International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 06 | June -2017 www.irjet.net p-ISSN: 2395-0072

Privacy Protection of Sensitive Microdata in Healthcare System using t-Closeness through Microaggregation

Sonu V. Khapekar1, Prof. Lomesh Ahire2

1 Sonu Khapekar, PG Scholar, Department of CSE, NMIET, Pune, Maharashtra.
2 Prof. Lomesh Ahire, Department of CSE, NMIET, Pune, Maharashtra.

Abstract- The preservation of privacy of published micro data is essential to prevent the sensitive information of individuals from being disclosed. Several privacy models are used for protecting the privacy of micro data. Micro aggregation is a technique for disclosure limitation aimed at protecting the privacy of data subjects in micro data releases. It has been used as an alternative to generalization and suppression to generate k-anonymous data sets, where the identity of each subject is hidden within a group of k subjects. Unlike generalization, micro aggregation perturbs the data and this additional masking freedom allows improving data utility in several ways, such as increasing data granularity, reducing the impact of outliers, and avoiding discretization of numerical data. k-Anonymity, on the other side, does not protect against attribute disclosure, which occurs if the variability of the confidential values in a group of k subjects is too small .

In this paper, the preservation of privacy of micro data released in healthcare system is focused through micro aggregation by using t-closeness which is a more flexible privacy model assuring strictest privacy. Existing algorithms to generate t-close data sets are based on generalization and suppression. This paper proposes, how to use micro aggregation in healthcare system to generate k-anonymous t-close data sets. The advantages of micro aggregation are analyzed, and the micro aggregation algorithm for k-anonymous t-closeness is presented. The micro aggregation by using t-closeness proves an effective tool for protecting the privacy of the sensitive attributes in the healthcare system.

Key Words: healthcare system, data privacy, microaggregation k-anonymity, l-diversity, t-closeness.

1.INTRODUCTION

The micro data such as medical data or census data is usually published by government agencies and other organizations for scientific, research and other purposes. Such data are stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes which are classified into three types[3],[8]:

1) Attributes that clearly identify individuals. These are known as explicit identifiers e.g., Social Security Number. 2) Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers e.g. Zip code, Birth-date, and Gender. 3) Attributes that are considered sensitive, such as Disease and Salary.

When releasing micro data, it is important to protect the sensitive information of the individuals from being disclosed. The information disclosure have two types : identity disclosure and attribute disclosure. When an individual is linked to a particular record in the released table, it causes identity disclosure. The attribute disclosure means when new information about some individuals is revealed, i.e., the released data make it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is identified. and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure. It has been recognized that even disclosure of false attribute information may cause harm . An observer of a released table may incorrectly perceive that an individual's sensitive attribute takes a particular value and behaves accordingly based on the perception. This can harm the individual, even if the perception is incorrect.

As the released table gives useful information to researchers [5], it presents disclosure risk to the individuals whose data are in the table. Thus the objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonym zing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge

(e.g., knowing a particular individual in person), or from other publicly available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. Generalization is the common anonymization approach. It replaces quasi- identifier values with values that are less-specific but semantically consistent. It causes more records will have the same set of quasi-identifier values. The equivalence class of an anonym zed table is a set of records that have the same values for the quasi-identifiers.

It is required to measure the disclosure risk of an anonym zed table to effectively limit disclosure . Thus k-anonymity is introduced [4], which defines the property tha each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier. In other words, k-anonymity requires that each equivalence class contains at least k records. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. To address this limitation of k-anonymity, new notion of privacy is introduced called as l- diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least l "well- represented" values.

The disadvantage of l-diversity is that it is limited in its assumption of adversarial knowledge. It is possible for an adversary to gain information about a sensitive attribute as long as it has information about the global distribution of this attribute. This assumption generalizes the specific background and homogeneity attacks used to motivate l-diversity. Another problem with privacy-preserving methods is that they effectively assume all attributes to be categorical. The adversary either does or does not learn something sensitive.

This paper proposes a novel privacy notion called "closeness." The idea of global back- ground knowledge is formalized and proposed the base model t-closeness. It requires that the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). It effectively limits the amount of individual-specific information an observer can learn. The analysis on data utility shows that t-closeness substantially limits the amount of useful information that can be extracted from the released data. Thus a more flexible model is proposed, called (n,t)-closeness. It requires that the distribution in any equivalence class is close to the distribution in a large-enough equivalence class (contains at least n records) with respect to the sensitive attribute. This limits the amount of sensitive information about individuals while preserves features and patterns about large groups. The analysis shows that (n,t)-closeness achieves a better balance between privacy and utility than existing privacy models such as l-diversity and t-closeness.

2. MICRODATA BACKGROUND

A micr odata is defined as a table where each row contains data on a different subject and each column contains information about a specific attribute. Let T (A1. ... Am) be a micro data set with n records r1 ..... rn , each of them with information about attributes A1 ..... Am.

The attributes in a micro data set are classified according to their dis closiveness into several classes [8] such as identifiers, quasi-identifiers, confidential attributes, and non-confidential attributes. The statistical disclosure control restricts the capability of an intruder with access to the released data set to associate a piece of confidential information to a specific subject in the data set. Thus a masked version T'(A1 , ... ,An) of the original data set T(A1 , ... ,An) is released. The term anonymized data set is referred to T'(A1 , ... ,An).

2.1 k-Anonymity

An intruder re-identifies a record in an anonymized data set when he can determine the identity of the subject to whom the record corresponds. In case of re-identification, the intruder can associate the values of the confidential attributes in the re-identified record to the identity of the subject, thereby violating the subject's privacy. k-Anonymity seeks [11]to limit the capability of the intruder to perform successful re-identifications.

Definition (k-Anonymity): let T be a data set and QIT be the set of quasi identifier attributes in it. T is said to satisfy k-anonymity if, for each combination of values of quasi-identifiers in QIT , at least k records in T share that combination.

In a k-anonymous data set , no subject's identity can be linked (based on the quasi-identifiers) to less than k-records. Hence the probability of correct re-identification is, at most ,1/k.

The protection k-anonymity provides is simple and easy to understand. If a table satisfies k-anonymity for some value k, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding

to that individual with confidence greater than 1/k. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. Two attacks were identified in this (i) Homogeneity attack and (ii) Background knowledge attack. The illustration of k-anonymity is shown by the following two tables which shows the original patients table and 3-anonymous version of the original patients records.

TABLE-1:Original Patients Table

	ZIP code	Age	Disease
1	35602	28	Cancer
2	35677	22	Cancer
3	35605	29	Heart Disease
4	35678	44	Heart Disease
5	35671	48	Heart Disease
6	35674	51	Cancer
7	35645	30	Flue
8	35652	36	Heart Disease
9	35602	32	Cancer

TABLE 2: A 3 –Anonymous version of Table-1

	ZIP code	Age	Disease
1	356**	2*	Cancer
2	356**	2*	Cancer
3	356**	2*	Heart Disease
4	3567*	>40	Heart Disease
5	3567*	>40	Heart Disease
6	3567*	>40	Cancer
7	356**	3*	Flue
8	356**	3*	Heart Disease
9	356**	3*	Cancer

2.2 l-Diversity.

The limitations of k-anonymity are addressed by a stronger notion of privacy known as l-diversity [14].

Definition (The l-diversity principle). An equivalence class is said to have 'l-diversity if there are at least l "well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. The various interpretations of the term "well represented" in this principle are listed below:

1.Distinct l-diversity: The simplest understanding of "well represented" would be to ensure that there are at least l distinct values for the sensitive attribute in each equivalence class. Distinct l-diversity does not prevent probabilistic inference attacks. An equivalence class may have one value appear much more frequently than other values, enabling an adversary to conclude that an entity in the equivalence class is very likely to have that value. This motivated the development of the following stronger notions of l-diversity.

2. Probabilistic l-diversity: An anonymized table satisfies probabilistic l-diversity if the frequency of a sensitive value in each group is at most 1/l. This guarantees that an observer cannot infer the sensitive value of an individual with probability greater than 1/ l .

3. Entropy l -diversity. The entropy of an equivalence class E is defined to be

Entropy(E) = $\sum p(E,s) \log p(E,s)$

in which S is the domain of the sensitive attribute and p(E,s) is the fraction of records in E that have sensitive value s. A table is said to have entropy l -diversity if for every equivalence class E, Entropy(E)>=log l .Entropy l -diversity is stronger than distinct l-diversity. In order to have entropy l- diversity for each equivalence class, the entropy of the entire table must be at least log (l). Sometimes, this may too restrictive, as the entropy of the entire table may be low if a few values are very common.

While the l-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure, it has several shortcomings. l-diversity is insufficient to prevent attribute disclosure. There are two attacks on l-diversity. (i) Skewnes attack and (ii) Similarity attack.

2.3 T-Closeness:

The k-anonymous data sets are vulnerable to attribute disclosure even though k-anonymity protects against identity disclosure. While the l-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure, it is difficult to achieve and may not provide sufficient privacy protection against attribute disclosure.

T-Closeness [1],[12] seeks to limit the amount of information that an intruder can obtain about the confidential attribute of any specific subject. To this end, t-closeness requires the distribution of the confidential attributes within each of the equivalence classes to be similar to their distribution in the entire data set.

Definition : An equivalence class is said to satisfy t-closeness if the distance between the distribution of the confidential attribute in this class and the distribution of the attribute in the whole data set is no more than a threshold t. A data set (usually a k-anonymous data set) is said to satisfy t-closeness if all equivalence classes in it satisfy t-closeness.

The specific distance used between distributions is central to evaluate t-closeness, but the original definition does not advocate any specific distance. The earth mover's distance (EMD) is the most common choice[10]. EMD(P,Q) measures the cost of transforming one distribution P into another distribution Q by moving probability mass. EMD is computed as the minimum transportation cost from the bins of P to the bins of Q, so it depends on how much mass is moved and how far it is moved. For numerical attributes the distance between two bins is based on the number of bins between them. If the numerical attribute takes values {v1, v2, ... vm}, where vi < vj if i < j, then ordered distance(vi,vj)=(i-j)/(m-1). Now, if P and Q are distributions over {v1, v2, ... vm} that, respectively, assign probability pi and qi to vi , then the EMD for the ordered distance can be computed as

$$\text{EMD (P,Q)} = \frac{1}{m-1}\sum_{i=1}^{m}\left|\sum_{j=1}^{i} pj - qj\right|$$

2.4 Micro aggregation

Micro aggregation is a technique for disclosure limitation [2] aimed at protecting the privacy of data subjects in microdata releases. It has been used as an alternative to generalization and suppression to generate k-anonymous data sets, where the identity of each subject is hidden within a group of k subjects. Unlike generalization, microaggregation perturbs the data and this additional masking freedom allows improving data utility in several ways, such as increasing data granularity, reducing the impact of outliers, and avoiding discretization of numerical data.

Microaggregation is a family of purturbative method for statistical disclosure control of microdata releases. One dimensional microaggregation was introduced in[2] and multidimensional microaggregation was proposed and formalized in[6]. The latter is the one that is useful for k-anonymity and t-closeness. It consists of the following two steps:

(1)Partition: The records in the original data set are partitioned into several clusters, each of them containing at least k records. To minimize the information loss, records in each cluster should be as similar as possible.

(2) Aggregation: An aggregation operator is used to summarize the data in each cluster and the original records are replaced by the aggregated output. For numerical data, one can use the mean as aggregation operator; for categorical data, one can resort to the median or some other average operator defined in terms of anontology.

The partition and aggregation steps produce some information loss. The goal of microaggregation is to minimize the information loss according to some metric. A common information loss metric is the sum of squared errors (SSE). When using SSE on numerical attributes, the mean is a sensible choice as the aggregation operator, because for any given partition it minimizes SSE in the aggregation step; the challenge thus is to come up with a partition that minimizes the overall SSE. Finding an optimal partition in multi- dimensional microaggregation is an NP-hard problem, therefore, heuristics are employed to obtain an approximation with reasonable cost.

The advantages of microaggregation [3] over generalization/recoding for k-anonymity mostly related to data utility reservation are listed as :

Global recoding may recode some records that do not need it, hence causing extra information loss. On the other hand, local recoding makes data analysis more complex, as values corresponding to various different levels of

generalization may co-exist in the anonymized data. Microaggregation is free from either drawback.

Data generalization usually results in a significant loss of granularity, because input values can only be replaced by a reduced set of generalizations, which are more constrained as one moves up in the hierarchy. Microaggregation, on the other hand, does not reduce the granularity of values, because they are replaced by numerical or categorical averages.

If outliers are present in the input data, the need to generalize them results in very coarse generalizations and, thus, in a high loss of information. For microaggregation, the influence of an outlier in the calculation of averages/centroids is restricted to the outlier's equivalence class and hence is less noticeable.

For numerical attributes, generalization discretizes input numbers to numerical ranges and thereby changes the nature of data from continuous to discrete. But, microaggregation maintains the continuous nature of numbers.

3. SYSTEM ARCHITECTURE

A healthcare system consists of four modules namely Doctor, Receptionist, Admin and Analyst which have interactions within them. Figure shows the Architecture diagram of the system.

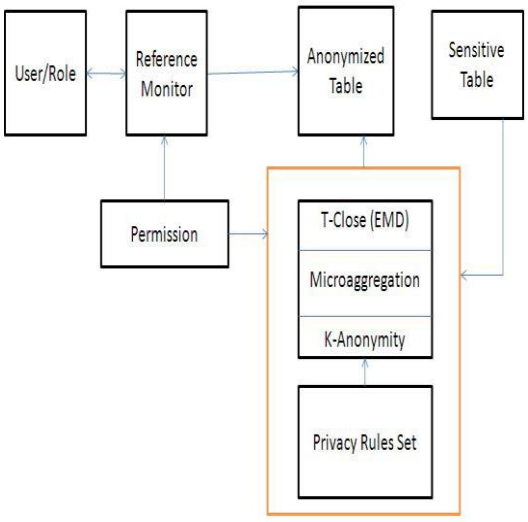

Figure 3.1 System Architecture Diagram

The Admin of the system is responsible for creating the login details of employees of hospital and has the rights to view the details of doctors, patients and employees.

Doctors can login in the system and can view the details of the patients whenever required.

Receptionist can login in the system for searching the particular patient and for enquiry of the patients. The patients data is searched by receptionist whenever required.

Analyst can login in the system. The analyst can access the full medical report. Analyst does the suppression of data. Here the constrains for privacy is implemented by t-closeness through microaggregation. The Figure 3.2 shows the module diagram of the healthcare system.

The healthcare system has microdata which has sensitive attributes such as patients disease, patients salary which are required to be protected when the medical data is released or published. Hence privacy preservation in this system is done by applying the strictest privacy preservation model of t-closeness through microaggregation.

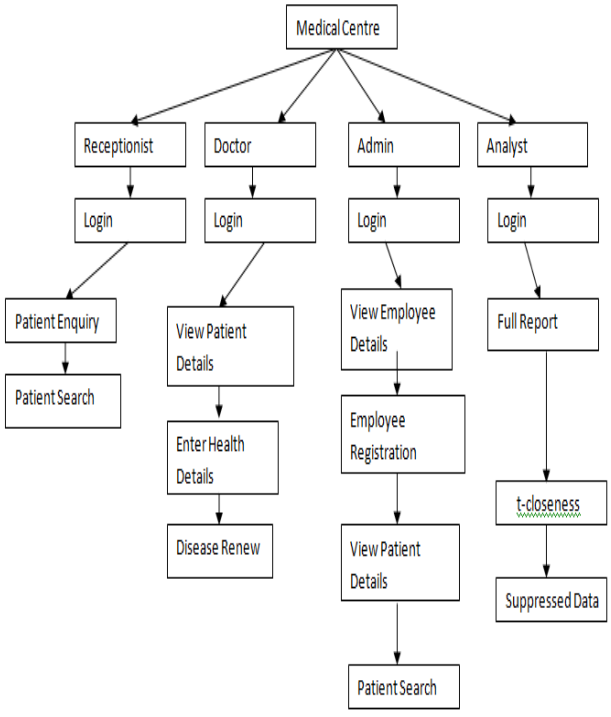

Figure 3.2 System's Module Diagram

4. SYSTEM DESIGN AND ANALYSIS

4.1 Mathematical Model

Considering a microdata set in the healthcare system with records (r1,….,rn) and the attributes (A1,…,Am)

1. Let (A1,….,Am) be a microdata set

2. Records r1,….,rn

3. Attributes A1,…,Am

4. The original data set T(A1,….,An)

5. k-anonymity is performed and T' (A1,….,An) is generated.

6. Let k be the size of cluster, the microaggregation is performed to form clusters.

7. If P,Q are bins of equivalence class, the records are moved from P to Q by calculating t-closeness to threshold value t. of the original data set which is released.. The term anonymized data set to refer to T'(A1,….,An) .

The detailed algorithm is given below which always returns a t-close data sets.

Algorithm : t-Closeness through Microaggregation and Merging of Microaggregated Groups of Records
Data: X: original data set
k: minimum cluster size
t: t-closeness level

Result Set of clusters satisfying k-anonymity and t-closeness

X'= microaggregation (X, k);
while EMD (X', X) > t do
C =cluster in X' with the greatest EMD to X
C'= cluster in X' closest to C in terms of QIs
X'= merge C and C' in X'
end while
return X' ;

The microdata set in the proposed healthcare system is subjected to generation of t-close data sets. First the microaggregation of data is carried out by the following two steps.

1. Partition: The records in the original data set are partitioned into several clusters, each of them containing at least k records.

2. Aggregation: An aggregation operator is used to summarize the data in each cluster and the original records are replaced by the aggregated output. For numerical data, one can use the mean
as aggregation operator.

The closeness of records is calculated by the specific distance used between distribution is central to evaluate t-closeness, but the original definition does not advocate any specific distance. The earth mover's distance (EMD) is the most common choice. EMD (P,Q) measures the cost of transforming one distribution P into another distribution Q by moving probability mass. EMD is computed as the minimum transportation cost from the bins of P to the bins

of Q, transportation cost from the bins of P to the bins Q, so it depends on how much mass is moved and far it is moved.

$$\text{EMD (P,Q)} = \frac{1}{m-1}\sum_{i=1}^{m}\left|\sum_{j=1}^{i} pj - qj\right|$$

4.2 Analysis of t-closeness with EMD

The TABLE 3 shows the Original salary/Disease table. The calculation of earth mover distance (EMD) for analyzing the t-closeness is carried out :

Q={2k,3k,4k,5k,6k,7k,8k,9k,10k}

P1={2k,3k,4k} and P2={5k,7k,10k} We calculate D [P1,Q] and D[P2,Q] using EMD.

Let v1=2k, v2=3k……..v9=10k, we define the distance between vi and vj to be |i-j| /8; thus the maximal distance is

1. We have D [P1,Q] = 0.325 and D[P2,Q] = 0.157

For the disease attribute, the hierarchy is used to define the ground distances. For example the distance between " Flue" and "Bronchitis" is 1/3, the distance between "Flue" and "Pulmonary embolism" is 2/3 and the distance between "Flue" and "Stomach cancer" is 3/3=1 then the distance between the distribution {gastric ulcer, gastritis, stomach cancer} and the overall distribution is 0.5, while the distance between the distribution { gastric ulcer, stomach cancer, pneumonia}is 0.265

The table 4 shows the anonymized version of table 3. It has 0.157 closeness w.r.t. Salary and 0.265 closeness w.r.t Disease. The Similarity Attack is prevented in this table.

TABLE 3:Original Salary /Disease Table

	ZIP code	Age	Salary	Disease
1	35677	31	2k	gastric ulcer
2	35602	24	3k	Gastritis
3	35678	29	4k	stomach cancer
4	35905	45	5k	Gastritis
5	35909	54	6k	Flue
6	35906	49	10k	Bronchitis
7	35605	32	8k	Bronchitis
8	35673	38	7k	Pneumonia
9	35607	34	9k	stomach cancer

TABLE 4:Table with 0.157 closeness w.r.t. Salary and 0.265
closeness w.r.t. Disease

	ZIP code	Age	Salary	Disease
1	3567*	<40	2k	gastric ulcer
3	3567*	<40	4k	stomach cancer
8	3567*	<40	7k	Pneumonia
4	3590*	>40	5k	Gastritis
5	3590*	>40	6k	Flue
6	3590*	>40	10k	Bronchitis
2	3560*	<40	3k	Gastritis
7	3560*	<40	8k	Bronchitis
9	3560*	<40	9k	stomach cancer

5. CONCLUSIONS

The preservation of privacy of sensitive attributes in healthcare system is securely and efficiently achieved by the proposed t-closeness model through microaggregation. The other privacy models, k-anonymity and l-diversity does not protect against attribute disclosure. The microaggregation perturbs the data and this additional masking freedom allows improving data utility in several ways, such as increasing data granularity, reducing the impact of outliers and avoiding discretization of numerical data. The proposed microaggregation algorithm to generate t-close data sets in microdata released in healthcare system stands out as providing one of the strictest privacy guarantees.

REFERENCES

[1] Jordi Soria-Comas,Josep Domingo-Ferrer,Fellow,IEEE,David Sanchez, and Sergio Martinez,"t-closeness through Microaggregation: Strict privacy with Enhanced utility Preservation" , IEEE Trans. Knowl. Data Eng. VOL.27,NO.11,November 2015.

[2] J. Domingo-Ferrer and J. Soria-Comas, "From t-closeness to differential privacy and vice versa in data anonymization," Knowledge based syst . vol. 74, pp. 151-158,2015.

[3] N. Li, T. Li, and S. Venkatasubramanian, "Closeness: A new privacy measure for data publishing," IEEE Trans. Knowl. Data Eng.,vol. 22, no. 7, pp. 943–956, Jul. 2010.

[4] J. Cao, P. Karras, P. Kalnis, and K.-L. Tan, "SABRE: A sensitive attribute Bucketization and Redistribution framework for t- closeness," VLDB J., vol. 20, no. 1, pp. 59–81, 2011.

[5] J. Soria-Comas and J. Domingo-Ferrer, "Differential privacy via t-closeness in data publishing," in Proc. 11th Annu. Int. Conf. Privacy, Security Trust, 2013, pp. 27–35.

[6] J. Soria-Comas and J. Domingo-Ferrer, "Probabilistic k-anonymity through microaggregation and data swapping," in Proc. IEEE Int. Conf. Fuzzy Syst., 2012, pp. 1–8.

[7] J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata," in Confidentiality, Dis- closure and Data Access: Theory and Practical Applications for Statistical Agencies. L. Zayatz, P. Doyle, J. Theeuwes, and J. Lane, Eds. Amsterdam, The Netherlands: North Holland, 2001, pp. 111–134.

[8] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," IEEE Trans. Knowl. Data Eng., vol. 14, no. 1, pp. 189–201, Jan./Feb. 2002.

[9] J. Domingo-Ferrer and U. Gonzalez-Nicolas, "Hybrid microdata using microaggregation," Inf. Sci., vol. 180, no. 15, pp. 2834–2844, 2010.

[10] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and hetero- geneous k-anonymity through microaggregation," Data Mining Knowl. Discovery, vol. 11, no. 2, pp. 195–212, 2005.

[11] M. Laszlo and S. Mukherjee, "Minimum spanning tree partitioning algorithm for microaggregation," IEEE Trans. Knowl. Data Eng., vol. 17, no. 7, pp. 902–911,

[12] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in Proc. 23rd IEEE Int. Conf. Data Eng., 2007, pp. 106–115.

[13] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubrama- niam, "l- diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, p. 3, 2007.

[14] P. Samarati, "Protecting respondents' identities in microdata release," IEEE Trans. Knowl. Data Eng., vol. 13, no. 6, pp. 1010–1027, Nov./Dec. 2001.

[15] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez and S. Mart ınez, "Enhancing data utility in differential privacy via microaggrega- tion-based k-anonymity," VLDB J. vol. 23, no. 5, pp. 771–794, 2014.

[16] W. E. Winkler, W. E. Yancey, and R. H. Creecy, "Disclosure risk assessment in perturbative microdata protection," in Inference Control in Statistical Databases. New York, NY, USA: Springer,2002, pp. 135–152.