(*l*, *e*)-Diversity – A Privacy Preserving Model to Resist Semantic Similarity Attack

Haiyuan Wang, Jianmin Han*, Jiyi Wang
Department of Computer Science and Technology
Zhejiang Normal University, Jinhua, 321004, Zhejiang, PRC
Email: hanjm@zjnu.cn

Lixia Wang
Xingzhi College
Zhejiang Normal University, Jinhua, 321004, Zhejiang, PRC

***Abstract*—Existing sensitive attributes diversity models do not capture the semantic similarity between sensitive values, so they cannot resist semantic similarity attack. To address the problem, we present a method to measure semantic similarity of a categorical sensitive attribute based on the attribute' semantic hierarchy tree. On basis of the measurement, the paper proposes a (*l, e*)-diversity model which has two constraints in each equivalence class: (1) there are at least *l* well-represented values; (2) any two sensitive values are not *e*-similar. Furthermore, the paper designs a liner-complexity maximum bucketization greedy algorithm to implement the model. Experimental results show that the anonymous data satisfied (*l, e*)-diversity has a higher diversity degree than that satisfied *l*-diversity, so (*l, e*)-diversity can protect privacy more effectively than *l*-diversity.**

***Index Terms*—Data privacy, *l*-diversity, (*l, e*)-diversity, anatomy, semantic similarity attack**

I. INTRODUCTION

Microdata, such as medical patient data, demographic data, and business data, play an increasingly important role in trend analysis, disease research, and market analysis etc, therefore many organizations are collecting and publishing microdata. However, the microdata contain private information, whose publishing or sharing will threaten individuals' privacy. How to effectively protect individual privacy on publishing microdata has become a hot topic in data mining area [1]. *K*-anonymity [2][3] has been widely concerned for its security and effectiveness. It requires that each tuple has at least *k*-1 indistinguishable tuples with respect to quasi-identifier in released data, so that adversaries cannot identify an individual from the released data. However, *K*-anonymity cannot resist homogeneity attack and background knowledge attack. So Machanavajjhala [4] proposed an *l*-diversity model, which requires that sensitive attributes of each equivalence class have at least *l* well-represented values. However, *l*-diversity does not control distribution of sensitive values in each equivalence class, so it cannot resist skewness attack and similarity attack. To address the problem, Li Ninghui [5] proposed *t*-closeness framework, which requires that the distribution of a sensitive attribute value in any equivalence class is close to the distribution of the attribute value in the overall table to resist skewness attack. However, *t*-closeness reduces the data utility greatly. To protect privacy more effectively, many other anonymization models have been proposed [6-9].

However, the anonymity models above-mentioned do not consider the semantic similarity between sensitive values, so they cannot resist semantic similarity attack. Semantic similarity attack is similar to homogeneity attack. When sensitive values in an equivalence class are so semantically similar, adversaries can infer individuals' privacy easily. For example, table I is an original table, table II and table III are both 3-diversity anonymous tables by anatomy techniques [10]. Table II is a quasi-identifier table (*QIT*) and table III is a sensitive attribute table (*SAT*). If the adversary can infer that Tom is in the 3rd equivalence class by his background knowledge, then he can infer that Tom suffers from stomach trouble. This is because the three values of *Disease* in the 3rd equivalence class all belong to gastropathy. Tom's privacy is compromised, though the released data may satisfy *l*-diversity or *t*-closeness.

TABLE I.
THE ORIGINAL TABLE

Tuple	Name	Age	Sex	Zipcode	Disease
t_1	Alice	23	F	13010	Flu
t_2	Bill	25	F	13050	penumonia
t_3	Bob	30	M	13020	Flu
t_4	Sophia	36	F	13220	Carcinoid
t_5	Lucy	39	M	13221	Cancer
t_6	Steven	42	M	13226	Cancer
t_7	Tom	52	F	14850	Gastric ulcer
t_8	Paul	53	M	14862	Dyspepsia
t_9	Ellen	61	M	14802	Gastritis

doi:10.4304/jcp.9.1.59-64

TABLE II.
A 3–DIVERSITY ANONYMOUS TABLE *QIT*

Age	Sex	Zipcode	Ground-ID
23	F	13010	1
25	F	13050	1
39	M	13221	1
30	M	13020	2
36	F	13220	2
42	M	13226	2
52	F	14850	3
53	M	14862	3
61	M	14802	3

TABLE III.
A 3–DIVERSITY ANONYMOUS TABLE *SAT*

Ground-ID	Disease
1	Flu
1	pneumonia
1	Cancer
2	Flu
2	Carcinoid
2	Cancer
3	Gastric ulcer
3	Dyspepsia
3	Gastritis

To resist similarity attack, (k, e)-anonymity [11], (ε, m)-anonymity [12] and multi-level distinct l-diversity [13] have been proposed. But these models are oriented to numerical sensitive attributes. They are not suitable to categorical sensitive attributes. To address the problem, we propose a (l, e)-diversity. Our main contributions are as follows: (1) we propose a semantic similarity measurement method oriented to categorical attributes based on semantic hierarchy tree; (2) we propose a (l, e)-diversity model based on the semantic similarity measurement; (3) we propose a maximal-bucket first greedy algorithm(MBF) to implement the (l, e)-diversity based on anatomy.

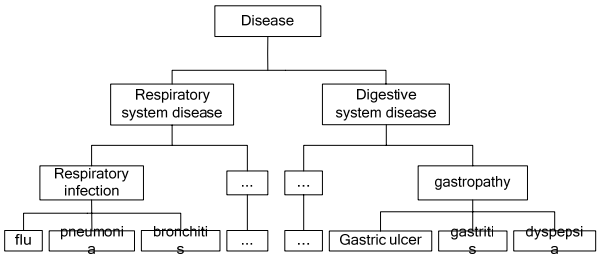

Figure 1. The semantic hierarchy tree of the disease attribute

II. THE (L, E)-DIVERSITY ANONYMITY MODEL

We use a semantic hierarchy tree of a sensitive attribute to describe the relationship between sensitive values. For example, Fig.1 is a semantic hierarchy tree of the *disease* attribute. The root of the tree is the whole set of the *disease* attribute, and the leaves are specific values. The level of root is 0, the son of the root is 1, and the rest can be done in the same manner. Semantic similarity between two values can be measured according to the path length between the two values on the semantic hierarchy tree. For example, flu and pneumonia have the common parent on the tree, so they are similar; however flu and gastritis have the common great-grandparent, so they are comparative dissimilar.

Definition 1 (Equivalence Class). A table can be horizontally partitioned into some subsets E_1, E_2, ... , E_m, s.t. $\bigcup_{i=1}^{m} E_i = D$ and $\forall i, j(1 \le i \ne j \le m)\ \ G_i \cap G_j = \varnothing$, then we say that E_1, E_2, ... , E_m is a partition of the table and E_i (i=1...m) is an equivalence class of the table.

Definition 2 (e-similar). Let h_1, h_2 be level of two sensitive values e_1, e_2 in their semantic hierarchy tree respectively, h_c be level of their most closest common ancestor, e=[(h_1-h_c)+(h_2-h_c)]/2, we can say that e_1 and e_2 are e-similar.

If two values are e-similar, we say that the similarity of the two values is e.

For example, in the semantic hierarchy tree of *disease*, seeing Fig.1, 'flu' and 'pneumonia' are 1-similar, 'flu' and 'gastritis' are 3-similar.

Definition 3 ((l, e)-diversity). Let *SAT* be a sensitive attribute table, E be an equivalence class of *SAT*. If E is l-diversity, and the similarity of any two values in E is more than e, we can say that E is (l, e)-diversity. If all of equivalence classes of *SAT* are (l, e)-diversity, *SAT* is (l, e)-diversity.

TABLE IV.
A (3, 1)–DIVERSITY ANONYMOUS TABLE *QIT*

Age	Sex	Zipcode	Ground-ID
23	F	13010	1
39	M	13221	1
52	F	14850	1
30	M	13020	2
36	M	13226	2
53	M	13482	2
25	F	13050	3
36	F	13220	3
61	M	14802	3

TABLE V.
A (3,1)–DIVERSITY ANONYMOUS TABLE *SAT*

Ground-ID	Disease
1	Flu
1	Cancer
1	Gastric ulcer
2	Flu
2	Cancer
2	Dyspepsia
3	pneumonia
3	Carcinoid
3	Gastritis

(l, e)-diversity model requires that sensitive attributes of each equivalence class have at least l well-represented values and similarity of any two sensitive values in an equivalence class is more than e. For example, e=1 requires that two sensitive values in any equivalence class cannot have a common father in semantic tree.

(l, e)-diversity model can preserve privacy more effectively than l-diversity model when e is larger than 1. Such as, table V is (3, 1)-diversity table. Even if attackers know that a patient belongs to the first equivalence class, but since the similarity of any two sensitive values is more than 1, it is difficult for adversaries to infer the category of disease that the patient is suffered from.

III. MAXIMAL-BUCKET FIRST ALGORITHM

We propose a maximal-bucket first (MBF) algorithm to achieve (l, e)-diversity. The idea is to partition an original table into several equivalence classes, and to make each equivalence class satisfied (l, e)-diversity constraint. Firstly, the MBF algorithm places all records with e-similar sensitive values into the same set respectively, which is called buckets. Secondly, the MBF algorithm selects records from different buckets to constitute an equivalence class sequentially according to the size of buckets, until the equivalence class is (l, e)-diversity. The algorithm recycles the process to construct equivalence classes until it cannot construct a new equivalence class satisfied (l, e)-diversity constraint. Then the algorithm adds the remaining records to the generated equivalence classes on the condition that the added records do not destroy the diversity of the equivalence class. Finally, the algorithm suppresses the remaining records which cannot be added to any equivalence class. The algorithm is described in algorithm 1.

Algorithm 1: Maximal-Bucket First (MBF) algorithm

Input:
the original data set D;
diversity parameter constraint l;
similarity constraint e;
the semantic hierarchy tree of a sensitive attribute;
Output:
the quasi-identifier attribute table *QIT*;
sensitive attribute table *SAT*.
Steps:
1. constitute bucket set $G(G_1, G_2, \dots, G_i, \dots, G_m)$;
   $k = 0$; $QIT = \emptyset$; $SAT = \emptyset$
2. while exists l buckets are not empty do
   (1) sort G in descending order based on the size of each bucket;
   (2) $E_k = \emptyset$;
   (3) for $i = 1 : l$
      (i) select an element t from the i-th bucket G_i in the bucket descending sequence;
      (ii) $E_k = E_k \cup \{t\}$, $G_i = G_i - \{t\}$;
   (4) $k=k+1$;
   (5) $T' = T' \cup E_k$;
   end while;
3. deal with the remain records :
   (1) for each remaining record r_t, find its' nearest equivalence class E, if $E \cup \{r_t\}$ satisfies (l, e)-diversity, $E = E \cup \{r_t\}$;
   (2) suppress all the remaining records which cannot be added to any generated group ;
4. assign Ground-ID for each equivalence class and output *QIT* table and *SAT* table.

In the MBF algorithm, the time complexity of step 1 is O(n); the time complexity of step 2 is O($n/l*m*logm$), where m is the number of bucket; the time complexity of step 3 is O(z), where z is the number of remaining records, generally z and l are relatively small and can be ignored. At the worst case, m is equal to n. So in that the worst case, the time complexity of the algorithm is: O(n)+O($n/l*m*logm$)+O(z)=O(n^2log$n$).

For example table I is an original table. Disease is a sensitive attribute. Fig.1 is a semantic hierarchy tree of Disease. 1-Similar sensitive values include: respiratory infection={flu, pneumonia, bronchitis}, tumour={Cancer, Carcinoid}, stomach={Gastric ulcer, Dyspepsia, Gastritis}. We will constitute a (3, 1)-diversity table from table I. At first the records in table I are divided into three buckets: the first bucket whose sensitive values are "Flu" or "pneumonia" has records { t_1, t_2, t_3}; the second bucket whose sensitive values are "cancer" or "Carcinoid" has records { t_4, t_5, t_6}; the third bucket whose values are "Gastric ulcer", "Dyspepsia" or "Gastritis" has records { t_7, t_8, t_9}.Then the MBF algorithm chooses records t_1, t_5, t_7 from the three buckets respectively to constitute the first equivalence class, then removes them from each bucket. After recalculating the record number of each bucket and sorting descend according to the size of each bucket, the algorithm chooses records t_3, t_6, t_8 from the above three buckets respectively to constitute the second equivalence class. Repeat the above process and finally release table IV and table V.

IV. QUALITY EVALUATION OF ANONYMOUS DATA

A. Metrics of Information Loss

We use the probability of data reconstruction [14] to measure data utility of anonymous data.

Each tuple t can be regarded as a point in a (d+1)-dimensional space *QS* (including all the *QI* and *SA*), i.e. t = (*QI*[1], *QI*[2], ... , *QI*[d], *SA*[d+1]). In *QS* space, probability that t may occur can be represented by a probability density function P_t, seeing formula (1).

$$P_t(x) = \begin{cases} 1 & if \ \ x=t \\ 0 & otherwise \end{cases} \tag{1}$$

where x is a random variable in *QS*.

For example, we assume that table I has two attributes {*Age*, *Disease*}, the first record t_1 corresponds to point (t_1[*Age*], t_1[*Disease*]), where t_1[*Age*]= 23, and t_1[*Disease*] = 'flu'. The possible probability density function of record t_1 can be described by formula (2).

$$P_{t_1}(x) = \begin{cases} 1 & if \ \ x=\left(t_1[Age], t_1[Disease]\right) \\ 0 & otherwise \end{cases} \tag{2}$$

where x is a random variable in the two-dimensional space QS_{AD} defined by *Age* and *Disease*.

Let v_1, v_2, ... , v_l be l distinct values of an equivalence class (as shown in table V, l=3), $c(v_i)(1<=i<=l)$ be the number of records in the *SAT* corresponding to v_i. The reconstruction probability of t based on anatomy can be described by formula (3).

$$P'_{t_1}(x) = \begin{cases} c(v_1)/|E| & if \ \ x=(t[1],\dots,t[d],v_1) \\ \cdots & \cdots \\ c(v_l)/|E| & if \ \ x=(t[1],\dots,t[d],v_l) \\ 0 & otherwise \end{cases} \tag{3}$$

where $|E|$ is the number of records in equivalence class E, the QI-values $t[1], \ldots, t[d]$ of t are directly released in the QIT.

For example, table IV shows that the first record belongs to the first equivalence class and the equivalence class has three corresponding records of the first equivalence class in table V, so the reconstruction probability of this record is formula (4).

$$P'_{t_1}(x)=\begin{cases}1/3 & if\ \ x=(23,flu)\ or\ x=(23,cancer)\ or \\ & \quad x=(23,gastrirticuler) \\ 0 & otherwise\end{cases} \tag{4}$$

Based on formula (2) and formula (4), information loss rate of record t can be also calculated, seeing formula (5).

$$IL(t)=\int_{x\in QS}(P'_t(x)-P_t(x))^2\,d(x) \tag{5}$$

For example, we can calculate that information loss rate of record t_1 is $IL(t_1)=(1/3-1)^2\approx 0.44$.

Information loss rate of table T can be calculated by formula (6).

$$IL(T)=\sum_{i=1}^{n} IL(t_i) \tag{6}$$

where n is the number of records in table T.

B. Diversity Metrics for Anonymous Data

Generally speaking, the larger the diversity of sensitive values in an equivalence class is, the greater the capability to thwart semantic similarity attack is. In order to measure the diversity of an equivalence class, we introduce the definition of difference matrix of an equivalence class.

Definition 4 (Distance of sensitive attribute values). Let v_i, v_j be two sensitive attribute values, q_i, q_j be their corresponding levels, p_{ij} be the level of their most closest common ancestor of v_i and v_j, the distance between the sensitive value v_i and v_j can be defined by (7).

$$Dist(v_i,v_j)=((p_{ij}-q_i)+(p_{ij}-q_j))/2 \tag{7}$$

Definition 5 (Difference matrix of an equivalence class). Let E be an equivalence class of table T, q be number of element in E, the difference matrix of E can be defined by a q-order square matrix denoted by C_E, seeing formula(8).

$$C_E=\begin{bmatrix} D_{11} & D_{12} & \cdots & D_{1q} \\ D_{21} & D_{22} & \cdots & D_{2q} \\ \vdots & \vdots & & \vdots \\ D_{q1} & D_{q2} & \cdots & D_{qq} \end{bmatrix} \tag{8}$$

where D_{ij} is the distance between the i-th element and the j-th element in the equivalence class, i.e. $D_{ij}=\mathrm{Dist}(v_i,v_j)$.

From definition of difference matrix of an equivalence class, we can see that $D_{ii}=0$, $D_{ij}=D_{ji}$.

Definition 6 (Diversity degree of an equivalence class). Let SAT be a sensitive attribute table, E be an equivalence class of SAT, q be element number of $E(q>1)$, the diversity degree $DOD(E)$ of an equivalence class E can be defined by formula (9).

$$DOD(E)=\frac{\sum_{i=1}^{q-1}\sum_{j=i+1}^{q} D_{ij}}{q} \tag{9}$$

In formula (9), the molecular is the sum of upper right triangular elements of difference matrix of an equivalence class E.

Definition 7 (Average diversity). Let T be a sensitive table, $T=\{E_1, E_2, \ldots, E_g\}(E_i(1\le i\le g)$ be an equivalence class of $T)$. The average diversity $DOT(T)$ can be defined by formula (10).

$$DOT(T)=\frac{\sum_{i=1}^{g} DOD(E_i)}{g} \tag{10}$$

where g is the number of equivalence classes in T.

V. Experimental Evaluation

Experiments are performed on PC with 3.3GHz Intel Core processor and 3.0G RAM. The operation system is Windows XP and the programming language is Java. Experimental data uses the Adult dataset from the UCI machine learning data warehouse downloadable at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult. The dataset have 32561 records and 15 attributes. Our experiments select 6 attributes. The experimental dataset is described in table VI.

TABLE VI.
THE DESCRIPTION OF ADULT DATASET

NO.	Attribute	Type	Distinct values	Height
1	Age	Numeric	74	5
2	Workclass	Categorical	8	3
3	Marriage	Categorical	7	3
4	Gender	Categorical	2	2
5	Race	Categorical	5	3
6	Education	Sensitive	16	/

The experiment compares (l, e)-diversity with l-diversity from the view of information loss, diversity degree, efficiency. Parameter e is set 1.

A. Information Loss

We use formula (6) to calculate the information loss of anonymous dataset. Fig.2 plots the performance curves of the information loss of anonymous data over various n with l = 4. Fig.3 plots the performance curves of the information loss of anonymous data over various l with n = 30000. Fig.2 and Fig.3 show that information loss of (l, e)-diversity model is higher than l-diversity model under the same condition. This is because (l, e)-diversity model enhances the diversity constraint of sensitive attribute so that the probability of data reconstruction reduces and information loss increases.

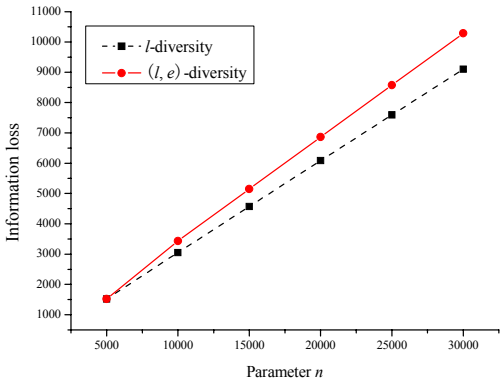

Figure 2. Information loss for various n

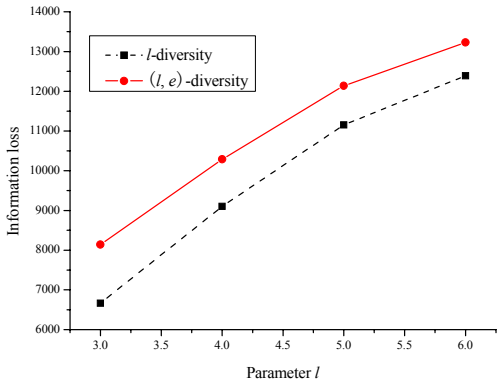

Figure 3. Information loss for various l

B. Diversity Degree

A formula (10) has been adopted to measure the diversity degree of anonymous data. Diversity degree reflects privacy preservation strength of anonymous data. The greater the diversity degree is, the stronger privacy preservation is. Because greater diversity means lower probability so that adversaries can infer individuals' privacy.

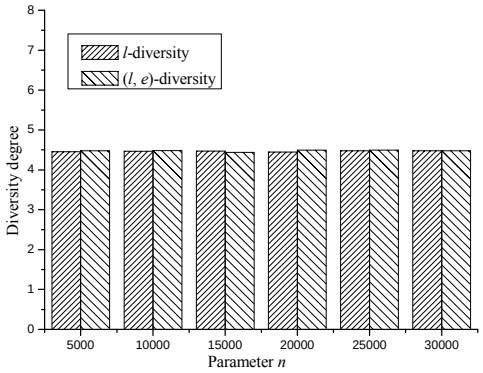

Figure 4. Diversity degree for various n

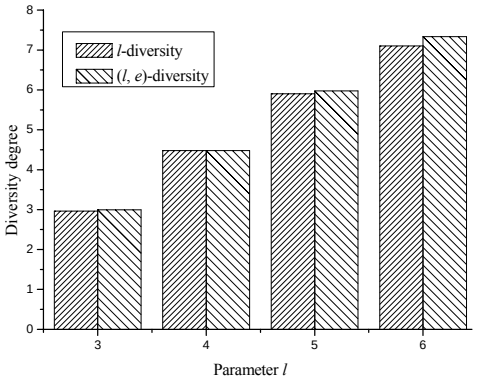

Figure 5. Diversity degree for various l

Fig.4 plots the performance curves of diversity degree over various n with l = 4. Fig.5 plots the performance curves of the diversity degree over various l with n = 30000. Fig.4 and Fig.5 show that diversity degree of anonymous data satisfying (l, e)-diversity is higher than that satisfying l-diversity under the same condition, this is because the (l, e)-diversity model adds sensitive attributes similarity constraint to anonymous data. So privacy preservation ability of (l, e) -diversity is stronger than that of l-diversity.

C. Algorithm Efficiency

Fig.6 plots the performance curves of the execution time of MBF algorithm over various n with l = 4. Fig.7 plots the performance curves of the execution time of MBF algorithm over various l with n = 30000. Fig.6 and Fig.7 show that the efficiency of MBF algorithm to achieve (l, e)-diversity model is lower than that to achieve l-diversity model. This is because achieving (l, e)-diversity needs to increase diversity of sensitive attributes for each equivalence class which needs more time.

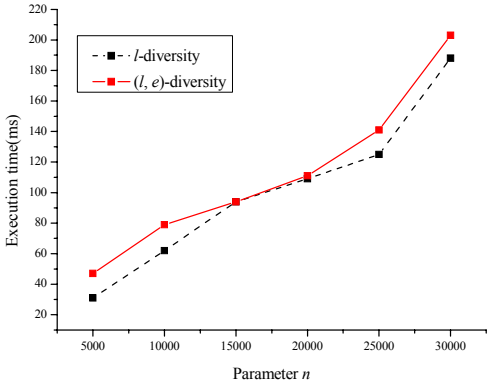

Figure 6. Execution time for various n

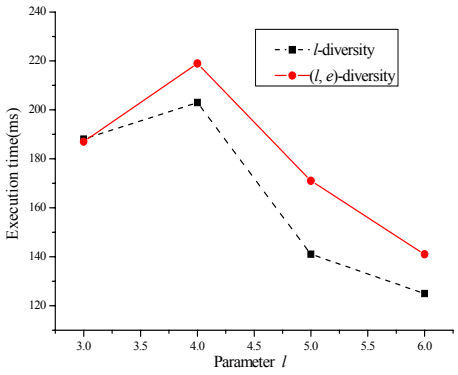

Figure 7. Execution time for various *l*

VI. CONCLUSIONS

The paper proposes a (*l*, *e*)-diversity model which requires that the sensitive values in publishing data have at least *l* well-represented and any two sensitive values are not *e*-similarity in each equivalence class. The paper also proposes a MBF algorithm to achieve (*l*, *e*)-diversity model. The experimental results show that the (*l*, *e*)-diversity has a higher diversity degree than *l*-diversity, so it can preserve privacy information more effectively.

Future research includes: (1) the work in this paper is oriented to single sensitive attribute microdata. Extending the work to multiple sensitive attributes microdata is a significant topic; (2) the work herein extends *l*-diversity to resist semantic similarity attack. Extending the method to improve other anonymity models to thwart semantic similarity attack is also an interesting work.

ACKNOWLEDGMENT

This work has been supported by the National Natural Science Foundation of China under Grant No. 61170108 and No. 6110019 and the National Natural Science Foundation of Zhejiang Province under Grant No. Y1100161 and No. Q13F020026.

REFERENCES

[1] Fung BCM, Wang Ke, Chen Rui, et al. Privacy-preserving data publishing: a survey on recent developments[J]. *ACM Computing Surveys*, 2010, 42(4): 1-55.

[2] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information[C]. *In: Proc. of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems*. 1998: 188.

[3] Samarati P. Protecting respondents' identities in microdata release[J].*IEEE Trans on Knowledge and Data Engineering*. 2001, 13(6) : 1010 -1027.

[4] Machanavajjhala A, Gehrke J, Kifer D. *l*-Diversity: Privacy beyond *K*-anonymity[C]. *In: Proc. of the 22nd International Conference on Data Engineering. Atlanta: IEEE Computer Society*, 2006: 24-35.

[5] Li Ning-hui, Li Tian-cheng, Venkatasubramanian S. *t*-Closeness: privacy beyond *k*-anonymity and *l*-diversity[C]. *In: Proc. of the 23rd ICDE*, 2007: 106-115.

[6] Wong CRW, Li Jiuyong, Fu A W C, et al. (*α*, *k*)-anonymity: an enhanced *k*-anonymity model for privacy preserving data publishing[C]. *Proc. of the 12th ACM SIGKDD Conference[C]*. Philadelphia, PA: ACM Press 2006: 754-759.

[7] Qian Wang, Cong Xu, Min Sun. Multi-dimensional *k*-anonymity Based on Mapping for Protecting Privacy [J]. *Journal of Software*, Vol 6, No 10 (2011), 1937-1944, Oct 2011.

[8] Md Enamul Kabir, Hua Wang. Microdata Protection Method Through Microaggregation: A Systematic Approach [J]. *Journal of Software*, Vol 7, No 11 (2012), 2415-2423, Nov 2012.

[9] Qian Wang, Cong Xu, Min Sun. Protecting Privacy by Multi-dimensional *K*-anonymity [J]. *Journal of Software*, Vol 7, No 8 (2012), 1873-1880, Aug 2012.

[10] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation[C]. *In: Proc. of the 32nd International Conference on Very Large Data Bases. Seoul: VLDB Endowment*, 2006: 139−150.

[11] Koudas N,Srivastave D, Yu t, et al. Aggregate query answering on anonymized tables[C]. *Proc of the 23th Int Conf on Data Engineering. IEEE Computer Soceity*, 2007: 116-125.

[12] Li J X, Tao Y F, Xiao X K. preservation of proximity privacy in publishing numerical sensitive data[C], *Proceedings of ACM Conference on Management of Data. Vancouver, BC, Canada*, 2008: 473 -486.

[13] Han Jianmin, Yu Juan, Yu Huiqun, Jia Jiong. A Multi-level *l*-Diversity Model for Numerical Sensitive Attributes[J]. *Journal of Computer Research and Development*, 2011.48(1):147-158.

[14] Xiao Xiao-kui, Tao Yu-fei, Anatomy .Simple and effective privacy preservation[C]. *Proc of the 32nd International Conference on Very Large Data Bases, Seoul, Korea*, 2006: 139-150.

[15] Li Jiuyong, Wong Raymond Chi-Wing, Fu Ada Wai-Chee, et al. Achieving *k*-anonymity by clustering in attribute hierarchical structure[C]. *DaWak. LNCS 4081, Springer-verlag, Berlins, Heidelberg*, 2006:405-416.