AN NEW ATTRACTIVE MAGE TECHNIQUE USING L-DIVERSITY

Anandhi G[1] and Dr.K.Saravanan[2]

[1]Assistant Professor, Department of Computer Science and Engineering, GSSS Institute Of Engineering & Technology For Women, Mysore

2Assistant Professor, Department of CSE, Erode Sengunthar Engineering College, Thudupathi, India

Abstract

Data that is published or shared between organizations contain private information about an individual. The concept of Privacy Preservation aims to preserve this sensitive information from various privacy threats that violate the privacy of an individual. Analysis of this private information could reveal information that can be used for malicious purposes by the attackers. Anonymization is a privacy preservation approach suitable for mixed data that contains both numerical and categorical attributes. In this paper a novel method called Micro-aggregation Generalization (MAGE) is used for anonymization of microdata that can retain more semantics of the original data. Here the Micro-aggregation is applied over the numerical data and Generalization is applied over the categorical data. Even though the MAGE approach preserves privacy it fails to address the homogeneity and background knowledge attacks. Later the l-diversity approach is applied to deal with homogeneity attack. In l-diversity, the anonymized records are reordered to satisfy a new privacy principle that removes homogeneity of sensitive information. The result shows that the MAGE approach suffers from homogeneity attack and applying l-diversity over MAGE prevents homogeneity attack and also provides better privacy and data utility.

Index Terms

Privacy, k-anonymity, l-diversity

1. INTRODUCTION

Privacy Preservation has been an important and necessary research topic since there is a need for protecting private information about individuals. The policy of privacy preservation is applied in data publishing where large amount of data are shared or published. Preserving the private information of individuals during data publishing is termed as Privacy Preserving Data Publishing (PPDP). Many approaches have been used to preserve the security of data over the years and the most prominent one is the anonymization.

The process of keeping an individual's information secure by modifying the data in such a way that an attacker will not be able to identify the individual's sensitive information is called anonymization. The data is modified in such a way that the privacy must be achieved and also the usefulness of the data is not degraded. Publishing the data in a manner that preserves the sensitive information of an individual and also provides enough data for decision making is the goal of PPDP.

The structure of the data that is published contains various elements such as: (1) Identifier attributes are those that are used to identify an individual uniquely from the data, (2) Quasi-identifiers are the attributes that depict the details of the individual and these attributes are published to the third party for some research purposes. (3) Sensitive attribute is the private or personal detail of the individual that should be kept secure. The identifier attribute is always kept secure and it is not included during data publication.

During anonymization the records in the data are split into many equivalence classes in such a way that all the records in a given equivalence class contains the same quasi-identifier values. In the Micro-aggregation Generalization (MAGE) approach [1] the equivalence classes are formed by using micro-aggregation for numerical data and generalization for categorical data. The MAGE approach provides k-anonymity to the data according to which each of the equivalence classes contains at least "k" records whose quasi-identifiers have been generalized.

Even though the MAGE approach hides the sensitive information of an individual in most of the cases, it is still prone to attacks such as homogeneity attack if all the sensitive attribute values inside an equivalence class are same. To avoid this, the concept of l-diversity [3] is used which states that the equivalence class should have "l" different sensitive attribute values.

1.1.Challenges and Our Contributions

Preserving the privacy of data is a challenging task since it deals with the privacy of many people and it could lead to harmful or malicious consequences if the secure information is breached by an attacker. So the data should be handled carefully when implementing privacy preservation such that the sensitive information should be kept secure from all means of attacks. Many such attacks are available that aims to breach the sensitive information of an individual by analyzing the published data.

The major challenges in Privacy Preserving Data Publishing (PPDP) are determined based on two factors such as: (1) Privacy level and (2) Data utility. The privacy level deals with the amount of information that can be made hidden from outside attacks and the data utility represents the data integrity of the original data. The PPDP approaches should always provide a high privacy level with very good data utility but there is always a tradeoff between these two factors.

In this paper, the l-diversity based PPDP approach provides a privacy model that has the following contributions:

- Anonymization of the original data is done before publishing.
- The sensitive attribute of individuals are protected by using the concept of diversity within each equivalence classes.
- The data utility of the original data is preserved as much as possible.

2.RELATED WORK

Privacy of published data is an important factor and many approaches have been used for this purpose. The research in privacy has been at its peak since years ago. An important approach to achieve privacy in data publishing is the concept of anonymization of data. In anonymization the data is changed in such a way that an attacker will not be able to identify an individual's

information by accessing the published data. This section talks about some of the related works that aim towards solving homogeneity attacks in the published data.

The first novel anonymization approach ever used is the k-anonymity [2] where the data is split into various list of equivalence classes in such a way that each equivalence class contains at least "k" records and the quasi-identifiers of all these records within an equivalence class are the same. To achieve this, technique like Micro-aggregation and Generalization is used for calculating the equivalence values within each class. The Micro-aggregation Generalization (MAGE) approach is a k-anonimization approach that makes use of the mixed distance to calculate the similar records and cluster them together as equivalence classes.

But the k-anonymization approaches including MAGE suffers seriously from homogeneity attacks where an attacker will focus on a particular equivalence class that contains all similar sensitive attribute values. To prevent this attack, an enhanced k-anonymity model named the (α,k)-anonymity [4] was proposed and here apart from k-anonymity each of the equivalence classes should satisfy another rule that states that the number of occurrence of particular sensitive value should always be less than "α".

The (α,k)-anonymity is still prone to homogeneity attacks in some of the cases and so a novel method called the l-diversity is used to completely remove homogeneity. In l-diversity all the k-anonymized equivalence classes are further diversified such that the number of unique sensitive attributes values in each equivalence class should be at least "l". In this way homogeneity is prevented in all the equivalence classes. This is further enhanced by using a different variation of the normal l-diversity model called as the distinct l-diveristy or the p-sensitivity k-anonimity. Here each equivalence class should contain at least "p" distinct sensitive values and the maximum allowed number of combinations of quasi-identifier values is kept minimal.

The l-diversity [3] and distinct l-diversity [5] models solve the homogeneity attacks but are still prone to another type of attack called the similarity attack. Also the information loss here is not handled that much. For these purposes special enhanced methods such as the (p,α)-sensitivity k-anonymity and the (p+,α)-sensitivity k-anonymity are used. The similarity attacks can be prevented here with only a minimum loss of information during anonymization.

Apart from these there are two other methods that address the problem of homogeneity in the published data. The t-closeness [6] apart from homogeneity also solves the proximity attacks. Here the distributions of sensitive attribute values in each equivalence class are kept similar to that of the whole table with a difference not more than "t". This is done by calculating the distance between the distribution in the equivalence class and the table.

Next in the (n,t)-closeness model [7] the equivalence classes are formed in such a way that there exist another equivalence class that contains at least "n" records and this class will be a natural subset of the previous equivalence class. The t-closeness distance between these equivalence classes should not be more than "t". Both t-closeness and (n,t)-closeness are more suitable for categorical sensitive values and in case of numerical sensitive values it leads to proximity attacks easily. But on overall situation these methods discussed here handles the problem of homogeneity attacks.

2.1.Preliminaries

A. K-Anonymity

Anonymization of data can be done using methods that protect the privacy of the data. The process of providing anonymization to the published data by splitting the data into many equivalence classes such that the number of records in each of the equivalence classes is at least "k" and they have the same quasi-identifier values. Since all the quasi-identifiers in an equivalence classes have the same values, the attacker will not be able to identify an individual's information using these values.

But in some cases if all the sensitive attribute values inside an equivalence class are same then the attacker can easily say that all the records within that equivalence class will have that sensitive value. This is called as homogeneity attack and to prevent this attack the concept of l-diversity [3] is used.

B. L-Diversity

The l-diversity is an advanced privacy preserving model that aims to prevent homogeneity attacks that prevail in many cases of k-anonymity. The principle behind l-diversity is that all the equivalence classes in the published data should have at least "l" different sensitive attribute values in them. That is no equivalence class should have all same sensitive attribute values and the distinct number of sensitive attribute values should be greater than or equal to "l".

The homogeneity attack is completely removed in l-diversity since all the homogeneities are removed from the equivalence classes. But there is a need for reordering of the records within various equivalence classes to achieve the l-diversity property. Due to this the equivalence classes should again be generalized.

C. Homogeneity Attack

The homogeneity attack is a type of malicious attack that violates the privacy of an individual in the published data. This attack is more prevalent in data that contains many repeated values of sensitive attributes. The attacker's goal will be to identify the equivalence classes that contain all similar sensitive attribute values. That is the equivalences classes where the unique sensitive attribute count value is one. This is one of the most common and general type of privacy attack and with further knowledge this will also lead to attribute disclosure attacks and background knowledge attacks.

3.MICRO-AGGREGATION GENERALIZATION METHOD

The most suitable k-anonymization method for mixed data that contains both numerical and categorical attributes is the Micro-aggregation Generalization (MAGE) approach [1]. The data to be anonymized is first preprocessed before applying the anonymization process. Invalid values, null values and assertives such as question marks and symbols should be removed during pre-processing step. The data after preprocessing is then clustered using the cluster partition algorithm where the given data is split into different "c" partitions.

The MAGE approach can anonymize small clusters of data faster compared to that of the whole dataset and for this purpose the cluster partitioning algorithm is implemented. At first the cluster partitioning algorithm selects "c" random cluster centers from the given dataset and then for each of the data values in the dataset, the mixed distance is calculated with each of the cluster centers. All the data values are assigned to the clusters with which they have minimum distance and the new cluster centers are calculated for all the clusters. The whole process is repeated until the stopping criterion is met. In this case, stopping criteria is met if two consecutive iterations of cluster partition algorithm produce the same cluster split.

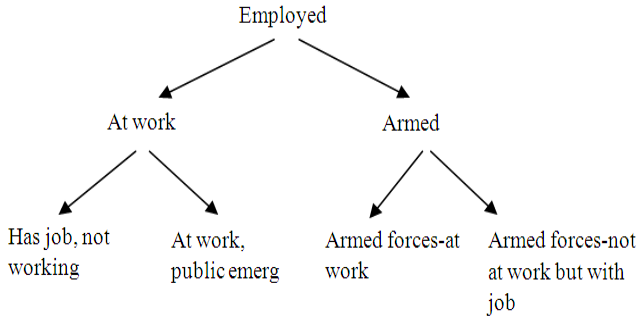

Figure. 1. Sample Value Generalization Hierarchical Tree

The mixed distance calculated in the cluster partitioning algorithm contains the distance calculated for numerical attribute and categorical attribute separately. For numerical attribute the distance is calculated using the Euclidean distance formula and for categorical attribute, the generalization distance is calculated by taking the position of the categorical attributes in the value generalization hierarchy tree (VGHT) that is built for that attribute. The quasi-identifier attributes "Age" and "Person weight" is taken as the numerical attributes and the "Employment status" is taken as the categorical attribute. A sample VGHT is given above in Figure 1 for the "Employment status" attribute. After calculating the numerical and categorical distance separately, the mixed distance is calculated using the formula given below.

$$d(t_i,t_j) = d_o(t_i,t_j) + f(d_c(t_i,t_j))$$

Where,

- $d_o(t_i,t_j)$ be the numerical attribute part distance of the tuples t_i,t_j,
- $d_c(t_i,t_j)$ be the categorical attribute part distance of the tuples t_i,t_j
- $f(.)$ is the mapping function which can control the proportion of numerical attribute distance and categorical attribute distance.

In the above equation the numerical distance is taken as it is where as the categorical distance is applied to a function f as given below.

$$f(d_c) = N_C / N_O \, (d_{omin} + (d_{omax} - d_{cmin} / d_{cmax} - d_{cmin}) * (d_c - d_{cmin}))$$

Where,

- N_C is the number of numerical attributes and N_O is the number of numerical attribute taken
- d_{omax} refers to maximum distance for numerical attributes
- d_{omin}refers to minimum distance for numerical attributes

The minimum and maximum distance for numerical and categorical attributes is identified by calculating the distances for all possible combinations of numerical attribute values and categorical attribute values separately.

After the cluster partitioning algorithm is implemented, each of the clusters are passed as input to the Clustering K-Anonymizing (CKA) algorithm that is in turn used to anonymize the given input cluster to get the anonymized data. To achieve this, first all the data values in the cluster partition is divided into various equivalence classes such a way that the data values in a particular equivalence class are similar.

An equivalence class is randomly selected from equivalence class set and the equivalence class that has the minimum mixed distance is anonymized with this equivalence class (randomly selected equivalence class). The categorical attribute values are replaced with the closest common generalization values and the numerical attribute values are replaced with the average values. The closest common generalization values (CCGV) are calculated using the value generalization hierarchy tree (VGHT). The final partitions represent the clusters that are obtained from the cluster partitioning algorithm and those partitions are given as input to CKA algorithm that implements the MAGE method.

D.Pseudo code

```
Clustering K-Anonymizing Algorithm(CKA)
Original Dataset: D
Input Dataset: P
Output: k-anonymized dataset
Quasi-identifiers: QI = [EMPSTAT, AGE, PERWT,
MARST]
Clusters: Ci with i = 1 to C
Number of clusters: C
Anonymization factor: k
Anonymized Dataset: T
begin process
    preprocess the dataset T
    get QI values alone and store in P
    begin cluster(P, C)
        select C random cluster centers from P
        calculate mixed distance of each row with all
        cluster
        allocate rows to cluster Ci with minimum
        distance
    end cluster
```

```
    begin anonymization(Ci, k)
        for each cluster Ci if cluster size < k then
            generalize all rows in Ci
            add generalized rows to T
            else implement CKA algorithm
            begin CKA(Ci, QI, k)
                form equivalence class with each row in
                Ci
                add equivalence classes to E
                randomly select Ei from E
                select Ej from E such that Ei and Ej
                have minimum mixed distance
                generalize rows of Ei and Ej combined
                add generalized rows to T
                remove Ei and Ej from E
                repeat until E is empty
            end CKA
        end for
        return T
    end anonymization
end process
```

3.L-DIVERSITY OVER MAGE

The MAGE approach was implemented for the dataset and from the experimental results it is evident that the MAGE method is less secure and is prone to many types of attacks such as homogeneity and background knowledge attacks. To handle these issues the k-anonymized dataset obtained from the MAGE approach can be further processed for l-diversity. The main requirement to satisfy l-diversity is that all the equivalence classes should have well represented or diversified sensitive attribute values in them. That is, the number of unique sensitive attribute values in an equivalence class should always be at least “l”. In the l-diversity approach the k-anonymization and l-diversity are both checked for and so the overall security of the data is preserved and it handles or prevents attacks such as homogeneity attack.

The overall process flow of the l-diversity over the MAGE approach is displayed below in Figure 3. The l-diversity approach can be applied as an extension of the MAGE approach. The system architecture aims to improve the security of the anonymized data when there is similarity in the sensitive attribute values in single or more equivalence classes. At first, each of the equivalence class obtained from the MAGE algorithm is tested for l-diversity in the sensitive values. If the equivalence class satisfies l-diversity then it is kept as it is or else the equivalence class is removed and all the tuples inside the equivalence class is taken for further processing.

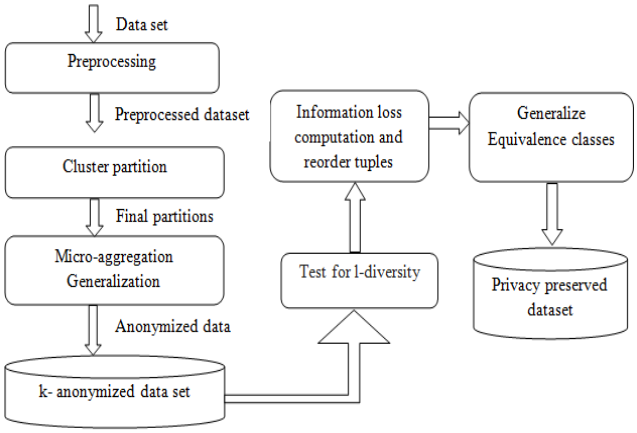

Figure. 3. Overall process flow of l-diversity method

In the next step, the tuples that do not satisfy l-diversity are merged with the remaining equivalence classes in such a way that the information loss is minimal. Information loss is calculated for each of these tuples with each of the equivalence classes that are satisfying l-diversity. These tuples are then put into the equivalence class in which they have minimum information loss. Every time an equivalence class is added with the new tuples which is again checked for l-diversity to see if the diversity has increased. In the last step, the final sets of equivalence classes that remain are generalized. The obtained results are then evaluated for privacy measures to make a comparison with the MAGE approach. The l-diversity approach will prevent the homogeneity attacks that are available in MAGE approach and the data utility can also be increased.

The two parameters that have been considered for privacy checking in l-diversity are the privacy measure and the utility measure. The information loss that is employed in l-diversity is discussed in detail further.

E. Privacy Measure

The privacy measure is calculated by using the query answering mechanism. After applying the l-diversity approach or MAGE approach to the data and obtaining the anonymized data, certain queries are executed to retrieve records from both the original dataset and the anonymized dataset. At first the query is executed for the original dataset and the number of records fetched is noted. The same query is executed for the anonymized dataset and the number of records fetched here is also noted down. If the number of records getting retrieved gets improved, it is meant that the privacy gets improved. That is, the records retrieved in the anonymized data must be higher than the number of records fetched in the original data. Certain queries are executed to check whether the anonymized table meets privacy or not. If the number of rows fetched in the original table is less than the rows fetched in the anonymized table, then theanonymized data is said to be more secure than the original data.

F.Utility Measure

Information loss factor is a very important parameter when merging two equivalence classes or when adding a tuple to another equivalence class. By calculating the Information loss before merging two equivalence classes or before adding a tuple into equivalence class we can identify the loss in amount of data after merging or adding. So it is advisable to add or merge data when the Information loss is minimal. If the characteristics of the data are similar then the Information loss will be minimal between them. The formula to calculate the information loss is given below.

$$IL(\Omega)=|\Omega|.\left(\sum_{i=1}^{r}\frac{N_{imax}-N_{imin}}{\eta N_{imax}-\eta N_{imin}}+\sum_{i=1}^{S}\frac{H(\Lambda(U_{C_j}))}{H(\tau_{C_j})}\right)$$

Where,

- $|\Omega|$ is the total number of records in Ω, where Ω represents a single equivalence class
- τ_{C_j} *is* the sub tree rooted at the lowest common ancestor of every value in U*Cj,*
- *H (τ)* is the height of tree *τ,* η is the records with "r" numeric quasi-identifiers $N_1,N_2,\ldots N_r$ and "*s"* categorical quasi-identifiers $C_1,C_2,\ldots,C_s$.
- N_{imax} be the minimum numerical attribute value where N_i = (i=1, 2….r)
- N_{imin} be the maximum numerical attribute value N_i = (i=1, 2….s)

The equivalence classes or tuples having minimum information loss are merged together to form new equivalence class. By keeping the information loss as minimum as possible the original data is not lose and the algorithm is said to be efficient.

4.DATASET DESCRIPTION AND EXPERIMENTAL RESULTS

The experimental results are evaluated by implementing the MAGE approach with the IPUMS dataset that is discussed below.

Integrated Public Use Microdata Series (IPMS) [29] is the world's most population database. IPUMS is a microdata that consists of historical data samples from United States census records and provides information about individuals and households. IPUMS dataset has data samples up to the year 2014 with more than 6 million instances and it also contains 100 or more attributes in it. The dataset provides different numerical attribute values such as Age, Person weight and various categorical attributes such as Occupation, Marital status, income, sickness etc. The attribute values "Age", "Person weight" and "Employment status" are taken as the Quasi-identifier attributes and the attribute "Marital status" is taken as the sensitive attribute which provides the personal detail of an individual that must be kept secure. The various numerical, categorical attribute and sensitive attributes values and their number of unique count values are calculated as given below in the Table 1.

Table I. Numerical and Categorical Attributes

Attribute type	Name	Data type	Values
Numerical attribute	Age	integer	0-135
	Person weight	integer	Upto 650
Categorical attribute	Employment status	string	16 unique values
Sensitive attribute	Marital status	string	6 unique values

Initially the records are converted into a persistent format before applying different anonymization operations over it. The anonymization operations are performed mainly to protect individual's personal information. At first the dataset is preprocessed to remove all the null values, assertion, invalid values, etc. After identifying the various numerical and categorical attributes in the dataset, only the attributes that are needed for anonymization process are taken. They are the quasi-identifiers and the other attributes can be neglected. For all the selected categorical quasi-identifier attributes, the Value Generalization Hierarchical Tree (VGHT) is implemented. In our case the VGHT is implemented only for the "Employment Status" attribute that is selected for the anonymization process. The dataset after preprocessing and selection of attributes will look as below in Table 2.

Table II. IPUMS dataset sample records

Employment Status	Age	Person Weight	Marital Status
Not in Labor Force	17	286	Never married/single
Not in Labor Force	23	159	Never married/single
Has job-not working	54	115	Married-spouse absent
Unemployed	19	33	Never married/single
Armed forces-at work	73	113	Divorced
At work	52	95	Married-spouse present

Next the mixed distance calculation process is executed that calculates the minimum and maximum distances for both numerical and categorical attributes. Using these values the mixed distance function can be used for calculating the distance between any two given vectors. Here a vector or record is represented as given below:

[Employment status, Age, Person weight, Marital status]

All the input records of the dataset are converted into records that contain only the selected quasi-identifier values and sensitive value. But the sensitive value is not used for any processing and remains unchanged till the end. This way the calculation of distance can be implemented easily. The preprocessed IPUMS dataset is given as the input to the MAGE approach that is discussed previously in chapter 3.

The obtained results show that some of the equivalence classes suffer from homogeneity attacks in MAGE approach. This can be explained using Table 3 below where two equivalence classes are formed using the dataset values given above in Table 2.

Table III. Homogeneity attack in MAGE

Equivalence Class	Employment Status	Age	Person Weight	Marital Status
EC1	Employed	20	159	Never married/single
	Employed	20	159	Never married/single
	Employed	20	159	Never married/single
EC2	Unemployed	60	108	Married-spouse absent
	Unemployed	60	108	Divorced
	Unemployed	60	108	Married-spouse present

From the table it is seen that the EC1 has homogeneity since all the rows have the same sensitive attribute value. But the EC2 does not have homogeneity. If an attacker knows the general details about an individual such as employment status, age group and weight range then the attacker can easily identify his/her marital status. From Table 3 if an individual is in the age group around 20 with person weight value of around 160 and is employed, then the attacker will easily know that he/she is not married since all values of marital status in that equivalence class is the same. From the obtained results the following inferences have been made:

- The obtained results from MAGE provide better privacy in terms of privacy measure (Query answering discuss in the previous chapter).
- The data utility of the MAGE approach is not considered in this implementation and so the overall efficiency cannot be identified.
- In some of the equivalence classes all the sensitive attribute values were the same and this led to the homogeneity attack in MAGE approach.

To avoid these drawbacks of MAGE algorithm discussed above, a novel method should be implemented in to avoid homogeneity attacks with a better data utility factor and still preserving the privacy measure.

5.CONCLUSION AND FUTURE WORK

Preserving the sensitive information about an individual in the published data is an important aspect in data mining. Data anonymization approaches are used to morph the data in such a way that it can be used for necessary analysis but also hides the private information in the data. The Micro-aggregation and Generalization (MAGE) approach makes use of the k-anonymity model but it suffers from homogeneity attack as all the sensitive attribute values in a particular equivalence class sometimes are same. The l-diversity method avoids the homogeneity attack by using the concept of diversity among the sensitive values in an equivalence class. The l-diversity based anonymization approach preserves security of the published data comparatively better than the MAGE approach. Though the l-diversity approach avoids the homogeneity attack in the equivalence classes by rearranging the records in such a way that the information loss is less, reordering the records leads to probability inference attack in the equivalence classes since the probability of the sensitive attribute increases in the equivalence class where the record is reordered. To avoid this problem a combination of MAGE and l-diversity model can be used in the future that can avoid both homogeneity and probability inference attack and it would also be interesting if the MAGE method can be adapted to other kinds of data such as set-valued data.

REFERENCES

[1] Jianmin Han, Juan Yu, Yuchan, Jianfeng Lu, Huawen Liu (2014), "MAGE: A Semantics Retaining k-anonymization method for mixed data", ELSEVIER Knowledge-Based Systems, Volume 55, pp.76-86.

[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, Volume. 1, No. 1, Article 3.

[4] Raymond, Jiuyong Li et al. (2006), "(α, k) Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 754-759.

[5] Traian Marius Truta, Bindu Vinay (2006), "Privacy Protection: p-Sensitive k-Anonymity Property", In proceedings of the 22nd IEEE International Conference on Data Engineering Workshops, pp.94.

[6] Ninghui Li, Tiancheng Li, Suresh Venkata Subramanian (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.

[7] Ninghui Li, Tiancheng Li et al. (2010), "Closeness: A New Privacy Measure for Data Publishing", Knowledge and Data Engineering, IEEE Transactions on, Volume: 22, Issue: 7, pp. 943 – 956.

[8] Enamul Kabir, Hua Wang, Elisa Bertino &Yunxiang Chi (2010), "Systematic Clustering Method for l-diversity Model", CRPIT– Database Technologies, Volume 104.

[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. pp. 1-53.

[10] Yang Xu,Tinghuai Ma et al."A Survey of Privacy Preserving Data Publishing using Generalization and Suppression", International Journal of Applied Mathematics & Information Sciences, Volume No. 3,pp. 1103-1116

[11] Q. Zhang, N. Koudas, D. Srivastava, T. Yu (2007), "Aggregate query answering on anonymized tables", in: Proceedings of International Conference on Data Engineering, pp. 116–125.

[12] J. Li, Y. Tao, X. Xiao (2008), "Preservation of proximity privacy in publishing numerical sensitive data", in: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 473–486.
[13] HuangXuezhen, LIU Jiqiang et al (2014), "A New Anonymity Model for Privacy-Preserving Data Publishing", IEEE journal of china communications, Volume 11, Issue 9,pp.47-59.
[14] Jianneng Cao, Panagiotis Karras (2012), "Publishing Microdata with a Robust Privacy Guarantee", Journal Proceedings of the VLDB Endowment, Volume 5 Issue 11,pp. 1388-1399.
[15] 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. pp. 1-53.
[16] Jiuyong Li, Chi-Wing Wong, Ada Wai-Chee Fu, Jian Pei (2008), "Anonymization by Local Recoding in Data with Attribute Hierarchical Taxonomies", IEEE Transactions on knowledge and data engineering, Volume.20, No.9.
[17] J. Domingo-Ferrer,V.Torra (2005),"Ordinal, continuous and heterogeneous k-anonymity through micro-aggregation", Journal of Data Mining and Knowledge Discovery Volume.11, No .2, pp. 195–212.
[18] K. Le Fevre, D.J. DeWitt, R. Ramakrishnan (2005), "Incognito: Efficient full-domain k-anonymity", in: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 49–60.
[19] X. Xiao, Y. Tao (2006), "Personalized privacy preservation", in: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 229–240.
[20] W. Jiang, C. Clifton (2006), "A secure distributed framework for achieving k-anonymity", International Journal on Very Large Data Bases 15 (4), pp.316–333.
[21] T. Li, N. Li(2008), "Towards optimal k-anonymization", Data & Knowledge Engineering 65 (1), pp. 22–39.
[22] V. Torra (2004), "Microaggregation for categorical variables: a median based approach", in: Workshop on Privacy in Statistical Database, pp. 162–174.
[23] A. Oganian, J. Domingo-Ferrer (2001), "On the complexity of optimal microaggregation for statistical disclosure control", Statistical Journal of United Nations Economic Commission for Europe 18 (4),pp. 345–354.
[24] C.C. Chang, Y.C. Li, W.H. Huang (2007), TFRP: "An efficient microaggregation algorithm for statistical disclosure control", Journal of Systems and Software 80 (11), pp. 1866–1878.
[25] Y. Tao, H. Chen, X. Xiao, S. Zhou, D. Zhang (2009), "ANGEL: enhancing the utility of generalization for privacy preserving publication", IEEE Transaction on Knowledge and Data Engineering 21 (7), pp.1073–1087.
[26] N.V. Mogre, G. Agarwal, P. Patil (2012), "A review on data anonymization technique for data publishing", International Journal of Engineering Research & Technology 1(10).
[27] J. Domingo-Ferrer, J.M. Mateo-Sanz (2002), "Practical data-oriented microaggregation for statistical disclosure control", IEEE Transactions on Knowledge and Data Engineering 14 (1),pp.189–201.
[28] D. Sacharidis, K. Mouratidis, D. Papadias (2010), "K-anonymity in the presence of external databases", IEEE Transactions on Knowledge and Data Engineering 22 (3), pp.392–403.
[29] https://usa.ipums.org/usa-action/variables/group