Indonesian Journal of Electrical Engineering and Computer Science
Vol. 16, No. 3, December 2019, pp. 1439~1446
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i3.pp1439-1446

Evaluation of proposed amalgamated anonymization approach

Deepak Narula, Pardeep Kumar, Shuchita Upadhyaya
Department of Computer Science & Applications, KU, India

Article Info

Article history:

Received Jan 15, 2019
Revised Jun 1, 2019
Accepted Jun 26, 2019

Keywords:

Anonymization
ATUS
CUPS
PPDP

ABSTRACT

In the current scenario of modern era, providing security to an individual is always a matter of concern when a huge volume of electronic data is gathering daily. Now providing security to the gathered data is not only a matter of concern but also remains a notable topic of research. The concept of Privacy Preserving Data Publishing (PPDP) defines accessing the published data without disclosing the non required information about an individual. Hence PPDP faces the problem of publishing useful data while keeping the privacy about sensitive information about an individual. A variety of techniques for anonymization has been found in literature, but suffers from different kind of problems in terms of data information loss, discernibility and average equivalence class size. This paper proposes amalgamated approach along with its verification with respect to information loss, value of discernibility and the value of average equivalence class size metric. The result have been found encouraging as compared to existing *k*-anonymity based algorithms such as Datafly, Mondrian and Incognito on various publically available datasets.

Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Deepak Narula,
Department of Computer Science & Applications,
KU, Kurukshetra, Haryana, India.
Email: dnarula123@yahoo.com

1. INTRODUCTION

In present era of information technology huge amount of data is collected day by day via different means such as online and offline. This huge collection will become a matter of concern when the matter of security occurs. As privacy to an individual is a crucial matter of concern, therefore PPDP is always a matter of notable research. A variety of techniques have been suggested in literature for anonymization. *k*-anonymity model was suggested by Sweeney and was the one basis of privacy protection model [1]. Kristen L. et al. named Mondrian multidimensional *k*-anonymity [2], It works by partitioning the domain space iteratively in to various regions where each of the region have to satisfy the condition of *k*-anonymity. Another algorithm of k-anonymization named Incognito was proposed by K. lefevre et al. [3]. This algorithm works on the approach of full domain generalization and based on the concept of single dimensional theory. It works by building a lattice and traverses the lattice in bottom up breath first search manner and returns the anonymized table. Sapana Anant patil et al. [4] made a comparative study of privacy preservation technique in data publishing where as Ram Mohan Rao P et al. [5] gave a study on privacy preservation technique. Erick et al. [6] explained the concept of improving clusters using fuzzy based approach. The concept of two level clustering approach is proposed to improve partition method using *k*-mean approach. Muhammad et al. [7] introduced the concept of generating clusters using fuzzy based approach. Jasmin Ilyani et al. [8] have described various methods for providing security to data as to restrict the unauthorized person to access the data along with advantages and disadvantages of the proposed technique. Manisha Sharma et al. [9] proposed approach for privacy preserving that allows publishing the data while retaining the seclusion of sensitive information about an individual. Kartik Patel et al. [10] given an approach for privacy preservation of data using randomization model and *k*-anonymity. The approach works by selecting key, quasi and sensitive attributes from given data set then out of the selected sensitive values and transfer tuples with the most sensitive values to another table followed by applying the process of *k*-anonymization. Fung et al. [11, 12] proposed an approach for classification of data using with an aim is to determine *k*-anonymization factor. The approach for classification is based on two observations. First observation is information specific to individuals while the second one is concerned with utility of data during classification. Thanveer et al. [13] described a novel holistic approach for achieving maximum privacy using fuzzy set with objectives of maintaining privacy preservation while revealing useful information for numerical and categorical values. Manikandan G. et al. [14] experimentally shows that the original data will be distorted when fuzzy logic is applied and given a complete analysis on different fuzzy based member functions. Katsuhiro Honda et al. [15] proposed another variant of *k*-anonymization using fuzzy based approach. The proposed method for anonymization is applied for collaborative filtering and does the estimation of applicability of unevaluated items. B. Karthikeyan et al. [16] proposed an approach for privacy preserving of sensitive data using the fuzzy logic. In this clustering on data set is made first, then noise is added to numeric data using fuzzy membership function that causes distorted data. A systematic comparison and evaluation of various *k*-anonymization techniques has been given by Vanessa et al. [17]. Deepak et al. [18-20] have given a complete analysis of information loss, discernibility and average equivalent class size metrics considering the characteristics of attributes on different data sets. Time to time various methods have been deployed to anonymize the data, out of these methods, *k*-anonymization is one of the fundamental model of anonymization and basis for the others. Algorithms such as Datafly, Mondrian, Incognito are the anonymization algorithms based on the concept of *k*-anonymization. Moreover, when these algorithms are applied on different datasets for anonymizing the data and after anonymization data utility metrics are applied to calculate various utility factors such as information loss [21], value of discernibility [22] and the value of average equivalence class size [23] their performance varies with the characteristics of data sets and these algorithms are not always showing consistent results.

1.1. Problem Formulation

This paper is an effort to perform verification of proposed approach with three *k*-anonymity algorithms such as Datafly, Mondrian and Incognito. The input is taken from three publically available data sets [24] and output will be anonymized data. The amalgamated approach is based on fuzzy approach and shuffling of tuples, which is why this approach can be applied to any type of data. Moreover, a proper verification of information loss [21], discernibility [22] value and the value of average equivalence class size [23] metric using amalgamated method have been given whereas the amalgamated method and metrics have been implemented by researcher.

The proposed algorithm has been implemented in Python as Python is having a rich set of libraries. Then UTD toolkit is to be used in the last step of the algorithm. In this toolkit the various parameters w.r.t. k-anonymization such as the name of algorithm to be applied e.g. Datafly, Mondrain, etc. and value of k, etc. will be given. On the basis of these parameters the toolkit provides the anonymized data.

1.2. Proposed Amalgamated Algorithm

The proposed amalgamated algorithm is a combination of three different approaches, based on shuffling the records, generation of fuzzy values for sensitive attribute and anonymizing only high sensitive records using *k*-anonymization. The main reason behind using shuffling of records is to distinguish the pattern of tuples as compared to original data set. Moreover, fuzzy is applied to sensitive attribute to preserve the privacy of an individual.

```
        Algorithm: Amalgamated_anonymization($DS_0$, $A_U$, $A_Q$, $A_S$ )
{                                  // $DS_0$=Original DataSet
                                              //$A_U$: Unique Attribute
// $A_Q$ : Quasi Attribute
// $A_S$: Sensitive Attribute
1.      Generate $DS_1$  where $DS_1 = DS_0 - \{A_U\}$ // {$DS_1$ :  Data set without  unique attributes}
2.      Generate $DS_2$ where $DS_2$= Sort($DS_1$,As) // {DS2: Sorted dataset on sensitive attribute}
3.      Generate $AS_{W=}$ Generate_Weight($A_S$) // {Generate weight for sensitive attribute}
4.      $DS_3$=SFuzzy($AS_W$) // { $DS_{3:}$ Data set containing fuzzy values for sensitive attribute.}
5.      Split($DS_{3\,,}T_H$)  to obtain $DS_H$ and $DS_L$  where $DS_H=\{X: 0<X<=T_H\}$, $DS_L=\{ X:  X>T_H\}$
6.       K_ANONYIZE($DS_H$)// Apply the process of k-anonymity only on High Sensitive dataset
7.      Obtain $DS_4$= Merge($DS_H$, $DS_L$) // Merge Low and high Sensitive data
return $DS_4$
}
```

2. VERIFICATION OF PROPOSED ALGORITHM

Verification of proposed algorithm is done by applying it on three publically available data sets [24]. Moreover, a systematic comparison have been made with existing *k*-anonymity algorithms such as Datafly [1], Mondrian [2] and Incognito [3] and proposed amalgamated algorithm in terms of data utility metrics such as information loss, discernibility and the value of average equivalence class size. For anonymizing the data set UTD [25] toolkit have been used.

3. EXPERIMENTAL ANALYSIS

3.1. Adult Data Set

Initially adult data set is taken to determine the amount of information loss, value of discernibility and the value of average equivalence class size metrics. The assessment was done on 5411 tuples with nine attributes. This process of assessment was done after discarding the tuples with blank values from the original data set. The attributes considered for this data set are:

Adult = {Age, Sex, Race, Marital Status, Education, State, Qualification, Designation, Salary}

3.1.1. Information Loss

To determine the value of information loss, the attribute Salary is considered to be sensitive attribute and process of anonymization is applied on quasi attributes such as Age, Marital Status, and Sex. After anonymizing the data set using existing *k*-anonymity algorithms and using proposed amalgamated algorithm, general information loss have been calculated. The results are shown in Figure 1.

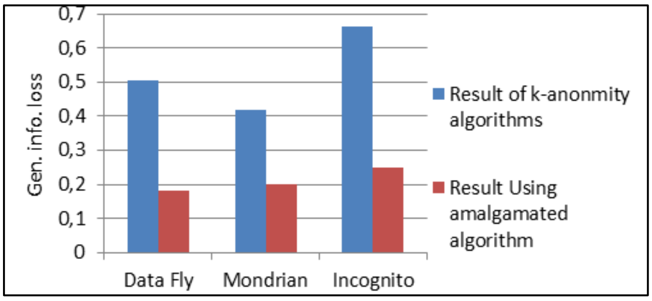

Figure 1. Comparative analysis of existing algorithms and proposed amalgamated algorithm on ADULT data set for general Information loss

It has been observed from Figure 1 that the result obtained in terms of information loss using proposed amalgamated algorithm is comparable very less as using existing algorithms.

3.1.2. Discernibility

To determine the value of discernibility, again the same parameters are considered as in case of general information. The only difference being that instead of general information loss the value of discernibility is calculated and is shown in Figure 2.

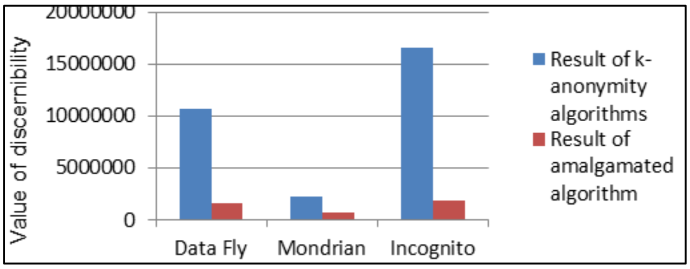

Figure 2. Comparative analysis of existing algorithms and amalgamated algorithm on ADULT data set for values of the discernibility

It has been interpreted from Figure 2 that the value of discernibility using proposed amalgamated algorithm is comparable lesser as compared with existing algorithms.

3.1.3. Average Equivalence Class Size

Similarly the average equivalence class size is evaluated and is shown in Figure 3.

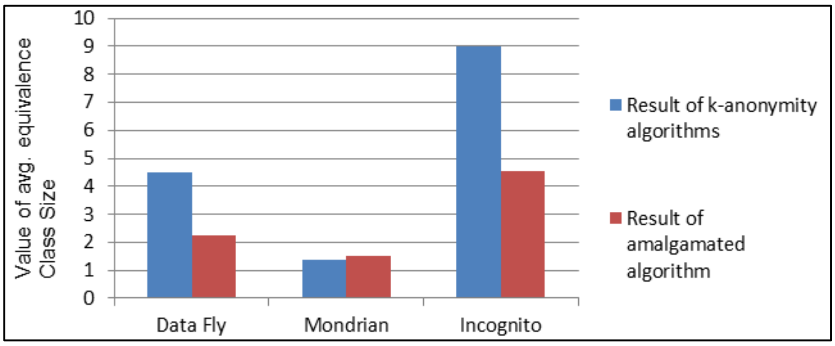

Figure 3. Comparative analysis of existing algorithms and amalgamated algorithm on ADULT data set for the values of average equivalence class size

From Figure 3, it has been observed that result obtained for the value of average equivalence class size metric using proposed approach is comparable good as in case of Datafly and Incognito algorithm. Whereas in case of Mondrian result obtained from existing and proposed are marginally comparable.

3.2. Cups Data Set:

Next CUPS data set is used for the purpose of evaluation. After eliminating the records containing NULL values the total number of attributes taken is five whereas the total number of tuples used with this data set are 62414. The attributes considered in this data set are:

CUPS = {Zip Code, Age, Sex, Salary, Qualification}

3.2.1. Information Loss

Now to determine the value of information loss Qualification is considered to be sensitive attribute and process of anonymization is applied on quasi attribute Age, Zip Code, Sex. After anonymizing the data set using existing *k*-anonymity algorithms and by using proposed amalgamated algorithm, general information loss has been calculated. The results are shown in Figure 4.

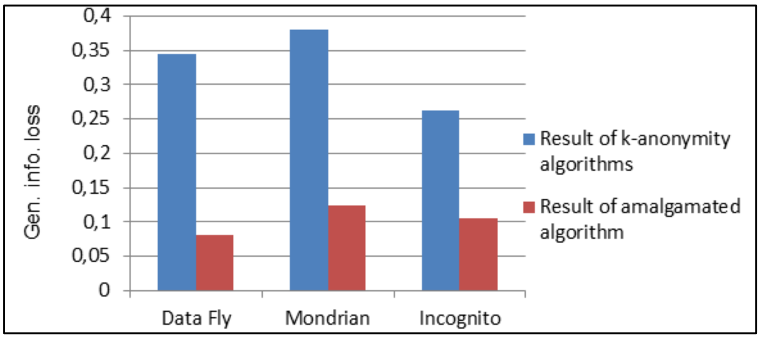

Figure 4. Comparative analysis of existing algorithms and amalgamated algorithm on CUPS data set for general information loss

It has been observed from Figure 4 that the result obtained in terms of information loss using proposed amalgamated algorithm is comparable very less as using existing algorithms.

3.2.2. Discernibility

To determine the value of discernibility, similar parameters are considered as in case of general information loss. After performing the anonymization process on the data set using existing *k*-anonymity algorithms and using proposed amalgamated algorithm, value of discernibility have been calculated, the results are shown in Figure 5.

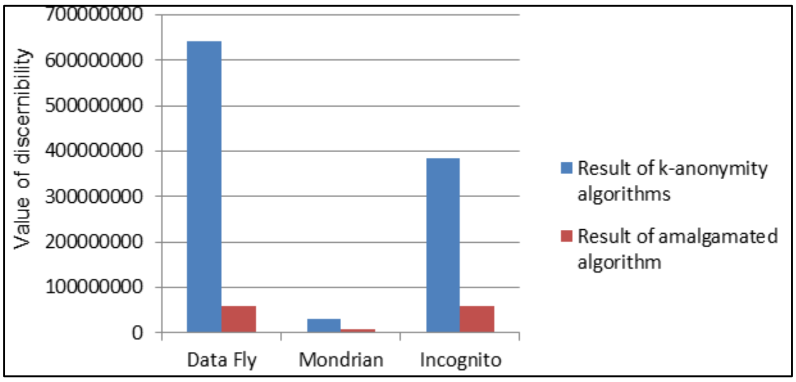

Figure 5. Comparative analysis of existing algorithms and amalgamated algorithm on CUPS data set for the values of discernibility

It has been interpreted from Figure 5 that the value of discernibility using proposed is comparable lesser as compared with existing algorithms.

3.2.3. Average Equivalence Class Size

Similarly parameters have been considered to calculate the value of average equivalence class size, the result is shown in Figure 6.

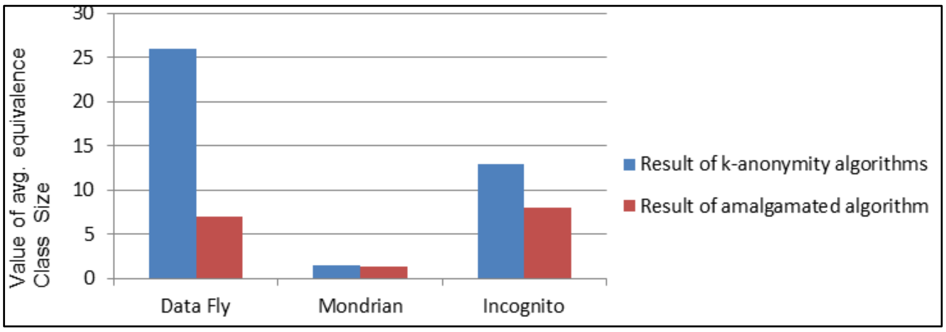

Figure 6. Comparative analysis of existing algorithms and amalgamated algorithm on CUPS data set for the values of average equivalence class size

From Figure 6, it has been observed that result obtained for the value of average equivalence class size metric using proposed is comparable good in case of Datafly and Incognito algorithm. Whereas, in case of Mondrian result obtained from existing and proposed are marginally comparable.

3.3. American Time Use Survey (ATUS) Data Set:

This is the next data set used for the purpose of evaluation. In this data set total numbers of tuples taken are 56663 with five attributes after deleting the records containing NULL values. The attributes considered in this data set are:

ATUS = {Age, Region, Race, Marital Status, Qualification}

3.3.1. Information Loss

To deduce the value of information loss Qualification is considered to be sensitive attribute and process of anonymization is applied on quasi attribute Age, Race, Marital Status. After anonymizing the data set using existing *k*-anonymity algorithms and using proposed amalgamated algorithm, general information loss have been calculated. The results are shown in Figure 7.

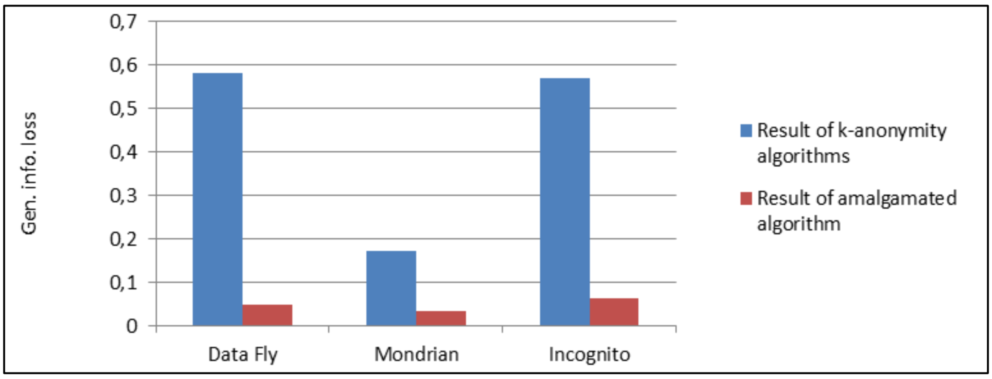

Figure 7. Comparative analysis of existing algorithms and amalgamated algorithm on ATUS data set for general information loss

It has been observed from Figure 7 that the result obtained in terms of information loss using proposed amalgamated algorithm is comparable very less as using existing algorithms.

3.3.2. Discernibility

To determine the value of discernibility, same parameters have been taken as in case of general information loss, the results are shown in Figure 8.

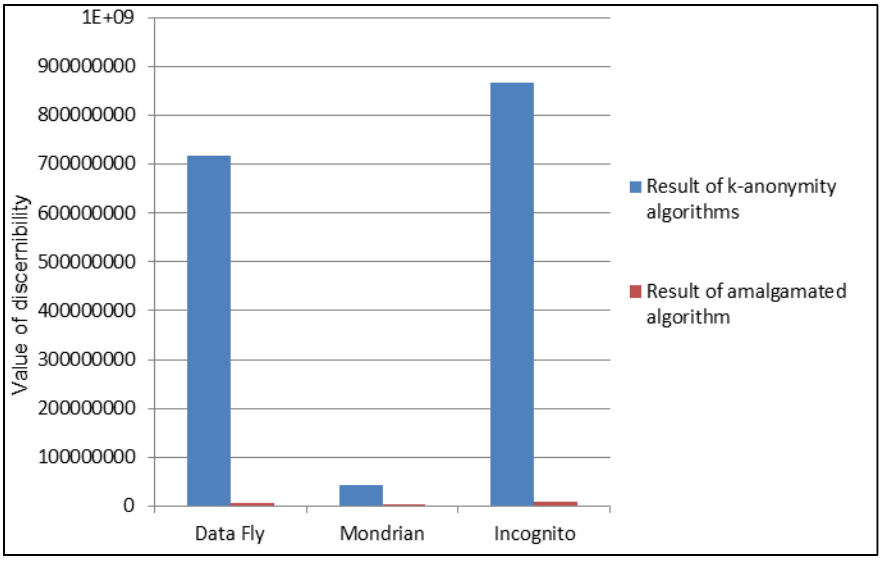

Figure 8. Comparative analysis of existing algorithms and amalgamated algorithm on ATUS data set for the values of discernibility

It has been interpreted from Figure 8 that the value of discernibility using proposed approach is comparable lesser as compared with existing algorithms.

3.3.3. Average Equivalence Class Size

On the basis of similar parameters, the value of average equivalence size has been calculated. The results are shown using Figure 9.

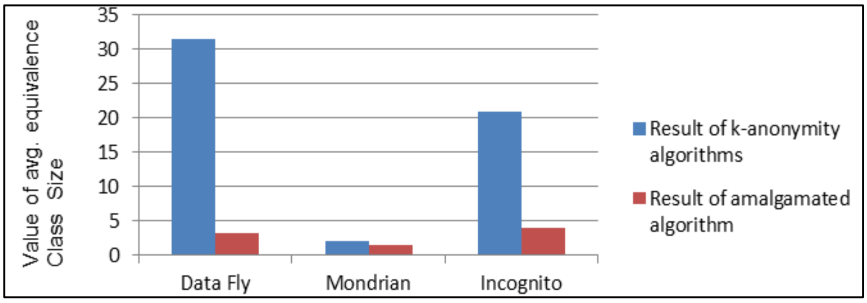

Figure 9. Comparative analysis of existing algorithms and amalgamated algorithm on ATUS data set for the values of average equivalence class size

From Figure 9 it has been observed that result obtained for the value of average equivalence class size metric using proposed amalgamated algorithm is comparable good in case of Datafly and Incognito algorithm whereas in case of Mondrian result obtained from existing and proposed are marginally comparable.

4. CONCLUSION AND FUTURE WORK

The results obtained for various data utility metrics using new proposed amalgamated algorithm for data anonymization are better as compared with existing *k*-anonymity algorithms such as Datafly, Mondrian and Incognito w.r.t. various metrics such as general information loss, discernibility and average equivalence class size. On the basis of these results, it is concluded that the proposed approach can be applied in the field of privacy protection data publishing for anonymizing any type of data sets. However, in future work can be extended in the direction of applying and verifying the proposed approach on various other data sets with varying characteristics.

REFERENCES

[1] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, vol. 10, pp. 571, 2002.
[2] K. LeFevre and D. J. DeWitt, “Mondrian Multidimensional K-Anonymity,” *Proceeding of 22nd International Conference on Data Engineering, ICDE’06*, pp. 25, 2006.
[3] K. LeFevre, et al., “Incognito: Efficient Full-Domain K-Anonymity,” *SIGMOD 2005, Baltimore, Maryland, USA Copyright 2005 ACM*, 2005.
[4] S. A. Patil and A. Banubakod, “Comparative Analysis of Privacy Preserving Techniques in Distributed Database,” *International Journal of Science and Research (IJSR)*, vol. 4, 2015.
[5] R. M. Rao P., “Comparative study of Privacy Preservation in Data Analytics,” *International Journal of Innovation in Engineering and Technology*, vol. 7, 2016.
[6] E. A. Lisangan, et al., “Two Level Clustering for Quality Improvement using Fuzzy Subtractive Clustering and Self-Organizing Map,” *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 15, pp. 373-80, 2015.
[7] M. A. Shaari, et al., “Forecasting Drought Using Modified Empirical Wavelet Transform-ARIMA with Fuzzy C-means Clustering,” *Indonesian Journal of Electrical Engineering and Computer Sc.*, vol. 11, pp. 1152-1161, 2018.
[8] J. I. Ahmad, et al., “Analysis Review on Public Key Cryptography Algorithms,” *Indonesian Journal of Electrical Engineering and Computer Sc.*, vol. 12, pp. 447-454, 2018.
[9] M. Sharma, et al., “An Efficient Approach for Privacy Preserving in Data Mining,” *International Conference on Signal Propagation and Computer Technology (ICSPCT 2014)*, 2014.
[10] K. Patel and T. Patel, “Privacy Preservation of Data in Data Mining using *k*-Anonymity and Randomization Method,” *International Journal for Innovative Research in Science and Technology*, vol. 2, 2016.
[11] B. C. M. Fung, et al., “Top-down Specialization for Information and Privacy Preservation,” pp. 205-216, 2005.
[12] B. Fung, et al., “Anonymizing Classification Data for Privacy Preservation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 711-725, 2007.
[13] G. T. Jahan, et al., “Data Perturbation using Fuzzy Based Logic to Preserve Privacy,” *Third International Conference on Computational Intelligence and Information Technology (CIIT 2013), Mumbai, India*, 2013.

[14] G. Manikandan, et al., "Survey on the Use of Fuzzy Membership Functions to Ensure Data Privacy," *Research Journal of Pharmaceutical Biological and Chemical Sciences*, vol. 7, pp. 344-348, 2016.
[15] K. Honda, et al., "A Fuzzy Variant of *k*-Member Clustering for Collaborative Filtering with Data Anonymization," *IEEE International Conference on Fuzzy Systems, Brisbane, QLD, Australia*, 2012.
[16] B. Karthikeyan, et al., "A fuzzy Based Approach for Privacy Preserving Clustering," *Journal of Theoretical and Applied Information Technology*, vol. 32, 2011.
[17] V. A. Rivera and P. McDonagh, "A Systematic Comparison and Evaluation of k-anonymization algorithms for practitioners," *Transactions on data privacy*, vol. 7 pp. 337-378, 2014.
[18] D. Narula, et al., "Performance Evaluation of k-Anonymization Algorithms for Generalized Information loss," *International Journal of Control Theory and Applications*, vol. 9, pp. 227-235, 2016.
[19] D. Narula, et al., "Performance Interpretation of k-Anonymization Algorithms for Discernibility Metric," *International Journal of Computer Science and Engneering*, vol. 5, pp. 74-78, 2017.
[20] D. Narula, et al., "Performance Explanation Of k-Anonymization Algorithms for Average Class Partitioning Metric," *International Journal of Advance research in Computer Science*, vol. 9, pp. 700-703, 2018.
[21] M. E. Nergiz, and C. Clifton, "Thoughts on k-Anonymization," *Data and Knowledge Engineering*, vol. 63, pp. 622-645, 2007.
[22] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," *Proceedings of the 21st International Conference on Data Engineering, ICDE '05*, pp. 217-228, 2005.
[23] K. LeFevre and D. J. DeWitt, "Workload-Aware Anonymization," *KDD'06, Philadelphia, Pennsylvania, USA. Copyright 2006 ACM*, 2006.
[24] Data Source. https://drive.google.com/open?id=0B1QMEQlbBZ9zMy1LU0FEaXprem8
[25] UTD Anonymization Toolbox. http://cs.utdallas.edu/dspl/cgi-bin/toolbox/