ISSN 2347 - 3983

Volume 8. No. 10, October 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter978102020.pdf

https://doi.org/10.30534/ijeter/2020/978102020

Achieving Quality of Anonymity and Data Privacy on released data of users by avoiding Data breaching using multidimensional k-Anonymity approach on large datasets

Abhijit J.Patankar[1], Dr.KshamaV.Kulhalli[2], Dr.S.Kotrappa[3]

[1]Research Scholar,Visveswaraya Technological University,Belagavi, Karnataka,India,
abhijitpatankarmail@gmail.com

[2]Dean I I I, Ex-Principal,Professor in CSE Department,D Y Patil College of Engineering & Technology
KasabaBawada, Kolhapur 416006,India, kvkulhalli@gmail.com

[3]Professor,CSE Dept,KLE's, Dr.M.S.Sheshgiri College of Engineering and Technology,
Belagavi,VTU
Karnataka, India, kotrappa06@gmail.com

ABSTRACT

In today's world data is collected and shared in large volume all public agencies like Insurance, Medical and Health sector etc. generates and share large volume of data in public domain but this data is released for various different purpose like for financial gain or for effective use of researchers or for the purpose to do compliance of government rules, but when such data is released publically on line then personal identity of user is on stake and any intruder can get critical personal information like salary, critical illness etc. about any individual user by linking this publically available data with other available dataset in public domain and due to these attacks on released data user's will get phone call about personal loan, treatment etc. these agencies obtained this personal confidential data by doing linking attack our proposed system successfully achieves data integrity and confidentiality while releasing data publically where multidimensional k-Anonymity approach is used to achieve anonymous datasets and to obtain equivalence classes in the datasets Nearest Neighbor strategy is used our obtained results on large volume of datasets proves that this method is doing better anonymization than traditional Encryption, Permutation or by Close-k and Mean methods using clustering approach

Key words : k-Anonymity, Close-k, Nearest Neighbour

1. INTRODUCTION

As new edge of Information technology is growing widely, there is wide requirement of data requirement and few of the organisations releasing the large volume of data for the purpose of public awareness, gaining money or for the purpose of researcher's benefit, also few of the organisations apply mining process on this released data to get some useful information which will be beneficial for them. But during this process of releasing, mining of data and gaining useful information out of this user's privacy is on the stake because from the released data some key identifiers called as Quasi Identifiers like Age, sex, zipcode should be common for individuals in the publically released records so there may be possibility of matching these key attribute values with other publically released records and if match is observed then there may be possibility of acquiring personal confidential information such as critical illnase,salary,mobile number etc. due to such attacks of linking information and already knowing some information about someone and by linking obtaining important confidential information takes place. This paper proposes K-anonymity [1] as a method for protection of data privacy during its release, during release to make anonymized dataset quasi-identifiers of published records must be same as of other k-1 records this will stop possibility of different linking attacks on the databases anonymity Sweeney L and Samarati proposed the k-anonymity approach [4][5][6], and describes algorithms. The representative task heuristic method Datafly [6] shows that k-anonymity implemented by generalization in full-domain also Incognito [7] recommended generalization by using full-domain approach where complete subset of records are taken for generalization , Reference [8] uses approach based on searching Ref [9] elaborates anonymity based on clustering approach, generally k-anonymity approach is divided in to two types top down and bottom up [25]and preference of the approach is decided by type of dataset and accordingly generalisation and suppression methods are used [28]to implement k-anonymity if generalized dataset is prepared for full domain sets it is called as full domain generalization and if some fields of data need to supress due to hiding confidential information it is called as full domain suppression these two methods are preferably used to obtained k-anonymity on data sets,

1.1 Multidimensional K-Anonymization

K-Anonymity[13][14] generally applied in single dimension where only single dimension of n dimensional datasets[23] is considered for anonymization and single quasi identifier will be either generalized or suppressed to obtain anonymous datasets but problem with single dimensional anonymization is anonymous [28]

datasets are easily able to recognised if other fields are known So multidimensional k-anonymity is used as a model for achiving quality of anonymity using different datasets

Table 1.1 Employee Data Table

Name	Zip	Age	Salary
Amit	411033	35	50000
Sachin	411038	36	35000
Kshitija	433110	25	34800
Sumit	411044	34	67000
Shravani	433155	24	27000
Revati	433122	28	23480

Table 1.2 Multi-dimensional k-Anonymous Table

Name	Zip	Age	Salary
Amit	411***	[34-38]	50000
Sumit	411***	[34-38]	67000
Sachin	411***	[34-38]	35000
Kshitija	433***	[24-28]	34800
Shravani	433***	[24-28]	27000
Revati	433***	[24-28]	23480

As shown in Table 1.1 Employee Table with salary, zipcode ad Age records of Employee and As shown in Table 1.2 Generated Table after applying Multidimensional K-Anonymity on Employee Data Table and which will generate 3 equivalence classes so called as 3-Anonymus Table and value of k is 2 and in generated 3-Anonymus Table if any intruder wants to know the salary of any employee using background knowledge and if Name filled is hidden it is difficult to guess the person and his/her salary.

2. RELATED WORK

During review of Literature various different papers are accessed for related work of the topic and I have compared two reference papers [11][12] in reference [11] privacy protection algorithm with different algorithms are discussed I have compared this work with my work for doing comparative study based on algorithm, methodology, Quality Matric, Information Loss Measures, Efficiency ,Testing and validation and ability to handle cross section databases and after comparative study it is observed that Clustering Method is used for finding the equivalence class in [11] Distance between the two points after clustering and Cost Metrics methods are used for anonymization, Distance and Cost Metrics are used to measure the quality of Anonymity Distance between 2 records defined by Formula

$$\Delta(r_1,r_2) = \sum_{i=1,\ldots,m} \delta_N(r_1[N_i],r_2[N_i]) + \sum_{j=1,\ldots,n} \delta_C(r_1[C_j],r_2[C_j]),$$

Efficiency of this algorithm depends on Measurement of Information Loss using IL Matric this (e) on MIN Ni and MAX Ni i.e MIN MAX value of e with attribute Ni we can calculate Total information loss after anonymization by

$$\text{Total-IL}(\mathcal{AT}) = \textstyle\sum_{e\in\mathcal{E}} IL(e).$$

Here testing and validation is not done on huge volume of data in public domain to calculate efficiency and effectiveness In Paper[12] OLA algorithm is used to achieve the k-Anonymity and Two times K-Anonymity algorithm is used for doing Anonymization here Quality Measure specified by this Paper is by using OLA Algorithm for Simple Code for Degree Priority Lattice Tree Visiting and Second K-Anonymity and Here dataset is taken of 5000 values and for large volume of data no measures specified. This algorithm not tested on Cross section like multiple databases by grouping them together

3. GAP ANALYSIS

As per reference [18][39]paper survey done for research work it is observed that gaps are identified from reference papers related to dealing with different attacks capability of algorithm for handling large volume of datasets etc. and multidimensional k-anonymity approach[35] must have following abilities for doing anonymization

1) It must be able to handle information loss during Online release of data.
2) It must be having Quality Parameter values less in nature as compare to other methods.
3) It should be able to avoid Linking attack, Homogeneity attack and background knowledge Attacks
4) It should be validated and tested on large volume of datasets and gives better performance

4. PROPOSED WORK

In the proposed model for Multidimensional k-Anonymity for protecting privacy Nearest Neighbor algorithm is used to achieve k-Anonymity after mapping of Multi-Dimension to single dimension, Here after finding PRO,PPA and K-Anonymity values mapping is done and using Nearest Neighbor algorithm closest nearby value is located to form equivalence classes, As per increase in value of k if obtained Less values of CDM then this algorithm obtained Good Quality of anonymity

In CAVG The closer that average number of records in equivalence is to k, the lower the information loss is CAVG is calculated by following formula where total number of n records divided by Total equivalence classes and whole answer is again divided by k to obtain CAVG

$$C_{AVG} = \left(\frac{Total\_records}{Total\_equivalence\_classes}\right) / (k)$$

Efficiency of this algorithm depends upon mapping Multi-dimensional data with single dimension and effective calculation of Neighbor value to get proper subset of equivalence class this we will obtained by Calculating Visual Measures CDM,CAVG and comparing Nearest Neighbor method with other similar methods such as Median and close-k Algorithm are Tested and verified for 54000 dataset and proves effective and efficient in execution , Algorithm is Tested and Validated on Intersection cross section databases like Adult and Voter data and LIC and Voter data processed, so as to reduce the probability that attackers This Method is well Tested on Linking Attack, Background Knowledge attack and Homogeneity attack and it can be used to avoid attack on real-time publically available datasets released online

4.1 SYSTEM ARCHITECTURE:-

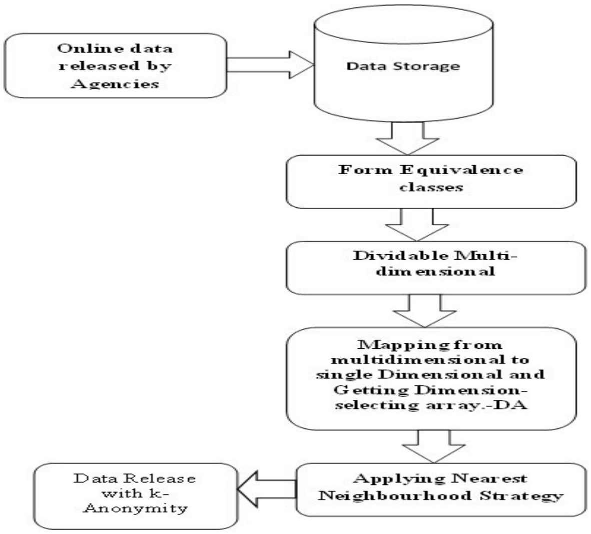
Online data released by Agencies
Data Storage
Form Equivalence classes
Dividable Multi-dimensional
Mapping from multidimensional to single Dimensional and Getting Dimension-selecting array.-DA
Applying Nearest Neighbourhood Strategy
Data Release with k-Anonymity

Figure 1: System Architecture of Multidimensional K-Anonymity

As shown in Figure1 the system architecture of Proposed multidimensional K-Anonimization Model once bulk amount of data like LIC ,Voter data is released by any organization data is temporarily stored in database for further processing and for doing pre-processing of data to remove any ambiguity or fill the missing data if any, after preprocessing of data Equivalence classes are generated by using generalization and suppression methods and checking is done whether divided multidimensional sets are available after that mapping is done from multi to single dimension to calculate PROS and PPAS and DA (Dimension Array)is generated for applying Nearest Neighbor method after that anonymized data is released with improved privacy and integrity and users confidentiality is maintained

5. RESULTS AND DISCUSSION

After doing the experimentation as per research objectives on very large volume of datasets results are obtained voter dataset, Adult dataset and LIC Dataset are taken as input and algorithm is applied step by step and obtained results are compared with results obtained in reference papers proves that Multidimensional k-Anonymity is the best and effective method for achieving user's data privacy

5.1 Comparative study of results obtained

As mentioned in paper[11] two types of Metrics used quality of Data and Efficiency Measures where 3 methods value is plotted with incrementing in k value, Greedy k-Member, Greedy K- Member CM and Median Partitioning where partitioning by median method provides optimal results as shown in figure 2 also it compares clustering , Partitioning and Discernibility methods and it shows that efficient method is clustering

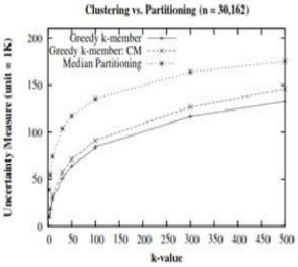

Figure 2: Information loss and Discernibility Metric

As mentioned in paper[12] Degree-priority algorithm is showing better performance as shown in figure on the number of computed nodes and computing time when we compare it to OLA algorithm Two Times Information loss experiment results of k-anonymity and degree priority showing that Prec is used as information loss measure and entropy is used as relative information loss measure after comparison conclusion is that two times anonymity method always has less information loss

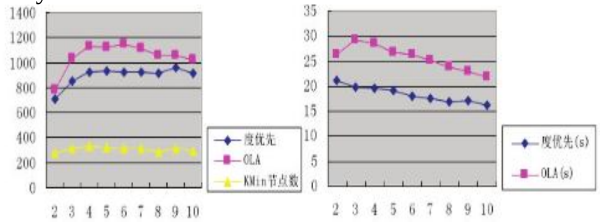

Figure 3 :-Efficiency Comparisons between Degree Priority and OLA

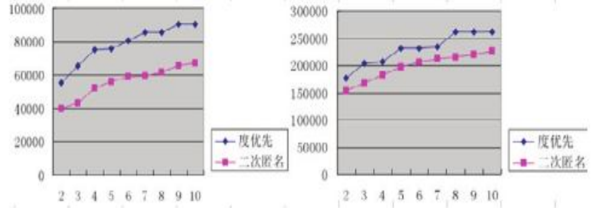

Figure 4:- Information Loss Comparisons between Degree Priority and OLA

also by using Multidimensional K-Anonymity to Protect Privacy by Nearest Neighbor Method results obtained are discussed as follows as shown in figure 4 and figure 5

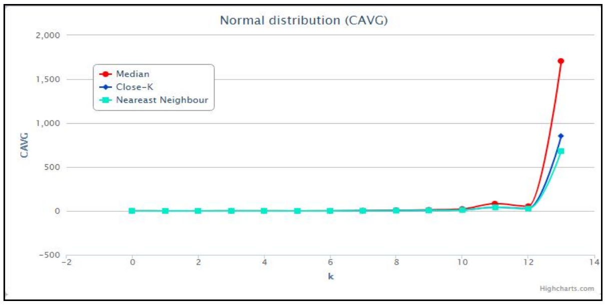

Figure 5:-. CAVG Adult dataset large size

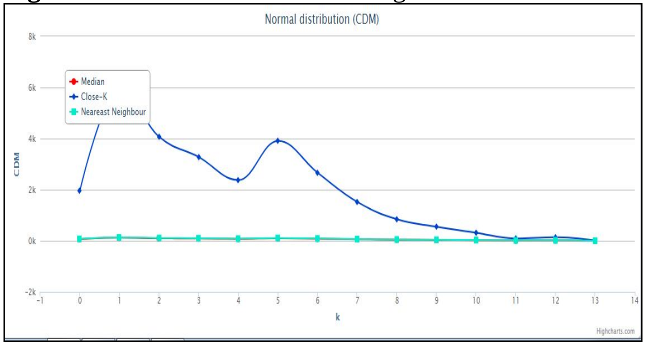

Figure 6:-. CDM Adult dataset large size

As shown in figure 5 Results of CAVG i.e. size of equivalence classes (when doing Normal distribution) of Adult database in this scenario increment in k value results in lower values of CAVG for Nearest Neighbor Method which proves the Effectiveness and Optimal Execution of Nearest Neighbor Algorithm as compare to Median as well as close-k methods

As shown in figure 6, Results of CDM called as Discernibility Matrix(when doing Normal distribution) of Adult database ,in this scenario increase in k value gives lower in order values of CDM as for lower C_{DM} values, the more even the partition of datasets is and anonymity also performs in better way and effectively it results in higher efficiency of Nearest Neighbor Algorithm.

6. CONCLUSION

As data release at large volume in public domain may cause user's privacy on stake due to liking available released information in public domain with any other publically available datasets to avoid that proposed methodology provides data confidentiality and also avoids different types of attacks on released data and after anonymization the resultant dataset is also effectively used by various agencies such as research organizations, banking sectors, insurance sector etc. and obtained results by our proposed approach when compared with other obtained results by different methods proves that they are efficient provides good quality of anonymity and also does protection of users privacy effectively and also obtained results shows consistency when tested on very large volume of data as well.

ACKNOWLEDGEMENT

At the outset, I would like to express deep gratitude to the various agencies for providing me required support to carry out my research work. These agencies helps me by providing large volume of datasets and all required tools for data analysis, I would like to express my deep regards to my research supervisors Dr.Kshama V.Kulhalli and Dr. S. Kotrappa sir for continuous motivation and timely guidelines during research journey my doctoral committee members also helps me a lot to get various dimensions of my research work.

REFERENCES

[1] Sweeney L. "K-Anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10 (5), pp.557-570.

[2] Rashid A.H. "Protect privacy of medical informatics using k-anonymization model," Hegazy A.F. Informatics and Systems (INFOS), the 7th International Conference on, 2010, pp. 1-10.

[3] Sacharidis Dimitris, Mouratidis Kyriakos, Papadias Dimitris. "K-anonymity in the presence of external databases," IEEE Transactions on Knowledge and Data Engineering, v 22, n 3, 2010, pp. 392-403.

[4] Samarati P. "Protecting respondents' identities in microdata release.,"Proc of the TKDE' 01, 2001, pp. 1010-1027.

[5] Samarati P, Sweeney L. "Generalizing data to provide anonymity when disclosing information," Proc of the 17th ACMSIGMODSIGACT - SIGART Symposium on the Principles of Database Systems, Seattle,WA, USA, 1998, pp. 188.

[6] Sweeney L. "Achieving k-anonymity privacy protection using generalization and suppression," International Journal of Uncertainty, Fuzziness and Knowledge2Based Systems, 2002, 10 (5), pp. 571-588.

[7] LeFevre K, DeWitt D J, Ramakrishnan R. "Incognito: Efficient full-domain k-anonymity," ACM SIGMOD International Conference on Management of Data Baltimore, USA: ACM, 2005, pp. 49-60.

[8] Iyengar V. "Transforming Data to Satisfy Privacy Constraints," Proc. of the ACM SIGKDD. USA: [s. n.], 2002, pp. 279-287.

[9] Byun Ji-Won, Kamra Ashish, Bertino Elisa, Li Ninghui. "Efficient k-anonymization using clustering techniques," Lecture Notes in Computer Science, 2007, pp.188-200.

[10] Bayardo R, Agrawal R. "Data privacy through optimal kanonymization," In ICDE, 2005.

[11] Efficient k-Anonymization Using Clustering Techniques
Ji-Won ByunAshishKamra 2Elisa Bertino and Ninghui Li, R. Kotagiri et al. (Eds.): DASFAA 2007, LNCS 4443, pp. 188–200, 2007. cSpringer-Verlag Berlin Heidelberg 2007

[12] Study on Privacy Protection Algorithm Based on K-Anonymity Zhao FeiFei Dong LiFeng Wang Kun Li Yang Published by Elsevier B.V. Selection and/or peer review under responsibility of ICMPBE International Committee. Doi: 10.1016/j.phpro.2012.05.093

[13] Accuracy-Constrained Privacy- Preserving Access Control Mechanism for Relational Data Zahid Pervaiz, Walid G. Aref, Senior Member, IEEE, Arif Ghafoor, Fellow, IEEE, and Nagabhushana Prabhu, IEEE ON KNOWLEDGE AND DATA ENGINEERING,VOL 26 NO. 4, APRIL 2014

[15].Privacy Preserving Data Publishing Basedon *k*-Anonymity by Categorization of Sensitive Values, Manish Sharma, Atul Chaudhary, Manish Mathuria, Shalini ChaudharyInternational Journal of Scientific & Engineering Research, Volume 5, Issue 4, April-2014 ISSN pp 2229-5518

[16]. Survey on Anonymization using k-anonymity for Privacy Preserving in Data MiningBinal Upadhyay*, Dr. Amit ganatra, International Journal of Scientific & Engineering Research, Volume 8, Issue 3, March-2017 ISSN 2229-5518 pp 376-380

[17] Clustering Based K-anonymity Algorithm for Privacy Preservation Sang Ni1, Mengbo Xie1, Quan Qian 1:2 International Journal of Network Security, Vol 19,No 6 PP.1062-1071, Nov. 2017 (DOI: 10.6633/IJNS 201711 .19(6).23) pp 1062-1071.

[18]. Preserving Privacy during Big Data Publishing using K-Anonymity Model – A SurveyDivya Sadhwani Dr. Sanjay SilakariMr Uday Chourasia International Journal of Advanced Research in Computer Science Volume 8, No. 5, May June 2017ISSN No. 0976-5697 pp 801-810

[19]. Liu, Kai-Cheng & Kuo, Chuan-Wei &Liao, Wen-Chiuan & Wang, Pang-Chieh. (2018). Optimized Data de-Identification Using Multidimensional Anonymity 1610-1614.10.1109/TrustCom/ BigDataSE 201800235

[20] The Five Safes of Risk-Based Anonymization, Published in: IEEE Security & Privacy (Volume: 17 , Issue: 5 , Sept.-Oct. 2019), Page(s): 84 - 89, Print 30 August 2019

[21] Aggarwal, C.C. and Yu, P.S. (2008) 'A framework for condensation-based anonymization of string data', Data Mining and Knowledge Discovery, 16(3), pp. 251–275. doi: 10.1007/s10618-008-0088-z.

[22] Aldeen, Y.A.A.S., Salleh, M. and Razzaque, M.A. (2015) 'A comprehensive review on privacy preserving data mining', SpringerPlus, 4(1).doi:10.1186/ s40064-015-148x

[23] Allard, T., Nguyen, B. and Pucheral, P. (2013) 'METAAP: Revisiting privacy-preserving data Publishing by secure devices', Distributed and Parallel Databases, 32(2), pp. 191–244. doi: 10.1007/s10619-013-7122-x.

[24]Ayala-Rivera, V., Mcdonagh, P., Cerqueus, T. and Murphy L .(2014) 'A systematic comparison and evaluation of k-anonymization Algorithms for practitioners TRANSACTIONS ON DATAPRIVACY 7(3),pp337–370

[25]Belsis, P. and Pantziou, G. (2012) 'A k-anonymity privacy Preserving approach in wireless medical monitoring environments', Personal and Ubiquitous Computing 18(1)pp.61–74.doi:10.1007/s00779-012-0618-y.

[26] Casas-Roma, J., Herrera-Joancomartí, J. and Torra, V. (2016) 'A survey of graph-modification techniques for privacy-preserving on networks', Artificial Intelligence Review,. doi: 10.1007/s10462-016-9484-8.

[27] Emam, K.E. (2007) Data Anonymization practices in clinical research. [Online] Available at: following URL http://www.ehealthinformation.ca/wpcontent/uploads4/2 107/2006-Data-AnonymizationPractices.pdf (Accessed:2 February 2017).

[28] Hussien, A.A., Hamza, N. and Hefny, H.A. (2013) 'Attacks on Anonymization-Based privacy-preserving: A survey for data mining and data publishing', Journal of Information Security, 4(2), pp. 101–112. doi: 10.4236/jis.2013.42012.

[29] Jain, P., Gyanchandani, M. and Khare, N. (2016) 'Big Data privacy: A technological perspective and review', JournalofBigData, 3(1). doi: 10.1186/s40537-016-0059-y.

[30] Li, F., Zou, X., Liu, P. and Chen, J.Y. (2011) 'New threats to health data privacy', BMC Bioinformatics, 12 (Suppl 12),p. S7. doi: 10.1186/1471-2105-12-s12-s7.

[31] Li, N., Li, T. and Venkatasubramanian, S. (2007) 'Tcloseness: Privacy beyond k-anonymity and l-diversity', ICDE 2007 IEEE 23rd International Conference on Data Engineering, doi: 10.1109/icde.2007.367856.

[32] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. (2007) 'L -diversity: privacy beyond k-anonymity', ACM Transactions on Knowledge Discovery from Data, 1(1) doi:10.1145/1217299.1217302 [33] Jaimain Han, Jaun Yu, Yuchang Mo, Jianfeng Lu, Huawen Liu, MAGE: A Semantics retaining k-anonymization method for mixed data. journal homepage: URL www.elsevier.com/locate/knosys 0950-7051/$ - see front matter 2013 Elsevier B.V. All rights reserved.

[34] Zhao FeiFei 1,Dong LiFeng2, Wang Kun2,Li Yang2 Study on Privacy Protection Algorithm Based on K-Anonymity 2012 International Conference on Medical Physics and Biomedical Engineering.

[35] 1School of Computer Science and Technology Tianjin University Tianjin, China 2Postgraduate Training BrigadeMilitary Transportation University Tianjin, China Study on Privacy Protection Algorithm Based on K-Anonymity 2012 International Conference on Medical Physics and Biomedical Engineering.

[36] Liu, Q.; Shen, H.; Sang, Y. Privacy-preserving data publishing for multiple numerical sensitive attributes. Tsinghua Sci. Technol. 2015, 20, 246–254.

[37]Sánchez, D.; Domingo-Ferrer, J.; Martínez, S.; Soria Comas, J. Utility-preserving differentially private Data releases via individual ranking microaggregation. Inf. Fusion 2016, 30, 1–14. [CrossRef]

[38]Hua, J.; Tang, A.; Fang, Y.; Shen, Z.; Zhong, S. Privacy- preserving utility verification of the data Published by non-interactive differentially private mechanisms. IEEE Trans. Inf. Forens. Secur. 2016, 11, 2298–2311.[CrossRef]

[39] Kavitha, S., S. Yamini, and Raja Vadhana. "An evaluation on big data generalization using k-Anonymity algorithm on cloud." Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference IEEE 15

[40] Basu, Anirban, et al. "k-anonymity: Risks and the Reality." *Trustcom/BigDataSE/ISPA, 2015 IEEE*. Vol. 1. IEEE, 2015.