DOI: 10.21276/sjet

Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2017; 5(6):238-245 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Review Article

A Review on K-Anonymization Techniques

Anisha Tiwari[1], Minu Choudhary[2]
[1] Dept. of Computer Science and Engineering, Rungta College of Engineering and Technology, Bhilai, Chhattisgarh, India
[2] Assistant Prof., Dept. of Information Technology, Rungta College of Engineering and Technology, Bhilai, Chhattisgarh, India

***Corresponding author**
Anisha Tiwari
Email: anishatiwari20@gmail.com

Abstract: Securing information protection is an imperative issue in microdata distribution. Anonymity strategies regularly mean to ensure singular security, with insignificant effect on the nature of the discharged information. As of late, a couple of models are acquainted with guarantee the security ensuring or potentially to diminish the data misfortune to such an extent as could be allowed. That is, they additionally enhance the adaptability of the anonymous system to make it all the more near reality, and after that to meet the various needs of the general population. Different proposition and calculations have been intended for them in the meantime. In this paper a review of anonymity techniques for privacy preserving. In this paper, the discussion about the anonymity models, the significant execution ways and the techniques of anonymity algorithm, and also analyzed their strength and limitations.
Keywords: Anonymity techniques, anonymity models, privacy preserving algorithm.

INTRODUCTION

Today's databases contain a considerable measure of delicate individual information. So it's significant to outline data frameworks which may confine the noteworthy of individual information. For instance, consider a healing center that keeps up patient records. The healing facility longings to unveil information to an organization in such some way that the organization can't deduce that patients have that diseases. One system to formally indicate protection arrangements is to particular sensitive information as inquiries and implements excellent security, a terribly durable thought of security that ensures that the other question replied by the data won't uncover any information with respect to the delicate information.

Security Preserving Data Publishing

Personal records of individuals are logically being gathered by various government and organization foundations for the requirements of information examination. The information examination is encouraged by these associations to distribute "adequately private" thoughts over this data that are collected. Privacy could be a twofold edged brand - there should be sufficient protection to ensure that touchy information concerning the general population isn't revealed by the perspectives and at a comparative time there should be sufficient data to play out the investigation. Besides, an enemy who needs to gather delicate information from the uncovered perspectives in some cases has some data concerning the general population inside the data. The principle goal is to change over the first data into some mysterious sort to prevent from inducing its record owner's sensitive information as examined in [9].

Information Anonymization

Information anonymization is the way toward expelling by and by identifiable data from informational indexes, to make the general population unknown about whom the information describe. It allows the exchange of information over a limit, as between two offices inside focus or between two offices, though lessening the peril of accidental uncovering, and in bound conditions in an exceedingly way that grants investigation and examination post-anonymization. This system is utilized as a part of undertakings to expand the security of the information while enabling the information to be broke down and utilized. It changes the information that will be utilized or distributed to keep the distinguishing proof of key data. Information

anonymization methods, for example, k-anonymity, l-diversity qualities what's more, t-closeness are broad.

k-Anonymity: The essential arrangement of k-anonymity is to shield a dataset against re-identified by summing up the characteristics that may be used in a linkage attacks (semi identifiers). A data set is considered k-anonymous if each data thing can't be recognized from at least k-1 elective information things.

l-Diversity: l-diversity qualities could be an assortment of group based for the most part anonymization that is wont to safeguard security in learning sets by decreasing the coarseness of a learning portrayal. This lessening might be an exchange off that winds up in some loss of adequacy of information administration or mining algorithm in order to accomplish some security. The l-differing qualities model is related degree expansion of the k-secrecy demonstrate that diminishes the harshness of data representation exploitation procedures and in addition speculation and concealment indicated any given record maps onto at least k elective records inside the information [26].

t-Closeness: t-closeness could be an extra refinement of l-assorted qualities bunch based generally anonymization that is acclimated safeguard security in learning sets by lessening the coarseness of a data portrayal. t-closeness could be an additional refinement of l-assorted qualities group essentially based anonymization that is wont to save protection in learning sets by decreasing the coarseness of a data delineation. This decrease could be an exchange off that winds up in some loss of viability of information administration or mining algorithm in order to understand a few protection [27].

K-ANONYMITY

k-Anonymity could be a formal model of protection [28]. The objective is to frame each record unclear from an illustrated variety (k) records if tries region unit made to detect the data. An arrangement of data is k-anonymized if, for any record with a given arrangement of characteristics, there square measure in any event k-1 elective records that match these traits. The properties can be any of the accompanying sorts.

Table-I: Dataset Description

Attributes	Description	Example
Explicit_identifier	Set of attributes	Name, Id
Quasi_identifier	Potentially identify record owners	Age, Sex, Zip
Sensitive attributes	Person's sensitive information that cannot revealed	Salary, Disease

The usage of k-anonymity needs the preparatory ID of the quasi identifier. The quasi identifier depends on the outer information accessible to the beneficiary, since it decides the connecting capacity (not all conceivable outside tables range unit open to every potential learning beneficiary); and diverse quasi identifiers will without a doubt exist for a given table [29].

Example

In the event that the previously mentioned table is to be anonymized with Anonymization Level (AL) set to 2 and the arrangement of Quasi identifiers as QI = {AGE, SEX, ZIP, PHONE}.Sensitive trait = {SALARY}. The quasi identifiers and touchy qualities are distinguished by the association as indicated by their rules and regulation.

Table-II: Table to be Anonymized

ID	Age	Sex	Zip	Phone	Salary (in Rs.)
1	24	M	641015	9994258665	78000
2	23	F	641254	9994158624	45000
3	45	M	610002	8975864121	85000
4	34	M	623410	7456812312	20000

Table: III: Anonymized Table

ID	Age	Sex	Zip	Phone	Salary (in Rs.)
*	20-50	ANY	641***	999*******	78000
*	20-50	ANY	641***	999*******	45000
*	20-50	ANY	612***	897*******	85000
*	20-50	ANY	623***	745*******	20000

This anonymization can be done by Generalization As well as Suppression

Generalization

Generalization is the way toward changing over an incentive into a less particular general term. For ex, "Male" and "Female" can be generalized to "Any". At the accompanying levels generalization procedures can be connected.

- Attribute (AG): Generalization is performed at the segment level; all the qualities in the section are generalized at a speculation step.
- Cell (CG): Generalization can likewise be performed on a solitary cell; at long last a summed up table may contain, for a particular section and values at various levels of generalization.

Suppression

Suppression comprises in averting delicate information by evacuating it. Suppression can be connected at the level of single cell, whole tuple, or whole segment, permits diminishing the measure of speculation to be forced to accomplish k-anonymity.

- Tuple (TS): Suppression is performed at column level; suppression operation evacuates entire tuple
- Attribute (AS): Suppression is performed at segment level; suppression operation shrouds every one of the estimations of a segment.
- Cell (CS): Suppression is performed at single cell level; at long last k-anonymized table may wipe out just certain cells of a given tuple/quality.

LITERATURE SURVEY

Xuyun Zhang *et al*. [16] proposes giving security and protection over the intermediate data sets become dispute problem since adversaries may retain micro data by identifying multiple data records. Encryption of all datasets in general society stage called cloud take in past systems may extremely tedious and exorbitant. So we give new novel upper bound protection spillage requirement based strategies to give which middle of the road information records request to be figured and which don't to guarantee a few information usage and security safeguarding.

Mohammad Reza Zare Mirakabad *et al*. [17] points giving protection over the information production. Under security information usage and aversion of divulgence of individual personality is more critical. One of the information anonymization methods called K-secrecy keeps the divulgence of individual character however it is for the most part neglected to accomplish. The other strategy called l-differing qualities will give the security of touchy data. So creator [2] recommends That L-Diversity is Lonely Enough to preserve Privacy. In this procedure the anonymization is performed by doling out people in the gathering of size more prominent than or equivalent to the estimation of semi identifier k.

Min Wu *et al*. [18] proposes saving security is most basic however a similar time it is inconvenience in arrival of small scale information discharge. In the perspective of trait disclosure K-namelessness is not well. So we propose new system called an ordinal separation based affectability mind full differing qualities metric model. The assorted qualities of touchy properties are done just on the K-anonymized table. To check the differences level of the bunch to start with we need to group the credits concerning to start with, second and third level. What's more, the assorted qualities degree is equivalent to the whole table. This strategy is basically concentrated on the categorical qualities.

Yunli Wang *et al*. [19] proposes k- anonymity neglects to accomplish qualities revelation however in l-assorted qualities plans to accomplish characteristic exposure. Second information anonymization procedure focus on cutting the illation from liberated miniaturized scale traits. Here we point another approach called a remarkable particular l-SR differing qualities to accomplish l-differences on the affectability degrees of delicate characteristics. These outcomes demonstrate that our calculation accomplished better execution on lessening illation of delicate data and accomplished the tantamount speculation information quality contrasted and other information distributing calculations. At long last we have two measures i.e Entropy Metric and Variance Metric to check the nature of distributed information.

Jordi Soria Comas *et al*. [20] points information anonymization strategies save protection, k- anonymity and €-differential security are two principle protection display. The t-closeness is the augmentation of k-obscurity, the development of private sensitive data depends on Bucketization algorithm. The fundamental point of t-closeness is to understand the doles out exposure issues. An information records is said to finish t-closeness if, for each horde of information managing a joining of semi identifier quality assess, the hole among the circulation of each private property in the group and the conveyance of the same secret appoint in the all information records is close as far as possible t. The Bucketization development is utilized to accomplish t-closeness. On the off chance that the calculation of Bucketization is excessively unforgiving the information misfortune in the classified appoint is expansive.

Table-IV: Shows comparison between various existing approaches and its limitation

S.No.	Ref.No.	Method Used	Data Source	Approach	Strength	Limitation
1	[1]	Systematic clustering method	Medical data	Author presents a clustering based *k*-anonymization technique to minimize the information loss while at the same time assuring data quality	Results show that our method attains a reasonable dominance with respect to both information loss and execution time.	Need to extend the systematic clustering algorithm to *k*-anonymity model
2	[2]	*k*-anonymization algorithm	*k*-anonymized dataset	Author proposed an efficient *k*-anonymization algorithm by transforming the *k*-anonymity problem to the *k*-member clustering problem.	Author develop a suitable metric to estimate the information loss introduced by generalizations, which works for both numeric and categorical data.	Additional performance measures are there.
3	[3]	Nearly-Optimal Anatomizing Algorithm	CENSUS dataset	Author develop a linear-time algorithm for computing anatomized tables that obey the l-diversity privacy requirement, and minimize the error of reconstructing the microdata.	experiments confirm that anatomy permits highly accurate aggregate information about the unknown microdata, with an average error below 10%	Need extend the technique to multiple sensitive attributes is an interesting topic
4	[4]	heuristic algorithm	Real world dataset	proposed an approach that identifies which part of intermediate data sets needs to be encrypted while the rest does not, in order to save the privacy preserving cost.	results demonstrate that the privacy-preserving cost of intermediate data sets can be significantly reduced	Need to investigate privacy aware efficient scheduling of intermediate data sets in cloud
5	[5]	sequential clustering algorithm.	US Census Bureau	proposed the private mutual information (PMI) utility measure that aims at maximizing the correlation between the generalized public data and the private data.	PMI measure is much more suitable when the goal is to achieve anonymizations	Additional performance measures are there.
6	[6]	apriori-based anonymization algorithm	Real world dataset	Author develop an algorithm which flnds the optimal solution, however, at a high cost which makes it inapplicable for large, realistic problems.	proposed algorithms are experimentally evaluated using real datasets	Need to consider sensitive values associated to set-valued quasi-identiflers.
7	[7]	LowCost algorithm	anonymous data record	propose the Efficiency metric □that represents both the utility and privacy of the anonymous data and can assess anonymization algorithms fairly.	performs the best both in terms of utility measure and privacy measure.	Additional performance measures are there.

8	[8]	Greedy *k*-member algorithm and Systematic clustering algorithm	UCI machine learning repository database	Author propose two approaches for minimizing the disclosure risk and preserving the privacy by using systematic clustering algorithm.	Author illustrate the effectiveness of the proposed approaches by comparing them with the existing clustering algorithms.	Ne4ed to evaluate the performance on a combination of multiple SA and QI attributes.
9	[9]	centralized and distributed anonymization algorithm	Health care data	PPDP Has Received A Great Deal Of Attention In The Database And Data Mining Research Communities	Degradation Of Data / Service Quality Loss Of Valuable Information Increased Costs	Additional performance measures are there.
10	[10]	Є-Differentially Private Algorithm	GWAS data	Author present methods for releasing differentially private minor allele frequencies, chisquare statistics and *p*-values.	Author provided a differentially private algorithm for releasing these statistics for the most relevant SNPs	Improvement required to get more accurate result
11	[11]	Generalization-based algorithm	BMS-WebView-1 and BMSWebView-2 datasets	Author develop PCTA, a generalization-based algorithm to construct anonymizations that incur a small amount of information loss under many different privacy requirements.	a clustering-based algorithm that can produce a significantly better result than the state-of-the-art methods in terms of data utility	Need to extend accommodate privacy and utility constraints that are common in real-world applications.
12	[12]	kACTUS algorithm	Adult dataset from the UC Irvine Machine Learning Repository	Author propose a new method for achieving k-anonymity named K-anonymity of Classification Trees Using Suppression (kACTUS)	Proposed method requires no prior knowledge regarding the domain hierarchy taxonomy	Need to extend the proposed method to other data mining tasks (such as clustering and association rules)
13	[13]	Privacy policy extraction (PPE) algorithm	GWAS data	Author proposed a method to anonymize patient-specific clinical profiles, which should be disseminated to support biomedical studies	approach automatically extracts potentially linkable clinical features and modifies them in a way that they can no longer be used to link a genomic sequence to a small number of patients	Algorithm does not guarantee that the anonymized clinical profiles will incur the least amount of information loss possible to satisfy the specified utility policy.
14	[14]	TDControl algorithm	BMS-POS, BMS-WebView-1, and BMSWebView-2. A	propose ρ-uncertainty, the first, to our knowledge, privacy concept that inherently safeguards against sensitive associations without constraining the nature of an adversary's knowledge and without falsifying data.	problem is solved non-trivially by a technique that combines generalization and suppression, which also achieves favorable results	Improvement required to get more accurate result

15	[15]	CLUSTERING-BASED ALGORITHM	synthetic and real-world data	presented a measure that can capture both data usefulness and privacy protection in this paper, and we developed a clustering-based algorithm that exploits this measure	algorithm is able to produce anonymizations of high quality, balancing usefulness and protection requirements	Improvement is require with respect to the performance.
16	[21]	Full-Domain Generalization Algorithms	Publicly available data sets.	Implementation Framework For Full Domain Generalization Using Multidimensional Data Model Together With Suite Of Algorithms	To Produce Minimal Full Domain Generalizations Perform Up To An Order Of Magnitude Faster Than Previous Algorithms On Two Real-Life Databases	Performance Of Incognito Can Be Enhanced By Materializing Portions Of The Data Cube, Including Count Aggregates At Various Points In The Dimension Hierarchies
17	[22]	Exhaustive Algorithm (Rothko-T)	Salary data	Anonymization Algorithms That Incorporate A Target Class Of Workloads, Consisting Of One Or More Data Mining Tasks As Well As Selection Predicates And The Datasets Much Larger Than Main Memory	High Efficiency And Quality Data Overcomes Problem Of Scalability	Problem Of Measuring The Quality Of Anonymized Data Fails To Work In The Top-Down Specialization (TDS) Approach
18	[23]	heuristic algorithms	Medical records	K-Anonymizing A Data Set Is Similar To Building A Spatial Index Over The Data Set Using R-Tree Index Based Approach	Achieve High Efficiency And Quality Anonymization Multidimensional Generalization, High Accuracy	More Compaction Is Needed To Achieve High Quality Anonymization Different Indexing Algorithm Provide Different Issues
19	[24]	MapReduce	Real time data	Mapreduce Is A Programming Model Implementation For Processing And Generating Large Data Sets Performs Map() And Reduce()	Allows Us To Handle Lists Of Values That Are Too Large In Memory The Model Is Easy To Use	Mapreduce Is Not Suitable For A Short On-Line Transactions
20	[25]	MapReduce	IDS data set	To Protect Data Privacy During Map-Reduce Programming	Sensitive User Data Protection High Privacy Assurance Ease To Use Fully Preserve The Scalability	Scalability Problem Occurs

CONCLUSION

In this paper, we talked about the Privacy preserving information distributing and information anonymization. We likewise talked about different anonymization strategies and for the most part focused on k-anonymity which involves both generalization and suppression. The last part is about the generalization algorithm and its execution for securing the protection of information utilized for the most part for data analysis.

REFERENCES

1. Kabir ME, Wang H, Bertino E. Efficient systematic clustering method for k-anonymization. Acta Informatica. 2011 Feb 1;48(1):51-66.
2. Byun JW, Kamra A, Bertino E, Li N. Efficient k-anonymization using clustering techniques. InInternational Conference on Database Systems for Advanced Applications 2007 Apr 9 (pp. 188-200). Springer, Berlin, Heidelberg.
3. Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation. InProceedings of the 32nd international conference on Very large data bases 2006 Sep 1 (pp. 139-150). VLDB Endowment.
4. Zhang X, Liu C, Nepal S, Yang C, Dou W, Chen J. Combining top-down and bottom-up: scalable sub-tree anonymization over big data using MapReduce on cloud. InTrust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on 2013 Jul 16 (pp. 501-508). IEEE.
5. Goldberger J, Tassa T. Efficient anonymizations with enhanced utility. InData Mining Workshops, 2009. ICDMW'09. IEEE International Conference on 2009 Dec 6 (pp. 106-113). IEEE.
6. Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data. Proceedings of the VLDB Endowment. 2008 Aug 1;1(1):115-25.
7. Huda MN, Yamada S, Sonehara N. On Enhancing Utility in k-anonymization. International Journal of Computer Theory and Engineering. 2012 Aug 1;4(4):527.
8. Bhaladhare PR, Jinwala DC. Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model. J. Inf. Sci. Eng.. 2016 Jan 1;32(1):63-78.
9. Mohammed N, Fung B, Hung PC, Lee CK. Centralized and distributed anonymization for high-dimensional healthcare data. ACM Transactions on Knowledge Discovery from Data (TKDD). 2010 Oct 1;4(4):18.
10. Fienberg SE, Slavkovic A, Uhler C. Privacy preserving GWAS data sharing. InData Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on 2011 Dec 11 (pp. 628-635). IEEE.
11. Gkoulalas-Divanis A, Loukides G. PCTA: privacy-constrained clustering-based transaction data anonymization. InProceedings of the 4th International Workshop on Privacy and Anonymity in the Information Society 2011 Mar 25 (p. 5). ACM.
12. Data Anonymization", ACM 2011
13. Kisilevich S, Rokach L, Elovici Y, Shapira B. Efficient multidimensional suppression for k-anonymity. IEEE Transactions on Knowledge and Data Engineering. 2010 Mar 1;22(3):334-47.
14. Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. Proceedings of the National Academy of Sciences. 2010 Apr 27;107(17):7898-903.
15. Cao J, Karras P, Raïssi C, Tan KL. ρ-uncertainty: inference-proof transaction anonymization. Proceedings of the VLDB Endowment. 2010 Sep 1;3(1-2):1033-44.
16. Loukides G, Shao J. Capturing data usefulness and privacy protection in k-anonymisation. InProceedings of the 2007 ACM symposium on Applied computing 2007 Mar 11 (pp. 370-374). ACM.
17. Zhang X, Liu C, Nepal S, Pandey S, Chen J. A privacy leakage upper bound constraint-based approach for cost-effective privacy preserving of intermediate data sets in cloud. IEEE Transactions on Parallel and Distributed Systems. 2013 Jun;24(6):1192-202.
18. Mirakabad MR, Jantan A. Diversity versus anonymity for privacy preservation. InInformation Technology, 2008. ITSim 2008. International Symposium on 2008 Aug 26 (Vol. 3, pp. 1-7). IEEE.
19. Wu M, Ye X. Towards the diversity of sensitive attributes in k-anonymity. InProceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology 2006 Dec 18 (pp. 98-104). IEEE Computer Society.
20. Wang Y, Cui Y, Geng L, Liu H. A new perspective of privacy protection: Unique distinct l-SR diversity. InPrivacy Security and Trust (PST), 2010 Eighth Annual International Conference on 2010 Aug 17 (pp. 110-117). IEEE.
21. Soria-Comas J, Domingo-Ferrer J, Sánchez D, Martínez S. t-closeness through microaggregation: Strict privacy with enhanced utility preservation. IEEE Transactions on Knowledge and Data Engineering. 2015 Nov 1;27(11):3098-110.
22. LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. InProceedings of the 2005 ACM SIGMOD

international conference on Management of data 2005 Jun 14 (pp. 49-60). ACM.
23. LeFevre K, DeWitt DJ, Ramakrishnan R. Workload-aware anonymization techniques for large-scale datasets. ACM Transactions on Database Systems (TODS). 2008 Aug 1;33(3):17.
24. Iwuchukwu T, Naughton JF. K-anonymization as spatial indexing: Toward scalable and incremental anonymization. InProceedings of the 33rd international conference on Very large data bases 2007 Sep 23 (pp. 746-757). VLDB Endowment.
25. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM. 2008 Jan 1;51(1):107-13.
26. Zhang K, Zhou X, Chen Y, Wang X, Ruan Y. Sedic: privacy-aware data intensive computing on hybrid clouds. InProceedings of the 18th ACM conference on Computer and communications security 2011 Oct 17 (pp. 515-526). ACM.
27. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. InData Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on 2006 Apr 3 (pp. 24-24). IEEE.
28. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. InData Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on 2007 Apr 15 (pp. 106-115). IEEE.
29. Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002 Oct;10(05):557-70.
30. Ciriani. "k-Anonymity", Springer US, Advances in Information Security (2007).