Wang R, Zhu Y, Chen TS *et al.* Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 33(6): 1231–1242 Nov. 2018. DOI 10.1007/s11390-018-1884-6

Privacy-Preserving Algorithms for Multiple Sensitive Attributes Satisfying t-Closeness

Rong Wang[1], *Student Member*, *CCF*, Yan Zhu[1,*], *Member*, *CCF*, Tung-Shou Chen[2], and Chin-Chen Chang[3], *Fellow*, *IEEE*

[1] *School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China*

[2] *Department of Computer Science and Information Engineering, "National" Taichung University of Science and Technology, Taichung 404, China*

[3] *Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, China*

E-mail: wangrong.kiko@qq.com; yzhu@swjtu.edu.cn; tschen@nutc.edu.tw; alan3c@gmail.com

Received September 18, 2017; revised July 23, 2018.

Abstract Although k-anonymity is a good way of publishing microdata for research purposes, it cannot resist several common attacks, such as attribute disclosure and the similarity attack. To resist these attacks, many refinements of k-anonymity have been proposed with t-closeness being one of the strictest privacy models. While most existing t-closeness models address the case in which the original data have only one single sensitive attribute, data with multiple sensitive attributes are more common in practice. In this paper, we cover this gap with two proposed algorithms for multiple sensitive attributes and make the published data satisfy t-closeness. Based on the observation that the values of the sensitive attributes in any equivalence class must be as spread as possible over the entire data to make the published data satisfy t-closeness, both of the algorithms use different methods to partition records into groups in terms of sensitive attributes. One uses a clustering method, while the other leverages the principal component analysis. Then, according to the similarity of quasi-identifier attributes, records are selected from different groups to construct an equivalence class, which will reduce the loss of information as much as possible during anonymization. Our proposed algorithms are evaluated using a real dataset. The results show that the average speed of the first proposed algorithm is slower than that of the second proposed algorithm but the former can preserve more original information. In addition, compared with related approaches, both proposed algorithms can achieve stronger protection of privacy and reduce less.

Keywords data privacy, k-anonymity, t-closeness, multiple sensitive attribute

1 Introduction

It is common for various organizations, such as government agencies and hospitals, to release their microdata (e.g., census data or medical records) for research and other purposes[1]. However, releasing the original data unavoidably will expose the privacy of the individuals from whom the data were obtained, and this could violate privacy laws[①] and reveal sensitive personal information to malicious adversaries. Thus, before the original data are released to the public, explicit identifier attributes, such as names, addresses, and social security numbers, should be erased or concealed to protect personal privacy. However, according to one study[2], approximately 87% of the population of the United States can be identified by their 5-digit zipcode, gender, and date of birth even in the absence of explicit identifier attributes. This means that people's identities still can be disclosed even if all explicit identifier attributes are removed from the original data. There are some major privacy models for preventing such disclosures, such as k-anonymity[2,3], l-diversity[4], and t-

Regular Paper

The work was supported by the Academic and Technological Leadership Training Foundation of Sichuan Province of China under Grant Nos. WZ0100112371601/004, WZ0100112371408, and YH1500411031402.

[*] Corresponding Author

[①] https://www.hhs.gov/hipaa/index.html, Mar. 2018.

closeness[5]. In this section, first, we provide a brief overview of these models, and second we describe the problem addressed by this paper.

As one of the oldest privacy models, k-anonymity[2,3] requires that the microdata be partitioned into a set of equivalence classes each of which contains at least k records, and all records within a class be assigned the same generalized value over each of their quasi-identifier attributes. Thus, each record in a k-anonymity model cannot be identified successfully with a probability greater than $1/k$. An example of the patients' original data is presented in Table 1, and the anonymized version of the data that satisfies 3-anonymity is shown in Table 2. The Zipcode and Age attributes are set as quasi-identifier attributes. From Table 2, it can be seen that the Name attribute in Table 1 has been erased and several equivalence classes have been created, each of which has three records. All records of each equivalence class have the same values for the Zipcode and Age attributes and thus are indistinguishable based on the two attributes. However, the 3-anonymity model shown in Table 2 cannot resist the disclosure of attribute. For example, assume that Alice knows that Bob is in his twenties and Bob's record is in Table 2. When the Disease attribute is sensitive, she can conclude that Bob must have pneumonia disease.

Table 1. Patients' Original Data

No.	Name	Zipcode	Age	Disease
1	Ackerley	47506	23	Pneumonia
2	Gael	47571	26	Pneumonia
3	Rehor	47575	21	Pneumonia
4	Jerzy	47603	34	Flu
5	Cade	47614	37	Colon cancer
6	Finley	47627	30	Bronchitis
7	Eartha	47709	45	Colitis
8	Keyon	47714	50	Colon cancer
9	Selby	47736	49	Stomach cancer

Table 2. A 3-Anonymity Version of Table 1

No.	Zipcode	Age	Disease
1	475**	2*	Pneumonia
2	475**	2*	Pneumonia
3	475**	2*	Pneumonia
4	476**	3*	Flu
5	476**	3*	Colon cancer
6	476**	3*	Bronchitis
7	477**	⩾45	Colitis
8	477**	⩾45	Colon cancer
9	477**	⩾45	Stomach cancer

The l-diversity model[4] extends k-anonymity. It requires that each equivalence class has at least l different "well-represented" values for the sensitive attribute, and it also implies l-anonymity. The simplest explanation of "well-represented" would be to make sure that each equivalence class has at least l distinct values for the sensitive attribute[5]. For example, Table 3 presents another anonymized version of Table 1 that satisfies 3-diversity. In Table 3, all records within each equivalence class have the same values for the Zipcode and Age attributes but different values for the Disease attribute. In this way, an attacker cannot exactly tell what disease some patient has even if he or she knows the equivalence class that contains the patient's record. However, l-diversity does not consider the rareness of each sensitive value. Suppose that Alice can make sure that Bob's record is in the second equivalence class of Table 3. Even if she cannot tell what specific disease Bob has, Alice can conclude that Bob has a respiratory infection.

Table 3. A 3-Diversity Version of Table 1

No.	Zipcode	Age	Disease
1	47***	[20, 45]	Pneumonia
5	47***	[20, 45]	Colon cancer
7	47***	[20, 45]	Colitis
3	47***	[20, 35]	Pneumonia
4	47***	[20, 35]	Flu
6	47***	[20, 35]	Bronchitis
2	47***	[25, 50]	Pneumonia
8	47***	[25, 50]	Colon cancer
9	47***	[25, 50]	Stomach cancer

To address these limitations of the k-anonymity and l-diversity models, Li *et al.*[5] introduced the concept of t-closeness, which requires that the distribution of the sensitive attribute values within each equivalence class of indistinguishable records be similar to the distribution of the sensitive attribute values in the entire data. For example, Table 4 presents a version of Table 1 that satisfies 0.33-closeness. In addition to the attacks mentioned above, t-closeness also can protect published data against the skewness attack and the similarity attack[5]. In this paper, we focus on the t-closeness model because it has the strictest privacy guarantee among the k-anonymity-like models. Most existing algorithms for t-closeness[6-11] in the literature deal with the original data that have only one single sensitive attribute; however, data with multiple sensitive attributes are more common in practice.

In this paper, we propose two algorithms that simultaneously can anonymize the original data with multiple sensitive attributes and make the anonymized version satisfy t-closeness. The two proposed algorithms are motivated by the observation that, if the values of the sensitive attributes in each equivalence class are spread to the maximum extent possible over all of the data, there is a higher probability of minimizing the distance between the distribution of the sensitive attribute values within each equivalence class and the distribution of the sensitive attribute values in the entire data, thereby meeting t-closeness. The first algorithm partitions all records of the original data into different clusters, and the records in these clusters are similar in terms of their multiple sensitive attributes and dissimilar to the records in other clusters. Then, records that are similar in terms of their quasi-identifier attributes are selected from different clusters to generate an equivalence class, so that the loss of information caused by anonymization can be minimized. Based on the same idea of spreading the values of each sensitive attribute in all equivalence classes, the second algorithm first reduces the multiple sensitive attributes to a one-dimensional data space, and then sorts the new data in ascending order and partitions them into different groups. The second algorithm also selects the most similar records in terms of the quasi-identifier attributes from different groups to generate an equivalence class.

Table 4. A 0.33-Closeness Version of Table 1

No.	Zipcode	Age	Disease
1	47***	[20, 45]	Pneumonia
5	47***	[20, 45]	Colon cancer
7	47***	[20, 45]	Colitis
3	47***	[20, 50]	Pneumonia
6	47***	[20, 50]	Bronchitis
8	47***	[20, 50]	Colon cancer
2	47***	[25, 50]	Pneumonia
4	47***	[25, 50]	Flu
9	47***	[25, 50]	Stomach cancer

The rest of the paper is organized as follows. Some general concepts used throughout the paper are presented in Section 2, and related work is addressed in Section 3. The proposed algorithms are described in Section 4, and the experimental results are analyzed in Section 5. Section 6 presents our conclusions.

2 Background

Some general concepts and definitions in the literature are reviewed in this section because they are used throughout the paper.

Microdata can be presented in a data table in which each record(row) corresponds to one person and each column to a specific attribute. Let $\boldsymbol{T}_{N\times M}$ be microdata with N records, i.e., $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, ..., $\boldsymbol{r}_N$, and each one of the N records has M attributes, i.e., $A_1, A_2, \ldots, A_M$. According to their different degrees of openness, attributes $A_1, A_2, \ldots, A_M$ in the original data can be divided into the following four categories:

- explicit identifier attributes, which provide information that clearly identifies individuals, such as name, address, and social security number;
- quasi-identifier (QI) attributes, which are often combined to determine the identification of the individual, such as zipcode, gender, and date of birth;
- sensitive attributes (SAs), which individuals are unwilling to release to the public, such as disease, salary, and job;
- non-sensitive attributes, which can be released to the public without causing individuals any concern.

When releasing the original data to the public, the explicit attributes that identify all records should be removed, and the non-sensitive attributes can remain. Therefore, the technologies of anonymization mainly are applied to the rest of the attributes, i.e., the QI attributes and the SAs.

2.1 Definitions

Definition 1 (Equivalence Class, EC)[5]. *An EC is a set of anonymized records that have the same values for all the QI attributes, i.e., all records in each equivalence class are indistinguishable in terms of their QI attributes*[5].

Definition 2 (k-Anonymity)[2]. *A data table satisfies k-anonymity if each record in any equivalence class is indistinguishable from at least another $(k-1)$ records with respect to the QI attributes. Hence, the probability of correct identification in a k-anonymity model is, at most, $1/k$.*

Definition 3 (t-Closeness)[5]. *An EC satisfies t-closeness if the distance between the distribution of a sensitive attribute in any equivalence class and the distribution of the same attribute in the entire data table is no more than a threshold, t. A data table satisfies t-closeness if all equivalence classes in it satisfy t-closeness*[5].

The distance between two distributions commonly is calculated by the earth mover's distance (EMD)[12], but other distances also have been studied[10,11]. Intuitively, $EMD(\boldsymbol{P}, \boldsymbol{Q})$ views one distribution $\boldsymbol{P}$ as a mass of earth piles spread over a space, and the other distribution, $\boldsymbol{Q}$, is viewed as a collection of holes over the same space. $EMD(\boldsymbol{P}, \boldsymbol{Q})$ is defined as the minimum work needed to fill the holes with earth, i.e., transforming one distribution $\boldsymbol{P}$ to the other distribution $\boldsymbol{Q}$.

For a numerical SA, if $\boldsymbol{P} = (p_1, p_2, \ldots, p_m)$ and $\boldsymbol{Q} = (q_1, q_2, \ldots, q_m)$ are probability distributions over $v_1, v_2, \ldots, v_m$, where $v_i < v_j$ $(1 \leqslant i < j \leqslant m)$, $EMD(\boldsymbol{P}, \boldsymbol{Q})$ between $\boldsymbol{P}$ and $\boldsymbol{Q}$ is calculated as:

$$EMD(\boldsymbol{P}, \boldsymbol{Q}) = \frac{1}{m-1} \sum_{i=1}^{m} |\sum_{j=1}^{i} (p_j - q_j)|.$$

For a categorical SA, first, a generalization hierarchy, H, over the domain of each QI attribute should be given (by the domain expert). For example, Fig.1 presents a hierarchy for digestive diseases. To calculate the distance between $\boldsymbol{P} = (p_1, p_2, \ldots, p_m)$ and $\boldsymbol{Q} = (q_1, q_2, \ldots, q_m)$ of the same categorical domain, a recursive function of the collective extra earth that should be moved into/out of node n first should be defined as:

$$extra(n) = \begin{cases} p_i - q_i, & \text{if } n \text{ is a leaf,} \\ \sum_{c \in child(n)} extra(c), & \text{otherwise,} \end{cases}$$

where $child(n)$ is the set of all leaf nodes below node n. Further, another two functions that accumulate amounts of earth to be moved in/out for an internal node of H are defined as:

$$negExtra(n) = \sum_{c \in child(n) \wedge extra(c) < 0} |extra(c)|,$$

$$posExtra(n) = \sum_{c \in child(n) \wedge extra(c) > 0} |extra(c)|.$$

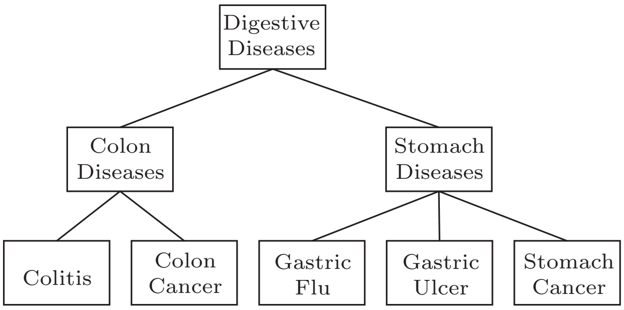


Fig.1. Example of a hierarchy of digestive diseases.

Thus, the minimum of the above quantities means the cost of all pending earth movements among the leaves under node n after their cumulative earth excess/deficit has been corrected:

$$cost(n) = \frac{h(n)}{h(H)} \times \min(posExtra(n), negExtra(n)),$$

where $h(n)$ is the height of n, and $h(H)$ is the height of H.

Then the EMD between $\boldsymbol{P}$ and $\boldsymbol{Q}$ is:

$$EMD(\boldsymbol{P}, \boldsymbol{Q}) = \sum_{n} cost(n),$$

where n is a non-leaf node in H.

2.2 Information Loss Metrics

We need an appropriate metric to measure the discrepancies between the original data and their anonymized outputs. Because a generalization method is used to anonymize the original data in the proposed algorithms, we use a generalized loss metric[8,13,14] to compute the information loss.

We denote $S = \{A_1, A_2, \ldots, A_d\}$ as a set of QI attributes. The range of the numerical attribute $NA_i \in S$ is $[L, U]$. Assume that v is one of the values of attribute NA_i and it is generalized as v' that has the range of $[L_{NA}, U_{NA}]$. Then, the information loss of v is defined as:

$$IL_{NA} = \frac{U_{NA} - L_{NA}}{U - L}.$$

For a categorical attribute $CA_i \in S$, it is assumed that H is a hierarchical tree of attribute CA_i and v is one of the values of attribute CA_i. After generalization, v becomes v', which corresponds to node n in the hierarchical tree, H. Then, the information loss of v is defined as:

$$IL_{CA} = \frac{LN_n - 1}{LN - 1},$$

where LN_n is the number of leaf nodes in the subtree of the root node, with n being the root node, and LN is the number of leaf nodes in the hierarchical tree, H.

For a record $\boldsymbol{r} \in \boldsymbol{T}$, the information loss after generalization of record $\boldsymbol{r}$ is defined as:

$$IL_{\boldsymbol{r}} = \sum_{i=1}^{d} IL_{A_i},$$

where IL_{A_i} is equal to IL_{NA} if A_i is a numerical attribute, or it is equal to IL_{CA} if A_i is a categorical attribute. As a result, the information loss of the entire data table, $\boldsymbol{T}$, after being generalized is defined as:

$$IL_{\boldsymbol{T}} = \frac{\sum_{\boldsymbol{r} \in \boldsymbol{T}} IL_{\boldsymbol{r}}}{|\boldsymbol{T}|}, \quad (1)$$

where $|\boldsymbol{T}|$ is the number of records in data table $\boldsymbol{T}$.

3 Related Work

Li *et al.*[5] first proposed the concept of t-closeness, which requires that the distance between the distribution of a sensitive attribute in any equivalence class and the distribution of the attribute in the entire data table should be no more than a threshold t. EMD[5,12] is used as a metric to measure the closeness between two distributions, and the Incognito method[15] is extended for k-anonymity to meet t-closeness. However, the method has low efficiency since its time complexity is exponential when increasing the number of QI attributes. To be more efficient, Li *et al.* proposed an improved algorithm[16] that recursively divides the combined domain of all QI attributes and carries out a split only if the resultant partitions satisfy t-closeness over the entire data table. Unfortunately, its data utility is low and it does not cater to some special features of t-closeness. In this paper, we extend the previous definition of t-closeness of a single attribute to a new definition of multiple attributes. The corresponding formula for calculating EMD also is extended to get the distance between two distributions with multiple SAs.

Cao *et al.*[8] proposed a sensitive attribute bucketization and redistribution framework for t-closeness, SABRE. First, the framework partitions the original data into a set of buckets of similar sensitive attribute values and then selects records from each bucket to generate equivalence classes. It guarantees the diversity of the values of the SAs in each equivalence class. However, the limitation of the framework is that it may create anonymized data with low quality because the buckets in SABRE are iteratively generated, which may lead to equivalence classes with more records by creating more buckets, thereby causing more losses of information. In this paper, we also partition the original data into different clusters/groups in terms of the SAs, but, to avoid generating more buckets, we use a point-assignment clustering method to get a modest number of clusters.

Similar to the idea mentioned above, Soria-Comas *et al.*[9] proposed two cluster-based algorithms using microaggregation to attain anonymized data that satisfied t-closeness. One algorithm initially generates a cluster in terms of the QI attributes and then checks to determine whether the cluster satisfies t-closeness. If that is not the case, it selects the closest record outside the cluster and swaps the record with a record in the cluster. However, it has a heavy cost resulting from the rearrangement of records required to fulfill t-closeness after the creation of each cluster. The other algorithm considers t-closeness from the very beginning by partitioning the ordered records in terms of the values of the sensitive attribute. This algorithm sorts all records first and then partitions them into different groups according to a value interval. Although the algorithm is suitable for anonymizing numerical values, it is difficult to apply it to categorical values because the ranking of categorical values is not straightforward. Because our work is designed for the original data with multiple SAs that may contain numerical and categorical attributes, one of our proposed algorithms uses the principal component analysis (PCA) to consider the properties of the two kinds of attributes comprehensively, rather than simply sorting records.

To the best of our knowledge, there are only a few papers that really make multiple SAs satisfy t-closeness because it is difficult to ensure strong closeness for every sensitive attribute. Fang *et al.*[17] introduced a method called Complete Disjoint Projections, CODIP, which deals with multiple SAs that may be multi-valued. CODIP replaces each multi-valued sensitive attribute with a mono-valued attribute first and splits all sensitive attributes into some disjoint subsets according to their associations. Then, CODIP deals with each subset, respectively. By contrast, our proposed algorithms publish the anonymous data in one table with a higher data utility. Sei *et al.*[18] assumed that several attributes have features of both QI attributes and SAs, and they proposed a privacy model that includes an anonymization algorithm. In order to satisfy t-closeness, the algorithm changes the original records with a fixed probability and adds some completely random records. Therefore, the reconstructed records are affected significantly by these random records, and the utility of the data is reduced.

4 Proposed Algorithms

Most existing t-closeness models[6-11] generate each equivalence class only by considering the QI attributes; however, keeping the focus on the QI attributes does not make it easier to refine the equivalence class to satisfy t-closeness. In this section, two algorithms are proposed to deal with multiple SAs and generate each equivalence class according to both QI attributes and SAs. We are motivated to develop these algorithms by the observation that the values of SAs in an equivalence class must be spread to the maximum extent possible over all of the data to make the class satisfy t-closeness. Based on this observation, the main aim of our proposed algorithms is to heterogenize the values of the SAs in different equivalence classes. And the more similar the QI attribute values of all records in an equivalence class are, the lower the information loss caused by anonymization should be.

Before we introduce the proposed algorithms, a definition of t-closeness of multiple SAs should be given. Note that there are some different considerations for t-closeness of multiple SAs. One approach is to consider each attribute separately, namely, if an equivalence class satisfies t-closeness, all sensitive attributes of it should satisfy t-closeness, respectively. Another is to consider the joint distribution of multiple sensitive attributes, which needs a more complicate EMD between two joint distributions. In this paper, the definition of t-closeness of multiple sensitive attributes is based on the former approach, and the latter will be handled in our future work.

Definition 4 (t-Closeness of Multiple Sensitive Attributes). *We denote EC = {QI attributes, SA_1, SA_2, ..., $SA_{M'}$} as an equivalence class with multiple SAs. The distance between the distribution of a sensitive attribute SA_i in this class and the distribution of the attribute in the entire data table is denoted as t_i ($i = 1, 2, \ldots, M'$). An $EC = \{QI\ attributes, SA_1, SA_2, \ldots, SA_{M'}\}$ satisfies t-closeness if $\max(t_1, t_2, \ldots, t_{M'}) \leqslant t$. A data table with multiple SAs satisfies t-closeness if all equivalence classes in it satisfy t-closeness.*

Because Definition 4 is defined for the multiple SAs in equivalence classes in each of which all records should have the same values for the QI attributes (see Definition 1), it is necessary to make the QI attributes of each equivalence class anonymized when satisfying the t-closeness principle.

4.1 Cluster-Based Algorithm for Multiple Sensitive Attributes Satisfying t-Closeness

The first proposed algorithm is based on a point-assignment clustering, and thus it is vital to define the similarity measure first. The distance between records is chosen to evaluate the similarity between them. Because the QI attributes or multiple SAs may contain numerical and categorical attributes, two different kinds of distance metrics are given. For a numerical attribute, first, the values of the attribute are sorted in ascending order. Let the attribute domain be $\{v_1, v_2, \ldots, v_m\}$, where v_i is the i-th smallest value. The distance between two values v_i and v_j is based on the number of values between them in the total order, and it is defined as[5]:

$$distNum(v_i, v_j) = \frac{|i-j|}{m-1}.$$

The domain hierarchy, H, is predefined for a categorical attribute. The distance between two leaf values, v_i and v_j, in H is defined as[5]:

$$distCat(v_i, v_j) = \frac{h(v_i, v_j)}{h(H)},$$

where $h(v_i, v_j)$ is the height of the lowest common ancestor node of v_i and v_j, and $h(H)$ is the height of the domain hierarchy H. As a result, the distances of two records,

$$\begin{aligned}\boldsymbol{r}_1 = (&QI_{num_{11}}, QI_{num_{12}}, \ldots, QI_{num_{1a}},\\ &QI_{cat_{11}}, QI_{cat_{12}}, \ldots, QI_{cat_{1b}},\\ &SA_{num_{11}}, SA_{num_{12}}, \ldots, SA_{num_{1c}},\\ &SA_{cat_{11}}, SA_{cat_{12}}, \ldots, SA_{cat_{1d}}),\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{r}_2 = (&QI_{num_{21}}, QI_{num_{22}}, \ldots, QI_{num_{2a}},\\ &QI_{cat_{21}}, QI_{cat_{22}}, \ldots, QI_{cat_{2b}},\\ &SA_{num_{21}}, SA_{num_{22}}, \ldots, SA_{num_{2c}},\\ &SA_{cat_{21}}, SA_{cat_{22}}, \ldots, SA_{cat_{2d}}),\end{aligned}$$

in terms of the QI attributes and multiple SAs are, respectively, defined as:

$$Dist_{QIs}(\boldsymbol{r}_1, \boldsymbol{r}_2) = \sqrt{\sum_{i=1}^{a} distNum(QI_{num_{1i}}, QI_{num_{2i}})^2 + \sum_{j=1}^{b} distCat(QI_{cat_{1j}}, QI_{cat_{2j}})^2}, \quad (2)$$

$$Dist_{SAs}(\boldsymbol{r}_1, \boldsymbol{r}_2) = \sqrt{\sum_{i=1}^{c} distNum(SA_{num_{1i}}, SA_{num_{2i}})^2 + \sum_{j=1}^{d} distCat(SA_{cat_{1j}}, SA_{cat_{2j}})^2}. \quad (3)$$

Then, an improved fuzzy c-means clustering (FCM) algorithm[19], called Equi-sized FCM, is used to partition the original records. The Equi-sized FCM algorithm yields approximately same-size clusters, with minimal sacrifice of heterogeneity between them. For a more detailed discussion of Equi-sized FCM, We refer to [19].

Based on the discussion above, the first proposed algorithm consists of three steps. First, we partition the original records into k (which is equal to the parameter k in k-anonymity) clusters of approximately the same size, so that records within the same cluster are as similar as possible to each other in terms of the multiple SAs but dissimilar to those records in other clusters. Second, we select one record from $cluster_i$ ($i = 1, 2, \ldots, k$), respectively, to generate an equivalence class and anonymize the class by generalization (see Pseudocode 1). In each selection iteration, if the size of any cluster is more than the minimum, we select one more record from it. Third, we check whether each equivalence class satisfies t-closeness; if not, we improve its level of t-closeness by merging its closest equivalence class in terms of the QI attributes (see Pseudocode 2). A brief description of this algorithm is depicted in Algorithm 1.

4.2 PCA-Based Algorithm for Multiple Sensitive Attributes Satisfying t-Closeness

Algorithm 1 proposes an approach of generating equivalence classes by selecting records from different clusters. The second algorithm (Algorithm 2) in this subsection is based on the same idea. That is because, instead of deferring the enforcement of t-closeness until later, taking the dispersion of sensitive attribute values into account at the time of the formation of equivalence classes will minimize their size and reduce the loss of information after anonymization as much as possible.

Inspired by [9], we assume that the values of the SAs can be ranked or ordered in some way and we could separate the original data into different groups based on their values. For a single sensitive attribute, there is only one ordering result according to the ascending order; however, if we consider multiple SAs separately, there will be multiple ordering results that may conflict with each other. Thus, the joint order of these multiple SAs should be specially designed. In our work, principal component analysis (PCA)[20] is used to get the joint-ordering result of multiple SAs. PCA has been used extensively as a factor analysis method, the aim of which is to change the representation of the data into a low-dimensional space while preserving the structure of the original data. For the second algorithm, the multiple sensitive attribute values are processed by PCA, and their first principal component is sorted. For example, suppose that each record that is represented by a hollow dot in Fig.2 has two SAs. After PCA processing, each hollow dot would be projected into a solid dot in a new one-dimensional space. We sort these original hollow dots according to the values of their corresponding solid dots.

Pseudocode 1 . *GenerateEquivalenceClasses*(C, k)

```
Input:
    $C$: clusters of the original data  /* $C = \{C_1, C_2, \ldots, C_k\}$ */
    $k$: $k$-anonymity level
Output:
    $C'$: set of equivalence classes
begin
    $C' = \emptyset$;
    while $|C_1| \neq 0$ do     /* Assume that $C_1$ is the minimum
    cluster among $C$. */
        $EC = \emptyset$;
        $r$ = randomly select a record from $C_1$;
        $EC = \{r\}$;
        $C_1 = C_1 \setminus \{r\}$;
        for $i = 2, 3, \ldots, k$ do
            $r'$ = the record in $C_i$ that is the closest to $r$ in terms
            of the QI attributes according to (2);
            $EC = EC \cup \{r'\}$;
            $C_i = C_i \setminus \{r'\}$;
            if $|C_i| > |C_1|$ then     /* If it is ture, take one more
            record from $C_i$. */
                $r'$ = the record in $C_i$ that is the closest to $r$ in
                terms of the QI attributes according to (2);
                $EC = EC \cup \{r'\}$;
                $C_i = C_i \setminus \{r'\}$;
            end if
        end for
        for each QI attribute in $EC$ do
            Replace attribute values with their lowest common
            ancestor node in the predefined hierarchical tree of
            the QI attribute;
        end for
        $C' = C' \cup EC$;
    end while
    Return $C'$;
end
```

Pseudocode 2 . *EnsureTCloseness*($\boldsymbol{T}$, $\boldsymbol{T}'$, t)

```
Input:
    $\boldsymbol{T}$: original data table
    $\boldsymbol{T}'$: anonymized data table
    $t$: $t$-closeness level
begin
    for each equivalence class $EC$ in $\boldsymbol{T}'$ do
        if $EMD(EC, \boldsymbol{T}) > t$ do
            $EC'$ = equivalence class in $\boldsymbol{T}'$ whose centroid is the
            closest to the centroid of $EC$ in terms of the QI
            attributes according to (2);
            /*A centroid of an equivalence class refers to a data
            record whose attribute value of each QI attribute is
            the lowest common ancestor node in the correspond-
            ing hierarchical tree.*/
            Merge $EC$ and $EC'$ in $\boldsymbol{T}'$;
        end if
    end for
end
```

Algorithm 1. Cluster-Based Algorithm for Multiple Sensitive Attributes Satisfying t-Closeness

```
Input:
    $\boldsymbol{T}$: original data table, containing multiple SAs
    $k$: $k$-anonymity level
    $t$: $t$-closeness level
Output:
    $\boldsymbol{T}'$: anonymized data table, satisfying $k$-anonymity and $t$-closeness
begin
    $C$ = partition $\boldsymbol{T}$ into $k$ clusters using Equi-sized FCM[19] according to (3);
    $\boldsymbol{T}' = GenerateEquivalenceClasses(C, k)$;
    $EnsureTCloseness(\boldsymbol{T}, \boldsymbol{T}', t)$;
    return $\boldsymbol{T}'$;
end
```

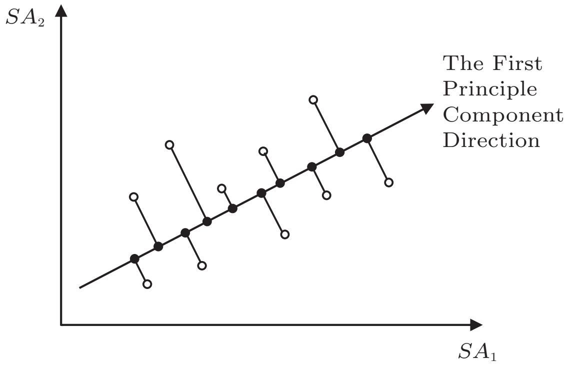


Fig.2. PCA example of one-dimensional projection of two-dimensional data points.

As mentioned above, the multiple SAs may contain both numerical and categorical attributes. To ensure the effectiveness of PCA working on the mixed data, we convert the categorical attributes to the numerical attributes first. Specifically, a binary attribute would be created for each value of a categorical attribute. For example, the Gender attribute is a categorical attribute with two values, i.e., male and female, then the pattern for a male instance will be "0 1", and "1 0" is for a female. Note that such conversion is only executed for the projection process, not the next processes.

All records are sorted in ascending order according to the PCA results and are partitioned into k sets with $\lfloor|\boldsymbol{T}|/k\rfloor$ records each, where $|\boldsymbol{T}|$ is the number of records in the original data $\boldsymbol{T}$, and k is the parameter of k-anonymity. The remaining records will be $r = |\boldsymbol{T}| \bmod k$, and the remaining r records will be assigned to one of the subsets. If $r \geqslant \lfloor|\boldsymbol{T}|/k\rfloor$, the number of one of the subsets will be large, and there will be records that are not assigned to any cluster (because only one or two records will be selected from each cluster in the second algorithm). To avoid this situation, k is adjusted as[9]:

$$k = k + \lfloor\frac{|\boldsymbol{T}| \bmod k}{\lfloor|\boldsymbol{T}|/k\rfloor}\rfloor.$$

A brief description of this algorithm is provided in Algorithm 2.

Algorithm 2. PCA-Based Algorithm for Multiple Sensitive Attributes Satisfying t-Closeness

```
Input:
    $\boldsymbol{T}$: original data table, containing multiple SAs
    $k$: $k$-anonymity level
    $t$: $t$-closeness level
Output:
    $\boldsymbol{T}'$: anonymized data table, satisfying $k$-anonymity and $t$-closeness
begin
    $k = k + \lfloor\frac{|\boldsymbol{T}| \bmod k}{\lfloor|\boldsymbol{T}|/k\rfloor}\rfloor$;
    $\boldsymbol{P}$ = the first principal component of $\boldsymbol{T}$ using PCA in terms of SAs;
    Sort all records of $\boldsymbol{T}$ in ascending order according to $\boldsymbol{P}$;
    $C$ = split $\boldsymbol{T}$ into $C_1, C_2, \ldots, C_k$ subsets each of which
    has $\lfloor n/k\rfloor$ records, with $n \bmod k$ records assigned
    to the central subset(s);      /* $C = \{C_1, C_2, \ldots, C_k\}$ */
    $\boldsymbol{T}' = GenerateEqivalenceClasses(C, k)$;
    $EnsureTCloseness(\boldsymbol{T}, \boldsymbol{T}', t)$;
    Return $\boldsymbol{T}'$;
end
```

5 Experimental Results and Analysis

In our experiments, the Adult dataset provided by the UC Irvine Machine Learning Repository[2] is used to test the performance of both algorithms. This dataset initially contains 48 842 census records with 15 attributes. After eliminating the records with missing values, there are 45 222 valid records remaining. The original attributes used in our experiments are shown in Table 5. We treat the Occupation and Education Num attributes as the SAs and the rest as the QI attributes.

We implement our Equi-sized FCM[19] for Algorithm 1 by Java programming language and the PCA[20] for Algorithm 2 with the help of Weka 3.6 tool[3]. In addition, according to Definition 4, the second algorithm in [9] is adjusted to anonymize the data with multiple SAs, and we refer to its adjusted version as the contrast algorithm in the following. The performance of

[2] http://archive.ics.uci.edu/ml/datasets/Adult, Mar. 2018.

[3] https://www.cs.waikato.ac.nz/ml/weka, Mar. 2018.

the contrast algorithm will be compared with that of our proposed algorithms based on the same dataset, experimental parameters, and measurements (i.e., equivalence class size, information loss, and speed). The experiments are performed on a PC with a 3.7 GHz @Intel core i7 CPU, 16 GB of RAM, running Windows 7 (64-bit).

Table 5. Attributes of the Adult Dataset Used in Our Experiments

Attribute	Type	Number of Values
Age	Numerical	74
Work Class	Categorical	7
Marital Status	Categorical	7
Race	Categorical	5
Sex	Categorical	2
Native Country	Categorical	41
Salary Class	Categorical	2
Occupation	Categorical	14
Education Num	Numerical	16

5.1 Equivalence Class Size

By applying the proposed and contrast algorithms to the Adult dataset for different values of k and t, we obtained a series of experimental results. The values of k are between 2 and 20, whereas the values of t are in the range of [0.1, 0.5], which cover the range of different privacy levels observed in existing studies. The results provided by these algorithms are shown in Tables 6–8. To minimize the loss of information, the closer the sizes of all of the clusters are to k, the better.

Tables 6–8 show that the sizes of equivalence classes become larger as the value of k is increased. When k is small or medium, such as 2, 5 or 10, the sizes of equivalence classes of Algorithm 1 and Algorithm 2 are larger than those of the contrast algorithm. Because the number of values of the SAs, Occupation and Education Num, is larger than k, Algorithm 1 has to partition similar records into different clusters (because all clusters are required to have the same size). Similar to Algorithm 1, Algorithm 2 always gets well-proportioned groups by sorting. The equivalence classes generated by records from these small clusters/groups have small sizes as well, which is difficult to meet the desired t-closeness principle. Thus, these classes have to be merged with their closest classes to meet the t-closeness principle, which in turn increases their sizes. On the other hand, the contrast algorithm does not have such limitation because it replaces clustered records by unclustered records, rather than merging clusters, if the equivalence class does not meet the t-closeness. Only if the replacement does not meet the condition, clusters would be combined. When k is sufficiently large, such as 15 or 20, the sizes of equivalence classes of Algorithm 1 and Algorithm 2 are equal to k, and the results of the contrast algorithm are inferior to those of our proposed algorithms. This is because our proposed algorithms analyze all records at first, instead of considering them one by one and because they generate clusters without discarding similar records in this case. It would be easier for both proposed algorithms to gather more records of high similarity into the same cluster when k is larger.

Table 6. Results of Algorithm 1 with Varying k and t (Minimum/Maximum/Average Size of Equivalence Classes)

k	t				
	0.1	0.2	0.3	0.4	0.5
2	8/14/12	8/10/9	8/10/9	6/10/8	6/8/7
5	10/25/18	10/20/16	10/20/13	5/15/10	5/15/8
10	10/30/25	10/20/18	10/20/14	10/20/12	10/20/12
15	15/15/15	15/15/15	15/15/15	15/15/15	15/15/15
20	20/20/20	20/20/20	20/20/20	20/20/20	20/20/20

Table 7. Results of Algorithm 2 with Varying k and t (Minimum/Maximum/Average Size of Equivalence Classes)

k	t				
	0.1	0.2	0.3	0.4	0.5
2	8/12/10	8/10/9	6/8/7	6/8/7	6/8/6
5	10/20/15	10/16/14	5/10/8	5/11/6	5/10/6
10	10/30/28	10/30/20	10/20/16	10/20/12	10/20/12
15	15/30/18	15/30/18	15/15/15	15/15/15	15/15/15
20	20/40/22	20/40/22	20/20/20	20/20/20	20/20/20

Table 8. Results of the Contrast Algorithm with Varying k and t (Minimum/Maximum/Average Size of Equivalence Classes)

k	t				
	0.1	0.2	0.3	0.4	0.5
2	8/10/9	8/10/9	6/10/8	6/8/7	6/8/7
5	10/20/14	10/20/12	5/10/8	5/10/8	5/10/6
10	10/30/25	10/20/15	10/20/14	10/20/12	10/10/10
15	15/30/23	15/30/21	15/15/15	15/15/15	15/15/15
20	20/40/27	20/40/24	20/20/20	20/20/20	20/20/20

5.2 Information Loss

The information loss represents the discrepancies between the original data and their corresponding anonymized versions. It is calculated by (1). We also set the values of k in the range of [2, 20] and the values of t in the range of [0.1, 0.5]. The performance of the proposed and contrast algorithms is presented in Figs.3–5, respectively. Fig.3 and Fig.4 show that the average information losses of Algorithm 1 and Algorithm 2 are approximately the same when k is 2, 5, or 10. This indicates that the earlier we consider the fulfillment of t-closeness during the formation of equivalence classes, the less the information will be lost in the anonymized output. Because the contrast algorithm is based on the same consideration, the above trend can also be obtained in Fig.5. However, Figs.3–5 also show that when k is 15 or 20, the performance of Algorithm 1 is better than those of the others. Because the number of values of multiple SAs in the Adult dataset is approximately equal to the parameter k, it is easier for Algorithm 1 to get pure clusters that contain homogeneous records in terms of SAs. It can be seen from Fig.4 and Fig.5 that Algorithm 2 is slightly inferior to the contrast algorithm. The reason is that Algorithm 2 uses PCA to rank all sensitive attributes by the total amount of variance that each sensitive attribute contributes, and some noisy attribute values overshadow the projection results. These projection results continue to influence the sorting order of multiple sensitive attributes.

5.3 Speed

Fig.6 shows the runtime of the proposed and the contrast algorithms, which consists of partitioning original records into clusters/groups, generating equivalence classes that satisfy t-closeness, and anonymizing these classes. To fairly and clearly compare these algorithms, we take k = 2 for k-anonymity with different values of t between 0.1 and 0.5 for t-closeness and force them to create the greatest number of clusters, which is the worst case from the perspective of runtime. Fig.6 shows that Algorithm 2 is more efficient than the other two algorithms in terms of runtime, because the Equi-sized FCM used in Algorithm 1 takes more time to allocate records to their corresponding clusters and gets the final result of the partition, and the contrast algorithm requires repeatedly much rearrangement of records. This figure also shows that the runtime of our proposed algorithms tends to decrease with the increase of parameter t because clusters are more likely to (nearly) fulfill t-closeness, thus requiring less rearrangement of records after each iteration. However, it is clear that the degree of data privacy is reduced when the parameter t is large.

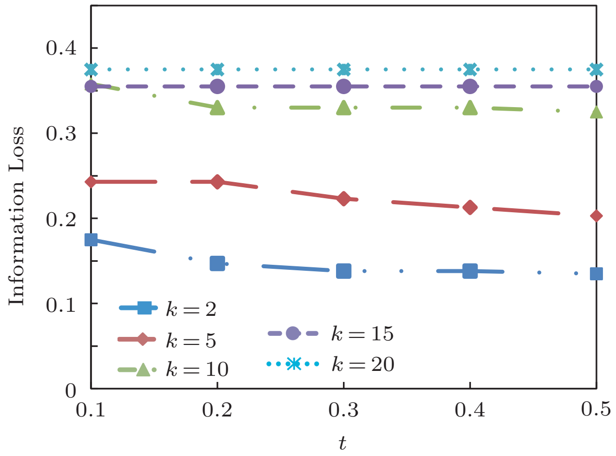


Fig.3. Algorithm 1: comparison of information loss at varying t.

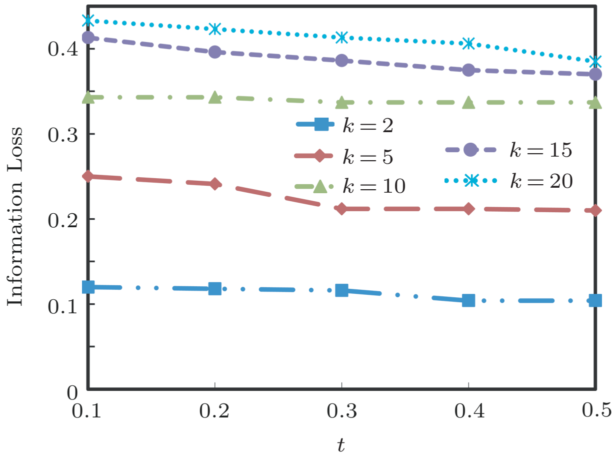


Fig.4. Algorithm 2: comparison of information loss at varying t.

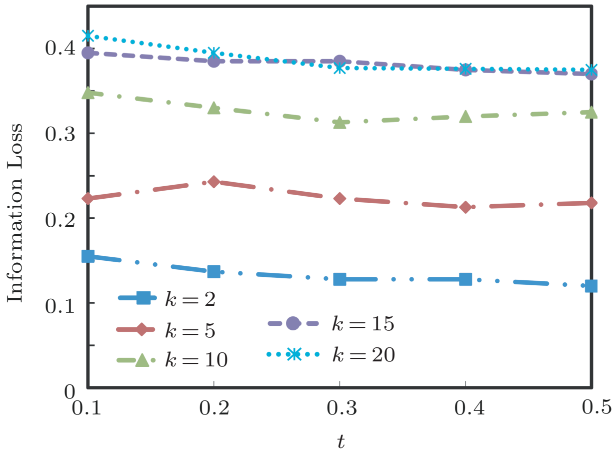


Fig.5. Contrast algorithm: comparison of information loss at varying t.

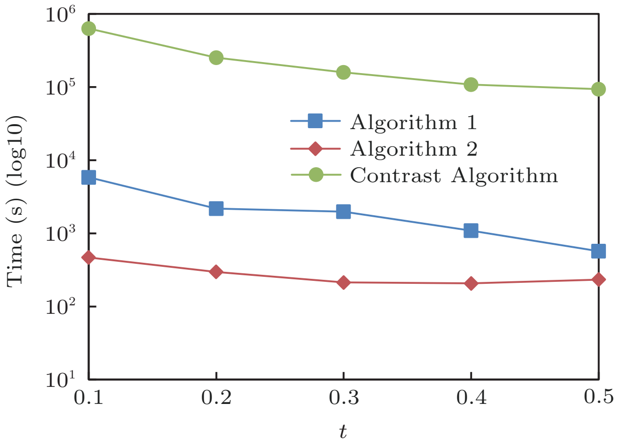


Fig.6. Comparison of execution time at varying t.

From the above comparisons, we can conclude that when the numbers of the values of SAs are approximately greater than or equal to the parameter k in k-anonymity, Algorithm 1 would be capable of generating homogeneous clusters so that the values of the multiple SAs are spread to each equivalence class. When the parameter k is small, Algorithm 2 is superior in generating homogenous clusters. The average speed of Algorithm 2 is much greater than that of Algorithm 1, but the average information loss of Algorithm 2 is greater than that of Algorithm 1.

6 Conclusions

While most existing k-anonymity-like models mask the original data that only have one sensitive attribute and protect their corresponding anonymous version from common attacks, they do not consider data with multiple sensitive attributes. Thus, we proposed two different algorithms to cover the gap between multiple sensitive attributes and the t-closeness principle. The first algorithm partitions all records into different clusters and generates equivalence classes by selecting records from these clusters separately. The second algorithm processes the multiple sensitive attributes by analyzing the principal components, sorts the original records according to the results of the projection, and partitions these records into different subsets by the sorting order. Both algorithms take the t-closeness principle into account when forming the equivalence classes. The experimental results demonstrated that the proposed algorithms achieved the purpose of data anonymity with low loss of information. Future work can investigate the techniques of distributed computing for dealing with t-closeness of multiple sensitive attributes more efficiently. In addition, other appropriate distance measures of multivariate distributions and other data mining algorithms are under development to facilitate data anonymization.

Acknowledgment We would like to thank two anonymous reviewers for their helpful comments that have led to an improved version of this paper.

References

[1] Sánchez D, Martínez S, Domingo-Ferrer J. Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata". *Science*, 2016, 351(6279): 1274.

[2] Sweeney L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 2002, 10(5): 557-570.

[3] LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional k-anonymity. In *Proc. the 22nd International Conference on Data Engineering*, April 2006, p.25.

[4] Machanavajjhala A, Gehrke J, Kifer D. Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. In *Proc. the 22nd International Conference on Data Engineering*, April 2006, p.24.

[5] Li N H, Li T C, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. the 23rd International Conference on Data Engineering*, April 2007, pp.106-115.

[6] Domingo-Ferrer J, Soria-Comas J. From t-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems*, 2015, 74: 151-158.

[7] Rebollo-Monedero D, Forne J, Domingo-Ferrer J. From t-closeness-like privacy to postrandomization via information theory. *IEEE Trans. Knowl. Data Eng.*, 2010, 22(11): 1623-1636.

[8] Cao J N, Karras P, Kalnis P, Tan K L. SABRE: A sensitive attribute bucketization and redistribution framework for t-closeness. *The VLDB Journal*, 2011, 20: 59-81.

[9] Soria-Comas J, Domingo-Ferrer J, Sánchez D, Martínez S. t-closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Trans. Knowl. Data Eng.*, 2015, 27(11): 3098-3110.

[10] Sha C F, Li Y, Zhou A Y. On t-closeness with KL-divergence and semantic privacy. In *Proc. the 15th International Conference on Database Systems for Advanced Applications*, April 2010, pp.153-167.

[11] Zhang J P, Xie J, Yang J, Zhang B. A t-closeness privacy model based on sensitive attribute values semantics bucketization. *Journal of Computer Research and Development*, 2014, 51(1): 126-137. (in Chinese)

[12] Rubner Y, Tomasi C, Guibas L J. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 2000, 40(2): 99-121.

[13] Xu J, Wang W, Pei J, Wang X Y, Shi B L, Fu A W C. Utility-based anonymization using local recoding. In *Proc. the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2006, pp.785-790.

[14] Ghinita G, Karras P, Kalnis P, Mamoulis N. Fast data anonymization with low information loss. In *Proc. the 33rd International Conference on Very Large Data Bases*, September 2007, pp.758-769.

[15] LeFevre K, DeWitt D J, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. In *Proc. ACM SIGMOD International Conference on Management of Data*, June 2005, pp.49-60.

[16] Li N H, Li T C, Venkatasubramanian S. Closeness: A new privacy measure for data publishing. *IEEE Trans. Knowl. Data Eng.*, 2010, 22(7): 943-956.

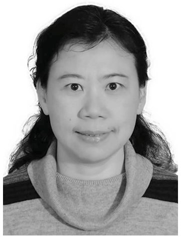
[17] Fang Y, Ashrafi M Z, Ng S K. Privacy beyond single sensitive attribute. In *Proc. the 22nd International Conference on Database and Expert Systems Applications*, August 2011, pp.187-201.

[18] Sei Y C, Okumura H, Takenouchi T, Ohsuga A. Anonymization of sensitive quasiidentifiers for l-diversity and t-closeness. *IEEE Transactions on Dependable and Secure Computing.* doi:10.1109/TDSC.2017.2698472.

[19] Höppner F, Klawonn F. Clustering with size constraints. In *Computational Intelligence Paradigms*, Jain L C, Sato-Ilic M, Virvou M, Tsihrintzis G A, Balas V E (eds.), Springer, Berlin, Heidelberg, 2008, pp.167-180.

[20] Jolliffe I T, Cadima J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2016, 374(2065): 20150202.

Rong Wang received her B.S. degree in computer science and technology from Mianyang Normal College, Mianyang, in 2011. She is now a Ph.D. candidate in the School of Information Science and Technology at Southwest Jiaotong University, Chengdu. Her current research interests include machine learning, data mining with big data, and privacy-preserving data mining.



Yan Zhu received her Ph.D. degree in computer science from Darmstadt University of Technology (TU Darmstadt), Darmstadt, in 2004. She also was a research staff at Department of Computer Science of TU Darmstadt from 1998 to 2004. She is now a professor in the School of Information Science and Technology at Southwest Jiaotong University, Chengdu. She has published two academic books in English and Chinese, and about 60 journal/conference papers. She is a reviewer of some SCI/EI indexed journals and a TPC member of some international conferences. Her current research interests include web data mining, privacy-preserving data mining, and big data management and analysis.

Tung-Shou Chen received his B.S. and Ph.D. degrees from "National" Chiao Tung University, Hsinchu, in 1986 and 1992, respectively, both in computer science and information engineering. Since August 2006, he has been a professor of the Department of Computer Science and Information Engineering at "National" Taichung Institute Technology, Taichung. His current research interests include data mining and data hiding.

Chin-Chen Chang received both his B.S. degree in applied mathematics in 1977 and his M.S. degree in computer and decision sciences in 1979 from the "National" Tsinghua University, Hsinchu, and his Ph.D. degree in computer engineering in 1982 from the "National" Chiao Tung University, Hsinchu. He is now a chair professor at the Department of Information Engineering and Computer Science, Feng Chia University, Taichung. Prior to joining the Feng Chia University, he was an associate professor in the Chiao Tung University, a professor in the "National" Chung Hsing University, and a chair professor in the "National" Chung Cheng University. He is also a fellow of IEEE and a fellow of IEE, United Kingdom. His current research interests include computer cryptography and information security, cloud computing, data engineering, and database systems.