International Journal of Advances in Computer and Electronics Engineering
Volume: 1 Issue: 1, May 2016, pp. 6–12

# Reduction of Data Loss and Privacy Disclosure using t-closeness for Multiple Sensitive Attributes

[1]S.Saraswathi, [2]K.Thirukumar
[1]PG Scholar, Department of Computer Science and Engineering,
Dr. Mahalingam College of Engineering and Technology, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Science and Engineering,
Dr. Mahalingam College of Engineering and Technology, Tamil Nadu, India
[1]suganthi.cse13@gmail.com, [2]thirukumar@drmcet.ac.in

**Abstract:** *Many Organizations distribute the individual's information in order to exploit the data for the research purpose. But the private information about the individual is exposed by the opponent by combining the various releases of the several organizations. This is called as linkage attacks. This attack can be avoided by the SLOMS method which vertically partitions the quasi identifier and sensitive attributes. The SLOMS method uses MSB-KACA algorithm to generalize the quasi identifier table in order to implement k-Anonymity and bucketizes the sensitive attribute table to implement l-diversity. But there is a chance of probabilistic inference attack due to bucketization. So, the method called t-closeness can be applied above MSB-KACA algorithm which compute the value using Earth Mover Distance(EMD) and set the minimum value as threshold in order to equally distribute the attributes in the table based on the threshold 't'. Particle swarm optimization (PSO) has been incorporated with EMD computation in order to equally distribute the data over entire table. Thus the probabilistic inference attack can be avoided. The performance of t-closeness gets improved and evaluated by Disclosure rate which becomes minimal while comparing with MSB-KACA algorithm.*

**Keyword:** *Privacy; k-anonymity; l-diversity; MSB-KACA, t-closeness; Particle Swarm Optimization;*

## 1. INTRODUCTION

Privacy is a major phase in data publishing. Privacy-preserving data publishing (PPDP) is a task to extend methods and tools to publish data in a destructive environment, so that the published data becomes useful while individual's privacy is preserved. Nowadays, many organizations are increasingly publishing the micro data-tables that contain the unaggregated information about the individuals. However, such publication may lead to privacy disclosure. To address this challenge, privacy preserving data publishing was proposed to protect the individual's sensitive data in the published table.

In general, a micro data table can contain three types of attributes:1) Explicit identifier attributes, (e.g., name , phone number) which allow direct linking of an instance to a person 2) Quasi-identifier (QI)attributes, (e.g., age, sex, zipcode) which are not explicit identifiers but, when combined together, can be used to reveal individual's identity, and 3) Sensitive attributes (SA) (e.g., Salary, Disease) each of which contains a sensitive data that must be protected.

Privacy preserving data publishing or PPDP method remove the explicit identifiers like name, phone number and generalize or suppress the quasi identifier attributes like gender, age, zip code in order to protect the individual's information. The information disclosure has been classified as identity disclosure and attributes disclosure. Identity disclosure uniquely identifies whether a particular individual is linked to a specified records. Attribute disclosure identifies the information about the individuals .i.e., the published data helps to identify the individual's information more accurately.

This information disclosure can be avoided by a method called k-Anonymization [2]. But the k-Anonymity does not prevent the individual's information from the background knowledge attack. So the next level of privacy has been provided using the method called l-diversity [3] which contains l well represented distinct values within an equivalence class. Though privacy is improved in the l-diversity method, it suffers from similarity attack and skewness attack. So a method called t-closeness [8] has been introduced to prevent the individual's information distress from skewness and similarity attack that were possible on l-diversity by equally distributing the data over the entire table.

The slicing on multiple sensitive method partitions the original table into single quasi identifier table and m-sensitive table where sensitive attributes are clustered based on mean square contingency formula which clusters the highly correlated sensitive attribute into single table and applies the MSB-KACA algorithm to the partitioned table.

MSB KACA generalize the quasi identifier table in order to implement the k-anonymity and bucketizes the sensitive attribute table to implement the concept called l-diversity. By applying the l-diversity concept to the sensitive attributes, it suffers from similarity attack and skewness attack where the individual's information can be exposed based on identifying the prospect within the equivalence class which is called as probabilistic inference attack or skewness attack. Hence, the t-closeness is applied over MSB-KACA algorithm to ensure the privacy protection and utility factors in this paper.

## 2. LITERATURE SURVEY

### 2.1 Anonymization

A released data is said to be a k-anonymized [1] data if the information about each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appear in the release. Anonymization applies generalization and suppression concept to achieve k-anonymity where generalization is defined by replacing the original value of the attribute with less specific but semantically consistent values. It splits the ordered-value domains into intervals and suppression is a special kind of generalization. It replaces some attribute values with special symbol which indicates that the value is suppressed or else the value of the attribute is not released at all but the k-anonymization suffers from background knowledge attack.

### 2.2 L-diversity

L-diversity is one of the methods which can be applied to the original data in order to protect the individual's information from background knowledge attack. According to Ashwin Machanavajjhala, an equivalence class is said to have l-diversity [3] if there are at least l "well represented" values for the sensitive attribute. But the l-diversity does not consider about overall distribution of sensitive values which is vulnerable to probabilistic inference attack or skewness attack.

### 2.3 Anatomization

Anatomization [4] is a technique that releases the data on quasi-identifier and data on the sensitive attribute in two separate tables. Both the quasi identifier and sensitive attribute table contains one common attribute known as Group Identifier (Group ID). All records in one equivalence class will have the same value of Group ID in both the tables. It is suitable for dealing with the high dimensional data but due to the direct publication of the data an adversary can identify the individual's sensitive information.

### 2.4 Slicing

Slicing [6] is a technique that partitions the data horizontally and vertically which is suitable to handle high dimensional data. It provides better data utility when compared to generalization and prevents the individual's sensitive information from membership disclosure. It preserves privacy by breaking the associations between uncorrelated attributes and provides utility by grouping highly correlated attributes together.

### 2.5 Probabilistic inference attack

When the overall distribution is skewed, that satisfies l-diversity does not prevent attribute disclosure [8]. An equivalence class can contain equal number of positive and negative records and it satisfies 2-diversity, hence a privacy risk occurs. Consider an equivalence class that has 49 positive records and only one negative record so the adversary can easily infer the sensitive attribute of the particular individual based on identifying the probability.

### 2.6 Discretization

Discretization [11] is the process of splitting the continuous attributes into intervals and reduces the number of values for the continuous attributes. It collects and replaces the lower level concept by higher level concept in order to reduce the data.

### 2.7 T-closeness

In 2011 Ninghui Li et al. Tiancheng Li and Venkatasubramanian, S [8] [10] stated the method called t-closeness in order to provide privacy for the published datasets. It requires that the earth mover's distance between the distribution of a sensitive attribute in any equivalence class does not differ from the overall distribution of the sensitive attribute with the threshold t. i.e., the distance between the two distributions should no more than a threshold t). The t-closeness method uses Earth Mover's Distance EMD [8], which is based on the minimal amount of work which has to be done to transform one distribution to another by moving distribution mass between each other. Table I represents original dataset and Table II represents t-closed dataset.

TABLE I ORIGINAL DATA SET

| ID | Name | Weight | Age | Disease |
|---|---|---|---|---|
| 1 | Mike | 60 | 40 | SARS |
| 2 | Alice | 70 | 50 | Intestinal Cancer |
| 3 | John | 60 | 60 | Pneumonia |
| 4 | Bob | 50 | 50 | Bronchitis |
| 5 | Beth | 80 | 50 | Gastric flu |
| 6 | Carol | 70 | 70 | Gastric ulcer |

TABLE II T-CLOSED DATA SET

| EC | Weight | Age | Disease |
|---|---|---|---|
| 1 | [50-60] | [40-60] | SARS |
|  | [50-60] | [40-60] | Pneumonia |
|  | [50-60] | [40-60] | Bronchitis |
| 2 | [70-80] | [50-70] | Intestinal cancer |
|  | [70-80] | [50-70] | Gastric flu |
|  | [70-80] | [50-70] | Gastric ulcer |

### 2.8 Particle Swarm Optimization (PSO)

Particle swarm optimization [14] is an optimization technique which optimizes the problem that has been modeled on evolutionary algorithm (EA). PSO optimizes an objective function using the population based search. PSO has a memory → not "what" that best solution was, but "where" that best solution was. Hence PSO optimizes the data based on objective function and equally distribute the data over original dataset.

### 2.9 Privacy and Accuracy constraints

Privacy and accuracy constraints [12] are based on protecting the individual's information without reducing the utility. It assigns the class label to each record and computes the information loss based on the adherence of a tuple to the majority class of its group. In case of categorical attributes NCP (Normalized Certainty Penalty)is defined with respect to the taxonomy tree of the attribute. For the set of all equivalence classes in the released anonymized table, a normalized formulation of the aggregate version of NCP, called the Global Certainty Penalty (GCP) is adopted.

## 3. SLOMS METHOD

### 3.1 Sensitive Attribute Partition

SLOMS first partitions the sensitive attributes into m parts based on the principle that highly correlated sensitive attributes are grouped into single table based on the mean square contingency coefficient. If there are some continuous sensitive attributes, that are considered as categorical attributes after being discretized.

The algorithmic strategy transforms the dataset to achieve SLicing On Multiple Sensitive i.e., SLOMS by implementing MSB-KACA algorithm [16].The following figure illustrates the methodology of the MSB-KACA algorithm. Figure 1 represents the work flow of the MSB-KACA algorithm. The following pseudo code provides steps involved in the MSB KACA algorithm.

*Input:* Pre-processed dataset T{a1, a2, ……, az, s1, s2, ……, sd}, parameter l and k,sensitive attributes classification table Y{y1,y2,……yd}
*Output:* Single Quasi Table QIT and m-Sensitive Table {ST1, ST2……STm}

**Description:**

[1] Vertically partitions dataset *T* into a quasi identifier table and *m* sensitive attribute tables.
[2] m sensitive attribute table are clustered based on finding the correlation using mean square contingency formula.
[3] Use MSB method to implement *l*-diversity for each sensitive table.
[4] KACA algorithm to implement *k*-anonymity for quasi-identifier table.
[5] Link Quasi identifier attributes and sensitive attributes for the data utility.

The Multi Sensitive Bucketization K-Anonymity Clustering Attribute Hierarchy (MSB-KACA) algorithm [16] has been applied to the sliced data where the MSB is applied to the sensitive attribute in order to implement *l*-diversity and KACA is applied to the quasi identifier in order to implement *k*-anonymity. The basic idea of MSB method [16] is: (1) assign each tuple to a bucket based on each sensitive attribute values of the tuple, each bucket has the same values on all sensitive attributes and the priority is assigned to each bucket according to the MMDCF (Maximal Multi Dimensional

Capacity First) method (2) tuple is chosen randomly in the highest priority bucket (3) based on satisfying l-diversity the tuples are grouped into the bucket.

The basic idea of KACA method which has been applied to the quasi identifier are to group the attributes based on the zip code and then generalize or suppress the value in order to provide privacy of an individual's sensitive information. Finally the quasi identifier and sensitive attribute table are linked by using the group ID. The resultant table suffers from the skewness or probabilistic inference attack.

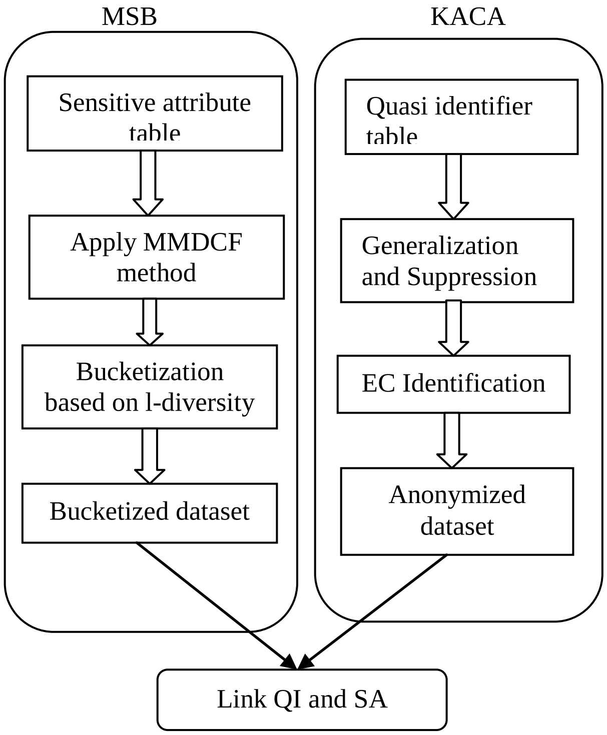

*Figure 1 MSB KACA algorithm*

## 3.2 Data set

The U.S. census data [17] is a collection of the real data set that is based on the U.S. Government census database. It is released on 01 may 1996 and contains 32570 instances which is of the size 3,913 KB. The U.S census database is a multivariate characteristic with the categorical and integer values.

The U.S census database contains 15 attributes such as age, work class, final weight, marital status, education, relationship, occupation, salary, capital gain, capital loss, education number, race, sex, native country, hours per week. These attributes can be divided into two categories they are: quasi identifier attributes and Sensitive attributes. Age, Gender, Zip code that uniquely identifies a person is considered as quasi identifier attributes. Occupation, Salary, work class, education, relationship which should remain confidential and should not reveal by an adversary is considered as sensitive attributes.

# 4. T-CLOSENESS OVER MSB KACA

## 4.1 T-Closeness over MSB KACA

Even though the published data is privacy protected by applying the MSB-KACA algorithm over the preprocessed data. But the privacy is restricted to the limited boundary. So the information can be revealed by the adversary. By applying MSB KACA algorithm, it satisfies the concept of l-diversity and k-Anonymity. Hence by satisfying the concept of l-diversity, the probabilistic inference attack occurs.

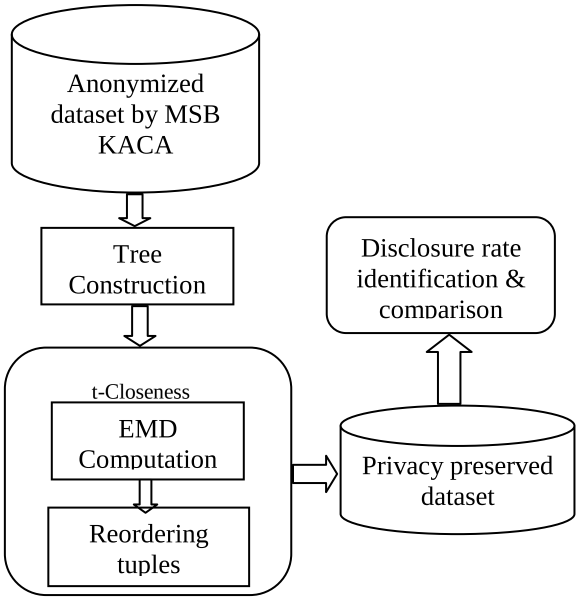
Anonymized dataset by MSB KACA
Tree Construction
t-Closeness
EMD Computation
Reordering tuples
Privacy preserved dataset
Disclosure rate identification & comparison

*Figure 2 T-closeness over MSB-KACA*

To solve the above problem, in this paper, the t-closeness is applied over the MSB-KACA algorithm. EMD, Earth Mover distance is calculated for the data set in which the privacy is preserved using MSB KACA algorithm. The t-closeness which contains the parameter t, for the sensitive attribute reorders the total dataset to ensure that the sensitive values are equally distributed within the equivalence class. Hence, the published data becomes unaffected by the similarity and probabilistic inference attacks. And the privacy to the published data is expected to increase as the reordering of tuples occurs. In case of a categorical SA, assumption is a generalization hierarchy H over its domain. For example, Figure 3 depicts a hierarchy of respiratory and digestive diseases. The distance between two (leaf) values $v_i$ and $v_j$ is defined as $h(v_i, v_j)/h(H)$, where h(H) is the height of H, and $h(v_i, v_j)$ that of the lowest common ancestor of $v_i$ and $v_j$ in H.

To define EMD, the following recursive function of the collective extra earth residing among the leaves under node n is defined

.

$$\text{Extra}(n)= \begin{cases} q_i - p_i, & \text{if n is a leaf } v_i \\ \sum_{c \sum child(n)}^{m} extra(c) \end{cases} \quad (1)$$

As mentioned in equation (1), The value of extra(n) [8] denotes the exact amount of earth that should be moved in/out of node n. Furthermore, accumulated amount of earth to be moved inwards and outwards for an internal node of H:

$$neg_e(n) = \sum_{c \in child(n) \wedge extra(c)<0} |extra(c)| \quad (2)$$

$$pos_e(n) = \sum_{c \in child(n) \wedge extra(c)>0} |extra(c)| \quad (3)$$

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

$$cost(n) = \frac{h(n)}{h(H)} \min(pos_e(n), neg_e(n)) \quad (4)$$

Then, the total EMD between P and Q is:

$$EMD= \sum_1^n cost(n) \quad (5)$$

where n is a non-leaf node in H.

Assume, SA distributions P=(1/6, 1/6, 1/6, 1/6, 1/6, 1/6) and Q=(1/3,1/3,1/3,0,0,0). Then extra (SARS) = extra(pneumonia) = extra(bronchitis) = 1/6. Thus extra(R) = 1/2, $pos_e(R)$ = 1/2 , and $neg_e(RD)$ = 0, hence cost(R) = 0. Likewise, extra(D) = -1/2, and cost(D) = 0. In effect, extra (RD) = 0, and $pos_e(RD)$ = $neg_e(RD)$ = 1/2 . Thus, cost(RD) = 1×min($pos_e(RD)$, $neg_e(RD)$) =1/2 , and EMD(P,Q) = cost(R) + cost(D) + cost(RD) = 0.5.Thus by calculating the Earth Mover Distance for categorical sensitive attribute and setting the threshold value, tuples can be reordered over the entire dataset to avoid the probabilistic inference attack.

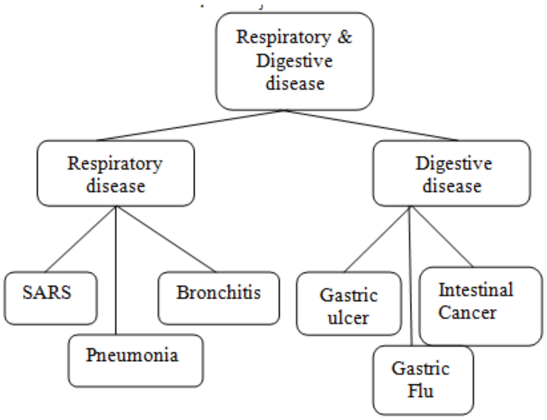

*Figure 3 Generalization Hierarchy*

## 4.2 Particle Swarm Optimization (PSO)

Particle swarm optimization [14] is an optimization technique which optimizes the problem that has been modeled on evolutionary algorithm (EA). PSO optimizes an objective function using the population based search. The population includes the possible solutions which are called as particles. These particles are randomly initialized across the multidimensional search space. The main idea of the particle swarm optimization is initialized with a population of random solutions and searches for optima by updating generations.

PSO primarily consist of two operators: 1.Velocity update 2.Position update. Each particle during its flight, updates its velocity and position based on experience of its own and the whole population. The velocity of the particle [15] is subjective to three components they are 1.inertial momentum 2.Cognitive and 3.social. The inertial parts simulate the inertial performance of the bird to fly in the previous direction. The cognitive part represents the memory of the bird about its previous best position, and the social component represents the memory of the bird about the best position among the particles.

Velocity of the ith particle ($P_i$) is given by

$$V_i = \omega V_{i-1} + c_1 r_1 (p_{best} - p_i) + c_2 r_2 (g_{best} - p_i) \quad (6)$$

Where, ω, $c_1$and $c_2$ → Constants; $r_1$ and $r_2$ → Random numbers,Best values for constants → ω = 0-1 (0.3 to 0.9); $c_1$and $c_2$= 2.$p_{best}$ → local best; $g_{best}$ → global best.
Position of the ith particle ($P_i$) is given by

$$P_i = P_{i-1} + V_i \quad (7)$$

Where, $P_{i-1}$ = Previous Position and $V_i$ = Particle's Velocity.

PSO by updating generations search for the optimal solutions which consist of the two best values pbest and gbest value that is updated for each iteration.

- Each particle keeps track of its coordinates in the solution space which are associated with the best solution (fitness) that has achieved so far by that particle. This value is called personal best , ***pbest***. pbest are used to update the velocity of any one of the particle and also it overcomes the problem of premature convergence for complex problems.

- Another best value that is tracked by the PSO is the best value obtained so far by any particle in the neighborhood of that particle. This value is called ***gbest***.

Steps involved in Particle Swarm Optimization:

[1] Initialize particles with random position and velocity vectors.
[2] For each particle's position (p) evaluate fitness
[3] If fitness(p) better than fitness(pbest) then pbest= p
[4] Set best of pBests as gBest
[5] Update particles velocity and position
[6] Stop: giving **gBest**, optimal solution.

## 4.3 Privacy and Utility Measure

Privacy and utility can be calculated for anonymized dataset. This system improves the privacy and utility depending on the number of records generalized or suppressed.

### 4.3.1 Information loss:

Information loss plays an important role on merging the equivalence class or on adding the number of tuple in one equivalence class with another equivalence class. Data utility can be evaluated by comparing the information loss with the original table and the anonymized table. Additional information loss [16] of entire table is as follows:

$$\text{AddInfoloss} = \frac{1}{m}\sum_{j=1}^{m}\ \sum_{i=1}^{b_j}\frac{|G_i|-l}{b\times l} \qquad (8)$$

Where, $G_i$ be a group in the *l*-diversity table, b be the number of groups in the *l*-diversity table and m be the number of sensitive attributes. Every tuple in the original table should be assigned to one group in anonymity table. Hence for the restriction of *l*-diversity some tuples cannot be assigned to any group. Such tuples must be suppressed. Suppression ratio[16] can be calculated as follows:

$$\text{Suppration=}\frac{n_s}{n} \qquad (9)$$

Where, $n_s$ be the number of suppressed tuples. Hence if the information loss is low more utility is guaranteed.

## 5. RESULT

The system explains about applying t-closeness over MSB KACA algorithm which includes EMD computation for the multiple sensitive categorical attributes. Hierarchical tree is constructed for the categorical sensitive attributes and Earth Mover Distance which has been used to find the distance between the sensitive values within the equivalence class and the entire table is computed for the sensitive table and finally obtaining the partitioned dataset which satisfies t-closeness principle.

The Data are distributed over the original table using Particle Swarm Optimization which optimizes and distribute the data based on objective function that has been declared. Hence it protects the data from the probabilistic inference attack. The privacy and utility is measured and compared with existing MSB KACA algorithm where the privacy measure is expected to be minimal and the utility measure is expected to be maximal for the proposed system. Hence the disclosure rate is low, so more privacy is guaranteed.

## 6. CONCLUSION

The evolution of Anonymization methods helps out to preserve privacy and satisfy utility criteria also. The MSB KACA method adopts the t-closeness to provide privacy to the individual's information without reducing the utility. Initially, the U.S. census dataset with a categorical sensitive attribute is subjected to the MSB KACA method which produces the anonymized dataset. But the improvement in the privacy level reduces the utility. Later, the t-closeness enhances the privacy, by reordering the tuples using Particle Swarm Optimization which means equally distributing the records over entire data. Hence, the privacy and utility are balanced for the published dataset.

## REFERENCES

[1] Pierangela Samarati, Latanya Sweeney (1998) 'Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression', Proceedings of the IEEE Symposium on Research in Security and Privacy, Technical Report, SRI International Computer Science Library-98-04.

[2] Latanya Sweeney (2002) ‘k-Anonymity: a model for protecting privacy’,International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp. 557-570.

[3] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkitasubramaniam (2007) ‘l-Diversity: privacy beyond k-anonymity’, ACM Transaction Knowledge Discovery, Voume. 1, No. 1, Article 3.

[4] Xiao X, Tao Y.(2006)’Anatomy: Simple and effective privacy preservation’,In: Proc. Of the 32nd International Conference on Very Large Data Bases. Seoul: VLDB Endowment,pp. 139−150.

[5] Yufei Tao, Hekang Chen, Xiaokui Xiao, Shuigeng Zhou, Member, IEEE Computer Society, and Donghui Zhang (2009),’ ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication’. *IEEE Transaction on Knowledge and Data Engineering,* vol. 21.No.7. pp.1073-1087.

[6] Li, Tiancheng (2012) ‘Slicing: A new approach for privacy preserving data publishing’, Knowledge and Data Engineering, IEEE Transactions on 24.3, pp. 561-574.

[7] Li Jiuyong, Wong Raymond Chi-Wing, Fu Ada Wai-Chee,et al.(2006),’ Achieving *k*-anonymity by clustering in attribute hierarchicalstructure[C]’. *DaWak. LNCS 4081, Springerverlag, Berlin, Heidelberg*, pp. 405-416.

[8] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian (2007) ‘t-closeness: privacy beyond k-anonymity and l-diversity’, in: Proceedings of the 23rd IEEE International Conference on Data Engineering, Istanbul, Turkey, pp.106–115.

[9] Benjamin C. M. Fung, Ke Wang, Rui Chen, Philip S. Yu(2010) ‘Privacy-preserving data publishing: A survey of recent developments’, Journal, ACM Computing Surveys, Volume. 42, No. 4, Article 14.

[10] Jianneng Cao, Panagiotis Karras, Panos Kalnis, Kian-Lee Tan (2011) ‘SABRE: a Sensitive Attribute Bucketization and Redistribution framework for t-closeness ’, The VeryLargeDataBase Journal, Volume 20, Issue 1, pp. 59-81.

[11] Jiawei Han and Micheline Kamber (2006). Data Mining: Concepts and Techniques.Department of Computer Science University of Illinois.

[12] Gabriel Ghinita, Panagiotis Karras,Panos Kalnis, Nikos Mamoulis (2009) ‘A framework for efficient data anonymization under privacy and accuracy constraints’, Journal, ACM Transactions on Database Systems (TODS), Volume. 34, No. 2, Article 9.

[13] Van den Bergh F. and Engelbrecht A.P.(2004), ‘A Cooperative Approach to Particle Swarm Optimization’, IEEE Transactions on Evolutionary Computation, pp. 225-239.

[14] Yamille del Valle Et. Al.(2008) “Particle Swarm basic Concepts, Variants and Applications in Power Systems”, IEEE Transactions On Evolutionary Computation, VOL. 12, NO. 2.

[15] Fangwei Luo, Jianfeng Lu and Hao Peng (2013), ‘SLOMS: A Privacy Preserving Data Publishing Method for Multiple Sensitive Attributes Microdata’, JOURNAL OF SOFTWARE, volume. 8, NO. 12,pp. 3096-3104.

[16] ‘https://usa.ipums.org/usa-action/variables/group’- U.S Census dataset for data collection accessed on 20/11/2014.