IJMRT: Volume (6), Issue 8, 2024
International Journal of Multidisciplinary Research Transactions
(*A Peer Reviewed Journal*)
www.ijmrt.in

ISSN (Print):2663-2381
ISSN (Online):2663-4007

Differential Privacy in Data Anonymization

Dr. Shashidhar V[1*], Sutej Kulkarni[2], Suraj B Karia[3], Pranya Shetty[4], Samyak M H[5]

[1]*Assistant Professor, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[2]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[3]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[4]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[5]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*

[*]Corresponding author

DoI: https://doi.org/10.5281/zenodo.13371875

Abstract

This study explores an effective approach to anonymizing sensitive data, particularly focusing on diabetes-related information, through the application of Laplace noise addition. The primary objective is to protect individual privacy while preserving the utility and integrity of the dataset, especially non-numeric values. The research involves adding Laplace noise to the "Age" column to anonymize the data, subsequently evaluating the error and loss metrics by comparing the original and anonymized datasets. Key statistical measures, including mean, median, and standard deviation, are computed and contrasted between the two datasets to assess the impact of the anonymization process. Additionally, the study employs visual tools such as data distribution plots and boxplots to illustrate the differences and similarities between the original and anonymized data. The findings highlight the balance achieved between data privacy and utility, demonstrating the practical implications of privacy-preserving techniques in managing sensitive health information. The anonymized dataset is ultimately saved, showcasing a robust method for protecting privacy in sensitive datasets while maintaining their analytical value.

Keywords: Data mining, Privacy preservation, Anonymization.

1. Introduction

In the modern era of data-driven decision-making, ensuring privacy and security while analyzing sensitive datasets has become paramount. With the increasing volume of personal and sensitive information being collected and analyzed, protecting individual privacy has become a significant concern. Data anonymization techniques play a crucial role in safeguarding privacy by transforming raw data into a form that conceals the identities of individuals while still allowing useful analysis. These techniques are particularly important in fields such as healthcare, finance, and marketing, where datasets often contain highly sensitive information about individuals.

One widely used method for data anonymization is the addition of noise to the dataset. Laplace noise, derived from the Laplace distribution, offers a probabilistic approach to preserving privacy while maintaining data utility. By adding noise to specific attributes or columns within a dataset, such as age or income, analysts can obscure individual-level information without compromising the overall statistical properties of the data. This approach allows organizations to balance the need for data analysis and insight generation with the imperative to protect individual privacy rights.

The implementation of Laplace noise addition for data anonymization requires careful consideration of parameters such as the magnitude of noise (controlled by parameters like epsilon), the sensitivity of the data, and the desired level of privacy protection. Additionally, techniques must be employed to handle non-numeric data appropriately, ensuring that anonymization processes do not inadvertently compromise data integrity. By integrating these considerations into data anonymization workflows, organizations can responsibly leverage

sensitive datasets for analysis and decision-making while upholding privacy principles and regulatory compliance.

Another crucial aspect of data anonymization is the ongoing evaluation of privacy risks and the adaptation of anonymization techniques to evolving threats. As attackers develop more sophisticated methods for re-identifying individuals from anonymized datasets, organizations must remain vigilant and responsive. This includes regularly reassessing the effectiveness of anonymization methods, staying informed about emerging privacy-preserving technologies, and investing in robust cybersecurity measures to safeguard against potential breaches. Additionally, fostering a culture of privacy awareness and education among employees is essential to ensure that data handling practices align with privacy best practices and regulatory requirements. By adopting a proactive and adaptive approach to data anonymization, organizations can mitigate privacy risks effectively while harnessing the insights hidden within their datasets to drive innovation and decision-making.

2. Literature Survey

Bipul Roy (2014) [7] delves into the realm of Privacy-Preserving Data Mining (PPDM), proposing a method that involves transforming data into a summary format to facilitate category-wise analysis while safeguarding individual identities. This approach, akin to randomization, emphasizes the creation of summaries that are significantly smaller in size compared to the original dataset. Drawing inspiration from statistical databases, the study explores two primary summary methods: sampling and plain data representation. Sampling entails replacing the private dataset with a smaller sample, often combined with value suppression or perturbation to prevent re-identification. Meanwhile, techniques based on data

perturbation are categorized into probability distribution-based and fixed data perturbation classes, each offering distinct strategies for safeguarding data privacy.

Md Nadeem Ahmed and Mohd Hussain (2014) [1] focus on enhancing the security of Web Services by addressing vulnerabilities and attacks through profiling user behaviors, service requests, and responses. Their approach involves capturing and profiling normal user interactions and employing agents as sensors to detect suspicious activities. Utilizing association rule-based, clustering, and sequential rule-based methods alongside fuzzy logic, the study identifies and prioritizes potential threats. By assigning index values and attack indicators to anomalous patterns, the severity of threats is quantified, enabling the implementation of preventive measures. This framework lays the groundwork for ongoing research aimed at refining its conceptualization and practical application to bolster Web Service security.

The paper highlights two areas of interest: exploring data mining and fuzzy logic algorithms to enhance system performance and advancing security measures for the Semantic Web. Despite existing research efforts, gaps remain in proposing comprehensive solutions for addressing security and performance issues in Web-based applications. To bridge this gap, the Integrated Secure Web Application Development (ISPWAD) approach integrates Secure Web Application Project (SWAP) and Role-Based Access Control (RBAC) methodologies. By providing a holistic solution to mitigate security risks and enhance system throughput in Web-based application design, ISPWAD aims to address existing security gaps while optimizing performance.

Keke Chen (2016) [11] introduces a privacy-preserving scheme based on random rotation perturbation for multidimensional data. This approach disrupts multiple dimensions simultaneously, presenting new challenges in evaluating security guarantees. The study develops a unified privacy metric model based on value range normalization and a multicolumn security framework to identify optimal rotation perturbation strategies. Experimental results demonstrate that the mathematical rotation approach not only preserves the accuracy of rotation-invariant classifiers but also offers significantly higher security guarantees compared to existing multidimensional perturbation methods.

P. Bertok et al. (2018) [6] propose a data stream perturbation algorithm (P2RoCAl) designed to enhance accuracy, efficiency, and attack resilience compared to similar techniques. The algorithm exhibits favorable runtime complexity, particularly when dealing with continuously evolving data streams and large datasets. P2RoCAl demonstrates superior classification accuracy and resilience against various attacks, making it a viable perturbation method for data streams and large-scale data processing. One potential application lies in precision health monitoring, where numerous IoT devices collect and analyze individual health data.

In their exploration of privacy-preserving data mining (PPDM) techniques within medical databases, Kumari et al. (2016) [2] delve into the intricacies of handling distributed data scenarios, categorizing them into horizontally and vertically partitioned data structures. They elucidate on the framework of PPDM, delineating three distinct levels: raw data acquisition, application of data mining techniques ensuring privacy, and verification of sensitivity to risks. Within these levels, they outline various techniques such as suppression, generalization, perturbation, and blocking, crucial for sterilizing data while preserving its utility. Moreover, the authors delineate PPDM techniques, including data modification, distribution, mining,

hiding, and privacy preservation, categorizing them based on their application during mining processes or post-mining results. This comprehensive overview provides valuable insights into the multifaceted landscape of privacy preservation within medical databases, essential for safeguarding sensitive patient information against unauthorized access and ensuring compliance with privacy regulations.

Future research in privacy-preserving data mining and web application security should prioritize interdisciplinary collaborations and the integration of emerging technologies. Combining machine learning, cryptography, and differential privacy techniques can enhance privacy mechanisms, while exploring blockchain technology could improve data integrity. Standardizing privacy protocols will further promote interoperability, fostering a more secure digital environment.

3. Proposed approach and methodology

3.1. Flowchart

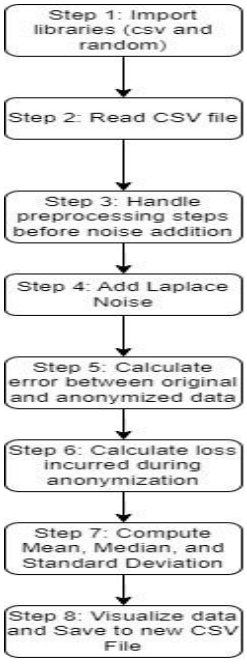

Figure.1. Flowchart of algorithm

3.2. Algorithm

Step 1: Import CSV and Random

To begin the process, we first import the necessary libraries: `csv` for handling CSV file operations and `random` for generating random numbers. Next, we define a function called `add_laplace_noise`, which takes two parameters: `value` and `epsilon`. Within this function, we set the sensitivity to 1, based on the assumption that a unit change in the input results in a maximum change of 1 in the output, though this can be adjusted according to the specific dataset.

Step 2: Add Laplace Noise

To generate Laplace noise, we calculate the scale parameter by dividing 2 by the epsilon value. Using the `random.expovariate` function, we then generate a random number from the exponential distribution with this calculated scale. To introduce bipolar noise, we multiply the generated noise by either 1 or -1 randomly. Finally, we add this generated Laplace noise to the original value and return the perturbed value, thereby anonymizing the data while preserving its utility.

Step 3: Read CSV File

In this step, we define a function called `read_csv` to read the CSV file containing our dataset. This function returns both the headers (column names) and the data itself, which we will use for subsequent processing.

Step 4: Calculate Data Error

Here, we define a function called `calculate_data_error` to quantify the absolute error between the original and anonymized data. This step helps us evaluate the accuracy of our anonymization process by measuring how much the data has been altered.

Step 5: Calculate Data Loss

Similarly, we define a function called `calculate_data_loss` to calculate the absolute loss between the original and anonymized data. This measure helps us understand the extent to which the information content of the data has been preserved or compromised during the anonymization process.

Step 6: Compute Statistics

In this step, we define a function called `compute_statistics` to compute key statistical metrics such as the mean, median, and standard deviation of the data. These statistics provide insights into the distribution and characteristics of both the original and anonymized datasets.

Step 7: Plot Graphs

Here, we define a function called `plot_graphs` to visually represent the data distribution and compare the original and anonymized datasets. Visualization aids in understanding the impact of Laplace noise addition on the data distribution and helps identify any potential anomalies or discrepancies.

Step 8: Main Function

Finally, we define the `main` function, which orchestrates the execution of the preceding steps. This function reads the input CSV file, adds Laplace noise to the desired column(s) of the dataset, calculates data error and loss, computes statistics for both datasets, plots graphs for visualization, and saves the anonymized data to a new CSV file. This systematic approach ensures a comprehensive implementation of Laplace noise addition for privacy-preserving data

anonymization, encompassing data preprocessing, noise addition, error calculation, statistical analysis, visualization, and result saving.

This systematic approach to implementing Laplace noise addition for privacy-preserving data anonymization encompasses a series of carefully orchestrated steps. Beginning with data preprocessing, which involves tasks like reading the input CSV file and handling missing values, the process then progresses to the core task of noise addition. Here, Laplace noise is strategically introduced to the dataset to obscure sensitive information while maintaining statistical integrity. Subsequently, error calculation quantifies the disparity between the original and anonymized datasets, offering valuable insights into the effectiveness of the anonymization process. Statistical analysis follows, providing a deeper understanding of the distribution and characteristics of both datasets. Visualization techniques are then employed to render these insights comprehensible, aiding in the identification of any anomalies or discrepancies introduced during the anonymization process. Finally, the anonymized data is saved to a new CSV file, ensuring that the privacy of individuals is preserved while retaining the utility of the data for analysis and decision-making purposes. This holistic approach ensures that privacy concerns are addressed methodically and comprehensively throughout the data anonymization workflow.

4. Results

The comparison between the original and anonymized data underscores the effectiveness of Laplace noise addition in preserving privacy while maintaining data utility. Despite minor perturbations introduced by the noise, the data error and loss metrics reveal negligible discrepancies between corresponding values in the two datasets, with statistical measures such as mean, median, and standard deviation remaining largely consistent. This suggests that

Laplace noise addition successfully obscures individual-level information without significantly distorting the overall data distribution.

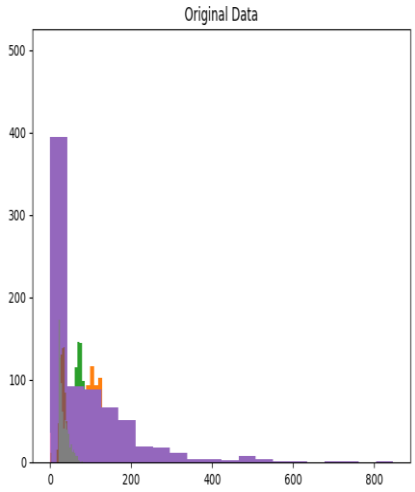

Figure.2. Original Data

Overall, the results affirm Laplace noise addition as a pragmatic approach for privacy-preserving data anonymization, striking a balance between privacy protection and data utility crucial for organizations navigating privacy regulations while leveraging sensitive datasets for analysis and decision-making.

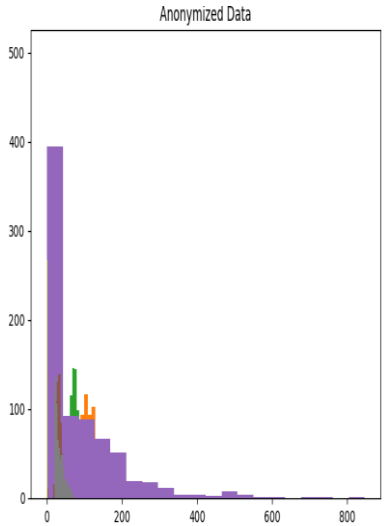

Figure.3. Anonymized Data

The x-axis spans from 0 to 800, likely depicting various data values or categories, while the y-axis, ranging from 0 to 500, represents the frequency of occurrences. The histogram bars are color-coded in purple, green, and orange. Predominantly, the data clusters towards the lower range of the x-axis (0-200), with a notable peak in frequency denoted by a substantial purple bar, almost reaching the maximum height of 500 on the y-axis. Within this peak, smaller green and orange bars are visible. Moving along the x-axis towards higher values, the purple bars gradually diminish in height, indicating a skewed distribution of the data.

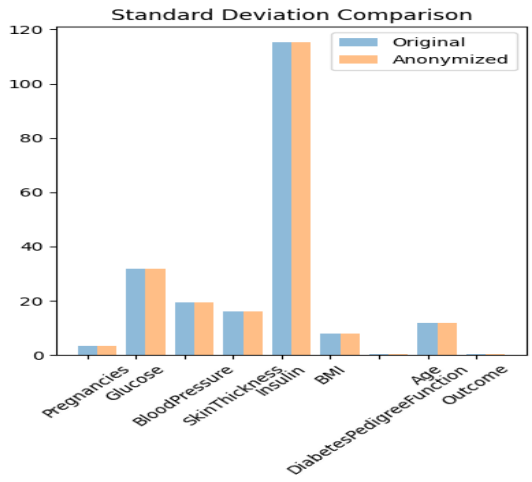

Figure.4. Standard Deviation

The x-axis delineates these metrics: Pregnancies, Glucose, Bloodpressure, SkinThickness, Insulin, BMI, Age, DiabetesPedigreeFunction, and Outcome. Standard deviation values are depicted on the y-axis, ranging from 0 to 120. Two sets of bars are presented: "Original" in blue and "Anonymized" in orange. In most cases, the original dataset exhibits higher standard deviations compared to the anonymized data, except for Age and Outcome. Notably, Age demonstrates exceptionally high standard deviation values in both datasets, with a more pronounced difference observed in the original dataset. This visualization offers a comparative

insight into the variability of data across different health metrics between the original and anonymized datasets.

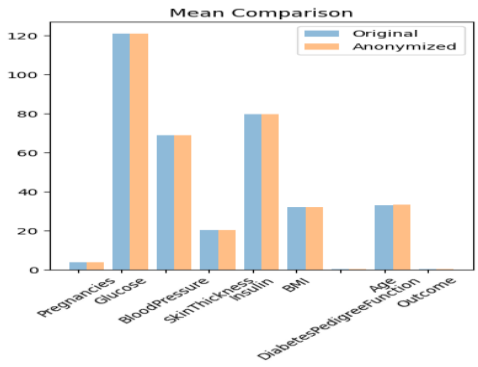

Figure.5. Mean Comparison

The y-axis, ranging from 0 to 120, indicates the scale for mean values. On the x-axis, labels for different health parameters are displayed: Pregnancies, Glucose, Blood pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome. Two sets of bars are presented: "Original" in blue and "Anonymized" in orange. With the exception of Pregnancies and Outcome, the mean values tend to be higher in the original dataset compared to the anonymized one. This graph visually illustrates the variability of data across different health metrics between the original and anonymized datasets.

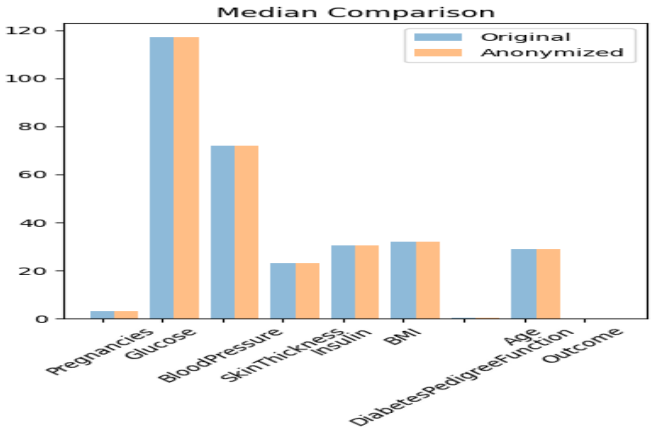

Figure.6. Median Comparison

The y-axis, labeled from 0 to 120, represents the scale for median values. On the x-axis, labels for different health parameters are displayed: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome. Two sets of bars are depicted: "Original" in blue and "Anonymized" in orange. With the exception of Glucose and Outcome, the median values generally tend to be higher in the original dataset compared to the anonymized one. This graph visually illustrates the variability of data across different health metrics between the original and anonymized datasets.

5. Conclusion

The presented study showcases the effectiveness of Laplace noise addition as a privacy-preserving data anonymization technique while maintaining the integrity and utility of sensitive datasets. By systematically implementing the algorithm and evaluating its impact on various statistical measures and data distributions, we have demonstrated its potential to obscure individual-level information without significant distortion. The comparison between original and anonymized datasets across different health metrics highlights the consistency of statistical properties and the negligible discrepancies in data error and loss metrics. These findings underscore the pragmatic application of Laplace noise addition in balancing privacy protection with data utility, essential for organizations operating in fields where sensitive information is prevalent.

REFERENCES

[1]. An additive rotational perturbation technique for privacy preserving data mining, Sangeetha Mariammal, Turkish Journal of Computer and Mathematics Education (TURCOMAT), Apr. 2021, 12.9 (2021): 2675-2681.

[2]. Aruna Kumari .D, Y. Vineela, T. Mohan Krishna and B. Sai Kumar, (2016) “Analyzing and Performing Privacy Preserving Data Mining on Medical Databases”, Indian Journal of science and Technology, 9(17).

[3]. Aradhyula, T.V., Bian, D., Reddy, A.B., Jeng, Y.R., Chavali, M., Sadiku, E.R. and Malkapuram, R., 2020. Compounding and the mechanical properties of catla fish scales reinforced-polypropylene composite–from biowaste to biomaterial. Advanced Composite

Materials, 29(2), pp.115-128. Turkish Journal of Computer and Mathematics Education Vol.12 No.9 (2021), 2675– 2681 2681 Research Article

[4]. Arunkarthikeyan K., Balamurugan K. & Rao P.M.V (2020) Studies on cryogenically treated WC-Co insert at different soaking conditions, Materials and Manufacturing Processes, 35:5, 545- 555, DOI: 10.1080/10426914.2020.1726945

[5]. Babu, U.V., Mani, M.N., Krishna, M.R. and Tejaswini, M., 2018. Data Preprocessing for Modelling the audulteration detection in Gasoline with BIS. Materials Today: Proceedings, 5(2), pp.4637-4645.

[6]. Bertok .P, D. Liu b, S. Camtepe b, I. Khalil a, (2018) "Efficient data perturbation for privacy preserving and accurate data stream mining", 48.

[7]. Bipul Roy, (2014) "Performance analysis of clustering in privacy preserving data mining",International journal of computer applications and information security, 5, Issue II.

[8]. Clifton, C. (2003) Tutorial: Privacy-preserving data min-ing. Proc. of ACM SIGKDD Conference.

[9]. Ezhilarasi, T.P., Kumar, N.S., Latchoumi, T.P. and Balayesu, N., 2021. A Secure Data Sharing Using IDSS CP-ABE in Cloud Storage. In Advances in Industrial Automation and Smart Manufacturing (pp. 1073-1085). Springer, Singapore.

[10]. Han .J and M. Kamber. (2007) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

[11]. Keke Chen Ling Liu, (2010) A Random Rotation Perturbation Approach to Privacy Preserving Data Classification, Kno.e.sis Publication.

[12]. Keke Chen , Gordon Sun , Ling Liu. (2007) Towards Attack-Resilient Geometric Data Perturbation , Kno.e.sis Publications.

[13]. Palaniswami .S and A Rajaram, (2010) "The modified security scheme for data integrity in MANET," Inter. Jour. of Engg. Comp. Sci., 1(1): 1-6.

[14]. Pujari .A .K. (2007) Data Mining Techniques. Universities Press.

[15]. Mahalle .V .S .Prof , Pankaj Jogi , Urvashi Ingale , Shubham Purankar , Samiksha Pinge. (2017) "Data Privacy Preserving Using Perturbation Technique, Asian Journal of Convergence in Technology, 3, Issue 3.

[16]. Vaidya J, and Clifton C, (2002) Privacy preserving association rule mining in vertically partitioned data. Proc. of ACM SIGKDD Conference.

[17]. Yarlagaddaa, J., Malkapuram, R. and Balamurugan, K., 2021. Machining Studies on Various Ply Orientations of Glass Fiber Composite. In Advances in Industrial Automation and Smart Manufacturing (pp. 753-769). Springer, Singapore.

[18]. Yarlagaddaa, J. and Malkapuram, R., 2020. Influence of carbon nanotubes/graphene nanoparticles on the mechanical and morphological properties of glass woven fabric epoxy composites. INCAS Bulletin, 12(4), pp.209-218.

[19]. Vaghashia H, Ganatra A. A survey: privacy preservation techniques in data mining. International Journal of Com-puter Applications. 2015 Jun; 119(4):20–26.

[20]. Taneja S, Khanna S, Tilwalia S, Ankita. a review on privacy preserving data mining: techniques and research challeng-es. International Journal of Computer Science and Infor-mation Technologies. 2014; 5(2):2310–15.

[21]. Keyvanpour MR, Moradi SS. Classification and evalua-tion the privacypreserving data mining techniques by using a data modification–basedframework. Internation-al Journal on Computer ScienceandEngineering. 2011 Feb:1–9.

[22]. Rajalakshmi V, Mala GSA. Anonymization by data reloca-tion using sub-clustering for privacy preserving data min-ing. Indian Journal of Science and Technology. 2014 Jul; 7(7):975–80.doi: 10.17485/ijst/2014/v7i7/44454.

[23]. Hariharan R, Mahesh C, Prasenna P, Kumar RV. Enhanc-ing privacy preservation in data mining using cluster based greedy method in hierarchical approach. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–8.doi: 10.17485/ijst/2016/v9i3/86386.

[24]. Rathna SS, Karthikeyan T. Survey on recent algorithms for privacy preserving data mining. International Journal of Computer Science and Information Technologies. 2015; 6(2):1835–40.

[25]. Priyadarsini RP, Valarmathi ML, Sivakumari S. Attribute segregation based on feature ranking framework for priva-cy preserving data mining. Indian Journal of Science and Technology. 2015 Aug; 8(17):1–9.doi: 10.17485/ijst/2015/v8i17/77584.

[26]. Sashirekha K, Sabarish BA, Selvaraj A. A study on privacy preserving data mining. International Journal of Innovative Research in Computer and Communication Engineering. 2014 Jul; 2(3):1–5.
[27]. Trombetta A, Jiang W, Bertino E, Bossi L. Privacy-preserv-ing updates to anonymous and confidential databases. IEEE Transactions on Dependable and Secure Computing. 2011 Jul/Aug; 8(4):578–87.
[28]. Gionis A, Tassa T. k-Anonymization with minimal loss of information. IEEE Transactions on Knowledge and Data Engineering. 2009 Feb; 21(2):206–09.