International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2019.33150

Volume 5, Issue 4
April - 2019

Sensitivity Based Data Anonymization Model with Mixed Generalization

Esther Gachanga[1], Michael Kimwele[2] and Lawrence Nderu[3]

[1-3]Jomo Kenyatta University of Science & Technology

School of Computing and Information Technology

Nairobi Kenya

ABSTRACT

Published micro-data may contain sensitive information about individuals which should not be revealed. Anonymization approaches have been considered a possible solution to the challenge of preserving privacy while publishing data. Published datasets contain sensitive information. Different sensitive attributes may have different levels of sensitivity. This study presents a model where the anonymization of tuples is based on the level of sensitivity of the sensory attributes. The study groups sensitive attributes into highly sensitive and non-sensitive attributes. Tuples with non-sensitive attributes are anonymized. The study conducts experiments with real-life datasets and uses naïve Bayes, C4.5 and simple logistic classifiers to assess the quality of the anonymized dataset. The results from the experiments show that by using the sensitivity based approach to anonymization, the quality of anonymized datasets can be preserved.

Key Words: *Sensitive, Anonymity, Privacy, Classification, Data Publishing.*

1. INTRODUCTION

The problem of privacy-preserving data publishing has become an important issue of concern because of the increasing ability to store and share personal data about users. The traditional approach of publishing data without breaching the privacy of individuals in the data is to de-identify the records in the data by removing the identifying fields such as name, address and social security number. However, joining the de-identified data with a publicly available database (like the voters database) on quasi identifier attributes such as race, age, and zip code may enable the re-identification of individuals [1], [2]. [3] and [4] define re-identification as *"the process of attempting to determine the identities of the data subjects whose identifiers have been removed from the dataset".* According to [5] re-identification combines datasets that were meant to be kept apart, and in so doing, gain power through accretion.

Anonymization methods and techniques are commonly used to address the challenge of re-identification. These techniques enable data owners to cautiously share sensitive information while preserving privacy [6]. The basic idea with anonymization is that an attacker cannot easily analyze the sensitive attribute of a tuple, from a transformed table, and therefore cannot identify a specific individual's sensitive information Wang et.al (2016). One the most commonly used approach to data anonymization is the k-anonymity model presented by [2]. To implement anonymization the k-anonymity model utilizes generalization and suppression [7]. Suppression is performed prior to generalization to reduce the amount of generalization required to generate the k-anonymous data set. Generalization reduces the granularity of the information contained in the quasi-identifier attributes, thereby increasing the chance of several records sharing the values of the attributes [8]. The rest of this work is organized as follows; in section 2 we review related work while section 3 presents our approach and the model. In section 4 we conduct experiments. In section 5 we present the results while Section 6 presents our discussions and finally, section 7 concludes the study.

2. RELATED WORK

Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy. Several anonymity models have been proposed to protect individual's privacy for micro-data publishing [9], among them, the k anonymity model [10], the l-diversity model [11], the t-closeness models [12] and differential privacy [13].

The k anonymity model requires that there are at least k people with the same quasi-identifier attributes such that the risk of identity disclosure is reduced to $1/k$ [2]. To determine how many individuals each released tuple actually matches requires combining the released data with externally available data and analyzing other possible attacks. The primary goal of k-anonymity

is to protect the privacy of the individuals to whom the data pertains. However subject to this constraint it's important that the data remains as useful as possible [14]. The l-diversity model by [11] addresses some of the shortcomings of the k anonymity model. The l-diversity model requires each equivalence group of released table to contain at least l-well represented records. It means that every sensitive attribute in each equivalence should have at least l different values. [11] gives three interpretations of the term "well represented"; first distinct ℓ-Diversity ensures that there exist at least ℓ- distinct sensitive values in each equivalence class, second Entropy ℓ-diversity. The entropy of an equivalence class E is defined as;

$$Entropy\ (E) = -\sum_{s\in S} P(E,s)logP(E,s) \qquad --------1$$

Where S , is the sensitive attribute domain and $P(E,s)$ is the fraction of records in E with sensitive value s, and third Recursive (c,l)- diversity, which ensures that the most frequent values does not appear too frequently.

t-closeness principle requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table as much as is possible (i.e., the distance between the two distributions should be no more than a threshold ***t*** [12]. Ninghui et al (2007), performed an experiment on *t*-closeness. In the experiment the distance in the distributions of the attributes in a class is measured using the Earth Movers Distance (EMD). Given two distributions P and Q; $p = (p_1, p2 \ldots . pm)$ $Q = (q_1, q_2 \ldots \ldots . . q_{m)}$ the variational distance in the distribution can be defined as;

$$D[P,Q] = \sum_{i=1}^{m} \frac{1}{2} |\text{pi} - \text{qi}| \qquad ----------2$$

Where D is the distance and P,Q are the distributions.

Differential Privacy (DP) is a mathematical framework that is widely accepted for protecting data privacy. It guarantees that the distribution of query results changes only slightly due to the modification of any one tuple in the database [13]. This allows protection, even against powerful adversaries, who may know the entire database except one tuple. In [15], a formal definition of differential privacy is given as: a randomized algorithm A satisfies ϵ - differential privacy if;

$$P(A(D_1) \in S \leq e^{\varepsilon} P(A(D_2) \in S \qquad \text{---------- [3]}$$

for any set S and any pairs of databases D_1, D_2 where D_1 can be obtained D_2 by either adding or removing one tuple or by changing the value of exactly one tuple.

2.1 Data Anonymization vs. Utility

Anonymization causes a decline of data utility. The main challenge with any anonymization process is to balance between the utility and the privacy of the data. Hiding data reduces the utility of the data, while disclosing the data reduces privacy. During the anonymization process there is a tradeoff between privacy and information loss [16].

Privacy preservation requires that data is protected with minimum impact on its accuracy and utility [17]. [18] argue that utility may be achieved at the expense of runtime since the anonymization process is a one-time process.

According to [17] excessive data anonymization can make the published data less useful and so it's important to measure the utility of anonymized data. During the data privacy process, the utility of datasets diminishes as sensitive information such as personal identifiable information (PII) is removed, transformed, or distorted to achieve confidentiality. Attaining an equilibrium between data privacy and utility, requires trade-offs, and this is further complicated by the fact that making such trade-offs also remains problematic. Even with the latest state of the art data privacy algorithms like differential privacy, confidentiality is guaranteed but at a major loss of data utility.

2.2 Sensitivity in Data Anonymization

[19] in rating the privacy preservation for multiple attributes with different sensitivity requirements present and explain that, different attributes may have different sensitivity requirements. [20] in their research controlled the frequency of the sensitive attribute to no more than α in order to achieve diversity of sensitive values in every equivalence class for an anonymized dataset.

DOI: 10.31695/IJASRE.2019.33150

[21] defined the sensitivity *S(f)* of a function *f* as the quantity inherent in f; it is not chosen by policy. S (f) is independent of the actual database. Dwork et. al (2012) continues to explain that some functions may have a low sensitivity while others may have a high sensitivity.

In the work of [22], it is well explained that the sensitivity of data tends to be connected with the potential harm of any confidentiality breach and that for a disclosure to be meaningful something has to be learned. These works further explain that personal data becomes sensitive according to its context and that the sensitivity of data can be reduced by removing the sensitive attributes. While there are several data release options available, the one you choose depends on the data you plan to release, the sensitivity of the data, and the proposed usage of the data.

2.3 Equivalence Class

An equivalence class of an anonymized table is defined to be a set of records that have the same values for the quasi-identifiers [23]. The number of equivalence classes C in anonymized data set is determined by the specified value of k in a k- anonymity model [24]. This is given by the following formulae;

$$C = M/k \quad \ldots\ldots\ldots\ldots\ldots\ldots [4]$$

Where given C is the equivalence classes, M is the number of records in a table and k is the value of k specified in k - anonymity. For instance, a table with 100 records with k specified as 3, will have 33 classes (100/3). The size of an equivalence class indicates the strength of identification protection of individuals in the class. If the number of tuples in an equivalence class is greater, it will be more difficult to re-identify individual [25].

2.4 Generalization & Suppression

Data publishers and Database experts have developed different anonymization techniques, which vary in their cost, complexity, ease of use, and robustness [5]. The k-anonymity model is a major technique to anonymizing a data set [26]. This work focuses on generalization and suppression approaches to data anonymization. Author [7] used generalization and suppression to implement k-anonymity model. Generalizing an attribute means substituting its values with corresponding values from a more general domain e.g. male and female can be generalized to person [27]. Generalization at the attribute level ensures that all values of an attribute belong to the same domain and this is achieved via a taxonomy tree [28]. Suppression involves removing data from a table so that the data is not released. This operation uses special symbolic character to replace its authentic value (e.g. *, &, #), and makes the value meaningless [7]. Increasing generalization may reduce the amount of suppression required thereby increasing the utility of data [29].

3 PROPOSED APPROACH

In this section, the study presents a sensitivity based anonymization model. The model first categorizes the different attributes in a dataset into quasi identifiers and sensitive attributes. Secondly the model identifies the distinct attributes within the sensitive attributes and groups them into two i.e. highly sensitive and less sensitive. Third the quasi identifier attributes for tuples with less sensitive attributes are generalized and finally the model is evaluated. The study uses the term mixed generalization to mean that the same attribute has been generalized to different levels within the same dataset.

3.1 The Model

The proposed model provides an outline of the various steps that one should follow in order to anonymize data while maintaining data quality and at the same time preserving privacy. The raw dataset is preprocessed. Preprocessing involves removing tuples with missing values. The next step involves categorizing the attributes in the dataset into; quasi identifiers (QIDs) and the sensitive attributes (SAs). The model then establishes the distinct attributes within the sensitive attribute and groups these distinct attributes into highly sensitive and less sensitive. The tuples with the less sensitive attributes are placed into table T1 while those with highly sensitive attributes are placed into table T2. The next step involves generalizing the QIDs in table T1. For table T2 only the attribute age is generalized. The two tables (table T1 and T2) are then merged together to form table T*. The model uses the k-anonymity and *l-diversity* model together. Finally the model tests the anonymized dataset for quality.

To implement the model experiments were conducted. An anonymization tool built on java platform was used to conduct the experiments. The tool provided a graphical user interface for the entry of the data and a display for the anonymized data. Figure 1 presents our model;

DOI: 10.31695/IJASRE.2019.33150

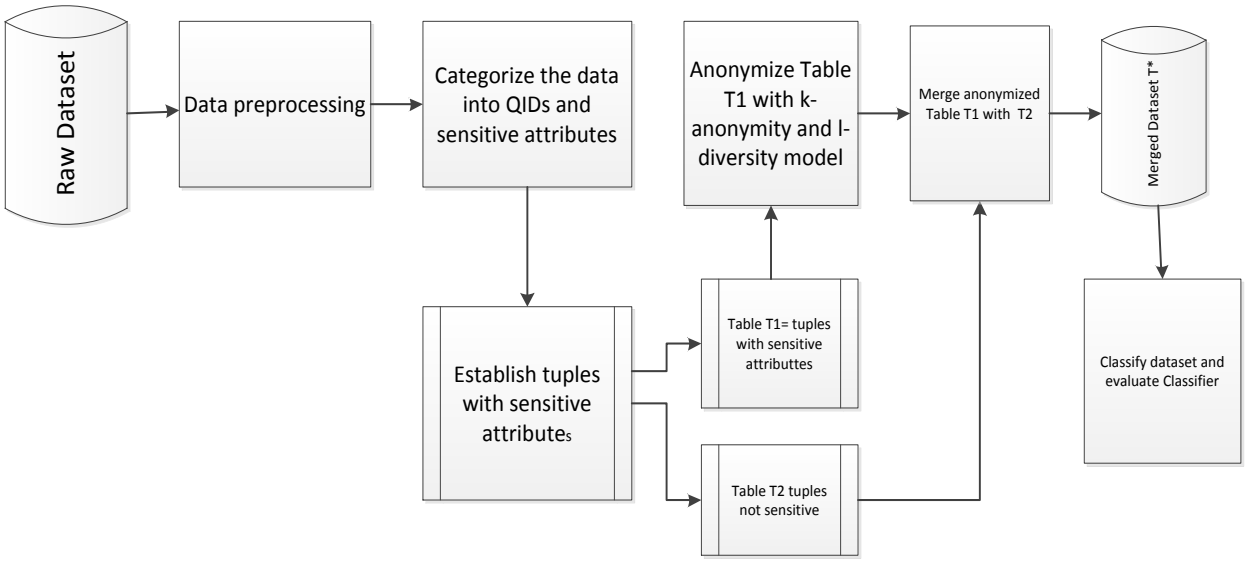

Figure 1: Sensitivity Based Data anonymization model.

4. EXPERIMENT

The study adopted the adult dataset [30]. Data cleaning was done by removing tuples with missing values. The dataset contains 30162 tuples after cleaning. Eight attributes among them; Age, Work class, Education, Marital status, Occupation, Relationship, Race and sex were used. The Dataset is conceptually organized as a table of rows (or records) and columns (or fields). Each row is termed a *tuple*. Tuples within a table are not necessarily unique. Each column is called an *attribute* and denotes a set of possible values within its domain. Each tuple is assumed to be specific to a person and no two tuples pertain to the same person. Occupation was used as the sensitive attribute while the other seven attributes were the quasi identifier attributes. The quasi identifiers of table T1 were generalized to enable anonymization. only the age attribute was generalized in table T2. All the experiments were conducted on java platform. We used the approach of [31] and [32]. For our experiments, among the wide variety of available anonymization methods, we used *ARX* because it is an open source tool that is readily available, combines most of the data anonymization algorithms, is suitable for anonymizing a wide variety of datasets, is an easy to use tool and its the simplicity of the privacy definition that relies only on *k*.

In our approach to data publishing, only tuples with values that are considered sensitive are generalization. The purpose for this is to reduce the amount of data distortion and thus increase the utility of published data. For the sensitive attribute, occupation, the values that were considered sensitive were; Protective-service, Farming-fishing, Priv-house-service and Armed-Forces. From the available dataset (a table T), the tuples with the sensitive attribute values are grouped together into a table T1 and separated/ removed from the table T. The remaining tuples are grouped together into a table T2. The table T2 is given by $T2 = T - T1$. The quasi identifiers for table $T1$ are generalized to give the table$T1^*$. For table $T2$ only the quasi identifier age is generalized to give table$T2^*$. For the generalization of table $T1$ &$T2$ we utilized the k- anonymity model with k=5. The l – diversity model is used with the parameter l = 3 for the sensitive attribute in table $T1$. Table $T1^*$ is then merged with table $T2^*$ to give table T^* the generalized table for publication. We utilized the k -anonymity and l – diversity model in our experiments.

5. Results

In this section the performance of the different classifiers build from the merged dataset (Table T*) was compared. The research used three classifiers namely, Naïve bayes, J48 and the simple logistic classifiers. The performance of these classifiers was compared with a classifier built from the original dataset which had not been anonymized. The results of the three classifiers are summarized in Table 1.

Table 1: Classifier Performance for Sensitivity Based Anonymization

	Naïve bayes	J48	Simple Logistic
original Data	82.8526	84.1688	83.9202
Table T k=5	82.5476	83.7975	83.8539
TableT with k=5 L=3	82.8957	84.1423	83.9832
T1&T2 with k=5	82.0801	83.5654	83.9964
Table T1&T2 with k=5 L=3	82.836	84.0627	84.2915

DOI: 10.31695/IJASRE.2019.33150

We observed that the naïve bayes classifier, built from the original dataset after data cleaning and before performing any transformations and modifications on it, the classifier accuracy was 82.8526%. This classifier was built before the anonymization process when the utility of the data is at its maximum 100%. When the data was anonymized with the value of k=5, the performance of the classifier was 82.5476% while introducing the *l-diversity* model with the parameter *l*=3 recorded a performance of 82.8957%. With the sensitivity based anonymization after merging table T1 and T2 with the value of k=5 the classifier accuracy was 82.0801% while by introducing the *l-diversity* model the classifier performance was 82.836%.

With the J48, with the original data the classifier accuracy was 84.1688%. This classifier was built before the anonymization process when the utility of the data is at its maximum 100%. When we anonymized the data with the value of k=5, the performance of the classifier is 83.7975% while introducing the *l-diversity* model with the parameter *l*=3 recorded a performance of 84.1423%. With the sensitivity based anonymization after merging table T1 and T2 with the value of k=5 the classifier accuracy was 83.5654% while by introducing the *l-diversity* model the classifier performance was 84.0627%.

The Simple Logistic, with the original data the classifier accuracy was 83.9202%. This classifier was built before the anonymization process when the utility of the data is at its maximum 100%. When we anonymized the data with the value of k=5, the performance of the classifier is 83.8539% while introducing the *l-diversity* model with the parameter *l*=3 recorded a performance of 83.9832%. With the sensitivity based anonymization after merging table T1 and T2 with the value of k=5 the classifier accuracy was 83.9964% while by introducing the *l-diversity* model the classifier performance was 84.2915%.

6. DISCUSSIONS

A summary of the results for the performance of the classifiers in Table 1 is presented in figure 2.

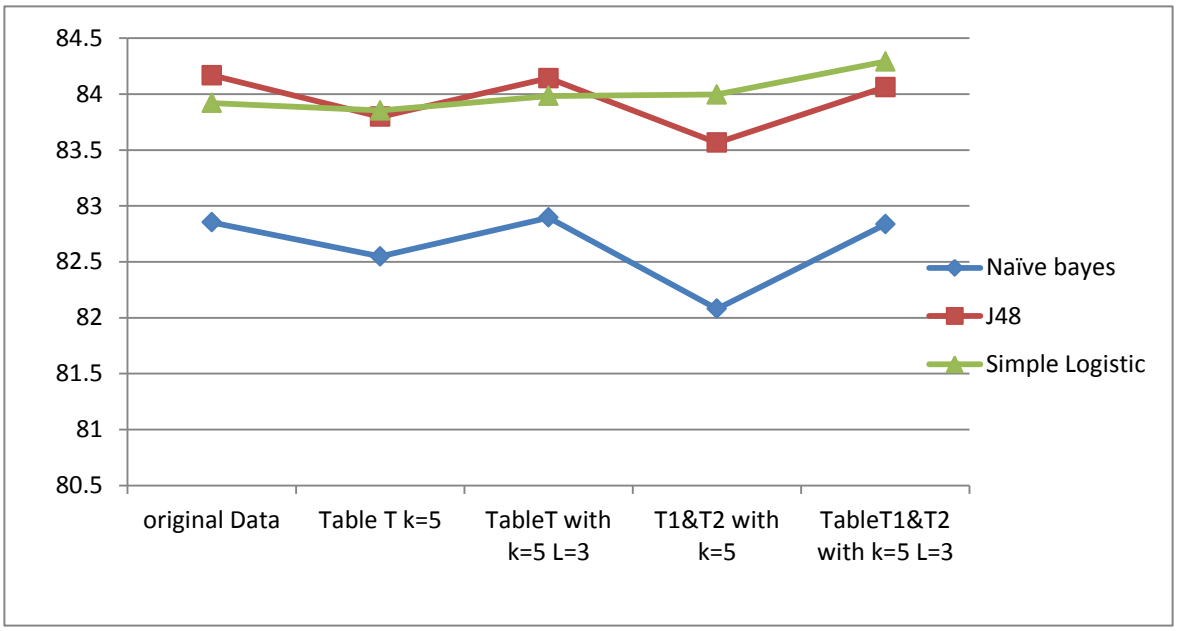

Figure 2: Classifier Performance for Sensitivity Based Anonymization

From figure 2 it was noted that with the Naïve bayes classifier the performance is highest for the dataset of Table T with k=5 and *l=3* while it was lowest with the dataset of Table T1&T2 with k=5. The J48 classifier performed best with the original dataset before anonymization while the classifiers lowest accuracy was with the dataset T1&T2 with k=5. The simple logistic classifier had highest performance with the dataset T1&T2 with k=5 *l*=3. The lowest performance for this classifier was noted with the dataset for Table T k=5.

7. CONCLUSION

Anonymization is a safe and effective method for data privacy protection, which can effectively balance the relationship between the efficiency and the security of the data. In this research we have implemented sensitivity based approach to anonymization and demonstrated that we can preserve the quality of data after the anonymization process. However in the face of an ever changing information environment and more data being made publically available, data anonymization remains to be/ is a complex issue.

DOI: 10.31695/IJASRE.2019.33150

REFERENCES

[1] X. Sun, H. Wang, T. M. Truta, J. Li, and P. Li, "(p+, α)-sensitive k-anonymity: A new enhanced privacy protection model," *Proc. - 2008 IEEE 8th Int. Conf. Comput. Inf. Technol. CIT 2008*, pp. 59–64, 2008.

[2] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002.

[3] S. L. Garfinkel, "NISTIR 8053 De - Identification of Personal Information NISTIR 8053 De - Identification of Personal Information," 2015.

[4] I. S. Rubinstein and W. Hartzog, "Anonymization and Risk," *Washingt. Law Rev.*, vol. 91, no. 2, pp. 1–54, 2016.

[5] P. Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Rev.*, vol. 57, no. 6, pp. 1701–1777, 2010.

[6] M. Al-Zobbi, S. Shahrestani, and C. Ruan, "Towards optimal sensitivity-based anonymization for big data," *2017 27th Int. Telecommun. Networks Appl. Conf. ITNAC 2017*, vol. 2017–Janua, pp. 1–6, 2017.

[7] L. Sweeney, "Achieving K -Anonymity Privacy Protection Using Generalization And Suppression," *Int. J. Uncertainty, Fuzziness Knowledge-based Syst.*, vol. 10, no. 5, pp. 1–18, 2002.

[8] J. Domingo-ferrer, S. David, and S. Hajian, *Privacy in a Digital, Networked World*. switzerland: Springer International Publishing, 2015.

[9] G. Yang, J. Li, S. Zhang, and L. Yu, "An enhanced l-diversity privacy preservation," in *Proceedings - 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2013*, 2013, no. 61170060, pp. 1115–1120.

[10] P. Samarati and L. Sweeney, "Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion.," *Proc IEEE Symp. Res. Secur. Priv.*, pp. 384–393, 1998.

[11] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "*L* -diversity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, p. 3–es, 2007.

[12] L. Ninghui, L. Tiancheng, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and ℓ-diversity," in *Proceedings - International Conference on Data Engineering*, 2007, no. 2, pp. 106–115.

[13] C. Dwork, "Differential Privacy," *Proc. Int. Colloq. Autom. Lang. Program. Part II*, pp. 1–12, 2006.

[14] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," *Proc. - Int. Conf. Data Eng.*, vol. 2006, p. 25, 2006.

[15] Y. Yang, Z. Zhang, G. Miklau, M. Winslett, and X. Xiao, "Differential privacy in data publication and analysis," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of data (SIGMOD)*, 2012, pp. 601–605.

[16] P. Bhaladhare and D. Jinwala, "A Sensitive Attribute Based Clustering Method for k-Anonymization," *Springer-Verlag Berlin Heidelb.*, pp. 163–170, 2012.

[17] T. Basso, R. Matsunaga, R. Moraes, and N. Antunes, "Challenges on anonymity, privacy, and big data," in *Proceedings - 7th Latin-American Symposium on Dependable Computing, LADC 2016*, 2016, pp. 164–171.

[18] K. Doka, M. Xue, D. Tsoumakos, P. Karras, A. Cuzzocrea, and N. Koziris, "Heterogeneous k -Anonymization with High Utility," pp. 1886–1890, 2015.

[19] J. Liu, J. Luo, and J. Z. Huang, "Rating: Privacy preservation for multiple attributes with different sensitivity requirements," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2011, pp. 666–673.

[20] H. Jin, S. Liu, and S. Ju, "Based on l -Coverage *," *Springer-Verlag Berlin Heidelb.*, no. 60773049, pp. 319–326, 2011.

[21] and A. S. Cynthia Dwork, Frank McSherry, Kobbi Nissim, "Calibrating Data to Sensitivity in Private Data Analysis," pp. 1–20, 2012.

[22] E. M. K. O. and C. T. Mark Elliot, "The Anonymisation Decision-Making Framework," *Univ. Manchester*, 2016.

[23] a. Sunitha, K. Venkata Subba Reddy, and B. Vijayakumar, "A Privacy Measure for Data Disclosure to Publish Micro Data using (N,T) - Closeness," *Int. J. Comput. Appl.*, vol. 51, no. 6, pp. 22–28, 2012.

DOI: 10.31695/IJASRE.2019.33150

[24] C. Chiu and C. Tsai, "A k -Anonymity Clustering Method for Effective Data," pp. 89–99, 2007.

[25] R. C. Wong, J. Li, A. W. Fu, and K. Wang, "(A, K)-Anonymity: an Enhanced K-Anonymity Model for Privacy Preserving Data Publishing," *Int. Conf. Knowl. Discov. Data Min. SIGKDD*, pp. 754–759, 2006.

[26] J. Li, R. C. Wong, A. W. Fu, and J. Pei, "Achieving \emphk-Anonymity by Clustering in Attribute Hierarchical Structures," *Data Warehous. Knowl. Discov.*, vol. 4081, pp. 405–416, 2006.

[27] F. Kohlmayer, F. Prasser, C. Eckert, and K. A. Kuhn, "A flexible approach to distributed data anonymization," *J. Biomed. Inform.*, vol. 50, pp. 62–76, 2014.

[28] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, 2002, p. 279.

[29] F. Prasser, R. Bild, J. Eicher, H. Spengler, and K. A. Kuhn, "Lightning : Utility-Driven Anonymization of High-Dimensional Data," *Trans. Data Priv.*, vol. 9, pp. 161–185, 2016.

[30] E. Dheeru, Dua , Karra Taniskidou, "{UCI} Machine Learning Repository," 2017.

[31] F. Prasser and F. Kohlmayer, "Medical Data Privacy Handbook," *Springer Int. Publ. Switz. 2015*, 2015.

[32] F. Prasser, J. Eicher, R. Bild, H. Spengler, and K. A. Kuhn, "A Tool for Optimizing De-identified Health Data for Use in Statistical Classification," *Proc. - IEEE Symp. Comput. Med. Syst.*, vol. 2017–June, 2017.

DOI: 10.31695/IJASRE.2019.33150