**Modrain Multidimenstional K-Anonymity**

**Team Members**

* Sai Krishna Kasarapu
* Sai Chaitanya Kollipara
* Janaki Ram Chimata
* Murali Venkateswara Gopi Krishna Ponnada

**Abstract:**

This project is all about protecting people's privacy in the world of data analytics, which is changing really fast. We're focusing on a method called Mondrian Multidimensional K-Anonymity, which is a strong way to keep sensitive information safe. This method is based on the idea of K-Anonymity, and the multidimensional part helps deal with the challenges that come with datasets that have a lot of dimensions (like a lot of different types of information).We're using this technique on a specific dataset. The dataset we used in this project is Employees dataset and our goal is to see how well this technique works and what its limitations are. We're going to look at the results very carefully and compare them with what other studies have found. The main thing we want to know is how good this method is at keeping data anonymous while still being useful.

This project aims to implement and evaluate a multidimensional recoding model for k-anonymity, drawing inspiration from the research paper titled "Multidimensional K-Anonymity" by Kristen LeFevre, David J. DeWitt and Raghu Ramakrishnan .The core objective is to enhance privacy preservation in datasets by extending the traditional k-anonymity framework to handle multidimensional attributes. The project involves the development and application of a greedy approximation algorithm for multidimensional partitioning, addressing the NP-hard nature of optimal multidimensional recoding. Key modifications and adaptations to the algorithm were made to suit the specific goals of the project. Using synthetic and real-world datasets, the implementation leverages tools and methodologies outlined in the research paper. Results are presented and compared with those obtained from single-dimensional recoding methods, highlighting the effectiveness of the multidimensional approach. Visual representations of anonymizations demonstrate the model's capability to capture the underlying multivariate distribution.

**Introduction**

In today's world, where decisions are often based on data, keeping people's privacy is a big deal. As organizations and researchers try to get useful information from really complicated datasets, there's a tough balance between wanting to know things and needing to keep people's identities private. This challenge has led to the creation of new ways to protect privacy, and one of them is called Mondrian Multidimensional K-Anonymity. It builds on the idea of K-Anonymity but is designed to handle the complexities that come with datasets that have a lot of different types of information.

The core principle of K-Anonymity involves anonymizing individual records within a dataset to ensure that no single entity can be uniquely identified. However, when dealing with multidimensional datasets, conventional K-Anonymity methods may face limitations in preserving privacy while maintaining data utility. Mondrian Multidimensional K-Anonymity emerges as a solution to this challenge, offering a nuanced approach that balances the competing demands of privacy and utility.

This project aims to explore, implement, and assess how well Mondrian Multidimensional K-Anonymity works in the context of privacy-preserving data mining. By considering the unique aspects of multidimensional datasets, our study aims to provide valuable insights into the ongoing conversation about privacy preservation. The following sections will cover the background of privacy preservation in data mining, explain the objectives of the project, and offer a comprehensive review of relevant literature. This sets the stage for a detailed exploration of Mondrian Multidimensional K-Anonymity and its application to a specific dataset.

**Add figures of the dataset interms of prem**

Generalize, modify, or distort quasi-identifier values so that no

individual is uniquely identifiable from a group of k

• In SQL, table T is k-anonymous if each

SELECT COUNT(\*)

FROM T

GROUP BY Quasi-Identifier

is ≥ k

• Parameter k indicates the “degree” of anonymity

The greedy technique works much better for single-dimensional models compared to the optimal k-anonymization algorithms that have been suggested. The time it takes for the greedy algorithm, which is O(nlogn), is faster than the worst-case exponential speed of the optimal methods. Besides the various random and heuristic single-dimensional search methods, the currently used greedy multidimensional method often gives better-quality results.

**Dataset:**

**Write about the data set**

**Methodology**

The basis of the applied Mondrian Multidimensional K-Anonymity lies in integrating a Multidimensional Global Recoding Model. This model is created to handle the difficulties posed by datasets with many dimensions, offering a systematic and thorough method to make sensitive information anonymous. By expanding the traditional principles of K-Anonymity to multiple dimensions, our project guarantees a careful and effective anonymization process that protects individual privacy across various aspects of the dataset.

**Single Dimensional Partitioning:**

Single-dimensional partitioning is a crucial aspect of Mondrian Multidimensional K-Anonymity, particularly when dealing with datasets that may not have high dimensionality. In single-dimensional partitioning, the dataset is divided based on the values of a single attribute at a time, aiming to achieve K-Anonymity within each resulting partition. This approach provides a simpler alternative when compared to multidimensional partitioning and is often employed when the dataset does not exhibit a significant number of dimensions.

**Attribute Selection**

In the context of Mondrian Multidimensional K-Anonymity, the process of single-dimensional partitioning begins with the careful selection of a specific attribute from the dataset. This attribute serves as the basis for partitioning and plays a pivotal role in determining the structure of the resulting anonymized groups. The choice of attribute may be influenced by considerations such as sensitivity, privacy impact, or relevance to the analytical goals.

**Sorting**

Once the attribute is selected, the dataset undergoes a sorting phase where records are arranged based on the chosen attribute's values. Sorting is a crucial step that groups similar values together, setting the stage for the subsequent partitioning process.

**Partitioning**

The sorted dataset is then divided into partitions, each corresponding to a distinct range of values of the selected attribute. These partitions are crafted with the objective of ensuring that each group adheres to the K-Anonymity requirement, ideally containing a minimum of K records.

**Anonymization within Partitions**

Within each partition, anonymization techniques are applied to further safeguard individual privacy. This may involve generalization or suppression of specific attribute values, striking a delicate balance between preserving the utility of the data and meeting the K-Anonymity criterion.The single-dimensional partitioning process can be iterated for additional attributes, creating a hierarchy of partitions. Each level in the hierarchy corresponds to a different attribute, allowing for a comprehensive approach to anonymizing the dataset.

**Pseudocode**

*function singleDimensionalPartition(dataset, k, attribute):*

*sortedDataset = sortDataset(dataset, attribute)*

*partitions = []*

*while not endOfDataset(sortedDataset):*

*currentPartition = createPartition(sortedDataset, attribute)*

*# Apply anonymization within the current partition*

*anonymizePartition(currentPartition)*

*partitions.append(currentPartition)*

*return partitions*

**Strict Multi-Dimensional Partitioning**

Strict multi-dimensional partitioning is a key element of the Mondrian Multidimensional K-Anonymity technique. It involves simultaneous consideration of multiple attributes during the partitioning process, ensuring a more comprehensive anonymization approach**.**

**Attribute Selection**

In the realm of Mondrian Multidimensional K-Anonymity, strict multi-dimensional partitioning introduces a more intricate approach by simultaneously considering multiple attributes. The initial step involves the judicious selection of these attributes, with each contributing to the formation of a multidimensional partition. Attributes are chosen based on their collective impact on privacy preservation and their relevance to the analytical goals.

**Sorting**

Similar to single-dimensional partitioning, the dataset is sorted based on the values of the selected attributes. This sorting process sets the stage for creating partitions that encapsulate multidimensional characteristics, enabling a more nuanced understanding of the data's structure.

**Multi-Dimensional Partitioning**

The dataset is divided into partitions, each corresponding to a distinct combination of attribute values. Unlike the single-dimensional approach, strict multi-dimensional partitioning concurrently considers multiple attributes, creating groups that align with the multidimensional nature of the data. Each resulting partition is designed to satisfy the K-Anonymity criterion across all selected attributes.

**Anonymization within Multi-Dimensional Partitions**

Within each multidimensional partition, anonymization techniques are applied to ensure the privacy of individual records. The anonymization process involves the careful treatment of values across all selected attributes, striking a delicate balance between maintaining data utility and achieving K-Anonymity within the multidimensional context. Strict multi-dimensional partitioning is an iterative process. After the initial partitions are formed, the algorithm may iteratively refine the groups, adjusting boundaries and reassessing the anonymization strategies to enhance both privacy and utility. This iterative refinement contributes to the adaptability of the method to different dataset characteristics.

A screenshot of a math problem

Description automatically generated

**Pseudocode**

***function strictMultiDimensionalPartition(dataset, k, attributes):***

***sortedDataset = sortDataset(dataset, attributes)***

***multidimensionalPartitions = []***

***while not endOfDataset(sortedDataset):***

***currentMultidimensionalPartition = createMultidimensionalPartition(sortedDataset, attributes)***

***# Apply anonymization within the current multidimensional partition***

***anonymizeMultidimensionalPartition(currentMultidimensionalPartition)***

***multidimensionalPartitions.append(currentMultidimensionalPartition)***

***return multidimensionalPartitions***

A diagram of a number

Description automatically generated

**NP-Hard**

**Greedy Partitioning Algorithm**

The Greedy Top-Down Partitioning Algorithm is a pivotal component of Mondrian Multidimensional K-Anonymity, designed to efficiently partition a dataset while optimizing both privacy and utility. Unlike traditional K-Anonymity approaches that focus on single-dimensional partitioning, the greedy algorithm extends its scope to multidimensional attributes. The algorithm operates in a stepwise manner, iteratively creating partitions and adapting boundaries to achieve the K-Anonymity requirement**.**

**A diagram of a generalization process

Description automatically generated**

**Adapting to Multidimensional Data**

The Greedy Top-Down Partitioning Algorithm excels in handling high-dimensional datasets by intelligently selecting attributes for partitioning. Unlike traditional methods that may struggle with multidimensional complexities, the greedy algorithm dynamically adjusts to the data's multidimensional nature. Within each partition, anonymization strategies are applied to ensure the privacy of individual records. The algorithm may employ techniques such as generalization or suppression to achieve K-Anonymity while maintaining the usefulness of the data**.** To enhance the quality of anonymization, the algorithm may include iterative refinement steps. These steps involve reassessing partitions, adjusting boundaries, and fine-tuning anonymization strategies to achieve an optimal balance between privacy and utility.

**Algorithm:**

A screenshot of a computer code

Description automatically generated

**Bounds on Quality**

A math equations on a white background

Description automatically generated

We check how well the Greedy Top-Down Partitioning Algorithm works by seeing if it can make partitions that keep information anonymous as per the K-Anonymity rule. The impact of the algorithm on the discernibility metric is important in figuring out the balance between keeping things private and still having useful data. The Greedy Top-Down Partitioning Algorithm is explained thoroughly, including its steps, how it deals with multidimensional data, the ways it anonymizes information, and how it goes through iterative improvement.

**Implementation**

* Language: Python
* Packages : Sys, Csv
* Tools: Visual Studio
* Input: **write dataset**
* Output: Anonymised Data

**Experiment Evaluation:**

**Results:**

**Conclusion:**

In data analytics, it's important to find useful information while keeping people's privacy safe. This project uses Mondrian Multidimensional K-Anonymity to do that. We've learned a lot about how to keep data private, especially when there's a lot of different information.

We looked at different methods, like single-dimensional and strict multi-dimensional partitioning, and tried out the Greedy Top-Down Partitioning Algorithm to see how well Mondrian Multidimensional K-Anonymity works. We used different techniques, like global recoding models and others, to make sure the privacy methods we chose are strong. Combining these with summary statistics helps us keep a good balance between privacy and useful data.We measured how well Mondrian Multidimensional K-Anonymity works by using metrics and carefully looking at results. By refining the process and picking attributes thoughtfully, the algorithm does a good job of keeping data private while still having useful information.

Mondrian Multidimensional K-Anonymity isn't just for making data anonymous; it's a good solution for dealing with data that has lots of different types of information. This project shows that privacy techniques are important, and it's always good to keep improving and finding the right balance between getting information from data and keeping it private. As privacy methods in data mining keep changing, what we learned in this project adds to the knowledge of making a safer and ethically right space for data analytics..

**Future work**

The research presented in the paper offers a promising advance in the field of data privacy, especially in the context of k-anonymity through a novel multidimensional recoding model. However, this study exposes several limitations and opens avenues for future research. The recognition that finding an optimal multidimensional partitioning is an NP-hard problem sets the stage for exploring heuristic or approximation algorithms that could achieve a balance between computational efficiency and partition quality. Additionally, the integration of anticipated workloads directly into the anonymization process represents an important frontier. Future work could focus on developing sophisticated models that account for the varying nature of anticipated queries, incorporating machine learning techniques to predict query patterns, and subsequently anonymizing the data in a manner congruent with these patterns**.**Furthermore, the scalability of the algorithms when applied to large-scale datasets remains a question. Future research should aim to address the challenges of scalability and optimize the algorithms for big data contexts, where the computational overhead can become a significant bottleneck. Examining the robustness of the greedy multidimensional partitioning algorithm in the presence of noise, outliers, and non-uniform data distributions would also be beneficial. This includes a deeper analysis of the trade-offs between data utility and privacy when employing relaxed versus strict partitioning approaches.

As we conclude this exploration, it is essential to consider potential avenues for future research. The ongoing evolution of data analytics and privacy preservation calls for continued exploration of novel techniques and enhancements to existing methodologies. Future work could delve into the integration of machine learning approaches, dynamic parameter tuning, and real-time applications of Mondrian Multidimensional K-Anonymity.

**Acknowledgements**

We express our gratitude to Professor Dr.Zeyu Ding for providing us with this fantastic chance to learn more about data privacy. As a result of working on this project, we gained deeper understanding of various data mining and privacy methods.

**References**

[1] K. Lefevre, D. Dewitt & R. Ramakrishnan, “Incognito: Efficient Full Domain KAnonymization”, SIGMOD 2006

K. Lefevre, D. Dewitt & R. Ramakrishnan,“Mondrian Multidimensional k-anonymity”, ICDE

2007

[2] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In ICDE,

2005.

[3] G. Ghinita, P. Karras, P. Kalnis& N. Mamoulis, “Fast Data Anonymization with Low

Information Loss”, VLDB 2007

[4] M. Terrovitis, J. Liagouris, N. Mamoulis& S. Skiadopolous, “Privacy Preservation by

Disassociation”, VLDB 2012

[5] K. Beyer, J. Goldstein, R. Ramakrishnan & U. Shaft, “When is “nearest neighbor”

meaningful?”, ICDT 1999

[6] C. Agarwal, “On K-Anonymity and the Curse of Dimensionality”, VLDB 2005