I J C T A, 9(28) 2016, pp. 225-234

MNQIA: A METHOD FOR MULTIPLE NUMERICAL QI ATTRIBUTE ANONYMIZATION

Abrar Ahmed K* Abdul Rauf H and Rajesh A.*****

Abstract: Modern improvements in Information Technology have increased the demand for collecting and sharing of data. However data may contain sensitive information about individuals. Usage of this data causes unexpected disclosure of private information. A common approach for privacy preserving data publishing (PPDP) is Anonymization, which perform generalization or suppression on single Quasi-Identifier (QI) attributes at once. However many real world applications data can contain multiple numerical QI attributes. In this paper we propose a privacy preserving data publishing methods called MNQIA, which uses the ideas of clustering, Bucketization and multiple multi-dimensional capacity first (MMDCF) to publish Anonymized micro data with multiple QI attributes at once. We take a model to represent the strength of our methods in terms of privacy protection and utility of data.

Keywords: Privacy Preserving Data Publishing (PPDP), Data anonymization, Quasi-identifier (QI) attribute, clustering, Bucketization.

1. INTRODUCTION

Micro data plays an increasingly important role in data analysis and scientific research.However, publishing and sharing of micro data will threaten to individual's privacy. **Privacy Preserving Data Publishing (PPDP)** is a research area that tries to speculate data before publishing in order to safeguard the sensitive information, while complete data is maintained on other hand for research purpose.

Usually Micro data consists of several attributes which may be Identifier (I), Quasi Identifier (QI), Sensitive attributes(S). Identifier attributes which carefully identify the records of owners and are consistently removed from the released data. **Quasi Identifier** attributes which could be linked with external information to re-identify individual's records of data owners. However sensitive attributes are protected. It is important to **anonymized** QI attributes, so that the individual's records can't be re-identified, by leveraging a trade-off in PPDP between data privacy and data utility.

Several approaches have been proposed to implement anonymity models. In 2002 Sweeney[1] proposed k-Anonymity model with the help of generalization technique whose idea is to replace real value of Quasi-Identifier with less specific but semantically consistent value. These methods are inefficient to protect attribute disclosure and anonymized only one QI attribute at once.

As far as data mining prospect is pertained, **clustering** is a valuable technique that partitions records into clusters such that records within a cluster are identical to each other, while records in different clusters are distinct from one to another. There are a few papers [2-6] that have used this technology to achieve *k*-Anonymization.

* Research Scholar,CSE,Manonmaniam Sundarnar University,Tirunelveli, kaa406@yahoo.co.in.
** Principal, Pinnacle School of Engineering and Technology,Kollam,harauf@yahoo.com.
*** Principal,C Abdul Hakeem College of Engg. And Tech., Melvisharam, amrajesh73@gmail.com

In 2006, Xiao and Tao[7] proposed Anatomy, which is a data anonymization approach that divides one table into two for release. One table includes the original quasi-identifier and a group id, and the other includes the association between the group id and the sensitive attribute values.

In this paper, we propose a method called MNQIA.The main idea of this method is to cluster the QI attribute records based on approximate degree and structure multi dimensional bucket and then apply multiple multi-dimensional capacity first (MMDCF) method to form a group tuples. Then from each group based on min and max values, generalization can be applied to achieve anonymized micro data.

2. RELATED WORK

In Paper (P. Kieseberg, S. Schrittwieser, M. Mulazzani, I. Echizen,and .Weippl,2014), Authors propose a technique which is based on the function of k-anonymity, whose objective is solving concerns such as one single step-anonymizations and finger printing of micro data. In addition to that, Authors builds conditions to find colluding attackers as well as anonymization approach that protect effects of colluding attackers on reducing anonymization level. Based on this outcomes they suggest an algorithm for creation collusion-resistant finger prints for micro data.

In paper (Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer,2007),The Authors propose a new and the powerful privacy description called l-diversity by reviewing with two attcks namely, Homogeneity attack and Background Knowledge attack that k-anonymized data usually suffer. The aim of l-diversity is to distribute sensitive attribute in each equivalence classes that has at least ‘l’ well represented values. Authors proves in an experimental assessment that l-diversities implemented practically.

In Paper (Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, 2008), The Authors proposes a new privacy notation called t-closeness by reviewing the drawbacks of l-diversity which is insufficient to avoid attribute disclosure. The objective of t-closeness is to distribute sensitive attributes in any equivalence classes is too close to t-value of the attributes.

In Paper (David J. Martin, Daniel Kifer, Ashwin Machanavajjhala, Johannes ehrke,2008), Authors begins with a study of worst case background knowledge. Based on the study, they propose a language that states any background knowledge about data. They propose the algorithm called Polymomial time algorithm to find the amount of disclosure of sensitive information in the worst case, given the attacker has atmost ‘k’ pieces of information in this language.

In paper (Xiaokui Xiao, Yufei Tao), Authors proposes a new technique called anatomy, whose objective is to release two separate tables, one for QI, other for sensitive attributes. They builds linear-time algorithm for computing anonymized table that satisfy l-diversity and reduce the errors in re-constructing micro data.

In Paper (Qing Zhang, Nick Koudas, Divesh Srivastava, Ting Yu),Author’s proposed a permutation based approach to anonymization significant benefit of this permutation based approach is that it will provide more accurate answer to aggregate queries. Author’s,further propose several criteria to optimize permutations for accurate answering of aggregate queries, and develop efficient algorithms for each criterion.

3. PROPOSED METHOD:

In this section, we will explain the proposed method for anonymization of multiple quasi identifiers at once.

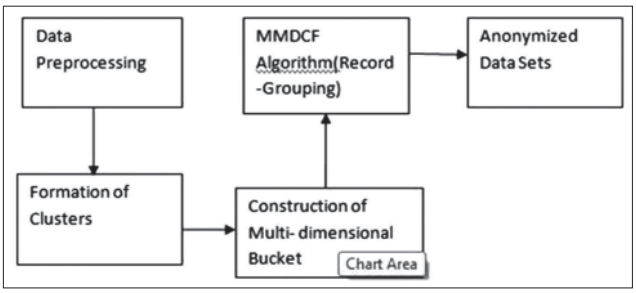

Figure 1. MNQIA Anonymization process

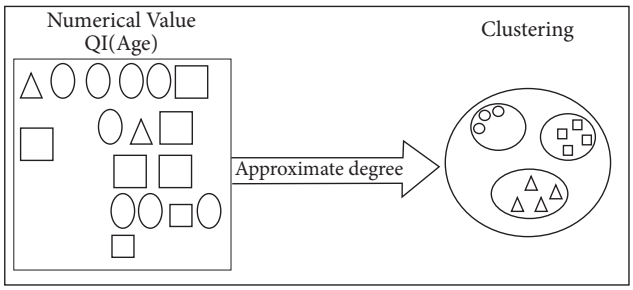

Figure 2. Clustering Process

Figure1 shows the MNQIA Anonymization process, we assume that the original dataset D has n number of records and each record has m numerical quasi-identifier (QI) attributes and x number of sensitive attribute. We mark the numerical QI attributes as Q-att$_1$,Q-att$_{2...}$ Q-att$_m$.

For each Q-att$_i$, 1≤ i≤ m we put its value into multiple group(Clusters) based on the appropriate degree, which are marked as $\{A_{i1,} A_{i2,}....A_{ij}\}$ and union of the group can cover all the values of S_i,$1 \le i \le m$ as shown in above figure1 .The intersection of any two groups is the empty set simultaneously. For instance, consider A_1 is age. This implies that there are n numerical cluster in A_1 on which we can use the numerical cluster methods to put the n numerical values into multiple groups, where the size of each group can be different.

Then we design the multi-dimensional bucket. Every QI attribute corresponds to a single dimensional of the multidimensional bucket. If D has m-numerical sensitive attributes, we structure an multidimensional bucket. Then n record of D is mapped into the corresponding according to their own Q-att attributes.

Once the multi dimensional bucket is structured, we select different record to form the corresponding Q-att group. For achieving this we use maximal multi dimensional capacity first (MMDCF). The basic idea of MMDCF is to choose different record to makeup the matching QI-group.

The section priority of maximal multi dimensional capacity first (MMDCF) is defined as Selection

$$(buk < s_0^1, s_0^2, ... s_0^d >) = \sum_{1<j<d} \text{capa}(s0j) + size(buk < s01, s02, \ldots s0d>)$$

Where $\sum_{1<j<d}$ capa($s0j$) is the maximal sum of number of tuples in each dimension bucket, size (buk < s_0^1, s_0^2,.... s_0^d >) is the size of bucket

Algorithm: Significant steps in MNQIA.

Input: DataSet T(Q-att_1,Q-att_2,….Q-att_z,SE$_1$,SE$_2$,…SE$_d$) , parameter t

Output: Anonymous table T `(Q-$att_{1,}$ Q-$att_{2,....}$ Q-$att_{z,,}$ SE$_1$,SE$_2$,…SE$_d$)

step1: get Quasi-identifier attribute Q-att_1,Q-att_2,….Q-att_z and all values;

step 2: for Each Q-att_i ($1 \le i \le z$);

step 3: cluster values into approximate groups, i.e. Q-att_{ij}.($1 \le i \le z$, $1 \le i \le y$)

step 4: end for;

step 5: For each Q-att_{ij} .($1 \le i \le z$, $1 \le i \le y$)

step 6: Mapped the records into their corresponding one dimensional bucket according to their own Quasi-Identifiers attributes and form m-Dimensional bucket.

step 7: calculate the capacity of approximate groups for each Q-att_i ($1 \le i \le m$);

step 8: while(can extract records constitute a group)

step 9: set unshielded marker for all buckets, Grouping $G_i = \Phi$, i←1;

step 10: Calculation selection of non empty bucket;

step 11: for (j=1; j≤1; j++)

step 12: if(there is non empty and unshielded bucket)

step 13: select a record t from the maximum selection bucket buk and add it into group Gi ;

step 14: delete t from buk, and size(buk)=size(buk)-1;

step 15: recalculate the capacity of buk for each dimension;

step 16: Shielding the bucket which has the same approximate group with t ;

step 17: }

step 18: else

step 19: if(there is no record can choose)

step 20: the end of the group process;

step 21: }

step 22: end for;

step 23: *i* ++;

step 24: end while

step 25: From each group G_i, take minimum-maximum value from Q-att_i and generalize the att .

step 26: return an anonymization table T`.

3.1 Information Loss Metric:

We utilize generalization on the values of *QIs* in order to modify data and form clusters. This anonymization process cause information loss because some original values of *QIs* in every sequence are either replaced with less specific values or are totally removed. In order to preserve data utility for data mining tasks, we should verify that anonymization cost is minimized. We consider the scenario where the data analysis task is unknown at the time of data publication. So, our goal is to anonymize sequence data to satisfy privacy while preserving data utility as much as possible. According to [16], Let *D** be an anonymization of sequence data *D*. *D** corresponds to a set of clusters $C = \{C1, C2,..., Cp\}$ which is a clustering of sequences in *D*. All sequences in a given cluster *Cj* are anonymized together.

We define the amount of information loss incurred by anonymizing *D* to *D** as

$$IL(D,D^*) = \frac{1}{|D|}\sum_{j=1}^{p} IL(C_j) \quad (1)$$

where $IL(C_j)$ is the information loss of the cluster C_j, which is defined as the sum of information loss of anonymizing every sequence *S* in C_j:

$$IL(C) = \sum_{i=1}^{|C|} IL(S_i, S_i^*) \quad (2)$$

where $|C|$ is the number of sequences in the cluster *C*, and IL() is the information loss of anonymizing the sequence *S* to the sequence *S** [17]. Each sequence is anonymized by generalizing or suppressing some of the *QIs'* values in some of its events. Let *H* be generalization hierarchy of the attribute *A*. We use the

Loss Metric (*LM*) measure [18] to capture the amount of information loss incurred by generalizing the value *a* of the attribute *A* to one of its ancestors, with respect to *H:[16]*

$$\mathrm{IL}(a,\hat{a}) = \frac{|\mathrm{L}(\hat{a})| - |\mathrm{L}(a)|}{|\Delta \mathrm{A}|} \tag{3}$$

where $|\mathrm{L}(x)|$ is the number of leaves in the subtree rooted at x.

The information loss of each event e is then defined as [16]

$$\mathrm{IL}(e,e^*) = \sum_{n=1}^{|\mathrm{QI}|} \mathrm{IL}(e(n),e^*(n)) \tag{4}$$

Where is the ancestor of the event e, $e(n)$ is the value of n^{th} *QI* of the event e and (n) is its corresponding value in the event . Hence, the information loss incurred by anonymizing each sequence is as follows:[16]

$$\mathrm{IL}(\mathrm{S},\mathrm{S}^*) = \sum_{m=1}^{|s|} \mathrm{IL}(em,em_m^*) \tag{5}$$

3.2 Example

In this section we explain our methods via real situation. Consider the following micro data.

Table 1.
Micro data

Id	*Age*	*Zip*	*Salary*	*Bonus*
T1	27	12,000	1000	1010
T2	22	22,000	2975	1010
T3	34	24,000	10,100	950
T4	26	17,000	1040	2000
T5	30	16,000	3050	2020
T6	32	14,000	5000	3035
T7	22	19,000	5120	2950
T8	37	26,000	7950	4100
T9	39	27,000	1050	6000

There are two quasi-Identifier Age and zip code and two sensitive attribute such as salary and bonus. We put the age cluster into four group: A11 = {26, 27}, A12 = {22, 22}, A13 = {30, 32, 34} and A14 = {37, 39}.Ultimately we put the zipcode into four cluster groups: A21 = {12,000, 14,000}, A22 = {17,000, 16,000, 19,000}, A23 = {22,000, 24,000} and A24 = {26,000, 27,000}, it is shown in table 2.

Table 2.
Two cluster Group

Age-Group(A1i)	*Zip-code Group(A2i)*
A11 = {22, 22}	A21 = {12,000, 14,000}
A12 = {26}	A22 = {17,000, 16,000, 19,000}
A13 = {32, 33, 34, 35}	A23 = {22,000, 24,000}
A14 = {37, 39}	A24 = {26,000, 27,000}

We make age and zip code to be the first dimension and second dimension respectively. Now check the tuple t1 values of age and zip code with two cluster group.

Table 3. Two Dimensional Cluster

	A11	*A12*	*A13*	*A14*
A21			{T1,T6}	
A22	T7	T4	T5	
	T2		T3	
A23				
				{T8,T9}
A24				

Then tuple *t*1 belongs to group A13, A21.Therefore we put *t*1 in the corresponding cell.Similarly, we place all the other records as well. We structure a two dimensional bucket as in above table.

According to MMDCF [15], we can choose different record to make up the matching QI-Group. For example Age and Zip code are choosen as QI-Attributes. Now according to the selection priority equation is as follows.

Group 1:

Iteration-1

Selection (buk <A21, A13> = 8 tuples → {T1, T6} to break the tie tuple T1 is selected

There are 4 tuples in A13, 2 tuples in A21 and 2 tuples in buk<A21,A13>, Totally 8 tuples. The Priority in buk<A22,A12> is 6 tuples, and in buk<A22,A13> is 7 tuples which is rejected because tuple from A13 already selected in Iteration 1. To break tie between buk<A22,A11> and buk<A22,A12>, we select buk<A22,A11> whose tuples is T4.Therefore the highest priority is buk<A22,A11> so tuples **T4** is selected, then we shield dimension <A11>.

Iteration-2

Slection(buk< A22,A11> = 6 tuples → T7 Selected

Selection(buk< A22,A12> = 5 tuples → T4

Slection(buk< A22,A13> = 8 tuples → Already tuple selected from A13

There are 2 tuples in A11, 3 tuples in A22 and 1 tuples in buk<A22,A11>, Totally 6 tuples. The Priority in buk<A22,A12> is 5 tuples, and in buk<A22,A13> is 8 tuples which is rejected because tuple from A13 already selected in Iteration 1.The highest priority is buk<A22,A11> so tuples T4 is selected, then we shield dimension <A11>.

Iteration-3

Selection(buk< A23,A11>= 5 tuples → Already tuple selected from A11

Slection(buk< A23,A13> = 6 tuples → Already tuple selected from A13

Iteration-4

Selection (buk< A24,A14> = 6 tuples → {T8 ,T9} to break the tie highest tuple T9 is Selected

Finally in all four iteration we selected 4 tuples in a first group (T1, T7, T9}

Group 2:

The tuples which is selected in group one should be removed from Two Dimensional Cluster table and repeat this procedure.

Table 4.
Two Dimensional Cluster

	A11	*A12*	*A13*	*A14*
A21			T6	
		T4	T5	
A22	T2		T3	
A23				T9
A24				

The same procedure has to be followed to obtain second group which contains tuples {T2,T4,T6} and Third group which tuples contain {T3,T5,T8}.From these group we perform generalization on multiple Quasi Identifier attributes to get 3-Anonymization with 3-diversity on micro data as shown above.

Table 5.
Anonymized Table

ID	*Age*	*Zip*	*Salary*	*Bonus*
T1	22-39	12,000-27,000	1000	1010
T7	22-39	12,000-27,000	5120	2950
T9	22-39	12,000-27,000	1050	6000
T2	22-32	14,000-20,000	2975	1010
T4	22-32	14,000-20,000	1040	2000
T6	22-32	14,000-20,000	5000	3035
T3	34-37	24,000-26,000	10,100	950
T5	34-37	24,000-26,000	3050	2020
T8	34-37	24,000-26,000	7950	4100

4. EXPERIMENTS

This section evaluates the effectiveness of our approach using Adult database from the UCI machine learning repository website. The experiments are conducted on a computer with Intel core i3-processor and 4 GB memory running the Microsoft window 7 OS. The Algorithm and Information Loss metric was implemented in java (JDK) using Eclipse juno and Datafly algorithm is implemented using weak 3.7

Below Table and graphs shows that the information loss calculated using above equation for both techniques on various values of K-level. It is cleared that information loss increases highly in Datafly algorithm as the value of K-level increases, whereas in our approach information loss is low when compared to Datafly method and also as the value K-level increases, there is a little variation in information loss.

Datafly method can able to anonymized only single QI attributes at once,but our approach can anonymized multiple QI attributes.

Table 6.
Information loss for Datafly and MNQIA Methods

Anonymization Level (K-Value)	*Informaton Loss*	
	Datafly Techniques	*MNQIA Technique*
3	0.25	0.68
4	4.25	0.80
5	4.25	1.2
6	7.5	1.35
7	7.85	1.5
8	11.5	1.7
9	11.5	2.2

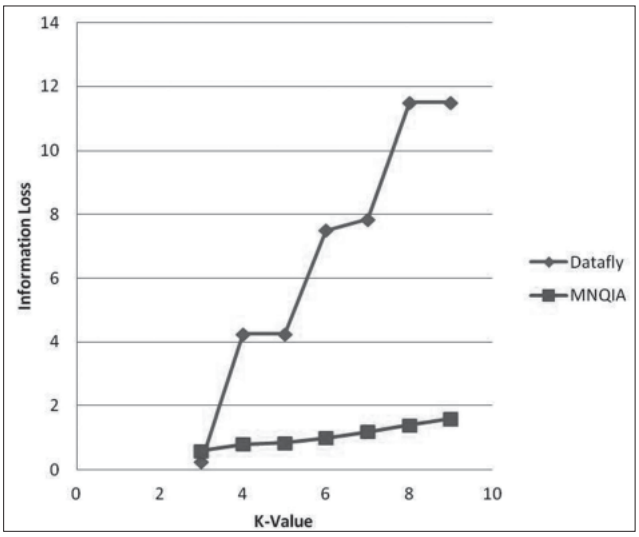

Figure 3. Representation of Information loss for Datafly and MNQIA Methods

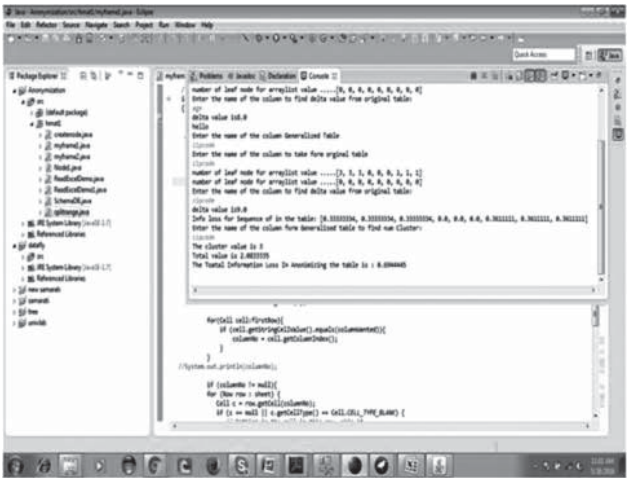

Figure 4. Information Loss of MNQIA when k = 3

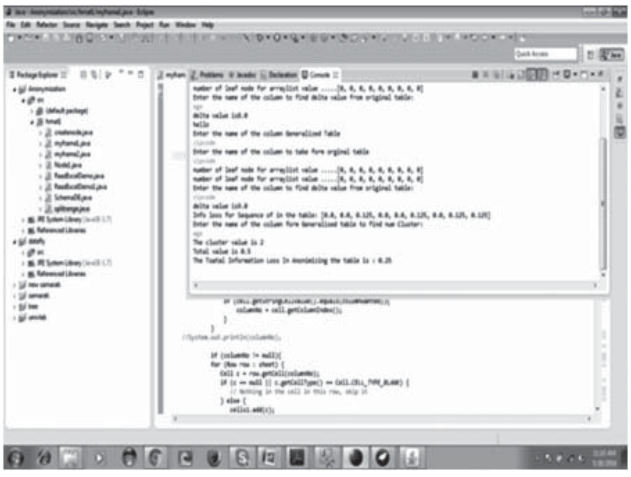

Figure:5 Information Loss of Datafly when k = 3

5. CONCLUSION:

In this paper, we propose an MNQIA method to anonymize multiple quasi-identifier (QI) attributes at once and analyze it in terms of utility. The experiment shows that the anonymization using MNQIA have low information loss compared to Datafly method.

However our approach achieves Anonymization on micro data, but lack to achieve l-diversity and t-closeness. Future research is to include methods in our approach that should achieve both anonymization along with l-diversity and t-closeness.

References

1. P. Kieseberg, S. Schrittwieser, M. Mulazzani, I. Echizen, and E. Weippl, An algorithm for collusion-resistant anonymization and fingerprinting of sensitive microdata, Electronic Markets, vol. 24, no. 2, pp. 113–124, 2014.

2. Y. Tang and C. Zhong, Probabilistic k-anonymity algorithm with multi-sensitive attributes based on Variable length clustering, (in Chinese), Computer Engineering and Design, no. 8, pp. 1–8, 2014.

3. Byun, A. Kamra, E. Bertino, and N. Li, Efficient k-anonymization using clustering techniques, in Proc. 12th International Conference on Database Systems for Advanced Applications (DASFAA 2007), Bangkok, Thailand, 2007, pp. 188–200.

4. C. Chiu and C. Tsai, A k-anonymity clustering method for effective data privacy preservation, in Proc. 3rd Intl. Con. on Advanced Data Mining and Applications, Harbin, China, 2007, pp. 89–99.

5. Z. He and G. Chen, Improvement of k-anonymity location privacy protection algorithm based on hierarchy clustering, Applied Mechanics and Materials, vols. 599-601, pp.1553–1557, 2014.

6. M. Verma, k-anonymity using two level clustering, Master degree dissertation, Dept. Computer Science and Engineering, National Institute of Technology Rourkela, India, 2013.

7. X. Xiao and Y. Tao, Anatomy: Simple and effective privacy preservation, in Proceedings of the 32nd International Conference on Very Large Data Bases, VLDB Endowment, Seoul, Korea, 2006, pp. 139–150.

8. P. Kieseberg, S. Schrittwieser, M. Mulazzani, I. Echizen,and E. Weippl, An algorithm for collusion-resistant anonymization and fingerprinting of sensitive microdata,Electronic Markets, vol. 24, no. 2, pp. 113–124, 2014

9. A. Machanavajjhala, D. Kifer, J. Gehrke, and M.Venkitasubramaniam, l-diversity: Privacy beyond kanonymity, ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 3, pp. 1–47, 2007.

10. N. Li, T. Li, and S. Venkatasubramanian, t -closeness: Privacy beyond k-anonymity and l-diversity, in Proc. IEEE 23rd International Conference on Data Engineering (ICDE 2007), Istanbul, Turkey, 2007, pp. 106–115.

11. D. J Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y Halpern, Worst-case background knowledge for privacy-preserving data publishing, in Proc. IEEE 23rd International Conference on Data Engineering (ICDE 2007), Istanbul, Turkey, 2007, pp. 126–135.

12. M. E. Nergiz, M. Atzori, and C. Clifton, Hiding the presence of individuals from shared databases, in Proceedings of the 2007 ACM International Conference on Management of Data (SIGMOD 2007), Beijing, China, 2007, pp. 665–671.

13. X. Xiao and Y. Tao, Anatomy: Simple and effective privacy preservation, in Proceedings of the 32nd International Conference on Very Large Data Bases, VLDB Endowment, Seoul, Korea, 2006, pp. 139–150.

14. Han,Luo, Lu and Peng, SLOMS:A privacy preserving data publishing methods for multiple attributes microdata, in Journal of Software Vol. 8, No. 12, PP: 3096-3104,2013.

15. Liu,Shen and Sang,Privacy Preserving Data Publishing for numerical sensitive attributes, in IEEE Transaction on Knowledge and Data Engineering ,Vol.20,No.3,PP:246-254,2015.

16. El Emam, K, Arbuckle, L., Koru,G., Gaudette, L., Neri, E., Rose, S., Howard, J., and Gluck, J., 2012. De-Identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Data Set. In Journal of Medical Internet Research, 14:1, DOI:10.2196/jmir.2001, 2012.

17. Kaufman L and Rousseeuw,P.J 1990.Finding grop in data:An Introduction to cluster Analysis.John Wiley 1990.

18. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito: Efficient full-domain k-anonymity,” in Proc. SIGMOD Conf., Baltimore, MD, USA, 2005, pp. 49–60.

19. K. Stokes and V. Torra, n-confusion: A generalization of kanonymity, in Proceedings of the 2012 Joint EDBT/ICDT Workshops (EDBT-ICDT’12), ACM, 2012, pp. 1–5.

20. R. Trujillo-Rasua and J. Domingo-Ferrer, On the privacy offered by (k,)-anonymity, Information Systems, vol. 38, no. 4, pp. 491–494, 2013.

21. P. Kieseberg, S. Schrittwieser, M. Mulazzani, I. Echizen, and E. Weippl, An algorithm for collusion-resistant anonymization and fingerprinting of sensitive microdata, Electronic Markets, vol. 24, no. 2, pp. 113–124, 2014

22. J. Liu and K.Wang, Enforcing vocabulary k-anonymity by semantic similarity based clustering, in Proc. 2010 IEEE 10th International Conference on Data Mining (ICDM), Sydney, Australia, 2010, pp. 899–904.

23. C. Wang, L. Liu, and L. Gao, Research on k-anonymity algorithm in privacy protection, Advanced Materials Research, vols. 756-759, pp. 3471–3475, 2013.