IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 22, 2021, accepted December 8, 2021, date of publication December 15, 2021,
date of current version December 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3135946

A Method for Solving Quasi-Identifiers of Single Structured Relational Data

YI HUA, ZHANGBING LI, BAICHUAN WANG, AND JINSHENG LI
School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
Hunan Key Laboratory for Service Computing and Novel Software Technology, Xiangtan 411201, China

Corresponding author: Zhangbing Li (lzb_xt@126.com)

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 62177014 and Grant 61370227, and in part by the Hunan Provincial Natural Science Foundation of China under Grant 2017JJ2100.

ABSTRACT Quasi-identifier is a set of attributes used to identify the specific entity in structured data, which can provide an inference path for query attacks. Improper selection of quasi-identifiers leads to the failure of current privacy-preserving data publishing. In this paper, we propose a method of solving quasi-identifiers based on functional dependency to ensure the accuracy and completeness of the selected quasi-identifiers for relational data publishing. First, we partition the identifying attributes and sensitive attributes in the relational scheme of relational data published according to the semantic relationship and publishing requirements. Second, we mine the dependencies on identifying attributes with other attributes in the relational schema according to semantics and instance data in relational data, subsequently we can obtain complete quasi-identifiers. Finally, we implement the algorithm for solving quasi-identifiers in Python language, and solve quasi-identifiers on three actual data sets of different sizes, and afterward use the model of 3-anonymity, 2-diversity, and 1-differential privacy for privacy protection experiments. The results demonstrate that the average group records of equivalent class divided on the solved quasi-identifier is 8% smaller than other five methods, and the probability of privacy disclosure is reduced by about 3%. So, the accuracy and completeness of our method are better than other five methods.

INDEX TERMS Quasi-identifier, data publishing, functional dependency, privacy preserve, relational data.

I. INTRODUCTION

At present, the privacy protection of publishing relational data with single structure has failed repeatedly, and cases of privacy protection being cracked have been reported from time to time. The frequent occurrence of privacy disclosure has attracted the attention of many researchers at home and abroad. In 2016, a group of researchers from the University of Melbourne accomplished the re-identification of some entities in medical records through the hospitalization information of some celebrities reported in the news and the de-identification medical information publicly released by the Australian government within merely six weeks [1].

To prove the seriousness of the privacy disclosure issue, Rocher *et al.* proposed a copula-based generation method, which can accurately estimate the possibility of personal re-identification in anonymous data [2]. Their experiments reveal that entity information can be re-identified with high reliability through the combinations of a few attributes such as postcode, date of birth, gender, and the number of children even in too incomplete data sets. The main reason for privacy disclosure is that the quasi-identifier in the published relational data provides a link and inference path to identify personal entities and sensitive information.

Quasi-identifier is a set of attributes that are combined to determine the identity of a specific entity in structured data. After the data set is published, attackers will carry out indifference attacks or query attacks through the data values of the quasi-identifiers. Subsequently, they can deduce the privacy of the unique entity, resulting in anonymous protection in failure and the information disclosure. As the main target object of anonymous model, whether quasi-identifiers are correct and complete has become key to success or failure of privacy protection of single structured relational data. Therefore, we would study how to accurately and completely determine quasi-identifiers in a publishing scenario of relational data with single structure in this paper.

Existing privacy-preserving data publishing mainly study how to publish data, so that it can provide data with high availability as much as possible while effectively

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

protecting personal privacy. Too many papers were about specific anonymization of different data subjects, such as classical anonymity model [3]–[6], personalized anonymity method [7], [8] and differential privacy model based on perturbation [9]–[13]. But few papers were directly related to the quasi-identifier solution in the domin of privacy-preserving data publishing. Omer *et al.* proposed a simple method of selecting quasi-identifiers based on the number of distinct attribute values [14]. Lee *et al*. mainly focused on the quasi-identifier determination of re-identify clinical data [15], [16]. Pastore *et al*. determined the quasi-identifier in the published data through the maximum unique column combination and the maximum singleton [17]–[19]. Soh *et al*. mined the critical features as the quasi-identifiers of the information system's rules by using the rough set theory [20]. Song and Yan solved the quasi-identifier by combining graph theory and the connection relationship of attributes [21], [22]. However, the above solution methods still have the defects of incomplete application scenarios and inaccurate solution results.

Few works with sufficient theoretical basis and practicability in the domain of privacy-preserving data publishing provided a strong motivation for the work done in this paper. So studying quasi-identifiers is of great significance and necessity for the success of privacy protection for publishing relational data with single structure. To solve the exact and complete quasi-identifiers in the data, we propose a method based on functional dependence.

Our contributions:

1. In order to facilitate the solution of quasi-identifiers, we partition the attributes of relational data with single structure to be published according to semantics and publishing requirements. We separate all attributes into identifying attributes, sensitive attributes, and non-sensitive attributes, where quasi-identifiers may contain sensitive attributes or non-sensitive attributes.

2. We propose a complete solution of quasi-identifiers based on functional dependence. Through min the functional dependencies between attributes according to semantics instance data, the minimum set of dependencies can be derived out for the relationship with other attributes and identifying attributes of entity, and subsequently we can get the quasi-identifiers. This method ensures the completeness and accuracy of the quasi-identifiers.

The rest of the paper is organized as follows. Related works are presented in Section II. Section III introduces the method of using database functional dependency to identify quasi-identifier and illustrates our proposed algorithm for solving quasi-identifiers. Experimental details and evaluation comes in Section IV. Section V concludes the paper and outlines the direction for future work.

II. RELATED WORKS

Dalenius first proposed the concept of quasi-identifiers and distinguished it from the explicit identifier [23]. He defined quasi-identifiers as attributes that data values can uniquely determine an entity record or a minor number of entity records. Subsequently he designed the matrix representation of all recorded values and the determination method of the unique value. But this study did not take into account the dependency relationship between quasi-identifiers and other attributes. Kaur *et al.* attached importance to the impact of the number and the classes of quasi-identifiers on the re-identified risk of published data [24]. But this study did not give the method of how to solve quasi-identifiers.

Omer *et al.* defined the quasi-identifiers as attributes containing the information that can be used to identify the data entity and proposed a simple method to select the quasi-identifiers based on the number of distinct attribute values [14]. This method first required nominating candidates for quasi-identifiers and did not correctly determine truly valid quasi-identifiers. Reference [15] and [16] first deduced the quasi-identifiers in medical data according to the relevant background knowledge. Afterward screened and supplement quasi-identifiers by calculating the re-recognition probability of records through simulation experiments. Their application scenarios were not universal, and the solution results were incomplete.

Pastore *et al*. proposed a method for detecting the quasi-identifiers by taking into account both single columns and their collections and counting the unique occurrence of values [17]. It chose the smallest combination of columns which enable the most extensive affecting singletons as the best Quasi-identifiers. Podlesny *et al*. proposed a greedy search algorithm to determine maximal partial unique attribute combinations [18], [19]. For each column combination, a SQL GROUP BY statement on the dataset for the particular combination of attributes could be applied to obtain the minimal set of quasi-identifiers. The limitation of the above studies was that quasi-identifiers cannot be found in extreme cases.

Soh *et al*. mined the critical features as the quasi-identifiers of the information system's rules by using the rough set theory [20]. Without an exhaustive search among data, this method could still extract the quasi-identifiers that achieve both the highest distinct ratio and the separation ratio. But it required searching all the equivalence classes divided by different combinations of attributes, and the algorithm was inefficient and time-consuming.

For finding the quasi-identifiers based on relevant view set with hypergraph, Song designed the mapping method from attributes in published view set to hypergraph nodes and defined the quasi-identifiers as the common attributes of the relevant view set [21]. On this basis, Yan proposed a quasi-identifiers partitioning algorithm based on cut-vertexes [22]. This method simplified the transformation process from hypergraph to simple graph and converts the quasi-identifiers determining issue into finding cut-vertexes for the attribute graph from the perspective of the independence of set. The above considered that all attributes of the same data set had an association relationship without considering the association's strength between attributes from their actual values.

Since the above study did not give the reasonable partition of attributes in published data, it was difficult to select the candidates for quasi-identifiers. They also did not take into account the relationship between quasi-identifiers and other attributes or identifying attributes of entity. So, it was impossible to solve the correct and complete quasi-identifiers of relational data with the single structure.

III. PRELIMINARIES

A. RELATED CONCEPTS

Structured relational data refer to the data that express the relationship between entities and entities and logically expressed by a two-dimensional table structure. Relational data with single structure refers to the published structured relational data based on the same relational model, mainly stored and managed through a relational database. In this paper, we mainly solve the quasi-identifiers of relational data with single structure when publishing. So, we will give some concepts about relational data and functional dependence in this section.

Relationship r is the subset of the Cartesian Product of all domains, and it is also a two-dimensional table. Each row of the table corresponds to a tuple or an entity, and each column corresponds to a domain or an attribute. From the tuple perspective, the relationship r is a set of m n-ary groups. Let $r = \{t_1, t_2, \ldots, t_m\}$, each tuple t_i corresponds an ordered list $< v_1, v_2, \ldots, v_n >$, where v_j is the value of the attribute A_j for the same tuple. We shall sometimes denote the value of the attribute A_j for the tuple t_i by $t_i\left[A_j\right]$, i.e., $t_i\left[A_j\right] = v_j$.

Projection is to select several attribute columns from relationship R to form a new relationship. Let $\pi_A(R) = \{t[A] \mid t \in T\}$ be a projection of R, where A is the attribute column in R, $t\,[A]$ is the value of tuple in R on A. Specially, the tuples with duplicate values in R will be deleted after projection.

Relational scheme is the description of the relationship. It just involves the relationship name, the attributes name, the domain name, and the mapping of attribute to domain. Let $R\,(A_1, A_2, \ldots, A_n)$ denote a relational schema composed of relational name R and attribute list $A_1, A_2, \ldots, A_n$. Relational model is a set of several relationships. In the relational model, real-world entities and various relationships between entities are represented by a single structure of data, namely relationship r.

Functional dependency refers to the dependency relationship between attributes or attribute sets in a relational model. Given a relational schema $R(U)$, A^X and A^Y are two nonempty subsets of U. If any possible relation r of $R(U)$, each pair of tuples t_1 and t_2 in r satisfies the following conditions: if $t_1\left[A^X\right] = t_2\left[A^X\right]$, then $t_1\left[A^Y\right] = t_2\left[A^Y\right]$. This means that the A^X function confirms A^Y or A^Y function depends on A^X. We can think that on the relation r satisfying the functional dependence. Generally, identifying attributes K^{ID} and all attributes have functional dependencies in a $R(U)$, i.e., $K^{ID} \to A$. Meanwhile, the K^{ID} can uniquely determine each tuple entity, i.e., $K^{ID} \to T$, where T is the set of all tuple entity.

B. ATTRIBUTE PARTITION BASED ON SEMANTIC MINING

To facilitate the solution of quasi-identifiers, we partition the relational data with single structure into identifying attributes, sensitive attributes, and non-sensitive attributes according to the attribute semantics of the given relational model. Nevertheless, in the scenario of the real world, the published data set does not directly give the functional dependency set. It is necessary to determine the identifying attributes and sensitive attributes through semantic mining.

According to the attribute semantics of a given dataset, if the value of the attribute can uniquely identify an entity, it can be determined that the attribute is an identifying attribute K. Identifying attributes are similar to keys in a database.

Sensitive information is recording information that may identify or track a special entity's identity, location, property, or other private information. As an example, sensitive information of individuals is likely to name, ID number, address, and so on.

Non-sensitive attribute is an attribute except for the identifying attribute and the sensitive attribute. Sensitive attributes and non-sensitive attributes are distinguished from the perspective of semantics.

C. DEFINITION OF QUASI-IDENTIFIERS

A quasi-identifier is a set of attributes that can uniquely determine an entity tuple in addition to identifying attributes. Quasi-identifiers also provide a route for attackers to obtain the privacy of entity in published data.

In relational schema $R(U)$, a set of attributes that can identify unique or partial tuple is called a quasi-identifier QI. Set $T/QI = \{T_1, T_2, \ldots T_n\}$, T/QI is a partition of tuples on the quasi-identifier QI. Each T_i is a set of records with the same values on the QI, and there exists only a unique tuple in T_i.

According to the above partition principles, the attribute partition diagram of the relational model shown as Fig.1. Based on the basic semantic or constraint conditions in the relational model, we can determine the identifying attributes and sensitive attributes. All identifying attributes can form a set $K^{ID} = \{K_1, \ldots, K_i\}$. The set of sensitive attributes A^S is all attributes with sensitive information $A^S = \{A_1, \ldots, A_j\}$. The rest of attributes is the non-sensitive attributes $A^{NS} = \{A_{j+1}, \ldots, A_n\}$. Based on the publishing requirements, we can solve all the quasi-identifiers. Quasi-identifiers may contain sensitive or non-sensitive attributes, so $A^{QI} = \{A_{i+1}, \ldots, A_m\}$ denote the set of attributes contained in all quasi-identifiers.

IV. ALGORITHM FOR SOLVING QUASI-IDENTIFIERS BASED ON FUNCTIONAL DEPENDENCE

In the relational model, the identifying attributes can be obtained according to the specified or attribute semantics.

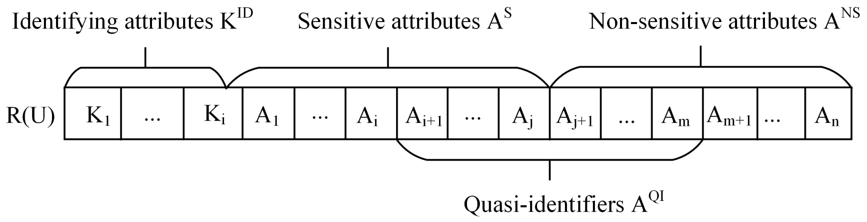

Identifying attributes K^{ID}
Sensitive attributes A^S
Non-sensitive attributes A^{NS}
R(U)
K_1
...
K_i
A_1
...
A_i
A_{i+1}
...
A_j
A_{j+1}
...
A_m
A_{m+1}
...
A_n
Quasi-identifiers A^{QI}

FIGURE 1. Attribute partition diagram.

Other attributes can identify tuple entities by generating functional dependencies with identifying attributes. Thus, we can obtain quasi-identifiers by mining the combinations of attributes that have dependencies on identifying attributes.

For example, given a relational scheme $R(A, B, C, D, \ldots)$, we can determine the following functional dependence:$B \rightarrow A$, $C \rightarrow A$, $K \rightarrow (B, C)$, and so on. If there is functional dependency $(D, E) \rightarrow K$, and the right side is the identifying attribute K, then the attribute set (D, E) on the left side of the functional dependency can determine the identifying attributes K through this dependency relationship, and then determine the entity tuple. So (D, E) can be used as a QI.

Nevertheless, this is not complete. The quasi-identifiers may be implied in the instance of published data. The correct range of quasi-identifiers should be any combinations of sensitive attributes and non-sensitive attributes, that is, all combinations of attributes except identifying attributes. Therefore, the solution requires to be divided into two parts.

A. QUASI-IDENTIFIERS SOLUTION BASED ON SEMANTICS FUNCTION DEPENDENCY

For a given relational model, we first require determining the functional dependency among the attributes. Nevertheless, relational schemes in data publishing are often not accompanied by an exact set of functional dependencies, which needs semantic mining. There has been a lot of related work on semantic search [27]–[30]. We mainly consider strengthening the efficiency of functional dependency determination through the semantic search in database.

According to the realistic semantics of attributes, such as A determines B, B is determined by A, and A is unique or has no repeat value, the functional dependency $A \rightarrow B$ can be obtained. It is also feasible to check the constraint of attribute A. If the constraint of "unique" is retained, attribute A is likely to be the identifying attribute. Subsequently, we can obtain the functional dependencies of the identifying attribute with all other attributes: $K \rightarrow B$, $K \rightarrow C$, and so on.

Given a relational schema $R(A, B, C, D, E, K)$, we obtain a functional dependency set $F = \{(B, D) \rightarrow C, (A, C) \rightarrow K, A \rightarrow E, (A, B, D) \rightarrow CE\}$ by mining semantics and reduce it to a minimal set of functional dependency $F = \{(B, D) \rightarrow C, (A, C) \rightarrow K, A \rightarrow E\}$. Subsequently we extend each functional dependency to obtain a complete set of functional dependencies. For example, $(B, D) \rightarrow C$ can be extended to the following functional dependencies: $(B, D, A) \rightarrow CA$, $(B, D, E) \rightarrow CE$, $(B, D, A, E) \rightarrow CAE$. Finally, we calculate the closure of attribute set on the left of each functional dependency.

Algorithm 1 Solving Quasi-Identifiers by Semantic Functional Dependency

```
Input: relational scheme $R(U)$; identifying attribute set $K^{ID}$;
    functional dependency set $F$.
Output: all quasi-identifier set $QIS$.
$F = Fmin()$; // solving minimum function dependency set
// extended function dependency set $F$
for each $X \rightarrow Y \in F$ do
    $F = F \cup \{(XZ) \rightarrow (YZ) \mid Z = (U - K^{ID} - Y)+\}$;
end for
// judge the attribute set on the left side of each function dependency in $F$
for each $X \rightarrow Y \in F$ do
    if $(X)_F^+ \supset K^{ID}$ then
        $QIS = QIS \cup X$;
    end if
end for
return $QIS$;
```

If the closure of the attribute set includes the identifying attribute K, then the attribute set can be used as a QI. For example, the closure of (BDA) on F is $ABCDEK$, $K \subset ABCDEK$, so (BDA) can be a QI. When all the attribute sets on the left of functional dependency have been calculated and judged, the algorithm terminates. The pseudo-code of the algorithm is as Algorithm 1.

B. QUASI-IDENTIFIERS SOLUTION BASED ON INSTANCE DATA FUNCTIONAL DEPENDENCY

In a given data instance to be published, instance dependency may be implied. Therefore, we consider a projection function dependency that satisfies the dependency on partial tuple and give the following definition.

Given a relational schema $R(U)$, A^X and A^Y are two nonempty subsets of U. Let $r_1 = \pi_{A^X}(R)$ denote the relation formed by the projection of $R(U)$ on A^X, $r_2 = \pi_{A^X \cup A^Y}(R)$ denote the relation formed by the projection of $R(U)$ on $A^X \cup A^Y$. If A^X and A^Y satisfies the following conditions:

TABLE 1. An example of projection functional dependency.

	Employee Code	Name
t_1	J02	Antony
t_2	J15	Taylor
t_3	J03	Rosa
t_4	J17	Taylor
t_5	J24	Eric
t_6	J28	Mark

if $u_1 \in r_1, t_1 \in r_2$, $T = \{t_1 \mid t_1[A^X] = u_1[A^X]\}$, then $\exists|T| = 1$. This means that the A^X projection function confirms A^Y or A^Y projection function depends on A^X. Different from functional dependency, as long as a value of $R(U)$ in U_i can determine the value of $R(U)$ in K, we can think K projection function depends on U_i.

Firstly, we will determine the scope of the candidates for quasi-identifiers. Since quasi-identifiers may contain sensitive or non-sensitive attributes, candidates for quasi-identifier should be any combination of these two types of attributes. Subsequently we will enumerate the instance data to mine the functional dependencies between candidate combinations and identifying attributes. Finally, we infer the quasi-identifiers according to the mining functional dependency set.

Given a relational scheme $R(A, B, C, D, E, K)$, where $K^{ID} = K$. Let $U = A^S \cup A^{NS} = \{A, B, C, D, E\}$ denote the range of attributes contained in candidates for quasi-identifiers, any combination of attributes in U are $U+ = \{A, B, C, D, E, AB, AC, \ldots, ABCDE\}$. Then, we will select each attribute combination U_i in turn, and create a projection view on U_i and the identifying attribute K^{ID}. By enumerating the instance data on the partitioned view, it is determined whether the attribute combination has functional dependency or projection functional dependency with the identifying attribute. For example, if we want to determine the functional dependency between the "name" attribute and the identifying attribute of "Employee Code", we first need to create a projection view as shown in Table 1. The partition $T/(name) = \{\{t_1\}, \{t_2, t_4\}, \{t_3\}, \{t_5\}, \{t_6\}\}$ is obtained, and then through $T_1 = \{t_1\}$ and $|T_1| = 1$ we can infer that (name) $\overset{P}{\rightarrow}$ (Employee Code). Finally, the left attribute set is determined according to the functional dependency that the right attribute set is K^{ID} and then all the quasi-identifiers are obtained. The pseudo-code of the algorithm is as Algorithm 2.

The inference path between quasi-identifiers and entity's information of individuals is established by the dependency between identifying attributes and quasi-identifiers. Thus, qualitative research of quasi-identifiers ensures the correctness of the solution of quasi-identifiers. The key point of solving the quasi-identifiers through functional dependency is to mine all the functional dependencies existing in the current relational model. We determine the functional dependency from the semantics of the relational model and the instance data respectively, to prevent the omission or neglect of the correlation between certain attributes, resulting in the incomplete solution.

Algorithm 2 Solving Quasi-Identifiers by Instance Data Functional Dependency

```
Input: relational scheme $R(U)$; identifying attribute set $K^{ID}$;
    functional dependency set $F$.
Output: all quasi-identifier set $QIS$.
// initialization parameters
$U = A^S \cup A^{NS}, Y = K^{ID}$
for each $X \in U+$ do
    // create projection view
    $r = \pi_{X \cup Y}(R)$
    $T_i = \{T_i \mid T_i \in T/X\}$
    if $|T_i| == 1$ then
        $F = F \cup \{X \overset{P}{\rightarrow} Y\}, QIS = QIS \cup X$;
    else
        next $X$;
    end if
end for
return $QIS$;
```

C. MERGING OF QUASI-IDENTIFIERS SOLUTIONS

A complete set of quasi-identifiers is solved by two algorithms respectively and then merged with the results. Finally, we optimize the set of quasi-identifiers according to lemma of Armstrong axiom: if $A^X \rightarrow A^Y$, then $A^X A^Z \rightarrow A^Y$. Since the superset of a quasi-identifier is still a quasi-identifier, the superset can be removed only to retain the minimum quasi-identifiers. As an example, $\{B, C\}$ is the superset of $\{B\}$ in QIS, remove $\{B, C\}$ and retain $\{B\}$, so as to obtain the minimum set of quasi-identifiers. The calculation formula is as follows:

$$A^{QI} = \bigcup_{i=0}^{n} QIS_i. \tag{1}$$

where $n = |QIS|$ is the total number of solved quasi-identifier sets.

V. EXPERIMENT

A. DATASET AND SETUP

To evaluate the effectiveness of the proposed approach, we test our method on three publicly available data sets: Estimation of Obesity Levels based on Eating Habits and Physical Condition, Bank Marketing, and Adult Dataset [29]. The data sets we selected are different in data scale and attribute categories. Continuous data publishing will have a corresponding impact on functional dependencies between attributes. For example, increasing data will destroy the one-to-one dependency of some attributes, which will lead to the change of solution results of quasi-identifiers. So the experiment will conduct comparative analysis on the static data set.

TABLE 2. Details of datasets.

Data Set	Estimation of obesity levels	Bank Marketing	Adult
Number of Records	2111	4521	30707
Number of Attributes	11	8	10
Identifying Attributes	{ID}	{ID}	{SSN}
Sensitive Attributes	{family, nobeyesdad}	{marital, loan}	{marital, income}
Non-sensitive Attributes	{age, gender, height, weight, CAEC, smoke, CALC, mtrans}	{age, education,job, housing, default}	{sex, age, zip, workclass, education, occupation, race}

TABLE 3. Solving quasi-identifiers by different methods.

Methods	Reference	Estimation of obesity levels	Bank Marketing	Adult
SQIA	[14]	{gender, height, weight, family, mtrans}	{age, job, marital, education}	{age, workclass, zip, education, occupation}
DPI	[17]	{age, gender, height, weight, family, CAEC, CALC, mtrans}	{age, job, marital, education, housing, loan, default}	{age, workclass, zip, education, marital, occupation, income}
F2CIC	[19]	{age, height, weight, CAEC, smoke, mtrans, nobeyesdad}	{age, job, marital, loan, default}	{age, workclass, zip, education, occupation, race, income}
QIA	[21]	{age,gender, height, weight, CAEC, smoke, CALC}	{age, education, job, housing, loan}	{sex, age, workclass, education, occupation}
FPCV	[22]	{age, gender, height, weight}	{job, housing, loan}	{workclass, education, occupation}
Our Method	/	{age, height, weight, CAEC, smoke, CALC, mtrans, nobeyesdad }	{age, job, marital, housing, loan, default}	{age, workclass, zip, education, marital, occupation, race, income}

We pre-process the data set and divided the main steps into the following two steps. First, remove invalid tuple and add or delete some attribute columns. Second, partition all the attributes in the data sets according to semantics and publishing requirements. Based on the relationship model given in the data set, we extract identifying attributes according to semantic and sensitive attributes according to sensitive information, including the disease history, medical diagnosis records, marriage, and personal property. There are only two value types "yes" and "no" of most attributes in the Bank Marketing data set, it will lead to the difference between the experimental results and other data sets. The details of the data sets are shown in Table 2.

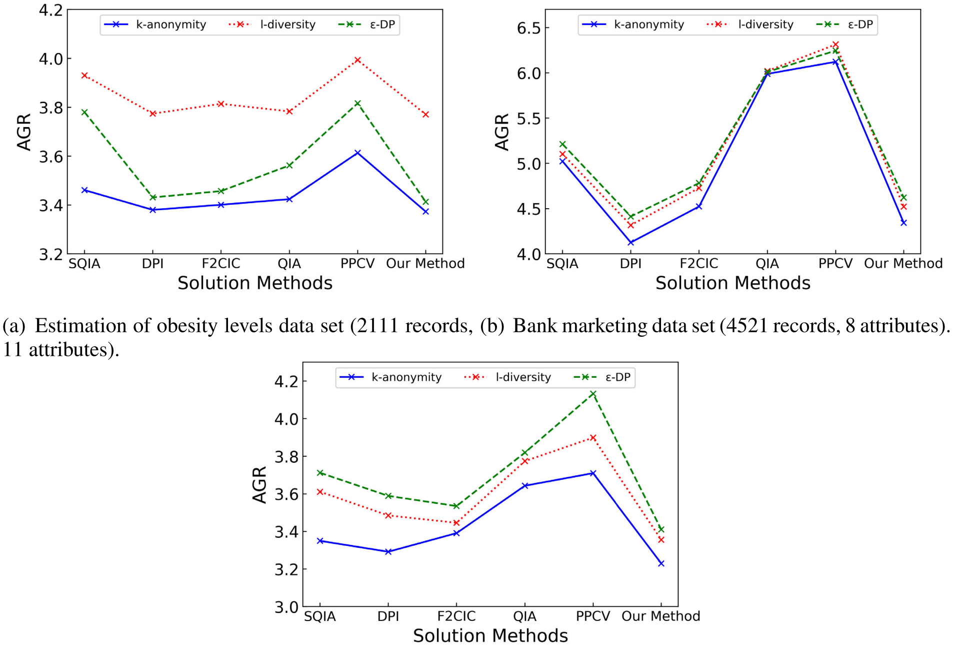
We use real data of different scales to compare our method with the following five quasi-identifiers solution method: *SQIA* [14], *DPI* [17], *F2CIC* [19], *QIA* [21], and *FPCV* [22]. The *SQIA* is mainly based on the number of categories that can be classified by attribute combination, which mainly focuses on the effect of discrimination ratio. Similar to the *F2CIC*, *DPI* detects the attribute combination with unique value, but *DPI* also considers selecting the minimum combination of attributes exposing the most affecting singletons as the best QID. The *QIA* and the *FPCV* use the form of graph to transform the relationship between attributes, which is mostly used to find the key attributes of linked tables. The *FPCV* simplifies the mapping process from hypergraph to ordinary graph and reduces the solution range of quasi-identifier. Based on the principle of experiments from different angles, we choose the above methods to compare with our proposed methods.

And we need to nominate the basic candidate for the *SQIA* as {gender, height, weight, family, mtrans},{age, job, marital,education}, and {sex, age, workclass, zip, education, marital, occupation}. Given the background knowledge and published information of the *QIA* and the *FPCV* as {ID, age, gender, height, weight}, {CAEC, smoke, CALC, family, nobeyesdad}, {ID, age, education}, {job, housing, loan, default}, {SSN, sex, age}, and {workclass, education, occupation, marital, income}. The data sets to be published are the original data sets that remove identifying attributes and sensitive attributes. We completed the experiments with Python language and jupyter notebook platform. The experimental results are shown in Table 3.

B. PRIVACY PROTECTION METRICS

To evaluate effect and utility of the method, we use standardized metrics to measure the data availability and the effect of privacy protection in published anonymous data.

We define the AGR(average group records) to evaluate data availability of published data. AGR measures the average number of records in each equivalence class. We take into account the two extreme cases. When AGR $= 1$, each record constitutes an equivalent class, and the value of each record is determined. So accurate information can be obtained by data mining. When AGR $= |T|$, as the degree of anonymity increases, there is merely one equivalence class in the entire anonymous data. So it is impossible to distinguish the each tuple and obtain effective information by data mining, and data availability is also reduced. The calculation formula is



(a) Estimation of obesity levels data set (2111 records, 11 attributes).

(b) Bank marketing data set (4521 records, 8 attributes).

(c) Adult data set (30707 records, 10 attributes).

FIGURE 2. Average group records charts.

as follows:

$$AGR = \frac{\sum_1^m |G_i|}{|G|}. \tag{2}$$

G is the set of equivalence groups of anonymous table, each G_i is an equivalence group. $|G|$ represents the number of equivalence groups, $|G_i|$ is the size of each equivalence group. The smaller the average group size is, the less records each equivalent group has, and the higher the value of the data is. Therefore, data availability is inversely proportional to the average group size.

And we define an another metric PDP(privacy disclosure probability) to measure the effect of privacy protection. The PDP of equivalence class is the probability that attackers correctly identify the tuple entities in each equivalence class divided by QI. The privacy disclosure probability of the original data table is the average of the privacy disclosure probability of all equivalence classes. The probability of privacy disclosure will decrease with the increase of privacy protection effect. The calculation formula is as follows:

$$PDP = \frac{\Sigma_1^{|G|} |p_i|}{|G|}. \tag{3}$$

p_i is the privacy disclosure probability of the ith equivalence class, which is also the mean of the privacy disclosure probability of all tuple in the equivalence class. $\frac{|T|}{|G_i|}$ is the privacy disclosure probability of each tuple in this equivalence class, where $|T|$ is the number of completely equivalent tuple and $|G_i|$ is the size of the ith equivalent class.

C. EXPERIMENTAL RESULTS AND DISCUSSION

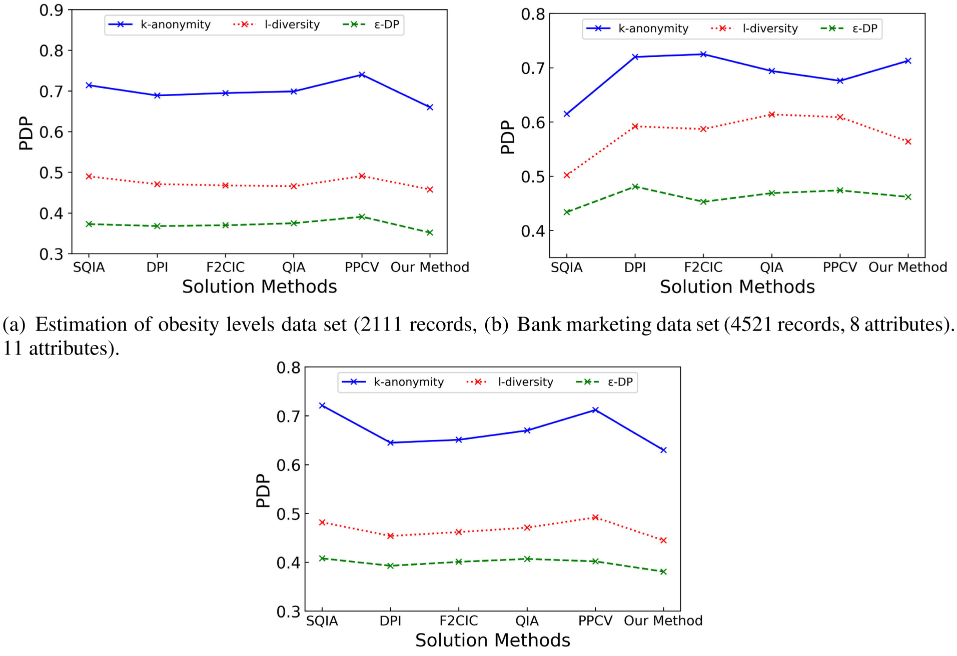
In the experiment, we study the impact of different quasi-identifiers on privacy-preserving data publishing when using the same anonymous models. And we choose three common anonymization models like k-anonymity, l-diversity, and ε-differential privacy model for experiments. Based on the principle of data security, we need to ensure that attacks cannot identify an entity in at least three records, so set $k = 3$. Since there are only two categories of data values in sensitive attribute of Bank Marketing dataset, so set $l = 2$. At this time, the table after using the l-diversity model needs to satisfy both 3-anonymity and 2-diversity. In the ε-differential privacy model, we add noise to original data by Laplace mechanism and Exponential mechanism and set the privacy budget $\varepsilon = 1$.

As to the same quasi-identifiers, the average group records of anonymous data using different models in three datasets is described in Table 4. To satisfy that there are at least two types of values for sensitive attributes in each equivalence class, the anonymous table of l-diversity model has a larger scale of equivalence class and a smaller total number of equivalence classes. So the data availability of the k-anonymity method is better than that of the l-diversity method. The differential privacy model adds a certain amount of noise to all data, resulting in a deviation between the original value of the and the processed value of data. Therefore, the data availability is worse than k-anonymity method.

Fig.2 shows the impact of quasi-identifiers solved by different methods on anonymous data. The average group

TABLE 4. Experimental results of average group records.

Methods	K-anonymity			L-diversity			ε-DP		
	Estimation of obesity levels	Bank Marketing	Adult	Estimation of obesity levels	Bank Marketing	Adult	Estimation of obesity levels	Bank Marketing	Adult
SQIA	3.461	5.023	3.350	3.930	5.103	3.611	3.780	5.213	3.712
DPI	3.380	**4.126**	3.292	3.774	**4.317**	3.485	3.431	**4.413**	3.589
F2CIC	3.401	4.523	3.391	3.814	4.728	3.446	3.457	4.782	3.535
QIA	3.424	5.988	3.643	3.783	6.120	3.774	3.562	6.010	3.819
FPCV	3.613	6.122	3.710	3.993	6.313	3.899	3.816	6.245	4.132
Our Method	**3.373**	4.343	**3.230**	**3.771**	4.521	**3.356**	**3.413**	4.622	**3.410**



(a) Estimation of obesity levels data set (2111 records, 11 attributes).

(b) Bank marketing data set (4521 records, 8 attributes).

(c) Adult data set (30707 records, 10 attributes).

FIGURE 3. Privacy disclosure probability charts.

records of the Bank Marketing Dataset is depicted in Fig.2(b), where the *DPI* is higher than other methods in data availability. Because all non-identifying attributes are quasi-identifiers according to the solving result of *DPI*, which leads to a larger number of equivalent classes and smaller AGR in anonymous data. On other data sets, our solution method based on functional dependence outperforms all other approaches. It correlates with the completeness and numbers of the quasi-identifiers. Our method is to find all quasi-identifiers as much as feasible without any negligence. With the increase in the number of quasi-identifiers, the number of equivalence classes will also increase. It means that the reduction of records in equivalence classes and the enhancement of data availability.

Table 5 shows the privacy disclosure probability of different anonymous data obtained using different models. As to the same quasi-identifiers, the l-diversity method performs better than the k-anonymity method. Since the l-diversity model takes into account the value of sensitive attributes, it further restricts the requirements of equivalence class partition in the k-anonymous model. The differential privacy model obtains the data deviation by adding noise to the data, so that the attacker cannot obtain the real data through query attack, which ensures the effectiveness of privacy protection.

The privacy disclosure probability of each data set obtained by different methods is depicted in Fig.3. Nevertheless, we note that our method performs not ideal on the Bank Marketing Dataset. Because some attributes in the Bank Marketing Dataset have merely two types of values, it leads to more equivalence classes being partitioned by our quasi-identifiers. This leads to a decrease in the number of tuples in the equivalence class and an increase in the probability of privacy disclosure. On other data sets, our solution method based on functional dependence outperforms all other approaches. This means that it is more difficult for an attacker to determine the identity of an entity. Therefore, the quasi-identifiers

TABLE 5. Experimental results of privacy disclosure probability.

Methods	K-anonymity			L-diversity			ε-DP		
	Estimation of obesity levels	Bank Marketing	Adult	Estimation of obesity levels	Bank Marketing	Adult	Estimation of obesity levels	Bank Marketing	Adult
SQIA	0.714	**0.615**	0.721	0.490	**0.502**	0.482	0.373	**0.434**	0.408
DPI	0.689	0.720	0.645	0.471	0.592	0.454	0.368	0.481	0.393
F2CIC	0.695	0.725	0.651	0.468	0.587	0.462	0.370	0.453	0.401
QIA	0.699	0.694	0.670	0.466	0.614	0.471	0.375	0.469	0.407
FPCV	0.740	0.676	0.712	0.491	0.609	0.492	0.391	0.474	0.402
Our Method	**0.660**	0.713	**0.630**	**0.458**	0.564	**0.445**	**0.352**	0.462	**0.381**

TABLE 6. Improvement ratio of AGR compared with our method.

Methods	K-anonymity	L-diversity	ε-DP
SQIA	6.55%	7.50%	9.73%
DPI	-1.06%	-0.31%	0.26%
F2CIC	3.18%	2.71%	2.72%
QIA	13.43%	12.51%	12.66%
FPCV	16.21%	15.88%	18.01%

TABLE 7. Improvement ratio of PDP compared with our method.

Methods	K-anonymity	L-diversity	ε-DP
SQIA	1.42%	0.62%	1.78%
DPI	2.50%	3.16%	3.78%
F2CIC	3.31%	3.24%	2.62%
QIA	2.94%	5.13%	4.67%
FPCV	5.62%	7.89%	5.91%

solution method proposed in this paper conforms to the privacy protection standard and ensures the availability of data while effectively protecting personal information.

We define the calculation formula of the IR(improvement ratio) between the two methods as follows:

$$IR = \frac{B - A}{B} \times 100\%. \tag{4}$$

where B is taken as the measurements of other methods, A is measurement value of the proposed method on the same metrics. We will calculate separately the average improvement ratio of AGR and PDP between our method compared with the baseline methods on three data sets applied in three anonymous models. The calculation formula is as follows:

$$\overline{IR} = \frac{\sum_1^n IR_i}{n} \tag{5}$$

where n is number of data sets, IR_i is the improvement ratio corresponding to the baseline method on the ith data set. The final results are shown in Table 6 and Table 7.

Obviously, we have reduced the number of average group records by about 8% on average, and the risk of privacy disclosure has been reduced about 3%. Although our average group record is slightly larger than the *DPI* method, it is a sacrifice to better reduce the risk of privacy disclosure. In fact, it is necessary to simultaneously consider both metrics in the real scene of anonymity protection. Because anonymous data should meet the publishing requirements that the data availability as high as possible and the probability of privacy disclosure as low as possible. In conclusion, our method can achieve better performance in effect and utility on different scale datasets than other methods.

VI. CONCLUSION

In privacy protection, only anonymizing sensitive information cannot completely cut off the inference path. It is essential to find and effectively anonymize all quasi-identifiers. Unquestionable, exact and complete quasi-identifiers are the key to the success of privacy protection. This paper presents a new quasi-identifier solution method based on the functional dependency in published data with the single structure. First, identifying attributes, sensitive attributes, and non-sensitive attributes are partition according to semantics and publication requirements. Second, we can get all quasi-identifiers by establishing the inference path through the dependency relationship among the quasi-identifiers and the identifying attribute. It requires mining functional dependencies from semantics and instance data respectively. Experiments on real data set based on different anonymous models demonstrate the effectiveness of our approach. Our method will also provide a reliable basis to further reduce the privacy disclosure risk of the anonymous model.

In the future, we would like to improve the efficiency of the algorithm and determination of functional dependence from the perspective of semantic analysis. And we would study the solution of quasi-identifiers for continuously published data with the same or different relational scheme. Furthermore, we would consider the privacy protection methods of structured relational data and unstructured data.

REFERENCES

[1] C. Culnane, B. I. P. Rubinstein, and V. Teague, "Health data in an open world," 2017, *arXiv:1712.05627*.

[2] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Commun.*, vol. 10, no. 1, pp. 1–9, Dec. 2019.

[3] L. Sweeney, "Achieving K-anonymity privacy protection using generalization and suppression," *Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, 2002.

[4] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

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

[6] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.

[7] D. R. Monedero, A. M. Mezher, X. C. Colomé, J. Forné, and M. Soriano, "Efficient k-anonymous microaggregation of multivariate numerical data via principal component analysis," *Inf. Sci.*, vol. 503, pp. 417–443, Nov. 2019.

[8] X. Wang, Y. Luo, Y. Jiang, W. Wu, and Q. Yu, "Probabilistic optimal projection partition KD-tree k-anonymity for data publishing privacy protection," *Intell. Data Anal.*, vol. 22, no. 6, pp. 1415–1437, Dec. 2018.

[9] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. (TAMC)*, Xi'an, China, 2008, pp. 1–19.

[10] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," in *Proc. NIPS*, 2012, pp. 2339–2347.

[11] 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.

[12] 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, 2017.

[13] J. Snoke and A. Slavkovic, "pMSE mechanism: Differentially private synthetic data with maximal distributional similarity," in *Proc. Int. Conf. (PSD)*, Valencia, Spain, 2018, pp. 138–159.

[14] A. M. Omer and M. M. B. Mohamad, "Simple and effective method for selecting quasi-identifier," *J. Theor. Appl. Inf. Technol.*, vol. 89, no. 2, p. 512, 2016.

[15] Y. J. Lee and K. H. Lee, "What are the optimum quasi-identifiers to re-identify medical records?" in *Proc. 20th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2018, pp. 1025–1033.

[16] J. Jung, P. Park, J. Lee, H. Lee, G. K. Lee, and H. S. Cha, "A determination scheme for quasi-identifiers using uniqueness and influence for de-identification of clinical data," *J. Med. Imag. Health Informat.*, vol. 10, no. 2, pp. 295–303, Feb. 2020.

[17] M. Pastore, M. A. Pellegrino, and V. Scarano, "Detecting and generalizing quasi-identifiers by affecting singletons," in *Proc. EGOV-CeDEM-ePart*, 2020, pp. 327–335.

[18] N. J. Podlesny, A. V. Kayem, S. von Schorlemer, and M. Uflacker, "Minimising information loss on anonymised high dimensional data with greedy in-memory processing," in *Proc. Int. Conf. (DEXA)*, Regensburg, Germany, 2018, pp. 85–100.

[19] N. J. Podlesny, A. V. D. M. Kayem, and C. Meinel, "Attribute compartmentation and greedy UCC discovery for high-dimensional data anonymization," in *Proc. 9th ACM Conf. Data Appl. Secur. Privacy*, Mar. 2019, pp. 109–119.

[20] C. Wafo Soh, L. L. Njilla, K. K. Kwiat, and C. A. Kamhoua, "Learning quasi-identifiers for privacy-preserving exchanges: A rough set theory approach," *Granular Comput.*, vol. 5, no. 1, pp. 71–84, Jan. 2020.

[21] J. L. Song, "Research on key issues related to anonymous data in the k-anonymous privacy protection model," Ph.D. dissertation, Dept. Comput. Sci. Eng., Yanshan Univ., Heibei, China, 2012.

[22] Y. Yan, "Research on privacy protection technology of big data release," Ph.D. dissertation, Dept. Elect. Eng., Univ. Lanzhou Technology, Gansu, China, 2018.

[23] T. Dalenius, "Finding a needle in a haystack or identifying anonymous census records," *J. Financial Statist.*, vol. 2, no. 3, p. 329, 1986.

[24] G. Kaur and S. Agrawal, "Differential privacy framework: Impact of quasi-identifiers on Anonymization," in *Proc. 12nd ICCCN*, Singapore, 2019, pp. 35–42.

[25] H. Bast and M. Celikik, "Efficient index-based snippet generation," *ACM Trans. Inf. Syst.*, vol. 32, no. 2, pp. 1–24, Apr. 2014.

[26] H. Bast and B. Buchhold, "An index for efficient semantic full-text search," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 369–378.

[27] J. Tekli, R. Chbeir, A. J. M. Traina, C. Traina, K. Yetongnon, C. R. Ibanez, M. Al Assad, and C. Kallas, "Full-fledged semantic indexing and querying model designed for seamless integration in legacy RDBMS," *Data Knowl. Eng.*, vol. 117, pp. 133–173, Sep. 2018.

[28] S. Kang, J. Shim, and S.-G. Lee, "Tridex: A lightweight triple index for relational database-based semantic web data management," *Expert Syst. Appl.*, vol. 40, no. 9, pp. 3421–3431, Jul. 2013.

[29] D. Dua and C. Graff, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Berkeley, CA, USA, Tech. Rep., 2007. [Online]. Available: http://archive.ics.uci.edu/ml

YI HUA received the bachelor's degree in computer science and technology from the Yunnan University of Finance and Economics. She is currently pursuing the master's degree with the Hunan University of Science and Technology. Her research interest includes privacy preserving data publishing.

ZHANGBING LI is currently an Associate Professor with the School of Computer Science and Engineering, Hunan University of Science and Technology. He has been engaged in computer teaching and scientific research for more than 30 years, and has published many articles and the textbook *Computer System Security*. He is mainly involved in the teaching of network security, information security management, and distributed database systems. His research interests include system security, network security, and cryptography.

BAICHUAN WANG received the bachelor's degree in accounting from Wuhan Polytechnic University. He is currently pursuing the master's degree with the Hunan University of Science and Technology. His research interest includes local differential privacy.

JINSHENG LI received the bachelor's degree in software engineering from the Nanjing University of Chinese Medicine. He is currently pursuing the master's degree with the Hunan University of Science and Technology. His research interest includes data security.