An Improved V-MDAV Algorithm for *l*-Diversity

Han Jian-min
Department of Computer Science and Engineering, East China University of Sci & Tech, Shanghai 200237, China
hanjm@zjnu.cn

Cen Ting-ting
Math, Physics and Information Engineering College of Zhejiang Normal University, Jinhua 321004, China
02190202@zjnu.net

Yu Hui-qun
Department of Computer Science and Engineering, East China University of Sci & Tech, Shanghai 200237, China
yuhuiqun@ecust.edu.cn

Abstract

V-MDAV algorithm is a high efficient multivariate microaggregation algorithm and the anonymity table generated by the algorithm has high data quality. But it does not consider the sensitive attribute diversity, so the anonymity table generated by the algorithm cannot resist homogeneity attack and background knowledge attack. To solve the problem, the paper proposes an improved V-MDAV algorithm, which first generates groups satisfying l-diversity, then extends these groups to the size between l and 2l-1 to achieve optimal k-partition. Experimental results indicate that the algorithm can generate anonymity table satisfying sensitive attribute diversity efficiently.

1. Introduction

Privacy preservation is an important issue in the release of data for mining purpose. The k-anonymity model, which was introduced for protecting individual identification by Samarati and Sweeney in 1998 [1], has been extensively investigated for its simplicity and effectiveness. *K*-anonymity requires that each record in the anonymized table be indistinguishable with at least *k*-1 other records within the dataset with respect to a set of quasi-identifier attributes. In this case, individuals cannot be uniquely identified by adversary, so the individuals' privacy can be preserved. In 2001, Samarati proposed an approach to achieve k-anonymization by generalization and suppression [2]. In 2002, Sweeney investigated how to thwart attacks on *k*-anonymity model [3]. In the same year, she presented Minimal Generalization Algorithm (MinGen) using generalization and suppression [4]. Later, Ninghui Li [5,6], Zude Li [7], Xiaochun Yang [8] et al. also did many research works in this area.

Homogeneity attack and background knowledge attack are two kinds of frequently-used attacks which compromise *k*-anonymous table. Machanavajjhala indicated that *k*-anonymity table without any constraint might lead to sensitive information disclosure, so introduced a solution——*l*-diversity model [9], which required the diversity of every equivalence class should be no less than *l* so as to enhance the difficulty to link a sensitive value to an individual. Chi-Wing introduced a (α,k)-anonymity model [10], which can prevent inference disclosure through the diversity of sensitive attributes in every equivalence class of anonymization dataset. This diversity is achieved by controlling the frequency of every sensitive value. Li Zude proposed a (*k*, *l*)-anonymity model [7] to support the individual-defined (*k*, *l*) mechanism for more flexible individual data anonymization.

All of the works mentioned above are based on generalization and suppression technique, which has some defects on efficiency, information loss and implementation. Microaggregation technique has been introduced to implement dataset *k*-anonymization recently, which remedies some defects of generalization and suppression. The core idea of microaggregation is: Dataset are partitioned into some clusters based on some heuristic methods. Each cluster should contain at least *k* records. The records in the same cluster are as similar as possible. Then the record values of each cluster are replaced by the cluster's centroid to implement *k*-anonymization. Microaggregation is originally designed for numerical data and recently extends for categorical data [11,12]. Multivariate microaggregation has been

Supported partially by the NFS of China under grants No. 60473055 and No. 60773094, and Shanghai Shuguang Program under grant No. 07SG32.

978-0-7695-3151-9/08 $25.00 © 2008 IEEE
DOI 10.1109/ISIP.2008.110

proved as an NP-hard problem [13]. Therefore, several heuristic methods have been proposed for multivariate microaggregation, which can be categorized into two classes: fixed-size microaggregation and variable-size microaggregation. Typical fixed-size microaggregation algorithms include: MD algorithm [15], MDAV algorithm [11], MDAV-generic algorithm [11]. Fixed-size microaggregation algorithms are efficient, but sometimes may lead to mistaken clustering. The clustering quality of variable-fixed microaggregation is better than that of the fixed-sized microaggregation. Some influential variable-size microaggregation algorithms include: MST algorithm [16], V-MDAV algorithm [17], TFRP algorithm [18], μ-Approx algorithm [19].

At present, there are less works about *l*-diversity implemented by microaggregation. Domingo-Ferrer proposed a microaggregation algorithm to satisfy *p*-sensitive *k*-anonymity constraint [20], and applied it to preserve location privacy. *p*-Sensitive *k*-anonymity requires that the size of each group is no less than *k* and the number of distinct values for each sensitive attribute is at least *p* within the same group. However, the method implements sensitive attribute diversity by enlarge the value of *k*, i.e., in order to increase the diversity of one group, the size of all of groups must be enlarged, so the solution is not optimal and the information loss is higher. The paper proposes an *l*-diversity microaggregation algorithm based on V-MDAV, which can implement *l*-diversity in multivariate microaggregation, so the anonymity table generated by the algorithm can resist homogeneity attack; background knowledge attack.

2. Microaggregation

2.1 Related concepts

Definition 1 (*k*-anonymity). Given a table $T(A_1, A_2,\ldots,A_n)$, and its quasi-identifier *QI*, *T* satisfies *k*-anonymity if and only if each sequence of values in $T[QI]$ appears with at least *k* occurrences in $T[QI]$, where $T[QI]$ denotes the projection, maintaining duplicate tuples, of attributes in *QI*.

Definition 2 (*k*-**partition).**The set of original records are partitioned into several clusters in such a way that records in the same cluster are similar to each other as possible and the number of records in each cluster is no less than *k*.

Definition 3 (**Aggregation).** An aggregation operator (for example, the mean for continuous data or the mode for categorical data) is computed for each cluster to get each cluster's centroid which is used to replace the original records, in other words, each record in a cluster is replaced by the cluster's prototype.

The optimal microaggregation is based on the optimal *k*-partition, which requires maximize within-group homogeneity. It has been proved in [13] that the size of optimal *k*-partition is between *k* and 2*k*-1.

Microaggregation can be operationally defined in term of the following two steps:

Step 1, *k*-partition: original records are *k*-partitioned into several clusters.

Step 2, aggregation: centroid is computed to replace the cluster's record.

The object of *k*-partition is to maximize within-group homogeneity. Within-group homogeneity can be measured based on distance of *p*-dimension(*X*,*Y*). The definitions of homogeneity and cluster's centroid are different for different type. The following part will discuss the measures for numerical data. The measures for categorical data are discussed in [12].

2.2 Distance definitions for numerical data

For numerical data, Euclidean distance is used to measure the distance of two records. The distance measure is defined as (1).

$$d(X,Y) = \| X - Y \| = \sum_{i=1}^{p} (X(i) - Y(i))^2 \quad (1)$$

where $X(i)$, $Y(i)$ is the *i*-th dimension attribute value of vector *X, Y*.

Let G_i be an equivalence class after *k*-partition, the homogeneity measure of G_i is defined as (2).

$$GSE(G_i) = \sum_{j=1}^{n_i} d(X_{ij} - \overline{X_i}) \text{ where } \overline{X_i} = (\sum_{j=1}^{n_i} X_{ij})/n_i \quad (2)$$

Where n_i is the record number of the *i*-th group G_i, $n_i \geqslant k$, X_{ij} is *j*-th record of G_i, $\overline{X_i}$ is centroid of G_i. The lower *GSE* is, the higher the within-group homogeneity will be.

The sum of all the within-group homogeneity is defined as (3)

$$SSE = \sum_{i=1}^{g} GSE(G_i) = \sum_{i=1}^{g} \sum_{j=1}^{n_i} d(X_{ij} - \overline{X_i}) \quad (3)$$

where *g* is the number of cluster, $n = \sum_{i=1}^{g} n_i$.

The sum of the table's squares is defined as (4).

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} d(X_{ij} - \overline{X}) \text{ where } \overline{X} = (\sum_{l=1}^{n} X_l)/n \quad (4)$$

where $\overline{X}$ is the average data vector over the whole data set, i.e., centroid of the data set.

For numerical data, the average data vector $\overline{X_i}$ over the *i*-th group can be used as *i*-th group's centroid.

2.3 Information loss measures for numerical data

For a data set *T*, the *SST* does not change, but different *k*-partition can lead to different *SSE*. Information loss *IL* can be defined as (5).

$$IL = SSE / SST \quad (5)$$

The lower *IL* is, the better the data utility will be.

2.4. Disclosure risk measures for numerical data

Disclosure risk measure is used to assess the security of anonymity table, which is based on the probability of inferring the original record from anonymity table. We adopt the Distance Linkage Disclosure risk (*DLD*) model proposed in [14] to measure the anonymity table's security property.

Definition 4 (linked-record). For any record t in anonymity table, computing the distance of t to other records in the table, we can get the nearest record set t' and the second nearest record set t''. If the set of t'(or t'') have the same record number of the corresponding original record, then the record t is linked-record.

Let *linked-records-num* be the number of linked-record in anonymity table, *total-records-num* to be the number of total records, and then *DLD* is defined as (6)

$$DLD=linked\text{-}record\text{-}num \,/\, total\text{-}record\text{-}num \quad (6)$$

3. *L*-diversity

3.1 *L*-diversity principle

Definition 3 (**Homogeneity attack).** In an anonymity table, if there exists an equivalence class in which all tuples share the same value of sensitive attributes, it will be exposed to homogeneity attack, i.e., adversary can easily infer individual's sensitive value by linking external table.

Definition 5 (Background knowledge attack). Adversary can discover sensitive information from anonymized table using his/her background knowledge.

Machanavajjhala indicates that k-anonymity can not resist homogeneity attack and background knowledge attack and then introduces *l*-diversity principle [9].

Definition 6 (*l*-diversity principle). An equivalence class is *l*-diverse if it contains at least *l* "well-represented" values for sensitive attribute. A table is *l*-diverse if every equivalence class is *l*-diverse.

Machanavajjhala [9] gives several interpretations of the term "well-represented": Distinct *l*-diversity, Entropy *l*-diversity, recursive (*c,l*)-diversity, positive disclosure recursive (*c*,*l*)-diversity and negative/positive disclosure recursive (c_1, c_2, *l*)-diversity [9].

3.2 *L*-diversity orienting numerical sensitive attributes

The model proposed in [9] has some defects in processing numerical sensitive attributes. For example, table 1 is a 2-diversity table. If adversary knows Andy's age, high, and weight, and he can knows from table 1 that Andy's record corresponds to the first equivalence class{25,165,70}. Although there are two different sensitive values {5500,5400} in this equivalence class, adversary can still conclude that Andy's salary is about 5450. Actually, sensitive information has been disclosed.

Table 1. **2-Diversity table**

Age(year)	High(cm)	Weight(kg)	Salary(yuan)
25	165	70	5500
25	165	70	5400

We can know that the cause of sensitive information disclosure is that the two sensitive values are too close. So we divide the numerical attribute domain into several levels, and let values in each level have the same diversity. Instead of calculating diversity according to original sensitive values, we calculate diversity according to the levels. For example, we classify salary domain into 5 levels showed in table 2, then we generalize 2-diversity table based on levels. Obviously, table 1 does not satisfy 2-diversity because 5500 and 5400 in the first equivalence class are in the same level-5.

Table 2. **Levels of salary**

Level	1	2	3	4	5
Range	[1000, 2000]	[2001, 3000]	[3001 4000]	[4001, 5000]	[5001, 6000]

3.3 Kinds of *l*-diversity

Definition 7 (Distinct diversity degree). Distinct diversity degree can be defined as (7):

$$D(E)=distinct(E) \quad (7)$$

where $distinct(E)$ denotes the number of different values in the equivalence class E,

For example, {5500,3200,2000}belongs to{5,3,2}, so its distinct diversity degree is: 3.

Definition 8 (Entropy diversity degree). Entropy diversity degree can be defined as (8):

$$D(E)=-\sum_{s\in S} p(E,s)\log p(E,s) \quad (8)$$

where $p(E, s)$ is the fraction of tuples in the equivalence class E with sensitive attribute level s.

Definition 9 (Diversity degree of table). Diversity degree of the table T can be defined as (9):

$$D(T)=\min(D(E_i)) \quad (9)$$

where E_i denotes the i-th equivalence class which obtained by anonymizing T.

4. Microaggregation algorithm for *l*-diversity

4.1 V-MDAV for *l*-diversity

V-MDAV algorithm [17] does not consider sensitive attributes diversity of the equivalence class. The paper proposes an *l*-diversity V-MDAV algorithm, See Figure 1. In this algorithm, $D(g)$ function returns diversity degree of equivalence class g.

Algorithm: *l*-diversity V-MDAV algorithm
Input: dataset *T*, diversity parameter *l*
Output: *l*-diversity table *T'*
Step: 1. Compute the distances between the records and store them in a distance matrix.
2. Compute the centroid *c* of the dataset.
3. Repeat, until diversity of remaining records < *l*
(1)Let *r* be the most distant record to *c*, and form a group $g_r=\{r\}$
(2)Size down ungrouped vectors of the dataset according to their distance to vector *r*, and form distance priority queue *Q*.
(3)Repeat, until equivalence class g_r satisfy *l*-diversity
{ diversity_first=$D(g_r)$
pop the head vector *v* from queue *Q*
diversity_Second=$D(g_r+v)$
if (diversity_second>diversity_first)
{ $g_r=g_r+v$
Mark *v* with assigned sign; }
}
(4)Extend group g_i
4. Assign the remaining records to their closest group.

Figure 1. ***l*-Diversity V-MDAV algorithm**

The policy of extending the group is: given the group's nearby records, if their distances to the group are smaller than their distances to the closest unassigned neighbor, and moreover, if their adding won't reduce the group's diversity, then add these records into the group.

Given a group *g* with *k* records, the record e_{min} among unassigned records outside *g* nearest to *g* and the minimum distance d_{in} between e_{min} and *g* are defined by (10) and (11):

$$d_{in} = \min_{j\in[1,Nun]}(d(e_i^g, e_j)), \forall i \in [1,k] \quad (10)$$

$$e_{\min} = \arg\min_{j\in[1,Nun]}(d(e_i^g, e_j)), \forall i \in [1,k] \quad (11)$$

where e_i^g denotes the *i*-th record in group *g*, e_j means the *j*-th record in the unassigned set of records and N_{un} is the number of unassigned records, i.e., the number of records which have not yet been assigned to any group. Next, the minimum distance d_{out} from the selected record e_{min} to any of the remaining unassigned records is found using (12):

$$d_{out} = \min_{j\in[1,Nun], e_{\min}\neq e_j}(d(e_{\min}, e_j)) \quad (12)$$

Supposing $D(g)$ returns the diverse degree of the equivalence class *g*, then (13) gives the decision criterion that whether e_{min} can add in group *g* or not.

$$ADD-RECORD = \begin{cases} YES & if\ (d_{in} < \gamma d_{out})\ and\ D(g+e_{\min}) \geq D(g) \\ NO & otherwise \end{cases} \quad (13)$$

where γ is a gain factor whose best value is due to space limitations. Group extension algorithm sees Figure 2.

Algorithm: Extension of the group
Input: equivalence class *g*, diversity constrain *l*, unsigned dataset *T*, gain factor γ
Output: extended equivalence class *g'*
Step: repeat
if ($d_{in}<\gamma d_{out}$) and ($D(g+e_{min})\geqslant D(g)$) then
{ $g=g+e_{min}$;
while Q is not null
{ e_{can}=Q.first();
if ($d_{in}<\gamma d_{out}$) and ($D(g+e_{can})\geqslant D(g)$) then
$g=g+e_{can}$;
else
Q1.add(e_{can});
Q.delete(e_{can}); }
Q=Q1;
}
else Q.add(e_{min});
until $D(g)=2l-1$ or e_{min} is the most distant record
return *g*

Figure 2. **Group extension algorithm**

4.2 . Algorithm analysis

Given a dataset with *n* records and its diversity constrain *l*. Taking no consideration of record dimension, the average time complexity of the algorithm is analyzed as follows: Computing the distances matrix can be done with $O(n^2)$. Computing the centroid has a cost of $O(n)$. Computing the distance from each record to the centroid so as to find the most distant record requires *n* distance computations. Thus, it has $O(n)$ cost. Supposing the size of each class is between $[l,2l-1]$, since $(3l-1)/2$ records are grouped on average, so about $2n/(3l-1)$ iterations are needed. Each iteration consists of: The cost of computing the diversity of group is $O(l)$; When extending the group, up to $l-1$ records can be added to the current group, say $(l-1)/2$ on average. And requiring averagely performing $n/2$ comparisons, so average time cost is $O(n(l-1)/4)$. Therefore the computational cost of the main loop is $O(2n/(3l-1)(l+n(l-1)/4))=O(n^2)$. Thus the overall complexity of *l*-diversity V-MDAV is $O(n^2)$.

5. Experimental results and analysis

5.1 Experimental datasets

The experiment used three real micro-datasets (Tarragona, Census and EIA) which have become the usual reference data sets for testing multivariate microaggregation [15,16,17,18,19]. The Tarragona dataset contains 834 companies' figure in the Tarragona area in 1995. Census and EIA were obtained from the Data Extraction System of the U.S. Bureau of the Census and the U.S. Energy Information Authority. All of the datasets

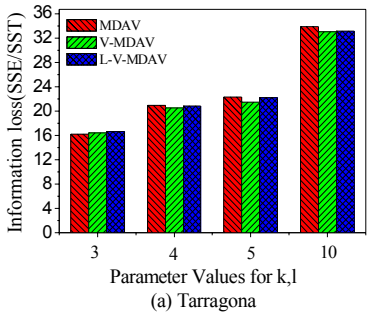
have only one sensitive attribute (*SI*). Table 3 gives the characteristics of above datasets.

Table 3. **Characteristics of experimental data**

Dataset	Record number	*QI* number	*SI* number
Tarragona	834	12	1
Census	1080	12	1
EIA	4092	10	1

5.2 Information loss comparison

Information loss caused by each algorithm is measured using (5). Figure 3a-c plots the performance curves of information loss over various k with three algorithms applied to Tarragona, Census, and EIA, respectively. In this experiments, we set γ=0.2 for V-MDAV and L-V-MDAV algorithms. Fig. 3a-c shows that the V-MDAV has lower information loss than the MDAV in most cases. This is because that the V-MDAV yields more homogeneous group, which implies lower information loss and clusters more reasonably than MDAV. However, L-V-MDAV has the highest information loss, because L-V-MDAV needs to satisfy sensitive attribute l-diversity which is a stronger constraint and causes more information loss.



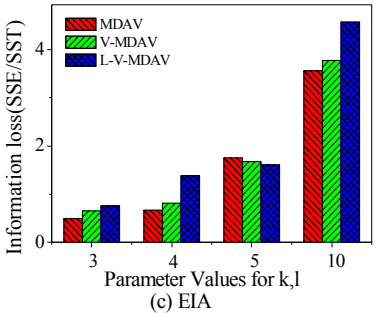
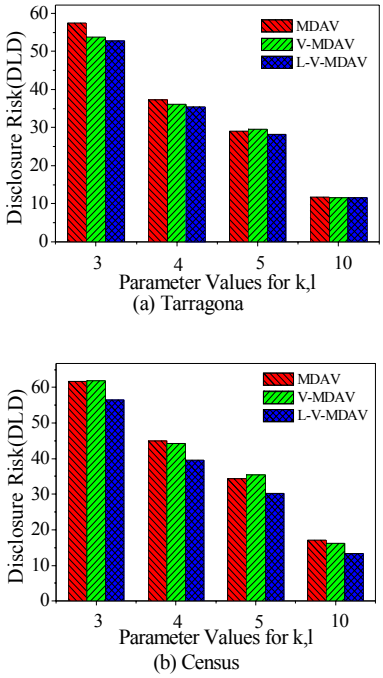


Figure 3. **Information loss comparison using three datasets**

5.3 Disclosure risk evaluate

In our experiments, Distance Linkage Disclosure risk (*DLD*) evaluation model is adopted. Fig. 4a-c plots the performance curves of disclosure risk over various k with different datasets. Figure 4a-c shows that *DLD* value will decrease with the k increasing. This is because that with the increase of k, records' distortion will increase, then the distances between the masked records and their original records will increase, so the probability of linked successfully will become small. Fig. 4 also indicates that for the same k, *DLD* values for L-V-MDAV、V-MDAV and MDAV have no distinct difference, which demonstrates that these three algorithms have no essential differences from the view of disclosure risk, but L-V-MDAV satisfies sensitive attribute diversity, and significantly outperforms other two algorithms from the view of resisting attack.



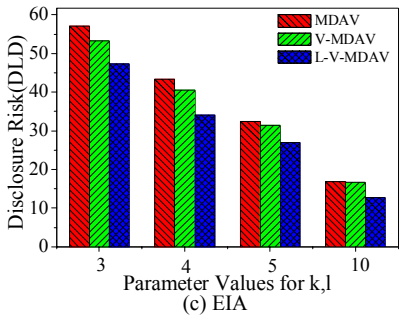
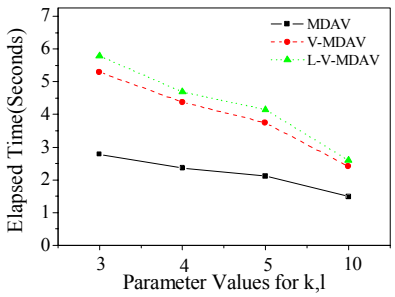
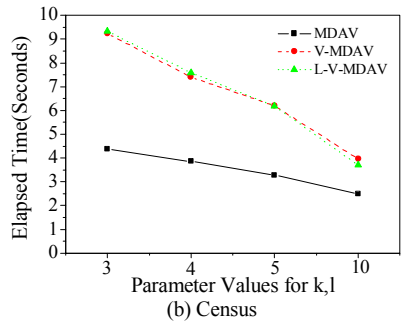


Figure 4. **Disclosure risk comparison using three datasets**

5.4 Running time comparison

Experiments are run on a 3.1GHz Intel Pentium 4 processor with 1.5G RAM, running the Window XP Professional Operating System. All algorithms are implemented in matlab 7.0. Figure 5 plots the execution time curves over various k values with the three algorithms applied to Tarragona, Census and EIA, respectively. The experimental results are average values of five times testing. Fig. 5a-c demonstrates that execution time of the three algorithms decreases with k increasing. That is because that with the k increasing, the clustering times will decrease. To the same k values, MDAV performs faster than all other methods; L-V-MDAV and V-MDAV's time cost are similar. It's because that both the L-V-MDAV and V-MDAV algorithm need to execute group extension after each group has formed, but MDAV need not. L-V-MDAV should satisfy the diversity of sensitive attributes, so its time cost is the highest, however the formed groups were larger than V-MDAV that makes the clustering times decrease. So L-V-MDAV sometimes executes faster, as shown in Fig. 5(a), sometimes executes slow, as shown in Fig. 5(c). But essentially, MDAV, V-MDAV, and L-V-MDAV's time complexity are same, all of them are $O(n^2)$.





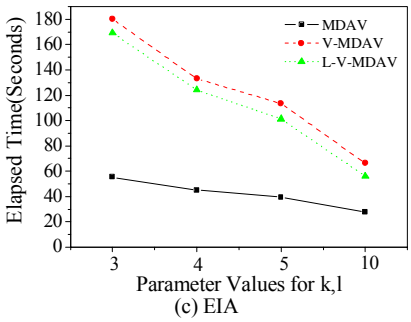


Figure 5. **Elapsed time comparison using different datasets**

6. Conclusion and future work

L-diversity V-MDAV algorithm improves V-MDAV algorithm on sensitive attribute diversity. From the view of safety, l-diversity V-MDAV is more secure. From the view of data utility, l-diversity V-MDAV generates more information loss for the enhanced constraint. From the view of time complexity, l-diversity V-MDAV has $D(g)$ function in the loop of V-MDAV algorithm. D(g) function has linear correlation with l, thus the total complexity of l-diversity V-MDAV still is $O(n^2)$.

Future researches are: (1) To develop l-diversity V-MDAV algorithms for categorical attribute. (2) To extend our ideas for handling multiple sensitive attributes.

References

[1] P. Samarati, L. Sweeney, "Generalizing Data to Provide Anonymity When Disclosing Information" (Abstract), *Proceeding of the 17th ACM-SIGMOD- SIGACT-SIGART Symposium on Principles of Database Systems*, IEEE press, Seattle, June 1998, pp. 188.

[2] P. Samarati, "Protecting Respondents' Identities in Microdata Release", *IEEE TKDE*, Nov./Dec. 2001, Vol 13, No. 6, pp. 1010-1027.

[3] L. Sweeney, "K-anonymity: a Model for Protecting Privacy", *Int'l Journal on Uncertainty Fuzziness and Knowledge Based Systems*, June 2002, Vol 10, No. 7, pp. 557-570.

[4] L. Sweeney, "Achieving k-anonymity Privacy Protection Using Generalization and Suppression", *Int'l Journal on Uncertainty Fuzziness and Knowledge Based Systems*, May 2002, Vol 10, No. 5, pp. 571-588.

[5] Tiancheng Li, Ninghui Li, "Towards Optimal k-anonymization", *Data and Knowl. Eng.*, July 2007.
[6] Ninghui Li, Tiancheng Li, Suresh. V., "t-Closeness: Privacy beyond k-anonymity and l-Diversity", *Proceeding of the 23rd ICDE*, Apr. 2007, pp. 106-115.
[7] Zude Li, Guoqiang Zhan, Xiaojun Ye, "Towards an Anti-inference (K, l)-anonymity Model with Value Association Rules", *DEXA*, Springer-Verlag Berlin Heidelberg, Krakow, Sep. 2006, pp. 883-893.
[8] Xiaochun Yang, Xiangyu Liu, Bin Wang, Ge Yu, "K-Anonymization Approaches for Supporting Multiple Constraints", *Journal of Software*, May 2006, Vol 17, No. 5, pp. 1222-1231.
[9] A. Machanavajjhala, J. Gehrke, D. Kifer, "L- Diversity: Privacy beyond k-anonymity", *Proceeding of the ICDE*, Atlanta, Apr. 2006, pp. 24-35.
[10] Wong R.C.W., Li J., Fu A.W.C., and Wang K., "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing", *Proceeding of the 12th ACM SIGKDD Conference on KDD*, PA: ACM Press, Philadelphia, Aug. 2006, pp. 754-759.
[11] J. Domingo-Ferrer, V. Torra, "Ordinal, continuous and heterogeneous k-anonymity through micro- aggregation", *Journal of Data Mining and Knowledge Discovery*, Sep. 2005, Vol 11, No. 2, pp. 195-202.
[12] V. Torra, "Microaggregation for categorical variables: a median based approach", *PSD*, LNCS, Barcelona, 2004, Vol. 3050, pp. 162-174.
[13] A. Oganian, J. Domingo-Ferrer, "On the complexity of optimal microaggregation for statistical disclosure control", *Statistical Journal of United Nations Economic Commission for Europe*, April, 2001, Vol 18, No. 4, pp. 345-354.
[14] J. Domingo-Ferrer, J.M. Mateo-Sanz, V. Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk", *Pre-proceeding of ETK-NTTS*, Luxemburg, 2001, pp. 807-826.
[15] J. Domingo-Ferrer, J.M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control", *IEEE Transactions on Knowledge and Data Engineering*, Jan. 2002, Vol 14, No. 1, pp. 189-201.
[16] M. Laszlo, S. Mukherjee, "Minimum spanning tree partitioning algorithm for microaggregation", *IEEE Transactions on Knowledge and Data Engineering*, July 2005, Vol 17, No. 7, pp. 902-911.
[17] A. Solanas, A. Martinez-Baslleste, J. Domingo-Ferrer, "V-MDAV: a multivariate microaggregation with variable group size", *Proceeding of COMPSTAT*. Springer, Rome, Italy, 2006.
[18] Chang Chin-chen, Li Yu-chiang, Huang Wen-huang, "TFRP: an efficient microaggregation algorithm for statistical disclosure control", *Journal of Systems and Software*, Nov. 2007, Vol 80, No. 11, pp. 1866-1878,.
[19] J. Domingo-Ferrer, F. Seb, A. Solanas, "A polynomial-time approximation to optimal multivariate microaggregation", *Computer and Mathematics with Applications*, Feb. 2008, Vol. 55, No. 4, pp. 714-732.
[20] J. Domingo-Ferrer, "Microaggregation for Database and Location Privacy", *NGITS*, LNCS 4032, 2006, pp. 106-116.