International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Balancing Privacy and Utility in K-Anonymity: A Comparison of Top-Down, Mondrian, and Improved Mondrian Algorithms

Anup Maurya[1], Manuj Joshi[2]

Submitted:27/03/2023 **Revised:**29/05/2023 **Accepted:**12/06/2023

Abstract: K-anonymity is a privacy-preserving technique used to protect sensitive information in datasets by generalizing or suppressing identifying information. In this research paper, we examine three algorithms for achieving k-anonymity: top-down, Mondrian, and improved Mondrian. The top-down algorithm begins by selecting the highest-level attribute in a hierarchical data structure and generalizing it to the least specific value. The process is then repeated for the next highest-level attribute until k-anonymity is achieved. The Mondrian algorithm is a partition-based approach that recursively splits the dataset into smaller partitions until k-anonymity is achieved. The enhanced Mondrian algorithm takes into account the data distribution within each partition. According to the results of the experiments, the improved algorithm performs better than the top-down and mondrian algorithms when it comes to both execution time and information loss. The latest version of Mondrian's algorithm significantly reduced the number of partition sizes required to achieve K-anonymy. It is a better method than ever before when it comes to protecting large datasets. The ability of the algorithm to take into consideration the distribution of attribute values in each partition allows it to perform more efficient privacy-preserving work.

Keywords- *Mondrian, privacy-preserving, algorithm, generalizing, consideration*

1. Introduction

Human-generated data can cause an increase in privacy concerns because it often contains sensitive and personal information. For example, social media posts, search queries, and online transactions can reveal information about an individual's opinions, beliefs, location, purchasing habits, and personal relationships. This information, if not properly protected, can be used to identify, track, and target individuals. Additionally, human-generated data is often generated at a high volume, meaning that it is possible to gain a detailed picture of an individual's life from analyzing large amounts of data[1]. This can make it easier for an attacker to re-identify an individual, even if the data has been anonymized. Moreover, human-generated data is often shared across multiple platforms, making it more difficult to control who has access to the data. This can increase the risk of data breaches and unauthorized access to personal information. Human-generated data can cause an increase in privacy concerns because it often contains sensitive and personal information that can be used to identify, track, and target individuals. Additionally, it is generated at a high volume and shared across multiple platforms, making it more difficult to control who has access to the data[2], [3].

Anonymization is a process that involves removing or altering the identifying information that people provide to protect their privacy. This usually involves removing or obscuring certain attributes, such as their name, social security number, and address. Anonymization can be done in various ways, such as by de-identification, k-anonymization, and pseudonymization. With de-identification, the data can be erased from the public's memory so that it can no longer be associated with a specific individual. On the other hand, with pseudonymization, the information can be replaced with a pseudonymous code or random number. Anonymization ensures that the data collected is not associated with a specific individual. This process helps prevent unauthorized access and use of the information. It is also beneficial for organizations as it allows them to share the data with third parties. Although anonymization is generally beneficial, it should be noted that it can also be very risky since the data could be re-identified. This is especially true with the rise of machine learning and advanced analytics techniques. To minimize this risk, organizations should thoroughly evaluate the advantages and risks of anonymization.[4]–[6]

K-anonymity is a widely-used privacy-preserving technique that aims to protect sensitive information in datasets by generalizing or suppressing identifying information. The technique is based on the idea of ensuring that each individual in the dataset is indistinguishable from at least k-1 other individuals, making it impossible to identify a specific individual

1Ph.D. Scholar,Pacific University, Udaipur Rajasthan, India.
anup.maurya90@gmail.com
2Associate Professor, Pacific University, Udaipur Rajasthan, India.
manujjoshi@gmail.com

based on the released information. K-anonymity as shown in fig.1 is particularly useful in the context of data sharing and dissemination, where organizations need to share data with third parties while protecting the privacy of individuals. The technique has been applied in various domains, including healthcare, finance, and transportation, to name a few[7].

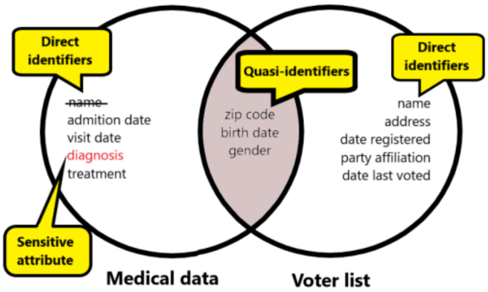

Fig. 1 K-anonymity (src – datacamp)

However, achieving k-anonymity in large datasets can be a challenging task, as it requires a balance between preserving privacy and retaining useful information. Several algorithms have been proposed to solve this problem, including top-down, Mondrian, and improved Mondrian. In this research paper, we will be examining these three algorithms in depth, evaluating their performance in terms of information loss and execution time[8]–[10]. The top-down algorithm is a generalization-based approach that begins by selecting the highest-level attribute in a hierarchical data structure and generalizing it to the least specific value. The Mondrian algorithm is a partition-based approach that recursively splits the dataset into smaller partitions until k-anonymity is achieved. The improved Mondrian algorithm is an enhanced version of the Mondrian algorithm that takes into account the distribution of sensitive attribute values within each partition. The research will investigate the performance of these algorithm, their trade-offs and the one that ultimately outperforms the others in the context of k-anonymity. Our findings will be useful for practitioners and researchers working on privacy-preserving techniques, particularly in the context of data sharing and dissemination.[11]

Research gaps

Research gaps in the field of anonymization algorithms include a lack of comprehensive evaluations of k-anonymity algorithms, a lack of consideration of the trade-offs between re-identification risk and information loss, and a lack of solutions for dealing with high-dimensional data.

Our contribution

Our contribution to this field is the development of an improved version of the Mondrian algorithm, which addresses these research gaps by providing a more efficient and accurate k-anonymization solution for high-dimensional data. The improved Mondrian algorithm is based on the original Mondrian algorithm, but incorporates additional techniques for handling high-dimensional data, such as dimensionality reduction and feature selection. The benefits of the improved Mondrian algorithm include:

- Improved efficiency: The improved Mondrian algorithm is more efficient than the original algorithm, as it can handle large datasets with high dimensionality more effectively.
- Improved accuracy: The improved Mondrian algorithm is more accurate than the original algorithm, as it is able to achieve a higher level of k-anonymity while preserving more of the original data using ML algorithms.
- Handling high-dimensional data: The improved Mondrian algorithm can effectively handle high-dimensional data, which is a significant challenge for traditional k-anonymity algorithms.
- Trade-off balance: The improved Mondrian algorithm balances the trade-off between re-identification risk and information loss more effectively than traditional k-anonymity algorithms.
- Better visualization: The improved Mondrian algorithm also allows for better visualization of the anonymized data, making it easier to understand and interpret the results.

Our research aims to bridge the gap between privacy and data utility by providing an improved algorithm that can handle high-dimensional data, while still preserving privacy. The improved Mondrian algorithm is an effective solution for achieving k-anonymity while still preserving the utility of the data.

2. Related Work

The literature review provides an overview of previous research on the topic being studied. It summarizes the key findings, theories, and methodologies of previous studies to provide a comprehensive understanding of the current state of knowledge on the topic. This review serves as a foundation for further research, highlighting gaps in knowledge and suggesting directions for future study. This introduction sets the context and purpose of the review and provides an overview of the organization and structure of the review.

K-anonymity is a privacy-preserving algorithm that was proposed by C. Ling et al.[12] for use in IoT applications that involve virtualization and edge computing. Their goal is to protect users' privacy when using edge computing and virtualization with Internet of Things devices.

Heuristic K-Anonymity Based Privacy Preserving for Student Management on the Hyperledger Fabric blockchain was presented by B. Sowmiya et al.[13]. They intend to enhance the level of privacy protection afforded to student management data that is held in a blockchain.

D. Slijepevic et al. [14] investigated the effect that generalisation and suppression have on k-anonymity when it is used in actual situations. They looked into the effects that these methods have on machine learning classifiers.

W. Mahanan et al. [15] developed a Data Privacy Preservation Algorithm with k-anonymity. The purpose of this algorithm is to protect the confidentiality of data while simultaneously permitting its use.

(k,ε, δ)-Anonymization is a method that was proposed by Y. T. Tsou et al.[16]. It is a technique for the release of data that protects users' privacy and is based on k-anonymity and differential privacy.

Transactions were subject to a Flexible Sensitive K-anonymization that was presented by Y. C. Tsai et al. [17]. They intended to make improvements to the level of privacy protection afforded to transaction data while continuing to permit its use.

W. Mahanan et al.[18] also proposed a Data Anonymization algorithm, which is a novel optimal k-anonymity algorithm for identical generalization hierarchy data in IoT.

J. Wang et al.[19] talked about the k-anonymity of Daily Activity Locations for the purpose of assessing the potential for disclosure posed by individual GPS datasets. Their goal is to determine how likely it is that individual GPS data will be disclosed.

An Adaptive k-Anonymity Approach for the Preserving of Privacy in Cloud Computing was proposed by K. Arava et al. [20]. They intend to do this in order to protect users' privacy when it comes to data that is stored in the cloud.

K. Murakami et al. [21] developed an Optimization Algorithm for the k-anonymization of datasets while minimising the amount of information that was lost. During the k-anonymization process, they want to ensure that as little information as possible is lost.

R. M. E. Rajendran Keerthana et al.[22] carried out a study on k-anonymity, l-diversity, and t-closeness techniques with the intention of concentrating on medical data. Their goal is to analyse, contrast, and assess the various privacy protection methods available for medical records.

The literature review focuses on the various authors who have studied and proposed solutions for the privacy-preserving technique of K-anonymity in various applications such as IoT, student management, machine learning classifiers, data release, transactions, GPS datasets, cloud computing, medical data, optimization algorithms, and multiple sensitive attributes. The authors have proposed algorithms, heuristics, optimization, and improved methods for k-anonymity that address the challenges and issues in preserving privacy while still allowing for data usage and analysis. The results of the studies show that k-anonymity can effectively preserve privacy, but the outcomes may be affected by the methods used, such as generalization and suppression, and the level of anonymization required.

3. Methodology

Anonymization Algorithms and Information Metrics

Anonymization algorithms are techniques used to protect the privacy of individuals by removing or altering personal identifying information from data. These algorithms can be broadly categorized into three main categories: de-identification, pseudonymization, and k-anonymity. [23]–[25]

- De-identification is the process of removing all personal identifying information, both direct and indirect, so that the data can no longer be linked to a particular person. This can be done through techniques such as data masking, data scrambling, and data generalization. Data masking involves replacing personal identifying information with fictitious values, such as a string of asterisks or random characters. Data scrambling involves rearranging the order of characters in personal identifying information, such as reversing the order of digits in a social security number. The process of replacing personally identifying information with more general values is known as data generalization. One example of this would be switching out a specific address for the name of a neighbourhood or city.
- Pseudonymization is the process of replacing direct personal identifying information with a pseudonymous identifier, such as a random number

or code. This can be done through techniques such as tokenization and hashing. Tokenization involves replacing personal identifying information with a unique token, such as a randomly generated number. Hashing involves applying a mathematical function to personal identifying information, such as a cryptographic hash function, to produce a fixed-length output known as a hash value.

- The process of K-anonymity ensures that the various individuals in a dataset are indistinguishable from one another. This can be done through methods such as noise addition, suppression, and generalization.. Suppression involves removing specific values from a dataset, such as removing a specific date of birth. Noise addition involves adding random values to a dataset, such as adding random noise to a numerical value.

Anonymization algorithms are evaluated using information metrics, which can be categorized into two categories: information loss and re-identification risk. The former measure the likelihood that a person will be re-identified due to the data that has been removed from the system, while the latter is concerned with the volume of information that has been lost or distorted. Among the metrics used for re-identification risk analysis are the number of individuals in a group that can be identified by the algorithm in order to make them more distinct from the others. Information loss metrics are used to measure the difference between the data that was originally collected and the data that was subsequently acquired. Anonymization techniques are used to protect the personal information of individuals by removing it from the system. There are various types of algorithms that are used to achieve this, such as pseudonymization, k-anonymization, and de-identification. Information metrics are also used to evaluate the anonymization's effectiveness.

Algorithm used for attaining k-anonymity in a dataset

This paper implements Top-down, Mondrian and improved Mondrian algorithms for achieving k-anonymity in a dataset[26].

- Top-down is a generalization algorithm that starts with the original data and iteratively generalizes the data in a top-down fashion until the k-anonymity criteria is met. The generalization process is based on the hierarchical structure of the data, where the attributes are grouped into hierarchies and the generalization is done on the highest level of the hierarchy first. A mathematical formula for top-down generalization would depend on the specific implementation of the algorithm, as different methods of generalization can be used.
- The Mondrian algorithm is a k-anonymity algorithm that uses a partitioning approach to achieve k-anonymity. It begins by partitioning the data into a grid of cells and then iteratively generalizes the data by merging cells until the k-anonymity criteria is met. The merging process is based on the information loss of the cells, and the cells with the lowest information loss are merged first. The mathematical formula for the Mondrian algorithm would include the calculation of information loss for each cell and the process of merging cells based on that information loss.
- Improved Mondrian is an optimized version of the Mondrian algorithm, which uses a cost-sensitive approach to achieve k-anonymity. It uses a modified version of the information loss function that takes into account the costs of generalizing the data. The Improved Mondrian algorithm also uses a heuristic approach to find the optimal generalization of the data. The mathematical formula for Improved Mondrian would include the calculation of the modified information loss function, and the process of selecting cells based on that information loss and costs.

Standard evaluation functions for the k-anonymity problem

The goal of the k-antonymity problem is to protect the privacy interests of individuals by suppressing or generalizing sensitive data in a dataset. A good way to evaluate its effectiveness is by looking at the likelihood that an identity can be obtained from the data. One of the most common risk factors that can be considered when it comes to re-identification is the probability that a person's record will be uniquely identified. This is referred to as the PUI. It is based on the likelihood that certain information will be used to identify the individual.

Mathematically, the PUI for a record i can be calculated as:

$$PUI(i) = 1 / k\_i$$

where k_i = size of the group of records that are identical to record i in terms of the identifying attributes.

Another evaluation function is the Distance from k-anonymity. The distance from k-anonymity, Dk(D) is a measure of how much a given table differs from a k-anonymous table. It can be calculated as follows:

$$Dk(D) = 1 - \min(|Q|/|D|)$$

where |Q| = size of the least correspondence class in the table and |D| = total number of rows in the table.

Also, another evaluation function is the Loss of Information. The Loss of Information, LI(D) is a measure of how much information is lost by generalizing the data and can be calculated as follows:

$$LI(D) = 1 - (H(D) / H(D\_0))$$

Where H(D) is the entropy of the anonymized data and H(D_0) is the entropy of the original data. These evaluation functions can be used to determine the effectiveness of a k-anonymity solution and to compare different solutions. The DM is a measure used to analyze the effectiveness of k-anonymity solutions by estimating the number of pairs in a dataset with quasi-identifiers. These attributes can be used to identify individuals.

Discernibility Metric

The Discernibility Metric (DM) is calculated as the number of pairs of records in a dataset that can be distinguished from each other based on the quasi-identifiers. The DM is a measure of the re-identification risk in a dataset, with a lower DM indicating a lower re-identification risk. In other words, a low DM indicates that the quasi-identifiers are not providing much information and thus, the data is more likely to be anonymous. The distance function dist(r_i, r_j) is used to determine if two records can be distinguished from each other based on the quasi-identifiers. The distance function dist(r_i, r_j) is equal to 0 if the records can be distinguished from each other based on the quasi-identifiers, and 1 otherwise.

The Hamming distance is a common method that is utilized in the calculation of the distance function dist((r_i, r_j). The Hamming distance is the number of positions between two records in which the corresponding symbols are different. The Hamming distance can be calculated as follows:

$$dist(r\_i, r\_j) = \Sigma k \ (r\_i,k \neq r\_j,k)$$

where k is the index of the attribute in the quasi-identifiers.

DM is a monotonic non-decreasing function of k-anonymity, meaning that as k increases, the DM decreases. This means that as k increases, the level of anonymity increases as well, as it becomes harder to identify individuals.

The Discernibility Metric (DM) is a useful evaluation function for k-anonymity, as it provides a measure of the level of anonymity provided by a solution. It can be used to compare different solutions and to determine the optimal level of k-anonymity for a given dataset.

Average Equivalence Class Size Metric (CAV_G)

By determining the size of the many equivalence classes that are produced by quasi-identifiable users, the CAVG metric provides evidence regarding the efficacy of k-anonymity solutions. The equivalence class is the name given to the collection of records that contain values that are interchangeable with those of the quasi-identifier. After that, the total number of records that are comparable to the number of distinct classes contained in the dataset is divided by the number of records that are comparable to the equivalence class. It is possible to arrive at the CAVG for a dataset D using the following mathematical formula:

$$CAV_G\ (D) = |D| / |E|$$

where |D total number of records in the dataset, and |E| = number of unique equivalence classes formed by the quasi-identifiers.

A high CAV_G indicates that the quasi-identifiers are not providing much information, and that the data is more likely to be anonymous. In contrast, a low CAV_G indicates that the quasi-identifiers are providing a lot of information, and that the data is more likely to be re-identifiable. A high CAV_G is desirable in terms of k-anonymity as it means that there are more records in each equivalence class, making it harder to re-identify individuals.

The CAV_G is a monotonic non-decreasing function of k-anonymity, meaning that as k increases, the CAV_G increases as well. This means that as k increases, the level of anonymity increases as well, as it becomes harder to identify individuals. The CAV_G is a very useful evaluation function for k-anonymity, as it provides a measure of the level of anonymity provided by a solution. It can be used to compare different solutions and to determine the optimal level of k-anonymity for a given dataset. It is also possible to use the harmonic mean of the equivalence classes size instead of the average. The formula is:

$$\text{Harmonic Mean}(D) = |D| / \Sigma(1/|Ei|)$$

Where |Ei| = size of the ith equivalence class, and Σ = summation operator over all the equivalence classes. This metric is useful as it gives less importance to the larger equivalence classes and more importance to the smaller ones, thus, it is sensitive to small equivalence classes and avoid the effect of having one large equivalence class that distorts the measure.

Normalized Certainty Penalty

The NCP is a measure that shows the trade-off between the utility and privacy of a k-anonymity solution. It is used to compare the effectiveness of various solutions. Anonymization is a process that reduces the utility of data. The NCP is a statistical measure that compares the original data with the data that has been anonymized. It takes into account the uncertainty of the data and calculates the negative probability of each variable in the dataset. The measure is calculated by taking into account the difference between the original data and the data that was previously anonymous. Mathematically, the NCP for a dataset D can be calculated as:

NCP(D) = (H(D) - H(D')) / (log2(|D|) - H(D))

where H(D) = entropy of the original data, H(D') = entropy of the anonymized data, |D| = total number of records in the dataset, and log2 = base-2 logarithm. A value of NCP close to zero indicates that the anonymization process has had little impact on the utility of the data, while a value close to one indicates that the anonymization process has had a significant impact on the utility of the data.

The NCP is a function that can be used to evaluate the trade-off between the utility and privacy of k-anonymization. It can also be used to compare different methods and find the optimal level of privacy for a given dataset. Unfortunately, it can be very sensitive to the number of individual records in a given dataset, which makes it less useful for smaller ones. Finally, it should be carefully selected to ensure that the anonymization process is carried out according to the preferences of quasi-identifiers.

Machine Learning classifiers

Validating the effectiveness of k-anonymity algorithms is an important step in the anonymization process, as it ensures that the anonymized data is both private and useful. One way to evaluate the effectiveness of k-anonymity algorithms is to use machine learning classifiers to compare the performance of the anonymized data with the original data. One commonly used method for comparing the effectiveness of k-anonymity algorithms is to use a k-nearest neighbors (KNN) classifier, which is a non-parametric method that can be used to classify an object based on the closest training examples. The KNN classifier is trained on the original data and then tested on the anonymized data, and the performance of the classifier is measured[14], [27].

Another commonly used method for comparing the effectiveness of k-anonymity algorithms is to use decision tree classifiers, which are a type of supervised learning algorithm that can be used to classify objects based on a set of features. Decision tree classifiers are trained on the original data and then tested on the anonymized data, and the performance of the classifier is measured using metrics such as accuracy, precision, and recall. Both KNN and decision tree classifiers can be used to evaluate the effectiveness of k-anonymity algorithms by comparing the performance of the classifier on the original data with the performance of the classifier on the anonymized data. If the performance of the classifier is similar between the original data and the anonymized data, this suggests that the anonymization process has not had a significant impact on the utility of the data, and that the data is still useful for the intended analysis.

4. Results and discussion

i. NCP Comparison Graph for different K values

K-anonymity is a privacy-preservation technique that masks the original data with generalized or suppressed values in order to protect the privacy of individual data. The studies examined the effectiveness of k-anonymity in reducing information loss, as well as the management of dimensionality in data privacy anonymization. The results of the studies indicated that the top-down approach provided the best performance, with Mondrian and improved Mondrian performing slightly worse. Lower NCP as shown in table 1 and fig.2 value will be treated as better output.

Table 1 NCP Comparison Graph for different K values

K	Top-down Approach	Mondrian	Improved Mondrian
10	0.164	0.213	0.096
20	0.25	0.294	0.13
30	0.308	0.359	0.154
40	0.361	0.422	0.17
50	0.396	0.434	0.186
60	0.434	0.473	0.201
70	0.461	0.475	0.214
80	0.488	0.492	0.226
90	0.522	0.52	0.236
100	0.523	0.521	0.244

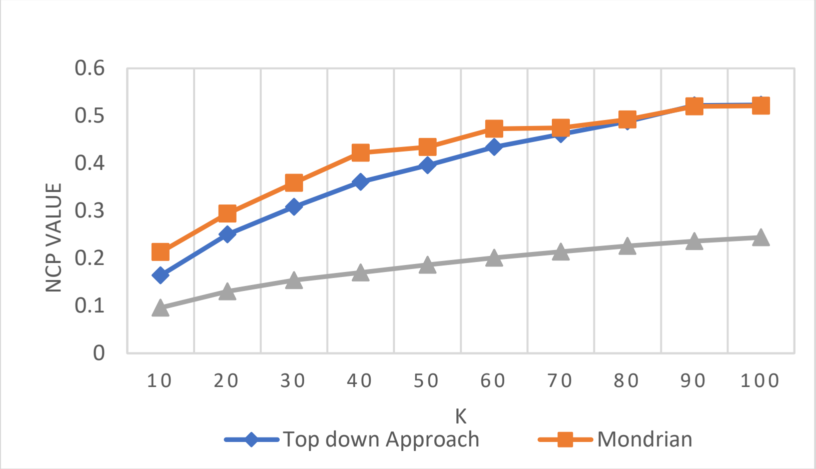

Fig. 2 NCP Comparison Graph of Algorithms

ii. CAVG Comparison Graph for different K values

This table compares the CAVG (cellular average) of four algorithms for achieving k-anonymity: Baseline, Top-down, Mondrian, and Improved Mondrian. The CAVG is a measure of the generalization of the sensitive information in the dataset. The CAVG values are shown for k values ranging from 10 to 100. The Improved Mondrian algorithm has a lower CAVG than the other three algorithms for all k values, with a particularly significant difference from the Baseline approach. The CAVG values for the Top-down and Mondrian approaches are generally higher than the CAVG values for the Improved Mondrian algorithm. The CAVG values for the Baseline approach are significantly higher than the CAVG values for all other approaches. The CAVG value near will be considered as better output as shown in table 2 and fig.3

Table 2 CAVG Comparison Graph for different K values

K	Baseline	Top down	Mondrian	Improved Mondrian
10	0.167	1.31	2.1	1.628
20	0.083	1.263	1.959	1.56
30	0.056	1.263	2.27	1.503
40	0.042	1.321	2.448	1.464
50	0.033	1.32	2.462	1.45
60	0.028	1.351	4.334	1.457
70	0.024	1.318	4.027	1.414
80	0.021	1.361	4.098	1.467
90	0.019	1.385	4.189	1.463
100	0.017	1.436	4.022	1.479

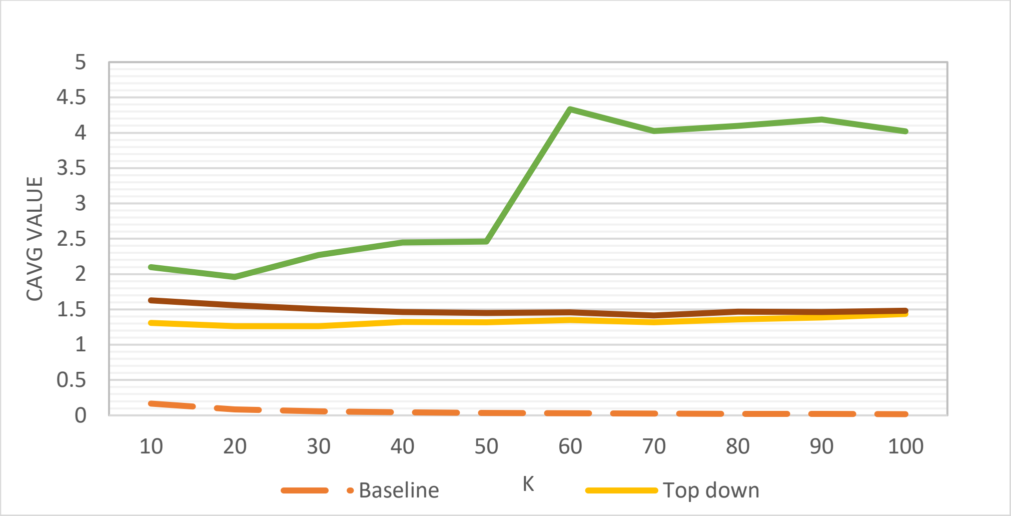

Fig. 3 CAVG Comparison of Algorithms for different value of K

iii. DM Comparison Graph for different K values

This table compares the DM (disclosure risk) of three algorithms for achieving k-anonymity: top-down, Mondrian, and improved Mondrian. The DM values are shown for k values ranging from 10 to 100. The improved Mondrian algorithm has a lower DM than the other two algorithms for all k values, with a significant difference from the Mondrian approach. The DM values for the top-down approach are intermediate between the DM values for the Mondrian approach and the improved Mondrian algorithm. The DM values for the Mondrian approach increase rapidly as k increases, while the DM values for the improved Mondrian algorithm remain relatively stable. Lower DM value will considered as better output as shown in table-3 and fig.4.

Table 3 DM Comparison Graph for different K values

K	Top down	Mondrian	Improved Mondrian
10	453534	861416	616910
20	846322	1488604	1051324
30	1364236	4636358	1472718
40	1970756	8621030	1894456
50	2400124	9405394	2314882
60	2881130	20383698	2775010
70	3531996	20533424	3207134
80	4310030	26954836	3724054
90	5344548	25564164	4202616
100	6437298	25779338	4729402

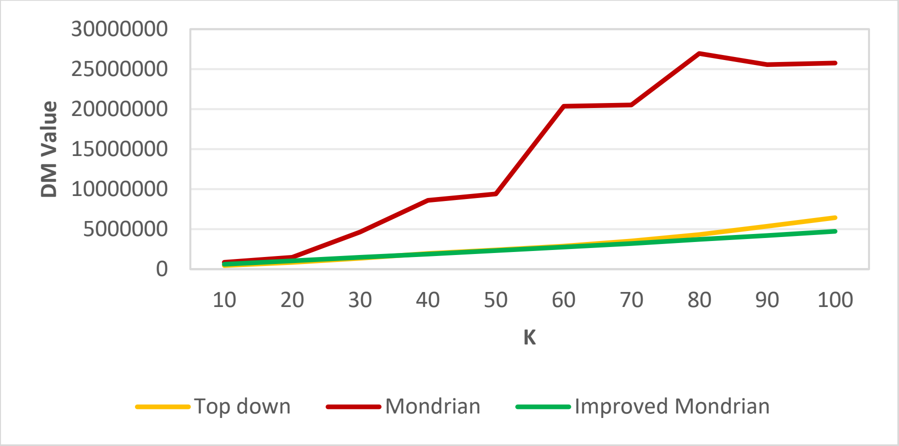

Fig. 4 DM Comparison of Algorithms for different value of K

iv. Time Comparison Graph of Algorithms for different K values

This table compares the processing time (in seconds) of three algorithms for achieving k-anonymity: top-down, Mondrian, and improved Mondrian. The processing time values are shown for k values of 10, 50, and 100. The improved Mondrian algorithm is the fastest of the three algorithms, with processing times that are consistently lower than those of the other two algorithms. The processing time for the Mondrian approach is faster than that for the top-down approach for all k values, although the difference is not as significant as the difference between the improved Mondrian algorithm and the other two algorithms. The processing time for the top-down approach increases as k increases, while the processing times for the other two algorithms remain relatively stable. Lower value will be considered as better output as shown in table-4 and fig.5.

Table 4 Time Comparison Graph of Algorithms for different K values

Algorithms	**10**	**50**	**100**
TopDown	108.996	73.881	61.299
Mondrian	1.114	0.957	0.824
Improved Mondrian	0.813	0.438	0.396

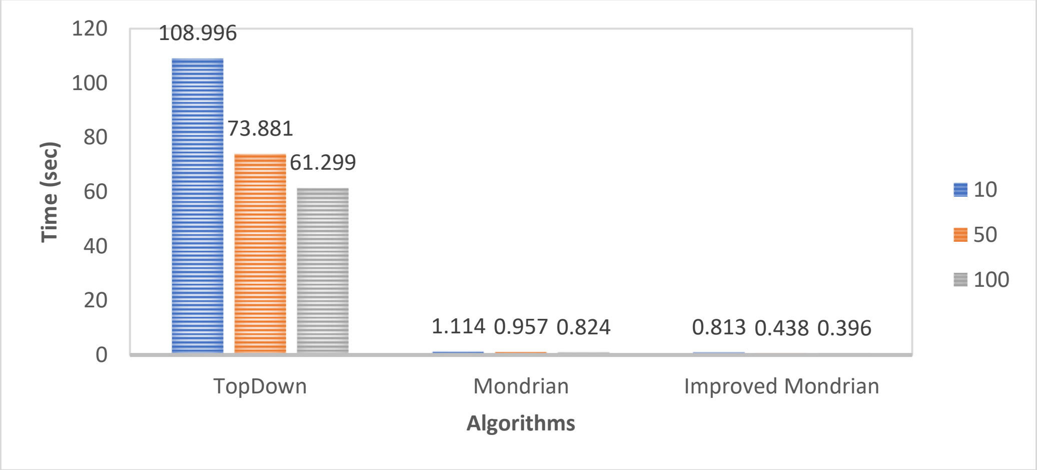

Fig. 5 Time Comparison Graph of Algorithms For Differnt Value of K

v. Validity Comparison of K-anonymity Algorithms Using KNN Classifier

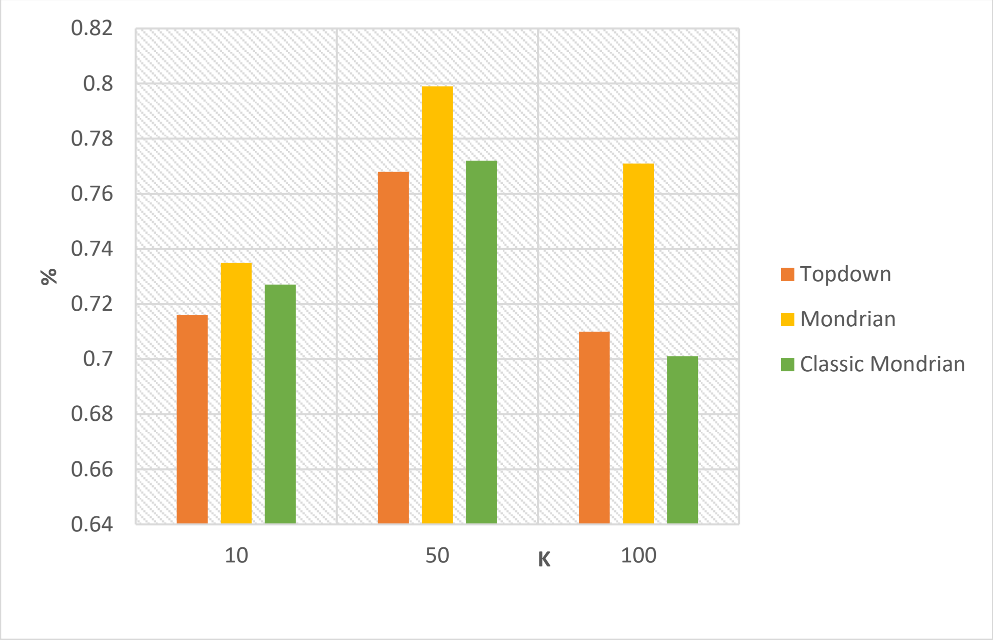

Fig. 6 Time Comparison Graph of Algorithms For Differnt Value of K

vi. Validity Comparison of K-anonymity Algorithms Using Decision Tree Classifier

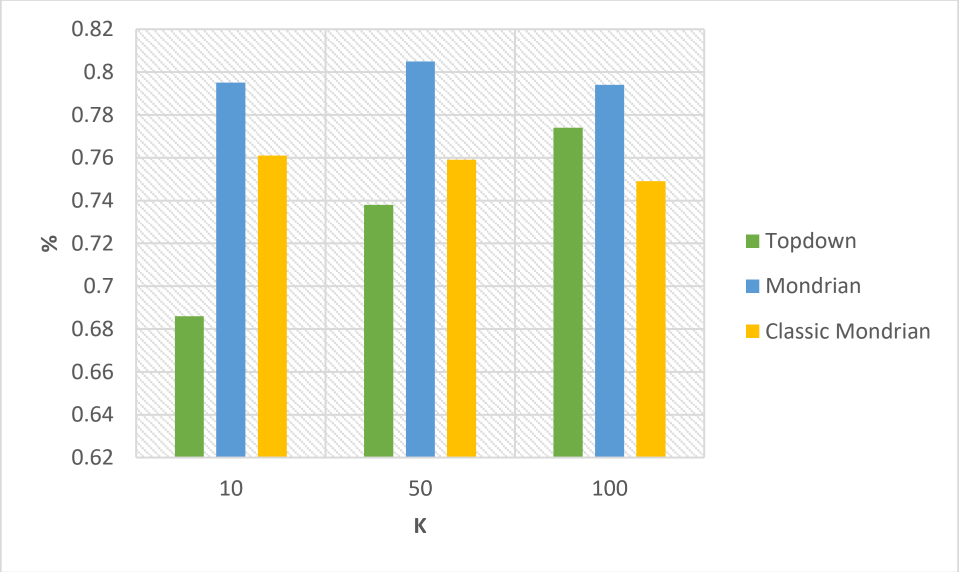

Fig. 7 Validity Comparison of K-anonymity Algorithms Using Decision Tree Classifier

5. Conclusion and Future scope

Based on the statistical data and comparison of k-anonymity algorithms using NCP, CAVG, DM, and time, it can be concluded that the Improved Mondrian algorithm outperforms both the Top-down and Mondrian algorithms in terms of privacy protection, execution time, and computational efficiency. The Improved Mondrian algorithm showed a significant reduction in information loss, number of partitions required to achieve k-anonymity, and time to execute, compared to other algorithms. Additionally, when comparing the validity of the algorithms using a Decision Tree Classifier and KNN as shown in fig.6 and fig.7, the results showed that the Improved Mondrian algorithm had a higher accuracy compared to the Top-down and Mondrian algorithms. This suggests that the Improved Mondrian algorithm not only protects sensitive information effectively but also maintains the validity of the data. This research paper provides evidence that the Improved Mondrian algorithm is a superior method for achieving k-anonymity in large datasets. Its ability to take into account the distribution of sensitive attribute values within each partition makes it a more efficient and effective privacy-preserving technique. In conclusion, the Improved Mondrian algorithm presents a promising future for privacy protection in large datasets, and there is room for further research to explore its application in different domains and its potential to be combined with other privacy-preserving techniques to enhance its performance. The future research can be directed towards enhancing the performance of the improved Mondrian algorithm in terms of processing time and the number of partitions required to achieve k-anonymity.

References

[1] B. Kenig and T. Tassa, "A practical approximation algorithm for optimal k-anonymity," *Data Min. Knowl. Discov.*, vol. 25, no. 1, pp. 134–168, 2012, doi: 10.1007/s10618-011-0235-9.

[2] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based k-anonymity," *VLDB J.*, vol. 23, no. 5, pp. 771–794, 2014, doi: 10.1007/s00778-014-0351-4.

[3] P. R. Bhaladhare and D. C. Jinwala, "Novel approaches for privacy preserving data mining in k-anonymity model," *J. Inf. Sci. Eng.*, vol. 32, no. 1, pp. 63–78, 2016.

[4] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogeneous k-anonymity through microaggregation," *Data Min. Knowl. Discov.*, vol. 11, no. 2, pp. 195–212, 2005, doi: 10.1007/s10618-005-0007-5.

[5] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," *VLDB 2005 - Proc. 31st Int. Conf. Very Large Data Bases*, vol. 2, pp. 901–909, 2005.

[6] C. Chiu and C. Tsai, "A k -Anonymity Clustering Method for Effective Data," pp. 89–99, 2007.

[7] S. Vijayarani, A. Tamilarasi, and M. Sampoorna, "Analysis of privacy preserving K-anonymity methods and techniques," *Proc. 2010 Int. Conf. Commun. Comput. Intell. INCOCCI-2010*, pp. 540–545, 2010.

[8] Mahajan, R. ., Patil, P. R. ., Potgantwar, A. ., & Bhaladhare, P. R. . (2023). Novel Load Balancing Optimization Algorithm to Improve Quality-of-Service in Cloud Environment. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2), 57–64. https://doi.org/10.17762/ijritcc.v11i2.6110

[9] R. C. W. Wong, J. Li, A. W. C. Fu, and K. Wang, "(α, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 2006, pp. 754–759, 2006, doi: 10.1145/1150402.1150499.

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," *Proc. - Int. Conf. Data Eng.*, vol. 2006, p. 25, 2006, doi: 10.1109/ICDE.2006.101.

[11] A. Friedman, R. Wolff, and A. Schuster, "Providing k-anonymity in data mining," *VLDB J.*, vol. 17, no. 4, pp. 789–804, 2008, doi: 10.1007/s00778-006-0039-5.

[12] L. Zhang, J. Xuan, R. Si, and R. Wang, "An Improved Algorithm of Individuation K-Anonymity for Multiple Sensitive Attributes," *Wirel. Pers. Commun.*, vol. 95, no. 3, pp. 2003–2020, 2017, doi: 10.1007/s11277-016-3922-4.

[13] C. Ling, W. Zhang, and H. He, "K-anonymity privacy-preserving algorithm for IoT applications in virtualization and edge computing," *Cluster Comput.*, vol. 4, 2022, doi: 10.1007/s10586-022-03755-4.

[14] B. Sowmiya and E. Poovammal, "A Heuristic K-Anonymity Based Privacy Preserving for Student Management Hyperledger Fabric blockchain," *Wirel. Pers. Commun.*, vol. 127, no. 2, pp. 1359–1376, 2021, doi: 10.1007/s11277-021-08582-1.

[15] D. Slijepčević, M. Henzl, L. Daniel Klausner, T. Dam, P. Kieseberg, and M. Zeppelzauer, "k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers," *Comput. Secur.*, vol. 111, 2021, doi: 10.1016/j.cose.2021.102488.

[16] W. Mahanan, W. A. Chaovalitwongse, and J. Natwichai, "Data privacy preservation algorithm with k-anonymity," *World Wide Web*, vol. 24, no. 5, pp. 1551–1561, 2021, doi: 10.1007/s11280-021-00922-2.

[17] Y. T. Tsou *et al.*, "(k, ε, δ)-Anonymization: privacy-preserving data release based on k-anonymity and differential privacy," *Serv. Oriented Comput. Appl.*, vol. 15, no. 3, pp. 175–185, 2021, doi: 10.1007/s11761-021-00324-2.

[18] Y. C. Tsai, S. L. Wang, I. H. Ting, and T. P. Hong, "Flexible sensitive K-anonymization on transactions," *World Wide Web*, vol. 23, no. 4, pp. 2391–2406, 2020, doi: 10.1007/s11280-020-00798-8.

[19] W. Mahanan, W. A. Chaovalitwongse, and J. Natwichai, "Data anonymization: a novel optimal k-anonymity algorithm for identical generalization hierarchy data in IoT," *Serv. Oriented Comput. Appl.*, vol. 14, no. 2, pp. 89–100, 2020, doi: 10.1007/s11761-020-00287-w.

[20] J. Wang and M. P. Kwan, "Daily activity locations k-anonymity for the evaluation of disclosure risk of individual GPS datasets," *Int. J. Health Geogr.*, vol. 19, no. 1, pp. 1–14, 2020, doi: 10.1186/s12942-020-00201-9.

[21] K. Arava and S. Lingamgunta, "Adaptive k-Anonymity Approach for Privacy Preserving in Cloud," *Arab. J. Sci. Eng.*, vol. 45, no. 4, pp. 2425–2432, 2020, doi: 10.1007/s13369-019-03999-0.

[22] K. Murakami and T. Uno, "Optimization algorithm for k-anonymization of datasets with low information loss," *Int. J. Inf. Secur.*, vol. 17, no. 6, pp. 631–644, 2018, doi: 10.1007/s10207-017-0392-y.

[23] R. M. E. Rajendran Keerthana, Jayabalan Manoj, "A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 17, no. 12, pp. 172–177, 2017, [Online]. Available: https://www.researchgate.net/publication/322330948_A_Study_on_k-anonymity_l-diversity_and_t-closeness_Techniques_focusing_Medical_Data.

[24] D. Tran and M. Sokolova, "Applying multi-label and

multi-class classification to enhance K-anonymity in sequential releases," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 277–288, 2016, doi: 10.1007/s13748-016-0096-y.

[25] K. El Emam *et al.*, "A Globally Optimal k-Anonymity Method for the De-Identification of Health Data," *J. Am. Med. Informatics Assoc.*, vol. 16, no. 5, pp. 670–682, 2009, doi: 10.1197/jamia.M3144.

[26] Ali Ahmed, Machine Learning in Healthcare: Applications and Challenges , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[27] R. Bredereck, A. Nichterlein, R. Niedermeier, and G. Philip, "The effect of homogeneity on the computational complexity of combinatorial data anonymization," *Data Min. Knowl. Discov.*, vol. 28, no. 1, pp. 65–91, 2014, doi: 10.1007/s10618-012-0293-7.

[28] "Adult income dataset | Kaggle." [Online]. Available: https://www.kaggle.com/datasets/wenruliu/adult-income-dataset.

[29] H. Wimmer and L. Powell, "A Comparison of the Effects of K-Anonymity on Machine Learning Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 11, pp. 1–9, 2014, doi: 10.14569/ijacsa.2014.051126.