IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 10, 2019, accepted May 20, 2019, date of publication May 27, 2019, date of current version June 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919165

A New Method of Privacy Protection: Random k-Anonymous

FAGEN SONG[1,2], TINGHUAI MA[1], YUAN TIAN[3], AND MZNAH AL-RODHAAN[4]

[1]School of Computer & Software, Nanjing University of Information Science and Technology, Nanjing 210-044, China
[2]Yancheng Institute of Technology, Yancheng 224-051, China
[3]Nanjing Institute of Technology, Nanjing 210-044, China
[4]Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh 11362, Saudi Arabia

Corresponding author: Tinghuai Ma (thma@nuist.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant U1736105 and Grant 61572259, in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant KYCX_0900, in part by the Natural Science Foundation of the Colleges and Universities, Anhui, China, under Grant KJ2017B016, and in part by the grant from the Research Center of the Female Scientific and Medical Colleges, Deanship of Scientific Research, King Saud University.

ABSTRACT A new k-anonymous method which is different from traditional k-anonymous was proposed to solve the problem of privacy protection. Specifically, numerical data achieves k-anonymous by adding noises, and categorical data achieves k-anonymous by using randomization. Using the above two methods, the drawback that at least k elements must have the same quasi identifier in the k-anonymous data set has been solved. Since the process of finding anonymous equivalence is very time consuming, a two-step clustering method is used to divide the original data set into equivalence classes. First, the original data set is divided into several different sub-datasets, and then the equivalence classes are formed in the sub-datasets, thus greatly reducing the computational cost of finding anonymous equivalence classes. The experiments are conducted on three different data sets, and the results show that the proposed method is more efficient and the information loss of anonymous dataset is much smaller.

INDEX TERMS Differential privacy, information security, k-anonymous, privacy protection, random k-anonymous.

I. INTRODUCTION

With the progress of technology, especially the emergence of smart phones, it is more and more convenient for people to collect, share and distribute information. Businesses can provide personalized services that are more suitable for user requirements by analyzing customer's data. Scientists can more easily obtain data for scientific research. The government can carry out a more scientific and effective social management based on the collected information. However, it also leads to a increasing threat to people's privacy information, and simply removing the identifier of record is not enough to protect the user's privacy. As described in [1] and [2], an electronic version of a city's voter list can be purchased for twenty dollars and can be used to re-identify the medical records. In addition to name and address, it also includes the date of birth and gender of more than five thousand voters. Of these, about 12% have unique birth date, 29% are unique with respect to birth date and gender, 69% with respect to birth date and a 5-digit ZIP. code, 97% are identifiable with just the full postal code and birth date [1]. In [3], Sweeney pointed out that about 87% of Americans can be uniquely identified through the combination of their gender, birth date and postcode [2]. Thus, It's easy for an attacker to get the user's privacy information if the dataset containing the above three attributes that was published directly.

In the field of medical research, some hospitals may share diagnostic records with relevant research institutions. The individual data is presented in the form of a tuple and has a fixed structure: including Name, Age, Zip, Code, and Disease Type. Even if the attribute ''Name'' is deleted, there is still a risk that personal privacy will be disclosed. A hacker knowing an individual's age and zip code may be able to conclude that a specific individual has dyspepsia [4]. But if there are more than K(the value of K is large enough) individuals have the same age and zip code, the hacker will not be able to guess out which one has dyspepsia confidently. K-anonymous can ensure that at least k records have the same age and zip code.

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Khurram Khan.

Researchers have proposed many privacy protection methods. Among these methods, k-anonymous [2] and differential privacy [4] are two most important privacy models. K-anonymous makes at least k records have the same quasi identifier by generalization or clustering. If k records in the data set have the same quasi-identifier, the attacker has only 1/k probability to guess out the correct result [32]. The method is easy to implement, and the leakage risk is measurable, so it is widely used [26]. However, because of the increase of the attacker's background knowledge, the effect of privacy protection is getting worse and worse [27]. In order to provide more protection of privacy, the differential privacy has been applied to privacy protection. Even if the attacker has all the information except the protected information, he is not able to obtain the privacy information confidently. But the cost of implementing differential privacy procedure is very expensive, and with the increasing of protection strength, the data availability will be badly damaged.

There are mainly three contributes in this paper. Firstly, a novel method which achieves k-anonymous by adding noise and randomness is proposed, and the drawback of traditional k-anonymous is overcome without increasing the loss of availability. The details of the method will be discussed theoretically and practically in the following sections. Secondly, an efficient method for generating anonymous classes is proposed. Last but not least, in order to verify the effectiveness of our method, experiments are conducted on three different real-world data sets. The results show that our method is more efficient, and the loss of availability of anonymous data sets is smaller.

The rest of the paper is organized as follows. Section 2 contains background on k-anonymous, l-diversity, t-closeness and differential privacy. In section 3, random k-anonymous is proposed. In order to achieve k-anonymous efficiently, a two-step clustering method is adopted. Experiments are conducted in section 4, and the experiment results are compared with others. Section 5 reviews related work. Conclusions are given out in section 6.

II. BACKGROUND

A. k-ANONYMOUS

k-anonymous was proposed by Samarati and Sweeney [6], and its implementation method was given in [2], [5]. The following is a brief introduction of its definition, implementation, and its improvements process.

Definition 1 (k-Anonymous [5]): Let T(A_1, A_2, A_n) be a table and QI. be a quasi-identifier associated with it. T is said to satisfy k-anonymity with respect to QI. if and only if each sequence of values in T[QI] appears at least with k occurrences in T[QI].

The k-anonymity property ensures protection against identity disclosure. However, it can not protect the data against sensitive attribute disclosure. For example, if the sensitive attribute has the same value, we know the sensitive value of the one who is in this anonymous group, though we do not know the QI. of the element. In order to resist Homogeneity attack and Background knowledge attack, Machanavajjhala proposed the l-diversity model in [7].

Definition 2 (l-Diversity [8]): A QI-group satisfies l-diversity if there are at least l distinct values for the sensitive attribute. A modified table satisfies l-diversity if every cluster of the table satisfies l-diversity.

L-diversity improved the security of k-anonymous, but it still has shortcomings. If the distribution of sensitive information in each anonymous group is quite different from the distribution of the whole data set, it still can leak private information. For instance, if the probability of a certain disease in somewhere is 0.01, and the probability is 0.5 in some anonymous group, though this anonymous group follows l-diversity, we still can get the information that the one who is in the anonymous group may be suffering from this disease with much higher probability. In order to overcome this drawback, Ninghui Li et al. proposed t-closeness which requires that the distance between the distribution of sensitive information in the anonymous group and the distribution in the whole data set is no more than t. The definition is as below.

Definition 3 (t-Closeness [9]): An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. The table is said to have t-closeness if all equivalence classes have t-closeness.

t-closeness proposed in [9] requires not only satisfying l-diversity, but also the distribution of sensitive attributes in the anonymous data set should be as close as possible to the distribution of sensitive attributes in the whole data set. k-anonymous, l-diversity, p-sensitive and t-closeness all require that all the records in the same anonymous group must have the same quasi identifier. We can't distinguish the records with each other. But, it is precisely because k records have the same quasi identifier that provides information for the adversary to successfully implement attacks. If the attacker has the quasi identifier of a record, he can judge the anonymous group that this record corresponds to, and with the help of auxiliary information, it is easy for him to carry out a successful attack by exhaustion. The method proposed in this paper does not require that all the elements in the same anonymous group have the same quasi identifier, and the attacker can not get the information which elements are in the same anonymous group. so if the attacker wants to finish an attack successfully by exhaustion, he must traverse all the records of the data set, which is clearly not possible. The specific algorithm will be discussed in detail in Section 3.

B. DIFFERENTIAL PRIVACY

Differential privacy was proposed by Cwork [4], [12], without making assumptions about the background knowledge of the attacker which can be proved mathematically. So it quickly became a hot spot of research in this field. The definition is given as follows.

Definition 4: A randomized function K gives ε-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq$ Range(K),

$$Pr[K(D_1) \in S] \leq exp(\varepsilon) \times Pr[K(D_2) \in S]. \quad (1)$$

There are two approaches of achieving differential privacy. One is Laplace mechanism, the other is exponential mechanism. The Laplace mechanism is more suitable for differential privacy transformation of numerical data [4], [11]. The exponential mechanism is more preferable for non-numeric data differential privacy transformation [13].

As soon as the differential privacy was proposed, it attracted a large number of scholars to carry out related research, which leads to getting a lot of important results. This paper adopts the idea of differential privacy adding noises to the quasi-identifier to achieve k-anonymous. The attacker cannot determine which records belong to the same anonymous group, which makes at least k records indistinguishable. The distribution of the sensitive attributes of the anonymous data set and the original data set is obviously the same, of course, satisfying t-closeness.

III. RANDOM k-ANONYMOUS

There are two ways to implement k-anonymous. One is clustering, and the other is generalization. Clustering classifies at least k nearest elements of original data into one subclass, and replaces the identifiers of other elements in the subclass with the quasi-identifier of the center element. This method is more suitable for numerical data. As shown in Figure 1 and Figure 2, generalization is a method of expanding the value of a specific quasi-identifier into a larger value range, so that it can no longer uniquely represent a unique record in the data set. For example, 'male' and 'female' can be generalized into 'gender unknown', 'married' and 'unmarried' can be generalized into 'unknown', age can be generalized to an age domain. This method is mainly suitable for data with hierarchical structure.

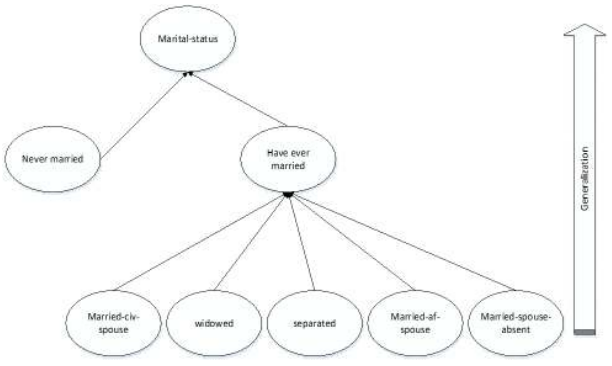

FIGURE 1. Generalization of non-numeric.

Regardless of whether clustering or generalization is used, the ultimate goal is to have at least k records in the data set with the same quasi identifiers, which makes these records

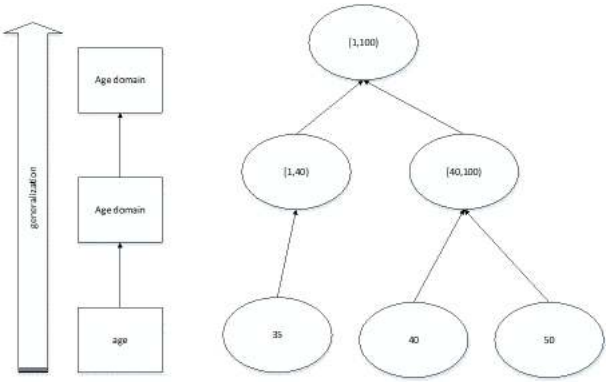

FIGURE 2. Generalization of numeric.

indistinguishable, and then the privacy of the user is protected. Since k records share the same quasi identifier, user's privacy is protected, but this also leads to the risk of privacy leakage. For example, if an attacker obtains a quasi-identifier of someone and wants to get the corresponding sensitive information, the only thing the attackers need do is to carry out an exhaustive attack with the help of auxiliary information.

For instance, there are ten records in the anonymous class. If the attacker gets the quasi-identifier of Bob. He has known that Bob's record is in the anonymous class. The attacker wants to know which particular record is Bob's. The only thing he needs to do is to test all ten records.

K records sharing the same quasi identifier is not our ultimate goal, and our goal is that k records are indistinguishable. In other words, as long as k records are indistinguishable, the value of their identifiers is not important. Based on this, we proposed the random k-anonymous, which is as follows.

Definition 5 (Random k-Anonymous): q is a random query on data set D, the probability that $q(D)$ is generated by $e_1, e_2, \ldots,$ or $e_{k'}$ is equal, where $e_i \in D, k' \geq k$. We say that for query q, D satisfy random k-anonymous.

Compared with definition 1, k records may have different quasi identifiers, which means that every record can have its own unique identifier. The only requirement is the probability that each record corresponding to the query results is equal with each other. The requirement that the quasi-identifiers must be have the same value. Definition 1 is only one of the methods in implementation of definition 5. If the one that satisfies the definition 1, it certainly satisfies the definition 5, and the definition 5 is an extension of definition 1.

As can be seen from the above discussion, different types of data have different methods for applying random k anonymous. There are two different methods to implement random k anonymous. The first method is to add uniform noise to numerical data, which makes more than k records indistinguishable and will be discussed in section 3.3. The second method is mainly for non-numerical data. The first step of the second method is to establish a generalizing tree, and find the anonymous equivalence subclass. The second step is to

output the value in the anonymous equivalence subclass with the same probability. The second method will be discussed in detail in section 3.4. Before giving the details of the two methods, we first give out the overall process of our method as **Algorithm 1**, and the execution flow of algorithm 1 and the sub-algorithm processes are shown in Figure 3. The key notations which will be used in the following section are summarized in Table 1.

Algorithm 1 The Entire Anonymous Transformation

Input: the original data set D and value of k
Output: the anonymous data set D'
Procedure:
1. $D' = \emptyset$, run the algorithm 2, divide the data set D into subclass $D = \{D_1, D_2, \ldots, D_c\}$
2. for all $D_i \in \{D_1, D_2, \ldots, D_c\}$
3. $E' = \emptyset$, run algorithm 3 on date set D_i and get
4. $E = \{E_1, E_2, \ldots, E_m\}$, and $k \leq |E_m| < 2 * k$ for all $E_j \in \{E_1, E_2, \ldots, E_m\}$
5. run Algorithm 4 or Algorithm 5 on E_j and get E'_j
6. add all elements of E'_j to E'
7. endfor
8. add all elements of E' to D'
9. endfor
10. return D'

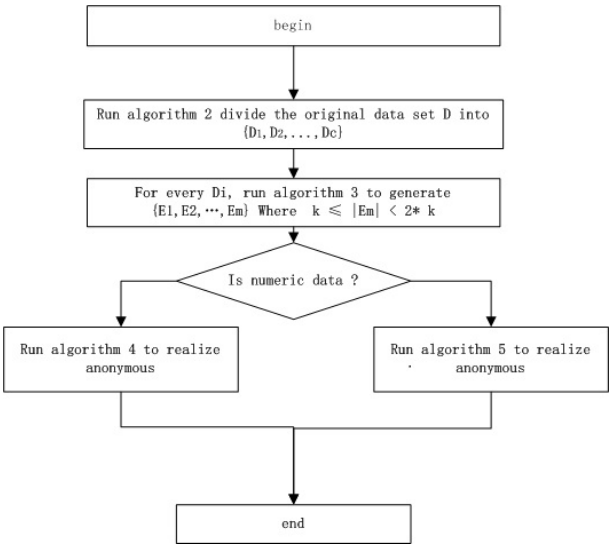

FIGURE 3. The workflow diagram of algorithm 1.

Algorithm 2, 3, 4 and 5 will be discussed in the following section. Algorithm 2 and Algorithm 3 give a method for efficiently searching for anonymous equivalence classes. Algorithm 4 and Algorithm 5 respectively give anonymous methods for numerical data and non-numeric data. As shown in algorithm 1, Line 1 by running Algorithm 2, the original data set is divided into several sub-data sets, so that the number of elements in each sub-data set is much smaller than the number

TABLE 1. Key notations.

Notations	Definition
D	The original data set
D'	The anonymous data set of D
D_i	The i-th partition of D
D'_i	The anonymous data set of D_i
E_j	The j-th partition of D_i
E'_j	The anonymous data set of E_j
e_i	The i-th element
$e_i.A_k$	The k-th attribute of element e_i
e_t^j	The t-th element of E_j
$e_t^{j'}$	the anonymous element of e_t^j
$dist1(e_1, e_2)$	Distance between e_1, e_2(numeric attribute)
$dist2(e_1, e_2)$	Distance between e_1, e_2(no-numeric attribute)
$dist(e_1, e_2)$	Distance between e_1, e_2(all attributes)
$dists(D_1, D_2)$	Distance between data set $D_1 and D_2$
$level(e.A_i)$	The level of $e.A_i$ in VGHT
$Pr(e)$	Probability of output e

Algorithm 2 Divide the Original Data Set Into Small Data Sets

Input: D: the original data set; c: the number of clusters
Output: $D_1, D_2, \ldots, D_c$: is the partition of D
Procedure:
1. randomly select c elements from D as the initial clustering center
2. repeat
3. for each tuple in D do
4. calculate the distance between each record and each cluster center
5. classify the record into the nearest cluster
6. endfor
7. update each cluster center
8. until the cluster center does not change or the number of iterations exceeds a certain value
9. return $D_1, D_2, \ldots, D_c$

of elements in the original data set, thereby greatly reducing the cost of finding equivalence classes. Line 3 generates the anonymous equivalence classes that satisfy the requirements on each sub-dataset, and in the line 5, the method of Algorithm 4 or Algorithm 5 is used to achieve anonymity of the data according to the data type in the equivalence class.

A. DISTANCE MEASUREMENT

In Algorithm 2 and Algorithm 3, the first step is to divide the original data set into several small subsets. After dividing, the records in the subset should be as similar as possible. Therefore, it is necessary to find a reasonable similarity measure method to evaluate the similarity between them. In this paper, distance is used as an assessment of similarity between two elements. For numerical attributes, Euclidean distance is used, and for non-numerical attributes, generalized distance is used.

Definition 6 (Distance Between Numerical Data): For element e_1, e_2, A_i is the ith attribute, and s is the number of attributes. The distance between e_1 and e_2 is defined as

Algorithm 3 Generate Equivalent Class Containing Elements More Than k and Less Than $2 * k$

Input: dataset D_l generated by algorithm 2; k: the k-anonymity constraint;
Output: $E' = \{E_1, E_2, \ldots, E_c\}$ is the partition of D_l
Procedure:
1. $E' = \emptyset$
2. $E = \{E_1, E_2, \ldots E_m\}$ is a partition of D_t(the records in the same E_i have the same QI values)
3. for all E_i in E do
4. if ($|E_i| \geq k$) than
5. add E_i to E', remove E_i from E
6. end if
7. end for
8. repeat
9. randomly choose E_i from E
10. calculate the distance between E_i and other sub-classes in E' and E
11. find out the closest sub-class E_j to E_i
12. if $|E_j| + |E_i| < k$
13. merge E_i and E_j into E^*
14. remove E_i, E_j, add E^* to E
15. else if $|E_j| + |E_i| < 2 * k$
16. merge E_i and E_j into E^*
17. remove E_i, E_j from E, add $E^* to E'$
18. else
19. remove $k - |E_i|$ elements from E_j to E_i
20. remove E_i from E
21. add E_i to E'
22. endif
23. until $E = \emptyset$
24. return E'

Algorithm 4 Anonymous Method for Numeric Data

Input: E_i: a data set generated by Algorithm 3
Output: E_i': anonymous of E_i
Procedure:
1. $a = max(e_k^i)$, $b = min(e_k^i)$, $k \in [1, |E_i|]$
2. for $j = 1$ to $|E_i|$
3. n is a random number that is uniformly distributed between intervals $[-\frac{a-b}{2}, +\frac{a-b}{2}]$
4. $e_j^{i'} = e_j^i + n$, $e^i{}_j$ represents the jth record in E_i
5. $if(e_j^{i'} > a)$
6. $e_j^{i'} = e_j^{i'} - (a - b)$
7. $if(e_j^{i'} < a)$
8. $e_j^{i'} = e_j^{i'} + (a - b)$
9. add $e_j^{i'}$ to E_i'
10. endfor
11. return E_i'

following:

$$dist1(e_1, e_2) = \sqrt{\sum_{i=1}^{s} (e_1.A_i - e_2.A_i)^2} \tag{2}$$

Algorithm 5 Random k-Anonymous Method of Non-Numeric Data

Input: E_i a dataset generated by algorithm 3, k: a number
Output: E_i' the anonymous data set of E_i
Procedure:
1. $E_i' = \emptyset$
2. all record $e_j^i, j = 1 \ldots k \ldots$ in data set E_i is generalized to e_a^i according to VGHT
3. replace all the record e_j^i ($j = 1 \ldots k \ldots$) in E_i with different elements of E_a with the same probability. (all the child nodes of e_a^i, which are leaf node of VGHT, are included in E_a)
4. move all the records from E_i to E_i'(the records have been anonymized)
5. return E_i'

For numerical attributes, the Euclidean function is used because it is one of the most widely used distance functions. Other distance function can also be used, such as cosine distance. However, algorithm 2 is a c-prototypes algorithm, which needs to calculate the mean value of the records. If other distances are used, the mean value will be meaningless. Here we focus on privacy protection methods. The discussion of which distance is more appropriate is beyond the scope of this paper.

Definition 7 (Distance Between Non-Numerical Attributes): For element e_1, e_2, A_i is the ith attribute, and s is the number of attributes. The distance between e_1 and e_2 is defined as:

$$dist2(e_1, e_2) = \sum_{i=1}^{s} \frac{w_{A_i} * (level(e'.A_i) - 1)}{h(VGHT_A)} \tag{3}$$

where $VGHT_{A_i}$ is the value generalization hierarchies tree (as shown in Figure 4), and $h(VGHT_{A_i})$ is the height of the tree. W_{A_i} represents the weight of attribute A_i, and e' is the first common generalization ancestor of e_1 and e_2.

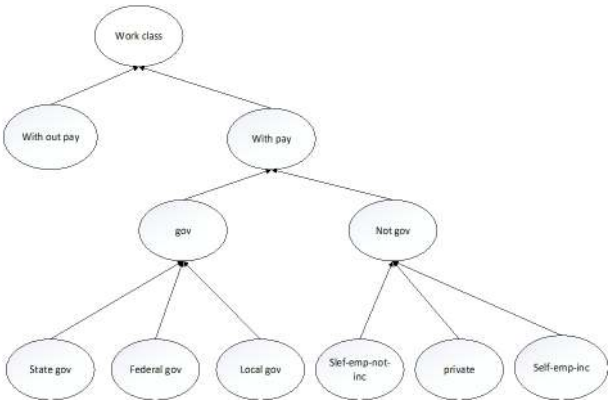

FIGURE 4. Value generalization hierarchies tree(VGHT).

For example, for the records e_1, e_2, the value of attribute A_i are 'State gov' and 'private', respectively, and the distance

between them is defined as follows:

$$\frac{level(e'.A_i)-1}{h(VGHT_{A_i})}=\frac{3-1}{4}=0.5 \tag{4}$$

In particular, if the non-numerical data does not have a VGHT structure, the user should construct a VGHT structure manually. For example, for the attributes ''male'' and ''female'', the user can construct a VGHT as shown in Figure 6, and the distance between them is (2-1)/2 = 0.5.

The records in the table often contain numerical attributes and non-numerical attributes. As a distance between records containing both numerical attributes and non-numerical attributes, a comprehensive measurement method is needed. The measurement method needs to consider numerical attributes and non-numerical attributes [31]. Here we use the mixed attribute distance, and the definition is defined as follow.

Definition 8 (Mixed Attribute Distance): Records e_1 and e_2 in data set D have both numerical and non-numerical attributes. The distance between e_1 and e_2 is defined as:

$$dist(e_1,e_2)=dist1(e_1,e_2)+f(dist2(e_1,e_2)) \tag{5}$$

Function f is used to adjust the influence of the number of non-numerical and numerical attributes, and the definition is as follow:

$$f(dist2)=\frac{N2}{N1}(dist1_{min}+\frac{dist1_{max}-dist1_{min}}{dist2_{max}-dist2_{min}}(dist2-dist2_{min})) \tag{6}$$

where $N1$ represents the number of numerical attributes, $N2$ represents the number of non-numerical attributes, $dist1_{min}$ and $dist2_{max}$ represent the minimum and maximum values of the numerical attribute distance, respectively. And $dist2_{min}, dist2_{max}$ represent the minimum and maximum values of the non-numerical attribute distance respectively.

In the process of generating an equivalent anonymous class, it is necessary to merge the sub-categories that do not meet the requirements. Therefore, the distance between two data sets is needed to be defined. The definition is as follow.

Definition 9 (Distance Between Two Data Sets): D_1, D_2 are two subsets of the original data set, $e_i \in D_1, e_j \in D_2$, then the distance of D_1, D_2 is defined as:

$$dists(D_1,D_2)=\frac{1}{2}\sum_{i=1}^{|D_1|}\sum_{j=1}^{|D_2|}dist(e_i,e_j) \tag{7}$$

where $|D_1|$ and $|D_2|$ represent the number of records in D_1 and D_2 respectively.

B. DIVIDING ORIGINAL DATA SET INTO SUB-CLASS

In order to achieve k-anonymous, the first work we should do is dividing the original data set into equivalent anonymous classes. Generating equivalent anonymous classes is the most time consuming process [29], [30], so the following focuses on the methods of efficiently generating equivalent anonymous classes.

In order to efficiently generate equivalent anonymous classes, this paper adopts a two-step clustering method. The first step is to divide the original data set into several different subsets, and the elements in the subset are as similar as possible (see Algorithm 2 for details). The second step is to form equivalence class on the same subset (see Algorithm 3 for details).

Algorithm 2 is used to divide the data set into several subsets. It is a c-prototypes algorithm [14], [15]. The original data set is divided according to the principle of minimizing the distance within the cluster. The detail will be shown in the following. The number of initial clustering centers is c, and how to determine the value of c will be discussed in section 4.2. Line 1 randomly select c initial clustering centers. Line 4-6 add other elements to different clusters according to the distance between the elements and the cluster centers, and then update the cluster centers. The process is repeated until the cluster center does not change or the number of iterations exceeds a certain value. Line 9 outputs the clustering result.

Algorithm 2 returns the cluster containing the number of elements much smaller than the number of elements in the original data set, thereby greatly reducing the cost of generating equivalent anonymous classes.

The anonymous equivalence class is generated by algorithm 3, and the number of records of the equivalence class must be larger than k and smaller than $2k$. The details of algorithm 3 are shown in the following section. The input data set of algorithm 3 is generated by algorithm 2. The elements of E' are the equivalent classes that satisfy k anonymity requirements. Line 2 classifies elements with the same quasi-identifier(QI) into one sub-class. The following operations merge these subclasses to obtain an anonymous equivalence class that satisfies the requirements. Lines 3-7, if the sub-class has more than k elements that with the same quasi-identifiers, this means that the requirement has been met, and the sub-class is directly added to E'. Lines 8-23, dealing with the case where the number of elements in the equivalence class is less than k. E_i is randomly selected in E, and the distance between E_i and other subsets included in E or E' is calculated, and the subset E_j closest to E_i is found. If $|E_j| + |E_i| < k$, this means that E_j is an element of E, and the new set E^* obtained by combining E_i and E_j still cannot meet the requirement of k anonymity, so E^* is added to E(lines 2-14). If $k < |E_j| + |E_i| < 2 * k$, the new cluster E^* obtained by combining E_i and E_j has satisfy the requirement of k anonymity, and this means that E_j may in E or in E', so remove E_j and E_i from E or E', and add E^* to E' (lines 15-17). If $|E_j| + |E_i| > 2 * k$, E_j must be an element of E'. If we move $(k-|E_i|)$ elements of E_j to E_i, then E_i and E_j are both satisfy the requirement of k anonymity. Then remove E_i from E' and add E_i to E'(lines 18-21). The elements of E' are all sub-class, and the number of records in each sub-class is greater than k less than 2*k. Finally E' was returned.

C. RANDOM K-ANONYMOUS METHOD FOR NUMERICAL DATA

Similar to the Laplace mechanism in differential privacy, random k-anonymous can be achieved by adding noise for numerical data. First, the whole data set is divided into several equivalent anonymous classes. The number of records in each anonymous equivalence class is larger than k less than $2 * k$, and the data in each equivalent class is as similar as possible. The range of the equivalent anonymous class is R. n is a random noise subject to uniform distribution with the mean 0. Finally n is added to the element of the equivalent anonymous class to achieve random k-anonymous. The detail of the algorithm is described in Algorithm 4. Line 1 calculates the maximum and minimum. Lines 3-4 produces the required noise, and adds it to the data set. Lines 5-8 ensure that the range of anonymous data set is equal to the range of the original data set, thus improves the utility of the anonymous data set.

The principle of Algorithm 4 is similar to the game of dialing the clock. If the hand is currently pointing to the 12 o'clock position. Select an angle r in [0,360] with the same probability, dial the hand r degree, then the probability that the pointer points to each moment on the dial is equal. Conversely, if it is known that the hand is now pointing to the 12 o'clock position, which is obtained by rotating r degree from a certain position, r is a random variable that is uniformly distributed between 0 and 360. The probability that according to the current position to guess out the position it originally pointed to is equal.

In order to verify the effect of Algorithm 4, we conducted an experiment. 100 artificial records which follows Normal Distribution $N(8, 2)$ were generated, and the uniform noises between -1 and 1 were added to the original record. The result was shown in Figure 5, where we can see that even we get value of the noisy record we also can not guess out which original record it was generated by confidently.

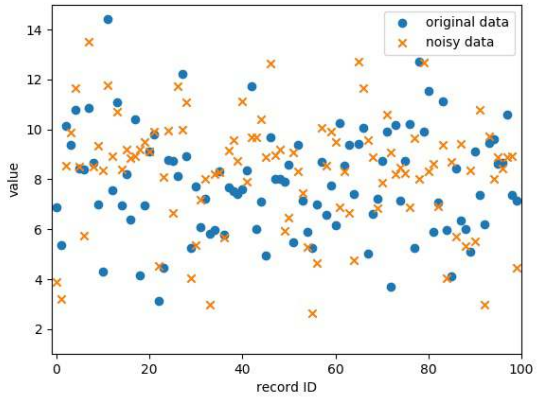

FIGURE 5. Records before and after adding noise.

Algorithm 4 satisfies the definition of random k-anonymous in definition 5. The following is a simple proof.

Proof: $e_i^{j'}$ is a noisy record in the output data set E, and e_i^j is the corresponding original record in data set E, n is a random number that is uniformly distributed in $[-\frac{a-b}{2}, +\frac{a-b}{2}]$.

$$Pr(e_i^j) = Pr(e_i^j = e_i^{j'} - n) = Pr(n = e_i^{j'} - e_i^j) = Pr(n) \quad (8)$$

There are at least k records in E, so the algorithm 4 satisfies the definition 5. □

The random k-anonymous which was implemented by Algorithm 4 has the following properties:

Property 1: the means of the anonymous data set and the original data set are equal.

Property 2: the ranges of the anonymous data set and the original data set are equal.

Property 3: the probability that two records have the same quasi identifiers in the same anonymous data set is almost zero.

The noise n is a random variable that is uniformly distributed in the interval $[-\frac{a-b}{2}, +\frac{a-b}{2}]$. The mean value of n is zero, and the mean value is additive. So the property 1 is obtained. Regardless of the original value of quasi identifier, the probability that there are two elements with the same quasi identifier is obviously low. The operation of line7-11 ensures that the range of E' and E is equal.

Property 1 and 2 ensure that the data set E' is as similar as possible to the data set E, so that the data set E' is highly available. Due to the property 3, the shortcomings which were overcome by l-diversity and t-closeness do not exist in this algorithm.

D. RANDOM K-ANONYMOUS METHOD FOR NON-NUMERICAL DATA

Compared with numerical attributes, non-numerical attributes have different properties. Random k-anonymous cannot be realized by adding uniform noise directly. Referring to the generalization method in traditional k-anonymous and the exponential mechanism in differential privacy, random k-anonymous is realized by the method of generalization first and then randomization(as shown in Figure 6). There are two main steps. The first step is to achieve k-anonymous through

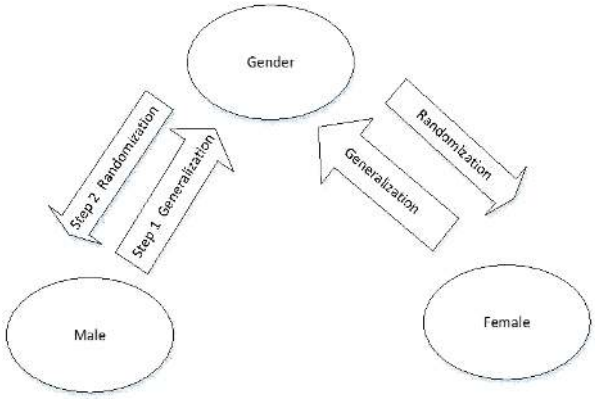

FIGURE 6. Non-numeric anonymous.

generalization, such as 'male' and 'female' can be generalized to 'gender unknown'. The second step is to randomize the generalized results, such as randomly outputting 'male' or 'female' with the same probability for 'gender unknown'. This process does not increase the risk of privacy breaches. The result of the randomized data is included in the value range of the generalization result. For example, even if the value of gender is 'gender unknown' after the generalization, the users know that the value of gender corresponding to the record is 'male' or 'female'. that is to say, users get the same information before and after randomization. If the output value of gender is unknown, the users know that the data has been anonymized by the data owner, but if 'male' or 'female' is output, the distribution of anonymous data set is closer to the original data set. The detail of this method is shown in Algorithm 5. Line 2, the records in the equivalence class are generalized(the generalization process is shown in Figure 1, Figure 2), and the different values in the equivalence class are generalized into the same ancestor. Line 3, the records are replaced randomly with the corresponding records with the same probability. e_a^i is the first common ancestor of all the records in E_i.

In order to explain the algorithm clearly, an example was proposed. e_1 and e_2 are two records in the same anonymous data set. One of the attribute is 'Occupation', and the values of $e_1.Occupation$ and $e_2.Occupation$ are 'private' and 'state gov' respectively. The VGHT is shown as Figure 4. The first common generalization ancestor of $e_1.Occupation$ and $e_2.Occupation$ is 'with pay'. 'with pay' contains six possible values, which are 'state gov', 'federal gov', 'local gov', 'self-emp-not-inc', 'private' and 'self-emp-inc'. In order to achieve anonymity the records e_1 and e_2, the 'Occupation' attributes of e_1 and e_2 are set to 'state gov', 'federal gov', 'local gov', 'self-emp-not-inc', 'private' or 'self-emp-inc' with the same probability of 1/6 respectively.

Compared with the traditional method of k-anonymous, Algorithm 5 has a randomized process. The traditional generalization method will increase the value range of the attribute. For example, if an anonymous group only contains 'male' and 'female', the attribute value 'unknown gender' will appear after generalization. The randomization process eliminates the new attribute value generated by the generalization, making the value range of the attribute in the randomized data be equal to the value range of the attribute in the original data set. Therefore, the anonymous data is more similar to the original data set, and the availability is higher.

In Algorithm 1, Algorithm 2 is called in line 1. Algorithm 2 is similar to k-means, and the time complexity is $O(c*n*t)$, where c is the initial number of divisions, n is the number of elements, and t is the number of iterations [28]. Original data set D is divided into $\{D_1, D_2, \ldots, D_c\}$. On average, each subset contains n/c elements. The complexity from line 3 to line 9 is $O((n/c)^2)$, and the time complexity from line 2 to line 10 is $O((n/c)^2)*c$. The total time complexity of algorithm 1 is $O(c*n*t)+O((n/c)^2)*c = O(c*n*t+n^2/c)$. Since c and t is much smaller than n, the time complexity of Algorithm 1 is $O(n^2)$.

IV. EXPERIMENTS AND DISCUSSION

A. EXPERIMENTS DESIGN

The experiment uses the Adult Data Set, the Default of Credit Card Clients Data Set(Credit Data Set) and the Iris Data Set respectively, which can be downloaded from the related website of UCI machine learning (http://archive.ics.uci.edu/ml/index.php). The Adult Data Set and the Credit Data Set contain both numeric and non-numeric attributes, and the Iris data set only contains numeric attributes. All the data set have been used many times in research fields such as machine learning, pattern recognition and privacy protection [2], [15], [16], [25].

As the records with unknown values are unlikely to reveal private information, the tuples with unknown values are eliminated. There are 45222 records without unknown values in Adult data set. Each record has 14 attributes. We use 8 attributes of the tuples including 4 numerical attributes and 4 non-numerical attributes, as shown in Table 2. These eight attributes are often used to infer user privacy, while the other 6 attributes are sensitive information that need to be protected.

TABLE 2. Description of the adult dataset.

NO.	Attribute	Type
1	Age	Numeric
2	Fhlweigh	Numeric
3	Education num	Numeric
4	Hours per week	Numeric
5	Work class	Categorical
6	Marital status	Categorical
7	Race	Categorical
8	Gender	Categorical

There are 30000 instances in the Credit Data Set. Six attributes, including three non-numerical attributes and three numerical attributes, are used in the experiments. The details of the attributes we used is shown in the Table 3. In the Table 3 record4-6 are three payment records.

TABLE 3. Description of the credit dataset.

NO.	Attribute	Type
1	sex	Categorical
2	Education	Categorical
3	Marriage	Categorical
4	Record1	Numeric
5	Record2	Numeric
6	Record3	Numeric

In the Iris data set, there are 150 records, including three types of data, each of which contains 5 attributes. The last attribute is a category identifier, and the first four attributes are numerical attributes. In the experiment, we used the first four attributes as the quasi-identifiers, as shown in Table 4.

TABLE 4. Description of the Iris dataset.

NO.	Attribute	Type
1	Sepal length	Numeric
2	Sepal width	Numeric
3	Petal length	Numeric
4	Petal width	Numeric

The information loss is measured by the Kullback-Leible Rdivergence Distance(KLD) between the original data set and the anonymized data set. KLD. distance is also called relative entropy [17], The definition is as follows:

$$D(p \parallel q) = -\sum_{x \in X} p(x) log \frac{q(x)}{p(x)} \tag{9}$$

where $p(x)$ and $q(x)$ are the frequencies of x in anonymous dataset and original dataset.

The greater the relative entropy, the greater the difference between the two distributions, which means that the amount of information loss is greater. In this paper, a novel k-anonymous is proposed. The strength of k-anonymous for privacy protection depends only on the value of k. So if the k-anonymous is achieved, the less information is lost, the better the performance of the method is.

B. EXPERIMENT ANALYSIS

1) DETERMINATION OF THE NUMBER OF INITIAL PARTITION

Using algorithms 2 and 3 can find equivalent anonymous classes more efficiently. Algorithm 2 and Algorithm 3 have no effect on improving the strength of privacy protection, however the initial partition number c has a great influence on the efficiency of Algorithm 1. As discussed in section 3, the time complexity of algorithm 5 is $O(c*n*t)+O((n/c)^2)*c = O(c*n*t+n^2/c)$. In order to minimize $f = c*n*t+n^2/c$, $f'(c) = n*t-n^2/c^2$ should be 0, so $c = (n/t)^{1/2}$.

2) TIME EFFICIENCY ANALYSIS

As the data set Iris is small, when verifies the efficiency of the algorithm, the experiment is only performed on the Adult Data Set and the Credit Data Set. Traditional k-anonymous and random k-anonymous methods are performed on the Adult Data Set and the Credit Data Set respectively. The experimental results are shown as Figure 7 and Figure 8. As can be seen from Figure 7 and Figure 8, most of the time, when implementing k-anonymous on both data set, the random k-anonymous method takes less time than the traditional k-anonymous method, and the difference is not too large. This demonstrates that the random k-anonymous method is more efficient than the traditional k-anonymous method, and this superiority is affected by the data set itself.

3) THE LOSS OF AVAILABILITY

We did comparative experiments on the Adult Data Set, the Credit Data Set and the Iris Data Set to analysis the loss of availability. The KLD. distances of different attributes are calculated respectively. The larger the value of KLD. the

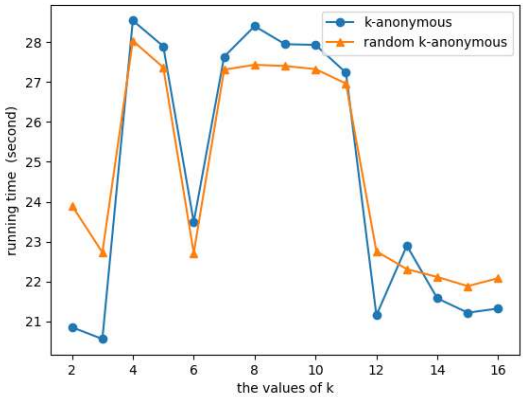

FIGURE 7. Run time with different k values on the Adult Data Set.

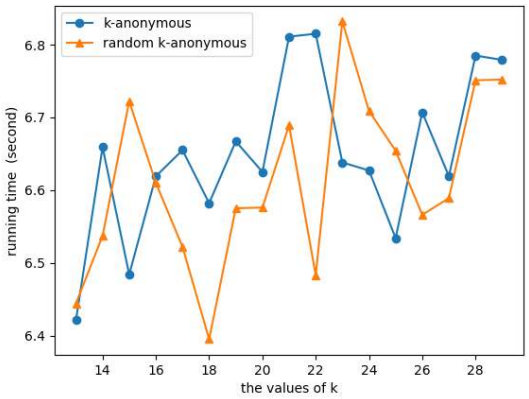

FIGURE 8. Run time with different k values on the Credit Data Set.

greater the difference between the distribution of the anonymous data set and the original data set, and information loss will be more. The results are shown in Figure 1-11. Figure 9 compares the loss of availability of the Credit Data Set. Figure 10 and Figure 11 compares the loss of availability of four non-numerical attribute information and four numerical attributes of the Adult Data Set. From Figure 9 and Figure 10, it can be seen that, for non-numerical attributes, the random k-anonymous method is obviously superior to the traditional k-anonymous method, and that with the increase of k, the advantage of random k-anonymous will be more obvious than that of the traditional k-anonymous. Figure 9 and

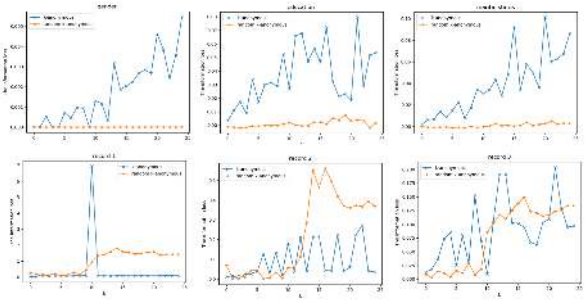

FIGURE 9. The information loss on the Credit Data Set.

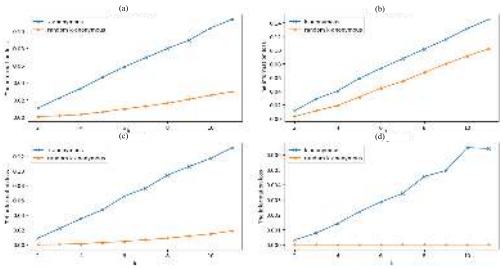

FIGURE 10. The information loss of non-numeric value on the Adult Data. Set.

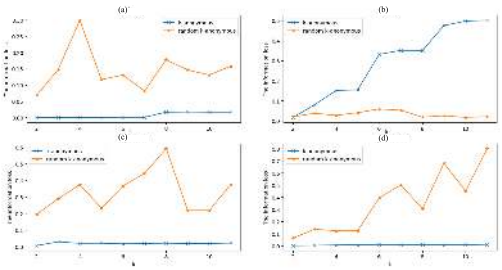

FIGURE 11. The information loss of numeric value on the Adult Data Set.

Figure 11 show that for numerical attributes, the superiority of random k-anonymous method is less obvious than the traditional k-anonymous method. For the Adult Data Set, Among the four attributes, the attribute 'flh' is the only one that the performance of random k-anonymous better than the traditional k-anonymous. For the Credit Data Set, only while k is smaller, the random k-anonymous is superior than traditional k-anonymous. This means that for numerical attributes, our method is sometimes superior to traditional methods, and sometimes inferior to traditional methods. This indicates that the performance of random k-anonymous method on the numeric data set is somewhat affected by the characteristics of the data itself.

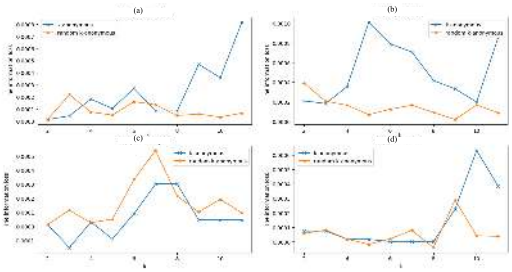

FIGURE 12. Information loss of iris dataset.

In order to further verify the applicability of random k-anonymous to numerical properties, we conducted comparative experiments on the Iris Data Set which have numerical property only. The experimental result is shown in Figure 12. For the attributes 'sepal width', 'sepal length', and 'petal width', random k-anonymous is obviously superior to the traditional k-anonymous method. Only the 'petal length' attribute is not as good as the traditional k-anonymous method. The ranges of attribute 'sepal length', 'sepal width', 'petal length' and 'petal width' are 3.6, 2.4, 5.9 and 2.4, respectively. But the range of 'petal length' is larger than the other three attributes, and this is the reason why our method does not perform well on the 'petal length' attribute. In general, on the Iris Data Set, the performance of random k-anonymous method is better than the traditional k-anonymous method.

C. DISCUSSION

From the above experimental results and analysis, we know that random k-anonymous method can provide better privacy protection, especially for non-numerical attributes, which is more efficient and less interfering with the original data. The performance of random k-anonymous method on numeric attributes depends on the data set itself. For the three numerical attributes of the Credit Data Set, the performance of this method is nearly similar to that of the traditional method, and the information loss is slightly higher than the traditional method. In the Adult Data Set, there is only one numerical attributes that the performance of random k-anonymous is better than the traditional method. The Iris Data Set has four numerical attributes. The performance of our method is better than the traditional one on three attributes. The reason for the different performance is that the ranges of numerical attributes in the three data sets are different. In the Adult Data Set, the maximum range of numerical attributes is 1476908. But the maximum range of numerical attributes in the Iris Data Set is only 5.9. So Random k-anonymous is not suitable for privacy protection of numeric attributes with very large range.

V. RELATED WORK

Researchers have done a lot of studies on privacy protection. There are three fields of work with seemingly different goals. The first field is about privacy by policy. The second field is about privacy by statistics. The third field is about privacy by cryptography [33]. Access control model is a privacy protection methods belong to the first field. [33] use a fully homomorphic encryption scheme to protect privacy. In this paper, The method we proposed belongs to the second field.

K-anonymous and differential privacy are two important privacy protection methods. K-anonymous was proposed by Samarati and Sweeney [6]. There are at least k elements with the same quasi-identifier. It can prevent the disclosure of quasi-identifiers, however it does not provide protection for the sensitive attributes. For instance, more than k elements in the anonymous equivalence class have the same quasi-identifier, but the sensitive attribute has the same value, such as 'cancer'. The attacker can still know that the users in this anonymous class are cancer patients. This shortcoming was overcome by l-diversity proposed Machanavajjhala *et al.* [7]. L-diversity requires not only that at least k records in the equivalent class have the same quasi-identifier, but also that sensitive information has at least l different values. t-closeness proposed in [9] requires that the distribution of sensitive attributes in the anonymous data set should be as close as possible to the distribution of sensitive attributes

in the whole data set. Both l-diversity and t-closeness can prevent the leakage of sensitive attribute. In [6], [7], and [9], k-anonymous can be achieved by generalization and suppression, but the authors did not give a way to find equivalence class efficiently. Micro aggregation method was used in [34], but if the data set is large, the computational cost will be high. In this paper, we propose a two-step clustering method, which can find equivalent classes efficiently, even if the original dataset is very large.

The k-anonymous methods proposed in [6], [7], [9], and [34] all require that the quasi-identifiers have the same value, which is the reason for the success of some attacks, such as exhaustive attacks. In this paper, the shortcomings were overcome by adding noise and randomization which were always used in differential privacy.

Researchers had done a lot of work on differential privacy. The optimal noise was introduced in [19]. Gaussian noise was used in [20]. Hsu *et al.* gave an economic method for choosing the value of epsilon in [21]. In [22], the global sensitivity was replaced by smooth sensitivity. Li *et al.* pointed out that k-anonymous, when preceded with a random sampling step, satisfied differential privacy in [23]. [24] pointed out that if a data set had been achieved k-anonymous, a differential privacy transformation was performed on the confidential attribute, then the result satisfied t-closeness. Although many achievements had been made in the research field of differential privacy. There are only two ways to realize differential privacy. One is by adding noise, the other is by randomization, and the both are all used to achieve k-anonymous in this paper.

Compared with the traditional methods, the method proposed in this paper mainly has two different points. First, this paper gives a method to quickly find anonymous classes, which greatly improves the efficiency. Second, by adding noise and randomization, the same quasi-identifier does not exist in the anonymized data set, which can effectively resist the exhaustive attack.

VI. CONCLUSION

A novel method is proposed in this paper. First of all, the original data set is divided into several relatively small data sets, and then the anonymous equivalence classes are found in the subset, which greatly reduces the computational cost of finding anonymous classes, and thus greatly improves the implementation efficiency of k-anonymous. Then, k-anonymous is realized by adding noise. As far as we know, it is the first time that k-anonymous is realized by adding noise. In this way, on the one hand, it can reduce the information loss of anonymous data sets. On the other hand, it avoids the problem that at least k records have the same quasi identifiers in the anonymized data set. So this method of anonymity can resist homogeneity attack, background attack and exhaustive attack.

REFERENCES

[1] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *J. Law, Med. Ethics*, vol. 25, nos. 2–3, pp. 98–110, 1997. doi: 10.1111/j.1748-720X.1997.tb01885.x.

[2] L. Sweeney, "Achieving K-anonymity privacy protection using generalization and suppression," *Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, 2002. doi: 10.1142/S021848850200165X.

[3] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proc. Int. Conf. Very Large Data Bases*, Trondheim, Norway, Aug./Sep. 2005, pp. 901–909.

[4] C. Dwork, "Differential privacy," *Lect. Notes Comput. Sci.*, vol. 26, no. 2, pp. 1–12, 2006. doi: 10.1007/11787006_1.

[5] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov./Dec. 2001. doi: 10.1109/69.971193.

[6] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in *Proc. PODS*, Jun. 1998, p. 188. doi: 10.1145/275487.275508.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, Apr. 2006, p. 24. doi: 10.1109/ICDE.2006.1.

[8] G. Yang, J. Li, S. Zhang, and L. Yu, "An enhanced l-diversity privacy preservation," in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discovery*, Jul. 2014, pp. 1115–1120. doi: 10.1109/FSKD.2013.6816364.

[9] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115. doi: 10.1109/ICDE.2007.367856.

[10] T. M. Truta and B. Vinay, "Privacy protection: P-sensitive k-anonymity property," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2006, p. 94. doi: 10.1109/ICDEW.2006.116.

[11] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.*, 2008, pp. 1–19. doi: 10.1007/978-3-540-79228-4_1.

[12] C. Dwork and G. N. Rothblum, "Concentrated Differential Privacy," 2016, *arXiv:1603.01887*. [Online]. Available: https://arxiv.org/abs/1603.01887

[13] F. Mcsherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2007, pp. 94–103. doi: 10.1109/FOCS.2007.66.

[14] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998. doi: 10.1023/a:1009769707641.

[15] J. Han, J. Yu, Y. Mo, J. Lu, and H. Liu, "MAGE: A semantics retaining K-anonymization method for mixed data," *Knowl.-Based Syst.*, vol. 55, pp. 75–86, Jan. 2014.

[16] G. Jennifer and E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, no. 4, pp. 845–889, Aug. 2004. [Online]. Available: http://www.jmlr.org/papers/v5/

[17] M. Li, L. Zhu, Z. Zhang, and R. Xu, "Achieving differential privacy of trajectory data publishing in participatory sensing," *Inf. Sci.*, vols. 400–401, pp. 1–13, Aug. 2017. doi: 10.1016/j.ins.2017.03.015.

[18] J. Soria-Comas and J. Domingo-Ferrer, "Optimal data-independent noise for differential privacy," *Inf. Sci.*, vol. 250, no. 11, pp. 200–214, Nov. 2013. doi: 10.1016/j.ins.2013.07.004.

[19] Q. Geng and P. Viswanath, "The optimal noise-adding mechanism in differential privacy," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 925–951, Feb. 2016. doi: 10.1109/tit.2015.2504967.

[20] F. Liu, "Generalized Gaussian mechanism for differential privacy," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 747–756, Apr. 2019. doi: 10.1109/TKDE.2018.2845388.

[21] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth, "Differential privacy: An economic method for choosing epsilon," in *Proc. IEEE 27th Comput. Secur. Found. Symp.*, Jul. 2014, pp. 398–410. doi: 10.1109/CSF.2014.35.

[22] K. Nissim and S. Raskhodnikova, and S. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proc. 39th Annu. ACM Symp. Theory Comput.*, Jun. 2007, pp. 75–84. doi: 10.1145/1250790.1250803.

[23] N. Li, W. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy," in *Proc. 7th ACM Symp. Inf., Comput. Commun. Secur.*, May 2012, pp. 32–33.

[24] J. Domingo-Ferrer and J. Soria-Comas, "From ιcloseness to differential privacy and vice versa in data anonymizatio," *Knowl.-Based Syst.*, vol. 74, no. 1, pp. 151–158, Jan. 2015. doi: 10.1016/j.knosys.2014.11.011.

[25] I.-C. Yen and C.-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2473–2480, Mar. 2009. doi: 10.1016/j.eswa.2007.12.020.

[26] T. Ma, J. Jia, y. xue, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, "Protection of location privacy for moving κNN queries in social networks," *Appl. Soft Comput.*, vol. 66, pp. 525–532, May 2018.
[27] T. Ma, W. Shao, Y. Hao, and J. Cao, "Graph Classification Based on Graph Set Reconstruction and Graph Kernel Feature Reduction," *Neurocomputing*, vol. 296, pp. 33–45, Jun. 2018. doi: 10.1016/j.neucom.2018.03.029.
[28] H. Rong, T. Ma, J. Cao, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, "Deep rolling: A novel emotion prediction model for a multi-participant communication context," *Inf. Sci.*, vol. 488, pp. 158–180, Jul. 2019.
[29] T. Ma, Y. Zhao, H. Zhou, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, "Natural disaster topic extraction in sina microblogging based on graph analysis," *Expert Syst. Appl.*, vol. 115, pp. 346–355, Jan. 2019.
[30] T. Ma, H. Rong, C. Ying, Y. Tian, A. Al-Dhelaan, and A. M. Al-Rodhaan, "Detect structural-connected communities based on BSCHEF in C-DBLP," *Concurrency Comput., Pract. Exper.*, vol. 28, no. 2, pp. 311–330, Feb. 2016. doi: 10.1002/cpe.3437.
[31] Y. Lv, T. Ma, M. Tang, J. Cao, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaa, "An efficient and scalable density-based clustering algorithm for datasets with complex structures," *Neurocomputing*, vol. 171, pp. 9–22, Jan. 2016. doi: 10.1016/j.neucom.2015.05.109.
[32] T. Ma, Y. Zhang, J. Cao, J. Shen, M. Tang, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, "KDVEM : A k-degree anonymity with vertex and edge modification algorithm," *Computing*, vol. 97, no. 12, pp. 1165–1184, 2015. doi: 10.17485/ijst/2016/v9i12/81982.
[33] Q. Zhang, L. T. Yang, A. Castiglione, Z. Chen, and P. Li, "Secure weighted possibilistic c-means algorithm on cloud for clustering big data," *Inf. Sci.*, vol.479, pp. 515–525, Apr. 2019. doi: 10.1016/j.ins.2018.02.013.
[34] Y. Shi, Z. Zhang, H.-C. Chao, and B. Shen, "Data privacy protection based on micro aggregation with dynamic sensitive attribute updating," *Sensors*, vol. 18, no. 7, p. 2307, Jul. 2018. doi: 10.3390/s18072307.

FAGEN SONG received the bachelor's degree, in 2006 and the master's degree, in 2009. He is currently pursuing the Ph.D. degree with the Nanjing University of Information Science and Technology, China. He is a Lecturer with the School of Information and Engineering, Yancheng Institute of Technology, China. His research interests include privacy preserving, machine learning, and network information security.

TINGHUAI MA received the bachelor's and master's degrees from the Huazhong University of Science and Technology, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Chinese Academy of Science, in 2003. He was a Postdoctoral Associate with AJOU University, in 2004. From 2007 to 2008, he visited Chinese Meteorology Administration. In 2009, he was a Visiting Professor with the Ubiquitous Computing Laboratory, Kyung Hee University. He is currently a Professor in computer sciences with the Nanjing University of Information Science and Technology, China. He has authored more than 100 journal/conference papers. His research interests are data mining, cloud computing, ubiquitous computing, and privacy preserving.

YUAN TIAN received the master's and Ph.D. degrees from Kyung Hee University. She is currently an Assistant Professor with the Nanjing Institute of Technology. Her research interests include privacy and security, which are related to cloud computing, bioinformatics, multimedia, cryptograph, smart environment, and big data. She is currently a member of technical committees of several international conferences. In addition, she is an Active Reviewer of many international journals.

MZNAH AL-RODHAAN received the B.S. degree (Hon.) in computer applications and the M.S. degree in computer science from King Saud University, in 1999 and 2003, respectively, and the Ph.D. degree in computer science from the University of Glasgow, Scotland, U.K., in 2009. She is currently the Vice Chair of the Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. Her current research interests include mobile ad hoc networks, wireless sensor networks, multimedia sensor networks, cognitive networks, and network security. She has served in the editorial boards for some journals, such as the *Ad Hoc Journals* (Elsevier) and has participated in several international conferences.