Semantic integrity and *K*-anonymity

Liming Huang[1*], Jinling Song[1], Yan Gao[2], Qianying Cai[1]

[1]*Hebei Normal University of Science & Technology, Qinhuangdao 066004, China*

[2]*Liaoning Institute of Science and Technology, Benxi 117004, China*

Received 23 November 2014, www.cmnt.lv

Abstract

The dataset in database have certain semantic commonly, and the semantic need to be satisfied with the form of some constrains, such as functional dependencies (*FD*s) and multivalued dependencies (*MVD*s). Nevertheless, the *k*-anonymity model may be destroyed the semantic integrity in the process of *k*-anonymization because of the incontinent generalizations. So, in this paper we address the issue of how to preserve the semantic integrity of dataset in the *k*-anonymization process. We define a new data dependency named *k*-multiset dependency (*K-MSD*), which can ensure a dataset satisfies *k*-anonymity constraint. In addition, we propose *K-MSD* algorithm to realize *k*-anonymization through constructing *K-MSD* between attributes, and propose *K-MSD-AG* algorithm to preserves *FD*s or *MVD*s as while as constructing *K-MSD*.

Keywords: *k*-anonymization, *k*-multiset dependency, *FD*s, *MVD*s

1 Introduction

K-anonymity [1] provides strong guarantees on the confidentiality of individuals in publishing data from databases, however, the soundness of the anonymized data is dissatisfactory. K-anonymity relies on generalizations to preserve privacy: attribute values are replaced with less specific information (for example, “state” may be replaced with “region” and “age” may be replaced with “age range”). The generalization can be considered as an update operator for the data. In the database systems, any update operation for data in the database should be satisfied with semantic integrity constraints [2] to ensure the soundness of the data. At present, the mechanisms to check semantic integrity constraints in the DBMS can only serve as the database, but they ignore the publishing datasets. In fact, k-anonymity may violate the semantic integrity constraints of the dataset, such as data dependencies.

Consider Tables 1 and 2. Table 1 is a teachers’ salary table T, if one department has only one telephone number, then there is a functional dependency FD: Department $\rightarrow$ Phone over this table. If the table is generalized on attributes {Country, Sex, Zip, Department, Phone} to protect the salary information of teachers. Table 2 is a 2-anonymized table generated by Incognito, a kind of global recoding algorithm [3]. Note that there is only one value “Teaching-Depart” on attribute Department, while there corresponds four values 85152**, 85154**, 85156** and 85153** on attribute Phone, so the FD: Department $\rightarrow$ Phone has been lost after 2-anonymization. If one receives Table 1 and make a query on the table, “the telephone of ‘Teaching-Depart’”, he will get four results 85152**, 85154**, 85156** and 85153**. Then he will think the view is incorrect if he has the knowledge of the Department $\rightarrow$ Phone. Thus, it is very important to preserve the original data dependencies over dataset in k-anonymization process.

K-anonymity constraint can be considered as a kind of data dependency, assume attributes (X, Y) satisfy k-anonymity constraint, then for each value x on attribute X, there corresponds one or more values on attribute Y and each value appears larger than or equals to k. Because of the corresponding values of Y is a special multiset (the appearance of each value is at least k), so the data dependency can be named as k-multiset dependency (K-MSD). We show that k-anonymization of the publishing data can be realized by constructing K-MSD among attributes. There are relationships between K-MSDs and FDs, K-MSDs and MVDs, so, by some special treatment, FDs can be satisfied and MVDs can be satisfied approximately when attributes satisfying K-MSDs as a precondition in the generalization process.

2 Related work

K-anonymity privacy protection model [1] got the wide attention of experts and scholars when it was presented by Sweeney. Previous studies mostly focus on k-anonymization algorithm under different scenarios. Datafly algorithm was adopted in [4], which have promoted the generation of k-anonymity model. To improve the data precision of the generated table, Mingen algorithm was adopted in [5]. Meyerson et al and Aggarwal et al proved respectively that obtaining optimal k-anonymous table was NP-hard in [6] and [7], and proposed the approximation algorithms of O (klogk) times and 1.5 times (k = 2) of the minimum generalization. In [3], the global Incognito algorithm was proposed, which

**** Corresponding author’s*** e-mail: huangliming99@126.com

generalizes all the domain values of attributes. In [8], multi-dimensional algorithm was proposed, which generalize multi-attributes at the same time.

TABLE 1 The original data of table *T*

Country	Sex	Zip	Department	Phone	Salary
USA	Female	02142	Maths	8515257	1,5000K
USA	Female	02139	Chemistry	8515226	2,6000K
Japan	Male	02138	Physics1	8515411	1,8000K
Japan	Male	02142	Physics2	8515412	1,1000K
Korea	Female	02138	Computer1	8515628	3,4000K
Japan	Female	02141	Computer2	8515629	2,8000K
Canada	Male	02142	Business	8515338	1,6000K
Canada	Male	02138	Management	8515326	1,2000K

TABLE 2 *K*-anonymized table generated by *Incognito* algorithm

Country	Sex	Zip	Department	Phone	Salary
North America	Female	021**	Teaching	85152**	1,5000K
North America	Female	021**	Teaching	85152**	2,6000K
Asia	Male	021**	Teaching	85154**	1,8000K
Asia	Male	021**	Teaching	85154**	1,1000K
Asia	Female	021**	Teaching	85156**	3,4000K
Asia	Female	021**	Teaching	85156**	2,8000K
North America	Male	021**	Teaching	85153**	1,6000K
North America	Male	021**	Teaching	85153**	1,2000K

Ren et al [9] proposed CBK(L,K)-anonymity algorithm which can make anonymous data effectively resist background knowledge attack and homogeneity attack, and can solve diversity of sensitive attribute, the main idea is anonymizing the data set by K-clustering based on influence matrix of background knowledge. Lv et al proposed m-threshold model to solve advanced attack and used GSSK (Generalization Step Safe of K-anonymity) [10] algorithm to deal with the model. The TDS (top-down specialization) algorithm [11] achieves the k-anonymity by gradual specialization from the most generalization state (attribute values are represented by the root nodes in classification tree). K-anonymization will affect the quality of publishing data, it not only reduces the precision of the data, but also violates the semantic integrity constraints on the dataset. Previous methods focus on how to improve the data precision, but ignore preserving the integrity constraints on the dataset. Our approach can not only preserve higher data precision with several metrics but also preserve original FDs or MVDs over dataset, so it can increase the utility of the anonymized datasets effectively. To preserve the clustering information of anonymous data, Fung et al [12] extended TDS algorithm. Liu et al [13] proposed a personalized privacy preserving parallel (alpha, k)-anonymity model based on k-anonymity to reduce high probability of the attributes in the equivalent group and reduce the probability of the likelihood of attack. In [14], a local coding anonymous algorithm was proposed based on the attribute hierarchy.

3 Basic definitions

$T(A_1,\ldots,A_n)$: a table with a finite number of tuples, where T is the name of the table, $A_1,\ldots, A_n$ are the finite set attributes of T.

<*entity*, *attribute*>: is a value associated with the entity, chosen from the domain of the attribute.

Domain: A set of possible values of one attribute, denoted as D, the domain of attribute A_i is D_i.

X-value: Let U is the attribute set of a relational schema, X-value is a mapping that assigns to each attribute $A\in X(X\subseteq U)$ a value from the corresponding domain of attribute A.

$t[A_i,\ldots, A_j]$: the sequence of the values $v_i,\ldots,v_j$ on the attributes $A_i,\ldots,A_j$ in tuple t.

$t_j[A_i]$: is the value of jth tuple on ith attribute.

$T[A_i,\ldots,A_j]$: projection of $T[A_1,\ldots,A_n]$ on the attribute set $\{A_i,\ldots,A_j\}$, where maintain duplicate tuples, namely, $T[A_i,\ldots,A_j]$ is a multiset of tuples.

Definition 1 (Quasi-identifier) Given a table $T(A_1,\ldots, A_n)$ that contains private information, we call $\{A_i,\ldots,A_j\}$ a *quasi-identifier* of T (written *QI*), if the set of attributes $\{A_i,\ldots,A_j\}\subseteq\{A_1,\ldots,A_n\}$ can be joined with other public information and re-identify individual tuples.

Definition 2 **(*K-anonymity Constraint*)** $T[A_1,\ldots,A_m]$ satisfies *k-anonymity constraint* on attributes $\{A_1,\ldots, A_m\}$, if each tuple in $T[A_1,\ldots, A_m]$ counts at least k ($k\geq 2$).

Definition 3 **(*Generalization*)** Given table $T(A_1,\ldots,A_m)$, if for any attribute $A_i\in\{A_1,\ldots,A_m\}(1\leqslant i\leqslant m)$. There is a many-to-one function f_i:$D_i\rightarrow D_i'$, where D_i is the domain of A_i, and D_i' contains more general values corresponding to D_i, then we call D_i' is the *generalization* of D_i, and f_i is the *generalization function* of attribute A_i, such that $f_i(t_j[A_i])$ is the *generalized value* corresponding to $t_j[A_i]$.

We can extend the *generalization function* of an attribute to a set of attributes, if $X=\{A_1,\ldots,A_l\}$, then call $f_X(X)=\{f_1(t_j[A_1]),\ldots, f_l(t_j[A_l])\}$ as the *generalization function* of A.

In fact, for attribute A_i, there are a sequence of *generalization functions* $f_i^1, f_i^2,\ldots, f_i^n$, where $f_i^n(f_i^{n-1}(\ldots(f_i^1$

$(t_j[A_i]))\ldots))$ is the final generalized value of $t_j[A_i]$. We call the number of generalization functions in the sequence is the *generalization distance* from $t_j[A_i]$ to f_i^n $(f_i^{\ n-1}(\ldots(f_i^1(t_j[A_i]))\ldots))$, denoted by $GD(t_j[A_i],\ f_i^n\ (f_i^{\ n-1}(\ldots(f_i^1(t_j[A_i]))\ldots)))$.

Obviously, the *generalization distance* can affect the data precision of the publishing table.

Definition 4 **(*Generalization Distance of Multiple Values*):** Given a table $T(A_1,\ldots,A_m)$. Let $V=\{v_1,\ldots,v_n\}$ is a set of values, where v_i may be an attribute value in $T(A_1,\ldots,A_m)$ or a generalized value. If we generalize $v_1,\ldots,v_n$ to the final generalized value s, then $MAX(GD(v_1, s), GD(v_2, s),\ldots, GD(v_n, s))$ is the *generalization distance of multiple values* $v_1,\ldots,v_n$. Denoted by $GD(V, s)$.

Example. If the values associated with attribute *Zip* are {02138, 02139, 02141, 02142}, and the generalization functions f_{Zip}^1, f_{Zip}^2, f_{Zip}^3 of *Zip* are shown in Tables 3-5. When we want to generalize the value 02138, for f_{Zip}^1(02138)= 0213*, so *GD*(02138,0213*)=1 ; for f_{Zip}^2 (f_{Zip}^1(02138))= 021**, so *GD*(02138 ,021**)=2; for f_{Zip}^3(f_{Zip}^2(f_{Zip}^1 (02138)))= *****, so *GD*(02138 ,*****)=3. When we want to generalize values 02138, 02139, 02141 and 02142 to the same value 021**, for *MAX*(*GD*(02138, 021**), *GD*(02139, 021**), *GD*(02141, 021**), *GD*(02142,021**))= *MAX*(2, 2, 2, 2) =2, so *GD* ({02138, 02139, 02141, 02142},021**)=2.

TABLE 3 The generalization function of attribute f_{Zip}^1:

Variable Value	Functional Value
02138	0213*
02139	0213*
02141	0214*
02142	0214*

TABLE 4 The generalization function of attribute f_{Zip}^2:

Variable Value	Functional Value
0213*	021**
0214*	021**

TABLE 5 The generalization function of attribute f_{Zip}^3:

Variable Value	Functional Value
021**	*****

Definition 5 (K-anonymization) Let the *quasi-identifier* of table *T* is *QI*. If the result table *T′*, produced by generalizing the values of *T* on *QI*, can satisfy *k-anonymity constraint*, then the generalization process from table *T* to *T′* is the *k-anonymization* of table *T*.

Definition 6 (Initial Maximal K-anonymited Attribute Set) Let the *quasi-identifier* of table *T* is *QI*, if there is an attribute set $\{A_i,\ldots,A_j\} \subseteq QI$ to make $T[A_i,\ldots,A_j]$ satisfy *k-anonymity constraint*, and for any other attribute $A_m \in (QI\text{-}\{A_i,\ldots,A_j\})$, $T[A_i,\ldots,A_j, A_m]$ does not satisfy *k-anonymity constraint*, then the attribute set $\{A_i,\ldots,A_j\}$ is the *initial maximal k-anonymited attribute set* (*IMKAS*) of table *T*.

The initial maximal k-anonymited attribute set of a table may be not unique.

4 Data dependencies over anonymized datasets

Data dependencies are a kind of database semantic tool [15]. In relational database, a data dependency is a proposition to show the relationships among data items. It represents the integrity constraint condition that any legal database must be satisfied. FDs and MVDs are two kinds of dependencies that appear naturally in the real world [16]. Assuming that the publishing data is a view by projecting on a relation R, then according to projective property [15], the FDs and MVDs in R also hold in the view, i.e. the view has satisfied integrity constraint conditions of original database automatically. Anonymized datasets are the output results of k-anonymization algorithms whose inputs are the publishing data, in order to keep the correct semantic, they should satisfy integrity constraint conditions of original database. Beside the FDs and MVDs over original database, anonymized datasets also satisfy the k-multiset Dependencies (K-MSDs) which can ensure the k-anonymization.

4.1 FD AND MVD

Let *A* and *B* be attributes in database *D*. We say that *B* is *functionally dependent* on *A* (in *D*) if, at every point of time, for a given value of $a \in DOM(A)$ there corresponds at most one value of $b \in DOM(B)$. Given a relation $R(U)$, a *Multivalued dependency*(*MVD*) on the set *U* is a statement $g{:}X \rightarrow\rightarrow Y$. Let *Z* denote the set $U\text{-}(X \cup Y)$. We say that the relation *R* obeys the *MVD g* if for every *XZ*-value *xz*, that appears in *R*, we have $Y_R(xz)= Y_R(x)$ (Y_R is a function that gives for each *X*-value the set of *Y*-values that appear with it in tuples of *R*).

4.2 K-MULTISET DEPENDENCIES(K-MSD)

Let's consider the relationship between the attribute values of *Race* and *Zip* in the 2-*anonymized* table T shown in Table 6. For T [*Zip*| *Race* = "Asian"]={0213*, 0213*,02141, 02141}, and both 0213* and 02141 appear twice, *T* [*Zip*| *Race* = "Black"]= {02138, 02138, 02138, 02138} and the value 02138 appears four occurrences. Therefore there is a dependent relationship among the 2-anonymited dataset on attribute set {*Race*, *Zip*}: in the tuples which have same values on attribute *Race*, there corresponds one or more values on attribute *Zip* and every value appears at least twice. The above dependent relationship can be generalized in the obvious way: In the dependent relationship among *k-anonymized* dataset, for the *Y*-values corresponding to any *X*-value *x*, may be multiple and each value counts *k*, that is *Y*-values are a special multi-set (each element occurs at least *k* times), so we name it as *k-multiset dependency* (*K-MSD*).

TABLE 6 Example of k-anonymized data, k=2

Race	Zip
Asian	0213*
Asian	0213*
Asian	02141
Asian	02141
Black	02138
Black	02138
Black	02138
Black	02138

4.2.1 Definition and inference rules of k-multiset dependencies

***Definition 7* (*K-multiset Dependency* (*K-MSD*))** Let $R(U)$ is a relational schema and $X,Y\subseteq U$. We say Y is *k-multiset dependent* (*K-MSD*) on X, if for each X-value x in any relation instance of R, each Y-value corresponding to x appears at least k ($k\geq 2$) occurrences. Denoted by $X \xrightarrow{k} Y$.

Example. We can see from table T (*Race*, *Zip*) in Table 1 that, *Zip* is 2-*MSD* dependent on *Race*.

It seems that the *K-MSD* is very similar to *MVD*, for they both have more than one value on attribute Y corresponding to the X-value x. But *MVD* relates not only to attributes X and Y but also to the other attributes Z in that relation, while *K-MSD* only relates to attributes X and Y. So *K-MSD* is more like a special trivial *MVD*.

Let $R(U)$ is a relational schema and $X,Y,Z\subseteq U$, the inference rules of *K-MSD* are as following:

Rule 1: (*Reflexivity*). If $Y\subseteq X$ and each X-value appears no less than k, then $X \xrightarrow{k} Y$.

Rule 2: (*Transitivity*). If $X \xrightarrow{k} Y$, $Y \xrightarrow{k} Z$, and for distinct X-values, there corresponds distinct values on attribute Y, then $X \xrightarrow{k} Z$ and $XY \xrightarrow{k} Z$.

Rule 3: (*Projection Rule*). If $X \xrightarrow{k} Y$, for $X'\subseteq X$, $Y' \subseteq Y$, then $X \xrightarrow{k} Y'$, $X' \xrightarrow{k} Y$, $X' \xrightarrow{k} Y'$.

4.2.2 Relationships between K-MSD and k-anonymity constraint

We can know there are following relationships between *K-MSD* and *K-anonymity constraint* from the definitions of *K-MSD* and *K-anonymity constraint*:

Theorem 1 Let $T(U)$ is a relational table, U is the set of attributes. If T satisfies *k-anonymity constraint*, then for any X, $Y\subseteq U$ ($X\cap Y=\Phi$), $X \xrightarrow{k} Y$.

Theorem 2 Let $T(U)$ is a relational table, $X,Y\subseteq U$ and $X\cap Y=\Phi$, if $X \xrightarrow{k} Y$, then $T[X, Y]$ achieves *k-anonymity*.

Theorem 3 The QI of table T is QI={ A_1 ,…, A_m }. If there is an attribute set { A_c ,…, A_d }$\subseteq QI$ makes { A_c ,…, A_d } $\xrightarrow{k}$ (QI-{ A_c ,…, A_d }) satisfied, then $T[QI]$ satisfies *k-anonymity constraint*.

From the above analysis of the relationships between *K-MSD* and *k-anonymity constraint*, we can get the following conclusion: *K-MSD* can represent *k-Anonymity constraint* in a better way.

5 K-anonymization algorithm based on K-MSD

It can be known from Theorem 3 that, if there is a *K-MSD* { A_c ,…, A_d } $\xrightarrow{k}$ (QI-{ A_c ,…, A_d }) among the *quasi-identifier* attributes of table T, then table T must satisfy *k-anonymity*. According to the theorem, if we can construct such a *K-MSD* among the *quasi-identifier* attributes, then table T realizes *k-anonymization*. So, we propose a *k-anonymization* algorithm based on *k-multiset dependency* (*K-MSD* algorithm), the basic idea of which is constructing *K-MSD* between attributes one by one until a partition of QI satisfying *K-MSD*. The *K-MSD* algorithm mainly includes two steps:

Let $QI = \{A_1,\ldots, A_m\}$ be the *quasi-identifier* of $T(U)$:

1) Selecting a certain attribute (set) satisfying *k-anonymity constraint* from the attributes of QI, and let A_{anony} represent the attribute (set).

2) Repeat the following step until $QI=A_{anony}$: select attribute $A_i \in (QI\text{-}A_{anony})$ and generalize it to satisfy $A_{anony} \xrightarrow{k} A_i$, $A_{anony}=A_{anony} \cup A_i$.

In order to realize *K-MSD* algorithm, we need to resolve the following problems:

1) To make the number of attribute (or attribute value) generalized later is minimal, the method to select the first attribute (set) as the basic of the *K-MSD* algorithm is: Choose a *IMKAS* as the basic of the *K-MSD* algorithm if there are *IMKASs* in table T, otherwise, select an attribute whose *k-violation* values (i.e. the values which appear less than k occurrences) are least and generalize *k-violation* values to satisfy *k-anonymity*.

2) The method to generalize *k-violation* values is: To any *k-violation* value v, select values v_i,…, v_j and generalize them to a same value s, so that GD ({v, v_i,…,v_j}, s) is minimal. Repeat this operation until the appearance of each value is no less than k. We call this generalization process as *minimal distance generalization* (*MDG*). *MDG* can reduce the generalization level of each value and decrease the generalized value as possible, so it can maintain higher precision of the publishing data.

3) To select the following attribute to construct *k-MSD*, we select an attribute whose *k-MSD-violation* values (the values which don't satisfy *K-MSD* with attributes having satisfy *k-anonymity*) are least from the remainder attributes every time.

4) To Generalize the values on attribute A_i to satisfy $A_{anon} \xrightarrow{k} A_i$, we can partition $T[A_{anony}]$ to several groups where each group has the same values, for every group, implement *MDG* to every *K-MSD-violation* value of attribute A_i and make its appearance is at least k.

The *K-MSD* algorithm is described as follows:

K-MSD (T (A_1,…, A_m), k, GF, G)

INPUT: The relational table $T(A_1,\ldots,A_m)$ where $\{A_1,\ldots,A_m\}$ is QI, the k value, $GF=\{f_1,\ldots, f_m\}$（where f_i (i=1,2,…,m) is the generalization function of attribute A_i), the power set of { A_1,…,A_m } is G.

OUTPUT: The result table GT through the *k-anonymization* of T

STEPS:

Initialization: $A_{anony}=\Phi$; /* A_{anony} denotes the attribute set that have satisfied *k-anonymity* */

If $|T|\geq k$, then

1. $A_{anony}\leftarrow$Select a subset G' from G where G' satisfies *k-anonymity* and $|G'|$ is largest;

If ($A_{anony}==\Phi$), then

$A_{anony}\leftarrow$select an attribute A_j whose *k-violation* values are least, and implement *MDG* to *k-violation* values to make A_j achieve *k-anonymity*;

2. While $A_{anony}\neq QI$, Do

2.1 Select an attribute $A_i\in(QI-A_{anony})$ whose *K-MSD-violation* values corresponding to A_{anony} are least, and implement *MDG* to *K-MSD*-violation values to satisfy $A_{anony} \xrightarrow{k} A_i$;

/*attribute set $\{A_{anony}, A_i\}$ satisfy *k-anonymity*/

2.2 $A_{anony}\leftarrow A_{anony}\cup\{A_i\}$;

3. $GT\leftarrow T$;

4. Return (GT);

Theorem 4 After *k*-anonymizing $T[QI]$ by *K-MSD* algorithm, $T[QI]$ satisfies *k-anonymity constraint*.

Example The result of performing *K-MSD* algorithm to table *T* in Table 1 is shown in Table 7.

TABLE 7 The *k*-anonymized table *GT* generated by *K-MSD* algorithm

Country	Sex	Zip	Department	Phone	Salary
USA	Female	021**	Teaching	85152**	1,5000K
USA	Female	021**	Teaching	85152**	2,6000K
Japan	Male	021**	Physics	851541*	1,8000K
Japan	Male	021**	Physics	851541*	1,1000K
Asia	Female	021**	Computer	851562*	3,4000K
Asia	Female	021**	Computer	851562*	2,8000K
Canada	Male	021**	Teaching	85153**	1,6000K
Canada	Male	021**	Teaching	85153**	1,2000K

6 K-anonymization algorithm preserving FDs or MVDs

6.1 K-ANONYMIZATION ALGORITHM PRESERVING FD

6.1.1 Relationships between K-MSDs and FDs

It can be known there are similarities between *K-MSDs* and *FDs* from the definitions of *K-MSD* and *FD*. For a given *X*-value *x*, each *Y*-value corresponding to *x* appears at least *k* occurrences in *K-MSD*, while the *Y*-values in *FD* are consistent. Thus *K-MSD* and *FD* can be transformed into each other under some cases. The relationships between *K-MSD* and *FDs* are as follows.

1) According to the definition of *K-MSD*, if $X \xrightarrow{k} Y$, and for a given *X*-value *x*, there only one *Y*-value corresponding to *x*, then $X\rightarrow Y$.

2) Basing on *FD*, if $X\rightarrow Y$ and each *X*-value appears at least *k* occurrences, then $X \xrightarrow{k} Y$.

Assume $R(U)$ is a relational schema, where $X,Y,Z\subseteq U$, the reference rules of *K-MSD* and *FDs* are:

***Rule* 1**: (*Transitivity*). If $X \xrightarrow{k} Y, Y\rightarrow Z$, then $X \xrightarrow{k} Z, Y \xrightarrow{k} Z$.

***Rule* 2**: (*Pseudotransitivity*). If $X \xrightarrow{k} Y$, $Y\rightarrow Z$, then $X \xrightarrow{k} YZ$.

***Rule* 3**: (*Union*). If $X\rightarrow Y, X\rightarrow Z, X \xrightarrow{k} Y, X \xrightarrow{k} Z$, then $X \xrightarrow{k} YZ$.

From these relationships we know that, two attributes can satisfy both *K-MSD* and *FD* if we perform special generalization when *k*-anonymizing the two attributes. That is, original *FDs* among attributes won't be violated during *k-anonymization* process.

Next, we will introduce the special generalization which can achieve *K-MSD* and *FD* between attributes, and name it as *association generalization*.

Definition 8 **(Association Generalization (AG))** Let $T(A_1,\ldots,A_m)$ be a table and $X,Y\subseteq\{A_1,\ldots,A_m\}$. A function $AG: D'_X\times D_Y\rightarrow D'_Y$ is an *association generalization function* of *X,Y*, if $f_X(t_l[X])=\ldots=f_X(t_d[X])$,where tuples $t_l,\ldots,t_d\in T$ and f_X is the *generalization function* of *X* , then $AG(f_X(t_l[X]), t_l[Y])=\ldots=AG(f_X(t_d[X]), t_d[Y])$.

AG means that if attribute set *X* has been generalized, then for each tuple group with same generalized values on *X*, there also corresponds same generalized value on attribute set *Y*.

Example Consider Table 8. If attribute *Zip* has been generalized firstly, which result is shown in Table 9. Note that, 02138, 02139 are replaced with 0213*, 02141,02142 are replaced with 0214*, thus *Zip* satisfy 2-*anonymity*. Then the *AG* of attribute *Race* corresponding to attribute *Zip* is: replace the values {15, 20} corresponding to *Zip*="0213*" with a same value 10-20, and replace the values {23, 28} corresponding to *Zip*= "0214*" with the same value 20-30.

Theorem 5 For attribute set *X,Y* of table *T*, if $X\rightarrow Y$ and *X* has been satisfied *k-anonymity* by generalization, then after the *association generalization AG* $(t[X], t[Y])(t\in T)$ to *Y*, let the generalized table is T', there must be $X\rightarrow Y$ and $X \xrightarrow{k} Y$ in table T'.

TABLE 8 Example before generalization

Zip	**Age**
02138	15
02139	20
02141	23
02142	28
02138	35
02138	35
02138	46
02138	46

TABLE 9 Example of association generalization

Zip	**Age**
0213*	10-20
0213*	10-20
0214*	20-30
0214*	20-30
02138	35
02138	35
02138	46
02138	46

6.1.2. K-anonymization algorithm based on K-MSD-AG

Basing on the concept of *AG* and Theorem 5 given in Section 6.1.1, *K-MSD* algorithm can be improved to preserve original *FDs* among attributes, named as *K-MSD-AG*. Main steps of *K-MSD-AG* algorithm are as follows:

Let $QI = \{A_1, \ldots, A_m\}$ be the *quasi-identifier* of table *T*, the set of *FDs* among *QI* is *F*.

Step 1. Finding out one attribute (set) from *QI* and make it satisfying *k-anonymity*, then let A_{anony} represent the attribute (set).

Step 2. Repeat the following operation until $QI=A_{anony}$: If exit $F|= X\rightarrow Y$ ($X\subseteq A_{anony}$, $(Y\text{-}A_{anony})\neq\Phi$), then $AG(X, Y)$ and let $A_{anony}=A_{anony}\cup Y$. Else select attribute $A_i\in(QI\text{-}A_{anony})$ and generalize it to satisfy $A_{anony}\xrightarrow{k}A_i$ and let $A_{anony}=A_{anony}\cup A_i$.

For selecting the first attribute (set) under the case of there are *FDs* among *quasi-identifier* attributes, we consider attributes with *FDs* firstly in order to preserve *FDs* well. We select attribute (set) that can functionally depend most attributes (i.e. the number of dependencies in which the attribute (set) on the left side is most in *F*), and generalize *k-violation* values to achieve *k-anonymity*. Now the attribute (set) is the basis of *K-MSD-AG* algorithm. Other solutions we use are same with *K-MSD* algorithm in Section 5.

The description of *K-MSD-AG* algorithm is as follow:

K-MSD-AG ($T(QI)$, k, GF, F)

INPUT: Table $T(A_1,\ldots,A_m)$ where the *quasi-identifier* is $QI = \{A_1,\ldots,A_m\}$, k value, $GF=\{f_1,\ldots, f_m\}$(where f_i (i=1,2,…,m) is the *generalization function* of attribute A_i), *F* is the set of *FDs* among *quasi-identifier* attributes

OUTPUT: The *k-anonymized* table *GT* of table *T*

STEP:

Initialization: $A_{anony}=\Phi$; /* A_{anony} denotes attribute set satisfying *k-anonymity* */

If $|T|\geq k$, then

1. $A_{anony}\leftarrow$select the attribute (set) on left side of functional dependencies of *F* which occurrences most frequently, and let it satisfy *k-anonymity* through *MDG* (or through *K-MSD* algorithm to attribute set);
2. While $A_{anony}\neq QI$, do

   2.1 for each $Y\in\{V|U=X, F|= U\rightarrow V, X\subseteq A_{anony}\}$, do

   { if $(Y\text{-}A_{anony})\neq\Phi$, then

   {$AG(X, Y)$;

   $A_{anony}\leftarrow A_{anony}\cup\{Y\}$;}}

   2.2 else

   select $A_i\in(QI\text{-}A_{anony})$ whose *K-MSD-violation* values are minimal and implement *MDG* to *K-MSD-violation* values to satisfy $A_{anony}\xrightarrow{k}A_i$;

   $A_{anony}\leftarrow A_{anony}\cup A_i$;
3. $GT\leftarrow T$;
4. Return (*GT*);

Theorem 6 *GT* satisfies *k-anonymity* and preserves original *FDs* among attributes through *k-anonymization* with *K-MSD-AG* algorithm.

Example. The result of performing *K-MSD-AG* algorithm to table *T* is showed in Table 10, where there exists a *FD*: *Department* → *Phone* in *QI* = {*Country*, *Sex*, *Zip*, *Job*, *Salary*}.

TABLE 10 The *k*-anonymized table *GT* generated by *K-MSD-AG* algorithm

Country	**Sex**	**Zip**	**Department**	**Phone**	**Salary**
USA	Female	021**	Teaching	8515***	1,5000K
USA	Female	021**	Teaching	8515***	2,6000K
Japan	Male	021**	Physics	851541*	1,8000K
Japan	Male	021**	Physics	851541*	1,1000K
Asia	Female	021**	Computer	851562*	3,4000K
Asia	Female	021**	Computer	851562*	2,8000K
Canada	Male	021**	Teaching	8515***	1,6000K
Canada	Male	021**	Teaching	8515***	1,2000K

6.2 K-anonymization algorithm preserving MVDs

Under the case of there are MVDs over the original dataset, we can preserve them using K-MSD-AG algorithm (i.e. replace the input FDs with MVDs) approximately. Because the MVDs over dataset can be converted into FDs by use AG in the process of constructing of K-MSDs, and FDs satisfy MVDs naturally. However, the data precision will be lower in the process of converting MVDs to FDs because some data will be generalized excessively.

7 Conclusion

K-anonymity constraint can be considered as a kind of data dependencies, defined k-multiset dependency (K-MSD) in this paper. So, there exit three data dependencies, such as

FDs, MVDs and K-MSDs over k-anonymized dataset. If these data dependencies are all considered in the k-anonymization process, then both semantic integrity and the privacy of the dataset can be guaranteed. We propose K-MSD algorithm and K-MSD-AG algorithm for k-anonymization.

Acknowledgments

This work is supported by National Natural Science Foundation (NO.60773100, NO.61070032) of China, Project of Science and Technology Office of Hebei Province (NO.13227427), Doctor Foundation (2013YB007) and Creative Research Groups Foundation (No. CXTD2012-08) of Hebei Normal University of Science and Technology.

References

[1] Sweeney L 2002 K-Anonymity: a model for protecting privacy *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**(5) 557-70
[2] Hammer M M, McLeod D J 1975 Semantic integrity in a relational data base system *Proceedings of the VLDB* 25-47
[3] Lefvre K, DeWitt D, Ramakrishnan R 2005 Incognito: Efficient full-domain k-anonymity *Proceedings of the International Conference on Management of Data* 49-60
[4] Sweeney L 1997 Guaranteeing anonymity when sharing medical data: the Datafly system *Proceedings of the 1997 AMIA Annual Fall Symposium* **4** (suppl) 51-5
[5] Sweeney L 2002 Achieving k-anonymity privacy protection using generalization and suppression *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**(5) 571-88
[6] Meyerson A, Williams R 2004 On the complexity of optimal k-anonymity *Proceedings of the ACM Symposium* 223-8
[7] Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigraphy R, Thomas D, Zhu A 2004 k-Anonymity: Algorithms and Hardness, *Stanford University* California USA *Technical Report* 2004-22
[8] LeFevre K, DeWitt D J, Ramakrishnan R 2006 Mondrian Multidimensional K-Anonymity *Proceedings of International Conference on Data Engineering* 25
[9] Ren X, Yang J, Zhang J, Wang K 2011 Research on CBK(L,K)-Anonymity Algorithm, *International Journal of Advancements in Computing Technology* **3**(4) 165-73
[10] Lv P, Wu Y 2010 Generalization Step Analysis for Privacy Preserving Data Publishing *International Journal of Digital Content Technology and its Applications* **4**(6) 62-71
[11] Fung B C M, Wang K, Yu P S 2007 *IEEE Transactions on Knowledge and Data Engineering* **19**(5) 711-25
[12] Fung B C M, Wang K, Wang L, Jung P C K 2009 Privacy-Preserving Data Publishing for Cluster Analysis *Data & Knowledge Engineering* 68(6) 552-75
[13] Liu Y, Yang B, Li G 2012 A Personalized Privacy Preserving Parallel (alpha, k) -anonymity Model *International Journal of Advancements in Computing Technology* 4(5) 265-71
[14] Li J Y, Wong R C W, Fu A W C, Pei J 2008 *IEEE Transactions on Knowledge and Data Engineering* **20**(9) 1181-94
[15] Beeri C, Bernstein P A, Goodman N 1978 A Sophsticate's Introduction to Database Normalization Theory *Proceedings of VLDB* 113-24
[16] Beeri C 1980 On the Membership Problem for Functional and Multivalued dependencies in Relational Databases *ACM Transactions on Database Systems* **5**(3) 241-59

Authors

Liming Huang, 1972, China.

Current position, grades: associated professor at Hebei Normal University of Science & Technology.
University studies: MS degree in information system and management at Beijing University of Technology in 2009.
Scientific interests: data privacy and security, big data.

Jinling Song, 1973, China.

Current position, grades: associated professor at Hebei Normal University of Science & Technology.
University studies: PhD degree in computer application technologies at Yanshan University in 2012.
Scientific interests: data privacy and security, big data.

Gao Yan, 1974, China.

Current position, grades: lecturer at Liaoning Institute of Science and Technology.
University studies: MS in system engineering at Northeastern University in 2005.
Scientific interests: software engineering.