology. His research interests include information security and data mining.

**RONGSHI WU** received the B.S. degree from the Anhui University of Science and Technology, in 2018, where he is currently pursuing the master’s degree. His research interest includes privacy preserving.

**LI WANG,** photograph and biography not available at the time of publication.