Journal of Theoretical and Applied Information Technology
15th February 2025. Vol.103. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE SCALABLE REVERSIBLE RANDOMIZATION ALGORITHM (SRRA) FOR BETTER PRIVACY PRESERVATION, IMPROVED FEATURE SELECTION STABILITY, AND HIGHER ACCURACY IN BIG DATA ANALYTICS

MOHANA CHELVAN P[1], Dr. RAJAVARMAN V N[2]

[1]Research Scholar, Department of Information Technology, Faculty of Engineering and Technology,
Dr. M.G.R. Educational and Research Institute, Chennai, India
[1]Senior Assistant Professor, School of Science and Computer Studies, CMR University, Bengaluru, India
[2]Professor, Department of Information Technology, Faculty of Engineering and Technology,
Dr. M.G.R. Educational and Research Institute, Chennai, India
Email: [1]pmohanselvan@rediffmail.com

ABSTRACT

In today's knowledge economy, every organization uses the enormous accumulated data for their development of business as well as for enhancing customer relationship management. However, most of the data produced is unstructured and semi-structured and comprises customers' private data. Data analysis is mostly done by a third party to get insights and hidden patterns or information. Preserving the individual's privacy is vital for organizations before analysing it for pattern mining. Because of the proliferation of electronic gadgets and technological progressions, the dimension of the collected dataset increases into the high dimensional dataset. This will lead to feature selection, which is a vital preprocessing step for big data analytics as a dimensionality reduction technique. Previously, researchers believed that the stability of feature selection depends generally on the algorithm but recently confirmed by scientists it depends on the dataset's statistical characteristics. Unstable feature selection results confuse researchers' minds about their research conclusions as the selection stability positively correlated with accuracy. Privacy preservation will disturb the statistical characteristics of a dataset, which will impact the stability of feature selection and accuracy. Hence, privacy conserving big data analytics techniques should protect the sensitive data of individuals with better stability of feature selection and accuracy. The research paper evaluates the methods of privacy conserving big data analytics and recommends a new reversible privacy conserving algorithm called Scalable Reversible Randomization Algorithm (SRRA).
Keywords: *Big Data, Data Analytics, Privacy Preservation, Feature Selection, Feature Selection Stability*

1. INTRODUCTION

Every day we produce massive amounts of data because of the revolution of digital devices and internet connectivity for most of our daily routine work in banks, e-commerce, supermarkets, hospitals, etc., resulting in the accumulation of mountains of transactional data. Numerous apps on our smartphones collect our personal data and transactional data, which are in multiple varieties. But most of the data formed from these electronic devices is in semi-structured and unstructured data like videos, text messages, etc., and also unceasingly arriving at each and every second. Hence, the gigantic accumulated dataset is coined by 'big data'. This accumulated dataset and its concealed information and patterns which are extracted by big data analytics can be used to improve customer relationship management and make strategic decisions for the welfare of the business organization. Also, the data used for big data analytics must be anonymized before it can be used for pattern mining.

There are numerous privacy issues in the healthcare sector and protecting sensitive patient data in transit and the data stored in systems is vital. The data are generally digitised unstructured data. In today's big data era of smart healthcare, new challenges have arisen to protect the sensitive electronic records of patients, and methods like blockchain and encryption are used to protect the privacy of sensitive clinical data [1]. However, implementing big data machinery generates openings in clinical decision support systems and treatment optimization. The necessary computing resources and machine learning algorithms scuffle to analyse the big data to get valuable insights from the data because of its massive volume and several varieties of data.

Conserving the individual's privacy is the most significant and privacy conserved datasets can only be used for big data analytics or other research purposes. Data custodians can be social networking applications, banks, hospitals, e-commerce sites, mobile apps, websites, etc. Safeguarding the privacy of the users' data from the third party involved in big data analytics is the accountability of the data custodian. Most countries implement stringent laws against privacy intimidations. However, lack of consciousness among people is the foremost reason for privacy breaches. Even though the data is held in the public domain, data leakage can be contributed by users themselves knowingly or unknowingly.

Examples of privacy intimidation are discrimination, surveillance, personal embarrassment, abuse, disclosure, etc. The business organization generally uses third parties for data analytics, and they map the existing data, in easily accessible public domains with records in the dataset like region, religion, occupation, zip code, and gender to get personal data like monetary value or disease, resulting in revelation leading to solemn privacy breaches and this is called as linkage attacks [2–4]. Most business organizations, including retail and e-commerce, expand their business through surveillance to improve customer relationship management through recommendation systems by providing offers to their clientele [5]. Some persons get medication and treatment that they don't want to reveal to others, including family members. If the medical shop sends a reminder or other offers to the person by smartphone, it can be noticed by his/her family members, leading to personal embarrassment and even abuse [6]. Discrimination which is prejudice or disparity towards a person or group of persons in a certain area happens when some private information of a person or group of persons is divulged to a third party. For example, electoral result analysis leaked information to third parties and people of one area or region were entirely against the party that formed the government leading to the government ignoring that community or even having prejudice over them.

The advancements of digital devices and information and communication technology increase the dimension of datasets to many folds. In big data analytics, all the columns of a dataset cannot be used, and unrelated and/or redundant and noisy columns must be eliminated for improved model comprehensibility, reduced processing time, enhanced accuracy of data analytics results, and improved privacy preservation, as the detached features may be quasi-identifiers. The answer to the 'curse of high dimensionality' [7] problem is dimensionality reduction methods like feature selection, as a vital preprocessing step for big data analytics. Feature selection is one of the most significant and essential data preprocessing steps, and is generally used to seek extremely correlated features and remove redundant or uncorrelated features from a feature set [8–11]. In big data scenarios, scalability is a foremost problem, and eliminating a massive amount of irrelevant or noisy data removes useless data and retains extremely correlated data with valuable clues, improving processing performance and computational efficiency.

Unstable feature selection results are because of trivial perturbations of data in the trial dataset, such as adding or deleting a few tuples to the dataset row-vice. Or else, it can be column vice alteration like the addition of calculated noise to feature values for privacy conserving perturbation. Stability of feature selection is a crucial area of research as unstable feature selection leads to untrustworthy research deductions as stability of feature selection is directly interconnected with the accuracy of data analytics results. For trivial perturbations of data, the feature selection algorithm chooses different sets of features, which will also be mirrored in the accuracy of big data analytics outcomes and will lead to confusion in the researcher's mind about his conclusion of research findings.

Privacy preservation methods safeguard the individual's privacy and also the perturbation of data to preserve privacy disturbs the statistical characteristics of a dataset. The researchers earlier

supposed that the stability of feature selection typically depends on algorithms rather than the statistical characteristics of the dataset. However, researchers later determined the stability of feature selection generally depends on the statistical characteristics of the dataset rather than feature selection algorithms [12–17]. Privacy conserving perturbation disturbs the characteristics of datasets, leads to unstable feature selection outcomes, and results in untrustworthy big data analytics result accuracy as feature section stability is directly interrelated with the accuracy of results.

Utility metrics are used to assess and measure the usefulness of the data in big data analytics after privacy safeguarding perturbation. The usefulness of the privacy conserved dataset depends on the modification in the collective statistical characteristics of the dataset, which must be small and will be reflected in the data analytics outcomes in terms of accuracy or error rate [18]. Utility metrics are based on the assessment of the usefulness of privacy conserved datasets to the original dataset for specific purposes like machine learning or data mining.

Hence, there is a trade-off between anonymization or perturbation of datasets for privacy conservation, stability of feature selection, and accuracy of outcomes of big data analytics. This research paper proposes a reversible privacy conserving algorithm called Scalable Reversible Randomization Algorithm (SRRA) to anonymize the dataset with improved privacy preservation, and minor effects on change in the statistical characteristics of the dataset, leading to better stability of feature selection with a lower error rate or higher accuracy.

2. RELATED WORK

Privacy preservation is typically, of two, interactive methods and also in this method the dataset is under the control of a custodian called PPDM (Privacy Preserving Data Mining). The second one is the non-interactive anonymization method and also in this method, the dataset is not under the control of a custodian called PPDP (Privacy Preserving Data Publishing). In PPDM like differential privacy techniques, the dataset is under the control of the custodian, and the big data analyst analyses the dataset. The PPDP methods like generalization, bucketization, perturbation, anonymization (t-closeness, l-diversity, and k-anonymity), and slicing, and randomization methods, the dataset is anonymized and published for any research works as it is not under the control of a custodian [19].

The differential privacy method [19] is interactive queries made to a database using a randomized response mechanism by adding calculated noise to the outcome of the query. This method shares data to permit inferences about groups of persons while preventing somebody from learning information about an individual through a rigorous mathematical foundation. It gives strong security for sensitive data even if the adversary has robust background knowledge, but less data utility because of added noise. Because of the three v's; volume, variety, and velocity of big data, the differential privacy technique is not appropriate for big data.

The important PPDP methods are slicing, t-closeness, l-diversity, and k-anonymity. The k-anonymity method anonymizes the data to eliminate re-identification as every quasi-identifier is anonymized with a k number of feature values by the generalization technique [20–22]. In the k-anonymity algorithm, the value of k of 10 means ensuring 10 blurry tuples when an effort is made to recognize a specific individual's private data. However, in the k-anonymity technique, homogeneity attacks and background knowledge attacks compromise the de-identification of an individual's record.

The l-diversity method was introduced to eliminate the attacks that compromise the k-anonymity model. There must be l well-represented sensitive feature values like disease, marital status, and monetary values in each equivalence class, in the l-diversity method [23]. Because of the variety of data in big data, l-diversity is not possible at all times, and implementing l-diversity in big data is very complex. The feature disclosure cannot be safeguarded when the global distribution of data in the dataset is tilted into a few equivalence classes. Nevertheless, de-identification is compromised by a similarity attack in the l-diversity model.

The t-closeness model is introduced to avoid the similarity attack in the l-diversity model. An equivalence class is measured to have 't-closeness' if the distance between the distributions of sensitive feature values in the class is no more than a threshold of the real skewness of the distributed data in the t-closeness model [24].

Slicing is the model of splitting the dataset column-wise or row-wise i.e., splitting feature values and/or tuples where extremely correlated features are as a group and dissimilar features are sliced into different groups [25]. Slicing models are more appropriate for big data datasets with high dimensionality. Nevertheless, it is very complex to implement and also has practical limitations.

The randomization method adds calculated noise to the particular feature values [26]. Randomization methods have disadvantages in big data due to processing time efficiency and reduction in data utility, i.e., the accuracy of outcomes of big data analytics. The randomization method can be applied in the preprocessing step itself and then it eliminates anonymization overhead.

Cryptographic methods such as secure multiparty computation (MPC) and homomorphic encryption are used to encrypt the data before the transfer of it for data analytics [27]. In this method of secure multiparty computing, multiple parties can analyse the dataset together without revealing their data. In this technique, data utility is considerably reduced by the application of encryption. Also, it is not easy in the big data scenario to encrypt all data.

All the privacy conserving methods like slicing, t-closeness, l-diversity, and k-anonymity have been practiced for a long time, but in the big data era, their applications have practical complications resulting in processing overheads because of the enormous volume, several varieties, and continuous velocity of arriving data. The tools for big data like Hadoop, MapReduce, Pig, and Spark work on distributed parallel processing, leading to complications in employing privacy conserving methods in terms of scalability and complexity. It is practically very complex to implement and time consuming in terms of processing efficiency and it is very difficult to implement these privacy conserving methods in terms of scalability and also results in diminution of the accuracy of outcomes of data analytics i.e., data utility.

Along with generalization, anonymization (l-diversity and k-anonymity), slicing, and a method analogous to a one-way hash function called privacy view generation are used in the proposed system in [28] to safeguard the sensitive data and to avoid privacy incursions. In feature partitioning using the slicing method, extremely interrelated features are sliced as a subset of records. An equivalence class is formed for each slice of records and is anonymised by k-anonymity. Quasi-identifiers are susceptible to background knowledge attacks and hence sensitive feature values are l-diversified. However, this model is very complex to implement as it is very time consuming in big data scenarios leading to increased processing time and also data utility drastically reduced.

The recent developments in privacy conserving methods in the big data era are Map-reduce-based anonymization (MRA) [29], Scalable k-anonymization (SKA) [30, 31], improved scalable l-diversity approach [32]. The privacy conserving big data analytics methods are not effective in terms of scalability and applying these methods in big data distributed frameworks like Hadoop, MapReduce, Pig or Spark is very intricate and time-consuming and data utility is considerably reduced. The proposed Scalable Reversible Randomization Algorithm (SRRA) in this research paper is suitable for the distributed parallel processing of big data frameworks as it can be applied in the preprocessing stage itself and so it is extremely scalable. The SRRA is practically not very difficult to implement in big data compared the similar methods and also it has better privacy preservation along with feature selection stability and higher data utility.

3. PROPOSED SCALABLE REVERSIBLE RANDOMIZATION ALGORITHM (SRRA)

There will be four types of features in the dataset, such as direct identifiers, indirect or quasi-identifiers, sensitive features, and non-sensitive features. Identifier attributes uniquely identify the record, like employee ID or social security number. Quasi-identifiers are features like date of birth, occupation, religion, location, and zip code that can be indirectly used for re-identification of a record of an individual using linkage attacks. Sensitive features are features like disease or monetary values targeted by the intruder. Non-sensitive features are the features other than the above three types that are not sought by the intruder. Identifier attributes are generally anonymized before being sent for data analytics. Hence, privacy preservation techniques should be applied to quasi-identifiers and sensitive features to safeguard the privacy of individual's records.

The proposed reversible privacy conserving algorithm i.e., Scalable Reversible Randomization Algorithm (SRRA) is shown in Fig. 1 for numeric attributes and in Fig. 2 for non-numeric attributes. The reversible algorithm to get the original dataset

from the privacy preserved dataset is shown in Fig. 3. The algorithm uses a randomization technique of calculating noise addition. The algorithm is applicable for both numeric and non-numeric data and hence suitable for big data. If the data is categorical or non-numeric, the attribute values are converted into equivalent numeric values before applying the algorithm. As the characteristics of the dataset are related to the stability of feature selection and accuracy, the statistical characteristics of the dataset have minimal modification because of the privacy conserving alteration by SRRA.

In the k-anonymity technique, quasi-identifiers are generalized to at least k number of alike attribute values for different records in each equivalence class. In the l-diversity technique, which is an improved model of the k-anonymity technique, the sensitive attributes are at least l distinct domain feature values for each equivalence class that includes both generalized k number quasi-identifier values and l distinct sensitive feature values to avoid both background knowledge and similarity attack. Applying these techniques is very difficult and complex in big data scenarios because of the huge volume and multiple varieties of data and also inefficient due to long processing time, resulting in very difficult to implement in big data frameworks like Hadoop, MapReduce, Pig, and Spark. However, in the proposed algorithm SRRA, both quasi-identifiers and sensitive features are anonymized by the randomization technique of calculating noise addition efficiently with minimal practical overheads in the preprocessing stage itself.

Big data analytics is challenging for computational resources in terms of storing and processing and also for machine learning algorithms in terms of efficiency and performance. If the reduced samples are selected by attribute selection optimization algorithms such as the Pareto-front, particle swarm optimization, ant-colony, and many others, their performance in terms of efficiency and accuracy of results of the analysis will be the same or closer because, with sampled versions, we should not be able to outperform the model of the 100% data. The privacy conserving alteration by the proposed algorithm SRRA is scalable and efficient along with better stability of feature selection and accuracy results.

As the proposed algorithm, SRRA is based on the randomization technique, it can be applied to the preprocessing stage itself, hence, it is highly scalable and efficient in terms of processing time for big data. At the preprocessing stage, the data are collected from different sources and go through data extraction, data cleansing, transformation procedures, and anonymization of sensitive and quasi-identifiers by SRRA to annihilate personal identifying information and safeguard private sensitive data, followed by feature selection dimensionality reduction technique.

The sensitive and quasi-identifiers of the experimental dataset are identified by the Information Gain algorithm. The selected attributes are perturbed by the proposed Scalable Reversible Randomization Algorithm (SRRA) for privacy preservation. The application of SRRA in big data, which contains numeric and non-numeric data converted into definite sizes or formats is suitable for the application of feature selection methods. The feature selection algorithm applied on the privacy conserved dataset, which is an important preprocessing step, will select an almost similar set of attributes as the privacy conserving perturbation does not much affect the statistical characteristics of the experimental dataset.

The privacy preserved dataset is used for data analytics to get insights and hidden patterns from it. The application of SRRA results in the change in statistical characteristics of attribute values of the experimental dataset is negligible, leading to better stability of feature selection and improved data analytics results in terms of accuracy and/or error rate along with robust privacy preservation of individual's records. The original dataset can be obtained from the privacy preserved dataset using a reversible privacy conserving algorithm of SRRA.

Input of SRRA: Dataset with Chosen Sensitive Feature values and Quasi-identifier Feature values

Output of SRRA: Put on Privacy Conserving Algorithm SRRA (Scalable Reversible Randomization Algorithm) and return Privacy Conserved Dataset with Privacy Conserved feature values

1. Select Feature
2. If the selected feature is not a sensitive feature or a quasi-identifier feature value
3. Do Nought
4. Else

5. If Feature value is Non-numeric Feature go to Non-numeric Privacy Conserving Algorithm Pseudocode
6. Else
7. If Feature is Numeric
8. Divide the domain values Feature into ranges of 100 records
9. For each group of ranges
10. Allocate Original value of Feature to K
11. Allocate the value of 1 to J
12. Allocate the number of digits of K to D
13. Calculate J multiplied by 10 to the power of D and allocate the value to J1
14. Divide J1 by K and allocate the value to N
15. If the N value is less than or equal to 0.5
16. Allocate N value to N1
17. Else
18. Assign 1 – N value to N1
19. Compute N1 multiplied by 10 to the power of D and allocate the value to N2
20. Assign the lowest value of the range to X
21. Assign the highest value of the range to Y
22. Add X and Y and assign the result to Z
23. Divide Z by 2 and roundoff the result and assign to M
24. If the value of K is greater than the M
25. Subtract the calculated noise value of N2 to M and allocate it to K1, which is the privacy perturbed value of the numeric feature value
26. Else
27. Add the calculated noise value of N2 to M and allocate it to K1, which is the privacy perturbed value of the numeric feature value

Figure 1: Pseudocode of proposed numerical privacy conserving algorithm of SRRA

1. Find all non-numerical values and replace them into corresponding numerical values for every non-numerical value.
2. Divide the domain feature values into ranges of 100
3. Allocate original value of feature to K
4. Allocate the value of 1 to J
5. Allocate the number of digits of K to D
6. Calculate J multiplied by 10 to the power of D and allocate the feature value to J1
7. Divide J1 by K and allocate the feature value to N
8. If the N value is less than or equal to 0.5
9. Allocate N value to N1
10. Else
11. Allocate 1 – N value to N1
12. Compute N1 multiplied by 10 to the power of D and assign the value to N2
13. Allocate the lowest feature value of the range to X
14. Allocate the highest feature value of the range to Y
15. Add X and Y and allocate the result to Z
16. Divide Z by 2 and roundoff the result and allocate to M
17. If the value of K is greater than the value of M
18. Subtract the calculated noise value of N2 to M and allocate it to K1, which is the privacy perturbed feature value of the numeric feature value
19. Else
20. Add the calculated noise value of N2 to M and allocate it to K1, which is the privacy perturbed value of the numeric feature value

Figure 2: Pseudocode of proposed non-numerical privacy conserving algorithm of SRRA

1. Divide the feature domain values into ranges of 100
2. The actual value of the privacy perturbed dataset is K1
3. Allocate the value of 1 to J
4. Allocate the lowest value of the range to X
5. Allocate the highest value of the range to Y
6. Add X and Y and allocate the result to Z
7. Divide Z by 2 and roundoff the result and allocate to M
8. If the value of the feature K1 is less than M
9. Subtract K1 from M and allocate it to N2
10. Else
11. Subtract M from K1 and allocate it to N2
12. Allocate the value of 1 to J
13. Allocate the number of digits of K1 to D
14. Calculate J multiplied by 10 to the power of D and allocate the feature value to J1
15. Divide J1 by N2 and assign the value to N1
16. Calculate N1 multiplied by 10 to the power of D and allocate the feature value to N, the original feature value of data
17. If it is non-numeric feature value
18. Convert N into corresponding non-numeric feature value

Figure 3: Pseudocode of proposed reversible privacy conserving algorithm of SRRA

4. CONCLUSION

Privacy breaches result in severe threats to the concerned individual. This paper delivers the importance of privacy preservation and recent progress in privacy conservation methods in big data analytics. This paper analyses feature selection in big data scenarios and the importance of stability of future selection. This paper also proposes a new reversible privacy conserving algorithm using a randomization technique called Scalable Reversible Randomization Algorithm (SRRA) which can be applied to both numeric and non-numeric data. This algorithm does not affect much on stability of feature selection and accuracy and/or error rate as the statistical characteristics of the trial dataset are not much affected by the privacy conserving perturbation. The proposed algorithm is scalable and efficient for big data scenarios and provides better privacy preservation.

REFERENCES

[1] Ramy Elnaghy and Hazem M. El-Bakry, "Studying the Security and Privacy Issues of Big Data in the Saudi Medical Sector" International Journal of Advanced Computer Science and Applications (IJACSA), 14(11), 2023. (DOI): 10.14569/IJACSA.2023.01411145

[2] Duncan GT et al. Disclosure limitation methods and information loss for tabular data. In: Confidentiality, disclosure and data access: theory and practical applications for statistical agencies. 2001. p. 135–166.

[3] Duncan GT, Diane L. Disclosure-limited data dissemination. J Am Stat Assoc. 1986;81(393):10–8.

[4] Lambert Diane. Measures of disclosure risk and harm. J Off Stat. 1993;9(2):313.

[5] Liu Y et al. A practical privacy-preserving data aggregation (3PDA) scheme for smart grid. IEEE Trans Ind Inf. 2018.

[6] Spiller K, et al. Data privacy: users' thoughts on quantified self-personal data. Self-Tracking. Cham: Palgrave Macmillan; 2018. p. 111–24.

[7] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer, 2001.

[8] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," Knowledge-Based Systems, vol. 86, pp. 33–45, 2015.

[9] "A review of feature selection methods on synthetic data," Knowledge & Information Systems, vol. 34, no. 3, pp. 483–519, 2013.

[10] Albattah, W.; Khan, R.U.; Alsharekh, M.F.; Khasawneh, S.F. Feature Selection Techniques for Big Data Analytics. Electronics 2022, 11, 3177. https://doi.org/10.3390/electronics11193177

[11] Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends, MIAO RONG, DUNWEI GONG, AND X.Z. GAO, IEEE Access, VOLUME 4, 2016, DOI: 10.1109/ACCESS.2019.2894366

[12] Salem Alelyani, Huan Liu., The Effect of the Characteristics of the Dataset on the Selection Stability, IEEE DOI 10.1109/International Conference on Tools with Artificial Intelligence, 2011.167, 1082-3409/11, http:// ieeexplore.ieee.org/document/ 6103458, 2011.

[13] Salem Alelyani, Zheng Zhao, Huan Liu., A Dilemma in Assessing Stability of Feature Selection Algorithms, IEEE DOI 10.1109/ International Conference on High Performance Computing and Communications, 2011.99, 978-0-7695-4538-7/11, http:// ieeexplore.ieee.org/document/6063062, 2011.

[14] Salem Alelyani, On feature selection stability: a data perspective, Doctoral Dissertation, Arizona State University, AZ, USA, ISBN: 978-1-303-02654-6, ACM Digital Library, 2013.

[15] Barbara Pes, Feature Selection for High-Dimensional Data: The Issue of Stability, Proceedings of the 26th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2017), June 21–23, 2017.

[16] Jundong Li, Huan Liu, Challenges of Feature Selection for Big Data Analytics, Special Issue on Big Data, IEEE Intelligent Systems, eprint arXiv:1611.01875, 2017

[17] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu, Feature Selection: A Data Perspective. ACM Comput. Surv, 50, 6, Article 94, 45 pages. DOI: https://doi.org/10.1145/3136625, 2018

[18] B. Fung, K. Wang, R. Chen, P. Yu, Privacy-preserving data publishing: A survey of recent

developments, ACM Comput. Surv. 42 (4) (2010).

[19] C. Dwork, Differential privacy, Proceedings of the 33rd international conference on Automata, Languages, and Programming - Volume Part II, Pages 1–12 https://doi.org/10.1007/11787006_1pages 1-12, July 2006.

[20] Bayardo RJ, Agrawal A. Data privacy through optimal k-anonymization. In: Proceedings 21st international conference on data engineering, 2005 (ICDE 2005). Piscataway: IEEE; 2005.

[21] Iyengar S. Transforming data to satisfy privacy constraints. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2002.

[22] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD international conference on management of data. New York: ACM; 2005.

[23] Machanavajjhala A et al. L-diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd international conference on data engineering (ICDE'06), 2006. Piscataway: IEEE; 2006.

[24] Rubner Y, Tomasi T, Guibas LJ. The earth mover's distance as a metric for image retrieval. Int J Comput Vision. 2000;40(2):99–121.

[25] Tiancheng Li, Jian Zhang, Ian Molloy.: Slicing: A New Approach for Privacy Preserving Data Publishing. IEEE Transaction on KDD (2012)

[26] Aggarwal CC, Philip SY. A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data mining. Springer: US; 2008. p. 11–52.

[27] Jiang R, Lu R, Choo KK. Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data. Future Gen Comput Syst. 2018; 78:392–401.

[28] Ganesh Dagadu Puri and D. Haritha, "Implementation of Big Data Privacy Preservation Technique for Electronic Health Records in Multivendor Environment" International Journal of Advanced Computer Science and Applications (IJACSA), 14(2), 2023. (DOI): 10.14569/IJACSA.2023.0140214

[29] Zakerzadeh, H., Aggarwal, C.C., Barker, K., 2015. Privacy-preserving big data publishing. In: Proceedings of the 27th International Conference on Scientific and Statistical Database Management, SSDBM '15. ACM, La Jolla, California. https://doi.org/10.1145/2791347.2791380, pp. 26:1–26:11.

[30] Mehta, B.B., Rao, U.P., 2017. Privacy preserving big data publishing: a scalable k anonymization approach using MapReduce. IET Software 11, 271–276. https://doi.org/10.1049/iet-sen.2016.0264.

[31] Mehta, B.B., Rao, U.P., 2018. Toward scalable anonymization for privacy preserving big data publishing. Recent Findings Intell. Comput. Tech. 708, 297–304.https://doi.org/10.1007/978-981-10-8636-6. Proceedings of the 5th ICACNI 2017, vol. 2.

[32] Brijesh B. Mehta, Udai Pratap Rao, Improved l-diversity: Scalable anonymization approach for Privacy Preserving Big Data Publishing, Journal of King Saud University – Computer and Information Sciences 34 (2022) 1423–1430