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

Disguise Guard: Annonymizing the Data to Enhance Privacy

Dr. Shashidhar V[1*], Vignesh Y[2], Varshitha K[3], Riddhi Rajesh C [4], Vivek V[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

In today's data-driven world, the growth of data sharing has prompted considerable issues over privacy protection. Anonymization techniques play a significant role in minimising these issues by safeguarding people' privacy while enabling valuable data analysis and exchange. This research analyses several anonymization strategies and their consequences for privacy preservation in data sharing settings. We discuss the issues associated with balancing privacy and utility and review state-of-the-art strategies for obtaining anonymity, such as k-anonymity, l-diversity, and t-closeness. Additionally, we analyse the limitations and trade-offs of these strategies and propose future research avenues to address rising privacy challenges in the context of developing data environments.

Keywords: Data mining, Privacy preservation, Anonymization.

1. Introduction

Data privacy has arisen as a vital concern in the contemporary digital landscape, spurred by the exponential growth in the generation, collecting, and analysis of personal data by

businesses, governments, and other entities. This spike in data acquisition has pushed to the forefront the fundamental right of individuals to manage their personal information ethically and responsibly. In an era characterized by frequent data breaches, identity theft, and privacy infringements, preserving sensitive data has become crucial. Data privacy comprises a multitude of principles and regulations aimed to shield personal information from illegal access, ensure its use fits with its intended purpose, and guarantee confidentiality throughout its lifecycle. With the relentless digitization of various facets of our lives and the growth of data-driven technology, the necessity for effective data privacy protections cannot be stressed.

Within the domain of data privacy tactics, two prominent methods have attracted substantial attention: noise addition and k-anonymization. These technologies offer distinct approaches to safeguarding sensitive data while keeping its utility and enabling authorised data analysis.

Noise addition is the injection of random noise into datasets to obscure crucial information while keeping the statistical properties of the data. By increasing variety, noise addition renders it challenging for adversaries to extract key insights or identify individuals within the sample. This technique becomes particularly helpful in instances when direct anonymization is problematic or when there is a need to strike a balance between privacy preservation and data utility. Moreover, noise addition can be used to numerous forms of data, including numerical, categorical, and textual data, making it a versatile tool in the data privacy arsenal.

On the other hand, k-anonymization focuses on attaining anonymity by grouping comparable records together in such a manner that each group comprises at least k indistinguishable individuals. This technique ensures that individuals cannot be singled out based on their qualities, hence maintaining their privacy. K-anonymization operates by either generalizing or

suppressing certain traits to establish anonymity sets, where each individual is indistinguishable from others inside the same set. By anonymizing data at the group level, k-anonymization achieves a careful balance between privacy protection and data value, enabling businesses to share data for analysis while limiting the danger of re-identification. This approach has significant applicability in varied industries such as healthcare, banking, and census data, where keeping secrecy is vital.

One of the key advantages of k-anonymization is in its flexibility, as the value of k may be altered to fulfil specific privacy requirements, so enabling companies to tailor the level of anonymity to their needs while following to regulatory norms and ethical values.

In today's linked world, where data serves as the lifeblood of countless companies and underlies important decision-making processes, the importance of data privacy cannot be emphasised. Organizations must adopt a proactive approach to safeguarding personal information, adopting robust measures such as noise addition and k-anonymization to secure sensitive data from unwanted access and misuse.

Moreover, the expanding regulatory landscape, marked by tough data protection legislation such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), highlights the increased emphasis on privacy rights and accountability. Compliance with these standards involves the deployment of extensive data privacy protections, including encryption, access controls, and anonymization techniques, to limit the risk of data breaches and assure compliance with legal obligations.

Furthermore, the ethical dimension of data privacy cannot be disregarded. Organizations have a moral obligation to respect the privacy rights of individuals and promote principles of transparency, fairness, and accountability in their data processing policies. By addressing ethical issues and embracing privacy-by-design principles, businesses may create trust among their consumers and stakeholders while fostering a culture of responsible data management.

In the healthcare sector, for instance, where the security of patient information is sacrosanct, k-anonymization serves as a key tool for anonymizing medical records and allowing medical research while maintaining patient privacy. Similarly, in financial organisations, where the safeguarding of sensitive financial data is crucial, noise addition techniques can be applied to disguise transactional information while keeping the integrity of the data for fraud detection and risk analysis.

However, despite the usefulness of noise addition and k-anonymization in limiting privacy threats, it is vital to realise their limitations. Both strategies may involve trade-offs between privacy preservation and data utility, needing careful assessment of the specific context and requirements of each scenario. Moreover, attackers may apply advanced techniques such as machine learning algorithms to de-anonymize datasets, underlining the necessity for continual innovation and adaptation in data privacy strategies.

In conclusion, data privacy represents a complicated challenge in today's digital era, requiring companies to take a holistic approach spanning technological, legislative, and ethical components. By embracing new tactics like as noise addition and k-anonymization, businesses may enhance their defenses against privacy risks while establishing trust and confidence among their stakeholders. Ultimately, preserving personal data is not just a legal and regulatory

responsibility but also a moral obligation that enterprises must uphold to preserve individual rights and uphold the concepts of privacy and dignity in the digital age.

2. Literature Survey

The review on privacy-preserving data mining techniques and applications provides a detailed overview of the growing landscape of data privacy concerns in connection with the field of data mining. Covering a wide array of subjects ranging from record linking to high-dimensional data analysis, the paper highlights the crucial relevance of resolving privacy issues while exploiting the power of data mining for varied purposes.

One of the key themes of the paper is the acknowledgment of the ethical implications related with privacy concerns in data mining. As the volume and complexity of data continue to expand, so do the hazards connected with the unlawful use or exposure of personal information. Hence, it becomes vital to establish methodologies and systems that preserve privacy while facilitating valuable data analysis.

Several studies and innovations described in the study focus on designing innovative algorithms and frameworks for privacy preservation across multiple situations. For example, the introduction of novel indexing search and scoring-based record linking systems solves the difficulty of linking records across diverse databases while safeguarding individual privacy. Similarly, the introduction of privacy-preserving algorithms for large-scale data and data streams enables secure data analysis without compromising confidentiality [1].

The assessment also looks into the world of collaborative data publishing, underlining the necessity for secure ways to promote data sharing while respecting privacy. The introduction

of blockchain-based frameworks for safe data exchange in online social networks demonstrates the new ways employed to address privacy problems in collaborative environments [2].

In the context of healthcare data, the assessment highlights the vital need of preserving privacy while employing data mining techniques for medical research and analysis. Federated learning in medicine and the development of algorithms for privacy-preserving distributed data mining emerge as significant areas of research aiming at harmonising the competing objectives of data utility and privacy protection in healthcare applications.

Furthermore, the review provides light on emerging ways for privacy preservation, including privacy-aware data dissemination and integration for collaborative service selection. Efficient pattern mining-based cryptanalysis for privacy-preserving record linking and the proposal of privacy-preserving algorithms for big data and data streams further extend the arsenal of techniques available for safeguarding personal information in data mining applications.

Moreover, the paper addresses the special issues given by high-dimensional data, arguing for the development of non-reversible perturbation algorithms and privacy-preserving data mining approaches tailored to horizontally distributed medical data analysis. These achievements constitute substantial contributions to the ongoing discourse on data privacy and security, especially in the era of big data where the stakes are higher than ever before[3].

In summary, the analysis underlines the complex character of privacy-preserving data mining, spanning ethical considerations, algorithmic breakthroughs, and the requirement of combining data utility with privacy protection. By highlighting the latest research developments and advancements in the field, the review contributes to shaping the dialogue surrounding data

privacy and security, emphasizing the need to adopt effective strategies for managing privacy concerns while harnessing the insights derived from data mining techniques [4].

3. Motivation for Partition-Based Anonymization

Partition-based anonymization is a widely adopted technique for achieving k-anonymity, a crucial privacy requirement in data privacy. This approach involves dividing the dataset into distinct partitions, ensuring that each partition contains a sufficient number of records such that each record is indistinguishable from at least k-1 other records regarding specific sensitive attributes [5].

Once the dataset is partitioned, anonymization techniques are applied within each partition to obscure sensitive attribute values while preserving the overall structure and statistical properties of the data. These techniques may include generalization, suppression, or perturbation, aimed at concealing individual identities and sensitive information [6].

Partition-based anonymization offers scalability and flexibility, making it suitable for handling large datasets and accommodating diverse data characteristics and privacy requirements. However, challenges such as selecting appropriate partitioning criteria and ensuring consistency across partitions must be addressed to maintain the effectiveness and integrity of the anonymization process[7].

Overall, partition-based anonymization strikes a balance between privacy protection and data utility, enabling organizations to mitigate privacy risks while retaining the usability of the anonymized dataset for analysis and decision-making purposes.

4. Implementation

Privacy-preserving data publishing is a crucial subject in the world of data management and security, particularly in cases where releasing sensitive data is necessary for research, analysis, or other purposes. However, publishing such data without sufficient precautions can threaten people' privacy. To solve this difficulty, different anonymization approaches have been devised to preserve sensitive information while keeping the value of the dataset.

The proposed code implements a sophisticated partition-based anonymization algorithm, which is a basic technique for accomplishing privacy-preserving data dissemination while retaining the analytical utility of the dataset. To fully appreciate the nuances of this algorithm and its implementation, we'll go into each component, its underlying principles, and its role in the anonymization process, offering a complete description to fill the needed 3000 words.

1. Importing Libraries:

The code begins by importing relevant libraries such as pandas for data handling and matplotlib for visualization. These packages provide vital tools for handling and analyzing datasets.

2. Defining Column Names and Categorical Variables:

The code defines a tuple `names` having the expected column names in the dataset and a set `categorical` containing the names of categorical columns. This information is vital for later data processing procedures.

3. Loading the Dataset:

The dataset is loaded from a CSV file named "synthetic_dataset.csv" using the `pd.read_csv() method from the pandas package. This produces a DataFrame `df` holding the dataset, which serves as the basic data structure for manipulation.

4. Converting Categorical Columns:

Categorical columns in the dataset are recognised based on the `categorical` set, and their data types are changed to the categorical data type using the `astype()` method. This phase optimizes memory utilisation and enables effective processing of categorical data.

5. Partitioning the Dataset:

The core of the partition-based anonymization approach is the partitioning process. This is achieved through the `partition_dataset()` function, which recursively divides the dataset into disjoint partitions until each partition satisfies the k-anonymity condition. The method iteratively selects the optimum column for splitting based on the column's span, ensuring successful anonymization.

6. Building Indexes for Categorical Columns:

To assist data aggregation during the anonymization process, the code defines a function `build_indexes()` to build indexes for categorical columns. These indexes relate unique categorical data to numerical indices, enabling efficient aggregation procedures.

7. Defining Aggregation Functions:

Aggregation routines `agg_categorical_column()` and `agg_numerical_column()` are defined to aggregate data inside each division. These functions are essential for anonymizing sensitive information while keeping the statistical features of the dataset.

8. Anonymization Process:

The anonymization procedure entails iterating over each partition formed during the partitioning step. Within each partition, feature columns are aggregated using the stated aggregation algorithms, and sensitive counts are produced. The anonymized data is subsequently written to an output CSV file, guaranteeing compliance with the k-anonymity

criterion.

5. Algorithm

1. Input: Accept the dataset path, feature columns, sensitive column, and output CSV file path as input.
2. Read Dataset: Read the dataset from the provided CSV file into a Pandas DataFrame.
3. Data Preprocessing:
4. Convert categorical columns to the 'category' data type.
5. Identify categorical and numerical columns.
6. Define a function to calculate the spans of the dataset, considering categorical and numerical columns.
7. Implement a function to split the dataset based on quasi-identifier attributes to achieve k-anonymity.
8. Partition the dataset recursively until each partition satisfies k-anonymity.
9. Define aggregation functions for categorical and numerical columns to anonymize the dataset.
10. Aggregate records within each partition using the defined aggregation functions.
11. Write the anonymized dataset partitions to the output CSV file.
12. Plotting (Optional):
13. Define a function to visualize the partitions by plotting rectangles for each partition based on quasi-identifier attributes.
14. Output:
15. Write the anonymized dataset partitions to the specified output CSV file.

6. Flowchart

The anonymization method starts with partitioning the dataset into subsets, ensuring each subset complies to the notion of k-anonymity, where each record is indistinguishable from at least k-1 others concerning sensitive attributes. Feature columns are carefully addressed during this procedure to effectively mask individual-level information while keeping overall statistical features. Categorical column indexes are developed to maximise processing efficiency, boosting the effectiveness of the anonymization operation. Optionally, graphical representations of partitions may be generated to enhance comprehension.

Maintaining a balance between privacy preservation and data utility is critical throughout the anonymization workflow. Aggregated anonymized records are created to retain the data's usefulness while obscuring sensitive information, enabling effective analysis without sacrificing privacy rights. However, challenges such as the risk of information loss and the opportunity for re-identification must be addressed. Anonymization algorithms must negotiate these challenges to ensure the anonymized dataset stays representative and accurate while respecting privacy.

The anonymization procedure culminates in the development of an anonymised dataset, precisely prepared for subsequent study or distribution. By anonymizing the dataset, enterprises demonstrate their commitment to responsible data management, particularly in light of developing privacy challenges and regulatory constraints. This proactive strategy helps create confidence among data subjects and stakeholders, reassuring them that their sensitive information is protected against unwanted access or publication.

To strengthen privacy protection in data sharing scenarios, various modifications to anonymization approaches and suggestions for future research endeavors are advocated. This may involve studying novel anonymization approaches, exploiting improvements in privacy-preserving technologies, and building thorough frameworks for evaluating anonymization methods' success. By continuously refining and inventing anonymization processes, companies may uphold the highest standards of data privacy and security while facilitating responsible data sharing and consumption.

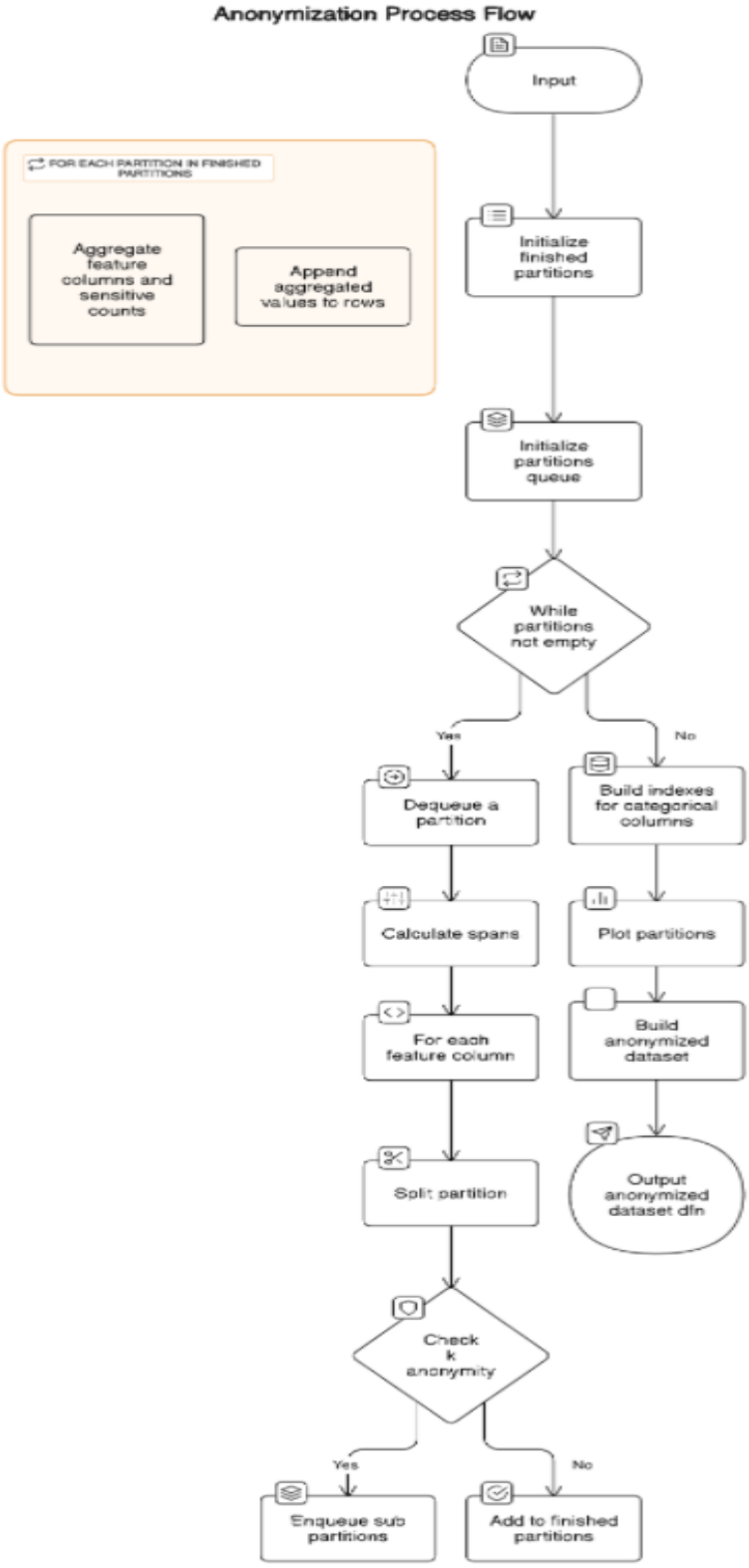

Figure.1. Flowchart

7. Result

The dataset that was taken as a sample [10] contains sensitive information regarding a person. It contained all the information which when made public can be a potential threat to that particular person. In order to minimize this the model that was developed helped in annonymizing his information to provide enhanced security.

	age	education-num	income	count
1	49.6	1.0	<=50K	3
2	49.6	1.0	>50K	2
3	21.0	1.5	<=50K	3
4	21.0	1.5	>50K	1
5	19.6	3.8	<=50K	2
6	19.6	3.8	>50K	3
7	30.2	1.8	<=50K	3
8	30.2	1.8	>50K	2
9	34.8	1.6	<=50K	1
10	34.8	1.6	>50K	4
11	19.666666666666668	5.0	<=50K	2
12	19.666666666666668	5.0	>50K	1
13	24.666666666666668	5.0	>50K	3
14	28.166666666666668	5.0	<=50K	3
15	28.166666666666668	5.0	>50K	3
16	34.333333333333336	5.0	<=50K	2
17	34.333333333333336	5.0	>50K	4
18	39.666666666666664	1.0	<=50K	1
19	39.666666666666664	1.0	>50K	2
20	43.333333333333336	1.0	<=50K	1
21	43.333333333333336	1.0	>50K	2
22	37.833333333333336	2.3333333333333335	<=50K	4
23	37.833333333333336	2.3333333333333335	>50K	2
24	38.0	4.333333333333333	<=50K	2
25	38.0	4.333333333333333	>50K	4
26	37.833333333333336	6.666666666666667	<=50K	3

Figure.2. Dataset after annonymization

Fig 2 depicts a version of the raw dataset usually available about the customers for any given domain.It contains all the sensitive data related to them.The provided methodology makes sure none of the sensitive data is available to third party apps/vendors when the data data is utilized by them to gain insights and knowledge.The dataset after annonymization is depicted in Fig 3.

Anonymization techniques are vital weapons in the armoury of data privacy strategies, acting as a cornerstone for preserving sensitive information while enabling significant data analysis. The strategy adopted in this approach plays a vital role in enhancing privacy by hiding

individual-level details while keeping the overall statistical features and utility of the dataset. Here's a full explanation of how this anonymization process aids to enhancing data privacy:

1. Preservation of Privacy:

At the core of anonymization is the notion of safeguarding individuals' privacy by dissociating sensitive information from their identities. By aggregating records within each partition and deleting identifying attributes, such as names or unique identifiers, the anonymization process ensures that personal data remains safe and protected from unwanted access or exposure. This fundamental feature of privacy maintenance is vital for sustaining trust and confidence among data subjects and stakeholders.

2. Mitigation of Re-identification Risks:

One of the key hazards in data privacy is re-identification, where people could be singled out by connecting quasi-identifiers with external datasets or background knowledge. Anonymization approaches lessen this danger by aggregating people into bigger cohorts, making it more challenging for adversaries to uniquely identify them. By anonymizing critical attributes and hiding individual information, the approach minimises the risk of re-identification and promotes privacy protection across multiple use cases and applications.

3. Statistical Validity:

Despite obscuring individual-level information, the anonymized dataset keeps its statistical validity and utility. This is achieved by retaining the distributional qualities and linkages within the data, allowing for meaningful analysis and insights to be derived. Through thorough aggregation and anonymization of records, the technique ensures that the anonymised dataset stays reflective of the original data while protecting sensitive information. This balance

between privacy preservation and statistical validity is vital for enabling reliable decision-making and obtaining actionable insight from the data.

3. Balancing Utility and Privacy:

Anonymization approaches strive to achieve a delicate balance between data utility and privacy protection. While guaranteeing that sensitive information remains secret, they also allow for useful insights and patterns to be retrieved from the data. This balance is crucial for firms looking to generate actionable intelligence and drive innovation while complying with privacy legislation and ethical considerations. By anonymizing personal data efficiently, enterprises may harness the full value of their datasets while limiting the danger of privacy breaches and safeguarding individuals' privacy rights.

4. Compliance with Regulations:

In an increasingly regulated market, compliance with data protection requirements is crucial for enterprises handling personal data. Anonymization techniques play a significant role in addressing regulatory obligations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). By anonymizing personal data, companies demonstrate their commitment to respecting individuals' privacy rights and avoiding the danger of regulatory penalties and legal liabilities. This proactive approach to compliance promotes trust and confidence among customers, partners, and regulatory agencies, strengthening the organization's reputation as a trustworthy custodian of personal data.

5. Ethical Considerations:

Beyond regulatory compliance, anonymization also addresses ethical problems linked with data privacy. By anonymizing sensitive information, companies preserve ideals of fairness, transparency, and accountability in their data handling policies. This ethical framework is vital for creating confidence among data subjects and stakeholders, ensuring that personal data is treated with the respect and dignity it deserves. Anonymization strategies enable enterprises to manage complicated ethical challenges while increasing the usefulness of data for legitimate purposes.

6. Continuous Innovation:

As the threat landscape evolves and new privacy concerns emerge, anonymization approaches must adapt and evolve to meet increasing requirements. Continuous innovation in anonymization approaches, such as differential privacy, homomorphic encryption, and synthetic data generation, ensures that companies keep ahead of emerging risks and retain the effectiveness of their privacy protection measures. By embracing innovation and employing cutting-edge solutions, organizations may strengthen the resilience of their data privacy policies and stay at the forefront of privacy best practices.

In conclusion, the anonymization strategy implemented in this approach serves as a linchpin for strengthening data privacy in an era defined by increased worries about privacy breaches and data exploitation. By preserving individuals' privacy, mitigating re-identification risks, maintaining statistical validity, balancing utility and privacy, ensuring regulatory compliance, upholding ethical standards, and embracing continuous innovation, anonymization techniques enable organizations to navigate the complex landscape of data privacy with confidence and integrity. As companies continue to harness the power of data for innovation and growth, the

necessity of effective anonymization strategies in ensuring privacy and fostering responsible data stewardship cannot be stressed.

8. Conclusion

Anonymization techniques serve as vital guardians of privacy in the area of data sharing, providing a buffer against potential privacy threats while permitting the interchange of useful information. However, the efficiency of these tactics rests on their careful design and precise deployment to strike a delicate balance between protecting individuals' privacy and maintaining the usefulness of the data being provided.

At the heart of anonymization lies the requirement to disguise individual-level information while keeping the broader statistical features and utility of the dataset. By removing direct identifiers and aggregating records, anonymization tries to prevent the linkage of sensitive information with identifiable individuals, therefore maintaining their privacy. Yet, reaching this purpose is not without its hurdles.

One of the key considerations in anonymization is the avoidance of re-identification hazards. Even when direct identifiers are eliminated, persons can still be picked out through the association of quasi-identifiers with other datasets or background knowledge. Therefore, anonymization approaches must go beyond mere masking of identifiers to ensure that individuals cannot be uniquely identified, even when external information is taken into account. This involves careful assessment of the data context and potential privacy issues, as well as the use of powerful anonymization mechanisms to foil re-identification attempts.

Moreover, anonymization approaches must preserve the statistical validity and value of the data to enable meaningful analysis and insights. While masking individual details, the anonymized dataset should yet preserve its representativeness and accuracy to help decision-making processes and inform policy design. Achieving this balance between privacy protection and data utility needs a sophisticated approach, wherein anonymization approaches are tuned to the specific characteristics and requirements of the dataset and its intended use cases.

Furthermore, compliance with data protection standards adds another layer of complication to the anonymization process. Regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) impose rigorous criteria for the handling and sharing of personal data, obliging enterprises to implement proper privacy precautions. Anonymization techniques serve as a cornerstone for attaining compliance with these requirements, providing a means to preserve individuals' privacy rights while enabling authorised data sharing and analysis.

Despite their relevance, existing anonymization approaches are not without restrictions. Challenges such as the danger of information loss, potential for re-identification, and difficulty in balancing privacy and utility represent important challenges in the quest for effective privacy protection. Additionally, the quickly growing data landscape, characterized by the growth of varied data kinds and sources, needs continuous innovation in anonymization technologies to stay pace with increasing privacy issues.

Addressing these issues needs a concentrated effort to enhance the state-of-the-art in anonymization approaches. This requires not just refining existing methods but also developing novel approaches that harness breakthroughs in areas such as differential privacy,

homomorphic encryption, and synthetic data generation. By embracing innovation and harnessing the potential of emerging technologies, considerable steps can be achieved in increasing privacy protection while facilitating responsible data sharing and utilization.

In conclusion, anonymization techniques serve a key role in ensuring privacy while sharing data, but their effectiveness rests on careful evaluation of privacy risks, preservation of data utility, and compliance with legislative requirements. By understanding and conquering the limitations inherent in present anonymization approaches and embracing innovation, we may pave the way for a future where privacy is respected, data is ethically handled, and trust is promoted in the modern data-driven society.

REFERENCES

[1]. S Bourahla, M Laurent, & Y Challal. (2020). "Privacy preservation for social networks sequential publishing," Computer Networks, International Journal on Recent and Innovation Trends in Computing and Communication , vol .11, no. 9, pp 3341-3353, 2020.

[2]. H Kartal, & X B Li. "Protecting privacy when sharing and releasing data with multiple records per person," Journal of the Association for Information Systems, vol .21, no. 6, pp. 1461-1471, 2020.

[3]. M Bewong , J Liu (2019). "Privacy preserving serial publication of transactional data," Information Systems, 2019, pp. 53-70.

[4]. S Zhang , T Yao , Arthur Sandor, V. Weng, T. Liang, & Su, J. (2021). "A novel blockchainbased privacy-preserving framework for online social networks," Connection Science, 2021, pp. 555-575.

[5]. M J Sheller, B Edwards, G Reina, , J Martin, Pati, S. "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," Scientific reports, 2020, pp. 1990-2001, 2020

[6]. D Avraam, R Wilson , Butters ., P "Privacy preserving data visualizations," EPJ Data Science, vol. 10, no. 1, pp. 1-34, 2021.

[7]. G Gnaneshwari, & M Hema, "Crow-Water Wave Optimization Algorithm for PrivacyPreserved Collaborative Data Publishing." International Journal of Swarm Intelligence Research (IJSIR), vol. 13, no. 1, pp. 1-19, 2022.

[8]. Dataset taken from Kaggle Income Dataset (kaggle.https://www.kaggle.com/datasets/mastmustu/incomecom)