PRIVACY PROTECTION USING T - CLOSENESS THROUGH MICRO - AGGREGATION

Final Project Report

Fall 2022-23

BCI2001 – DATA PRIVACY

Guided by:

Prof. Jasmin.T.Jose

Submitted by:

Yenigandla Venislaus Ashish - 20BCI0016
Chatakondu Naga Sai - 20BCI0089
Esikela Shanmuka Sainath - 20BCI0095
Kanugo Krishna Ganesh - 20BCI0149
Ranjana Tarini R - 20BCI0160
Abhishek Raj Chauhan - 20BCI0161

SLOT: B2+TB2

B.Tech.

in

Computer Science and Engineering with Specialization in Information Security

School of Computer Science & Engineering

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

ABSTRACT:

Preserving privacy includes limiting disclosure and protecting data subjects' privacy. Our project aims to improve this through Microaggregation. It has been used to create k-anonymous data sets, where each subject's identity is concealed inside a group of k individuals, as an alternative to generalization and suppression. Microaggregation disturbs the data in a different way than generalization, and this extra masking freedom enables improving the value of the data in a number of ways, including enhancing data granularity, minimizing the influence of outliers, and avoiding discretization of numerical data. On the other hand, attribute disclosure is not protected by k-Anonymity and happens when there is insufficient variation in a set of k individuals' secret values. Several improvements to k-anonymity, including t-closeness, have been proposed to address this problem.

Key Words: MDAV, K-ANONYMIZATION, T-CLOSENESS, MICROAGGREGATION.

INTRODUCTION:

A file with a number of records and variables specific to each record about a respondent, who can be a person or an organization, makes up a microdata collection. Many organizations are increasingly publishing microdata – tables that contain unaggregated information about individuals. These tables can include medical, voter registration, census, and customer data. Microdata is a valuable source of information for the allocation of public funds, medical research, and trend analysis. However, if individuals can be uniquely identified in microdata, then their private information (such as their medical condition) would be disclosed, and this is unacceptable. To avoid the identification of records in microdata, uniquely identifying information like names and social security numbers are removed from the table. However, this first masking still does not ensure the privacy of individuals in the data. So we go for the data privacy mechanisms such as K-Anonymization, L-diversity and T-closeness etc., Only by using those mechanisms we cannot provide enough balance between privacy and utility for the data. So we combine the perturbative and non-perturbative mechanisms to enhance privacy which includes combining MDAV based micro-aggregation and anonymization techniques such as K-anonymization and T-closeness

Literature Survey Table:

YEAR OF PUBLICATION	AUTHORS ANDTITLE	CONTRIBUTIONS	LIMITATIONS
2019	Josep Domingo-Ferrer and Jordi Soria-Comas. Steered Microaggregation: A Unified Primitive for Anonymization of Data Sets and Data Streams	1. They have presented the idea of steered microaggregation, which is a general way to deal with guide standard microaggregation algorithms.This is finished by adding fake credits with suitable loads that impact the microaggregation. 2. Requirement of t-closeness in a static informational index. This is finished by presenting a fake characteristic that controls the inside group inconstancy of the delicate quality. 3. Explained the best way to utilize steered microaggregation to work on the safeguarding of the first request of the tuples in the anonymized information stream.	1. An order preservation guarantee is not provided by the microaggregation algorithm for stream k-anonymity, but it satisfies the maximum delay constraint.
17 October 2019	David Sánchez, Sergio Martínez , Josep Domingo-Ferrer, Jordi Soria-Comas and Montserrat Batet. µ-ANT: Semantic Microaggregation-based Anonymization Tool.	1. In this paper, authors introduced and discussed about µ-ANT, a useful and flexible tool for (healthcare) data anonymization. 2.It uses several cutting-edge techniques to provide strong privacy assurances and keep the anonymised data's usefulness as intact as you can. 3. Many developers interested in data anonymization can benefit from µ-ANT, which also supports the heterogeneous attribute types frequently found in electronic medical records.	

2000	J.M. MATEO-SANZ J and DOMINGO-FERRER. A COMPARATIVE STUDY OF MICROAGGREGATION METHODS.	1.In this papers, authors discussed about variant microaggregation methods and their performances 2. Compared to univariate methods, multivariate methods exhibit much superior behaviour. 3. For the data set used, UFS_{FPC} performs marginally better than the other univariate approaches. 4. For the data set used, the new approach MFS_{MD} outperforms the competition among multivariate algorithms. 5. The optimal sorting criterion for multivariate microaggregation appears to be the MD sorting criterion. It is extremely reliable, produces the greatest outcomes, and takes minimum calculation.	1. The cause is that microaggregation and one-dimensional data projection are both sources of information loss in univariate microaggregation. 2. The only information loss in multivariate microaggregation comes from the microaggregation process itself. 1. 3. MFS typically will in general make a greater number of gatherings than MDO, and consequently requires more distance calculations to finish the microaggregation cycle.
2010	David Rebollo-Monedero, Jordi Forne, and Josep Domingo-Ferrer. From t-Closeness-Like Privacy to Postrandomization via Information Theory.	1.In this paper, authors given a definition of the privacy-distortion trade-off in information-theoretic terms is given for applications like microdata anonymization and location privacy in location-based services. 2.Comparing experimental findings with discretized statistics to well-known deterministic aggregation techniques reveals superior performance. 3.The usage of PRAM for characteristics with finite alphabets and noise addition for continuous cases is supported by this research. 4.The use of deterministic aggregation, like that employed by MDAV and -Approx, cannot theoretically result in inferior performance than randomised perturbation rules because they are more broad.	1. Even in this straightforward scenario, the QGLB cannot be reached by the k-anonymity methods. 2.If a critical attribute became multidimensional, both algorithms performances would likely depart more from ideality.

2018	Shi,Yancheng &Zhang, Zhenjiang & Shen, Bo. “Data Privacy Protection Based on Micro Aggregation with Dynamic Sensitive Attribute Updating”	1. In this study, the subject of data privacy protection is investigated, and a dynamic updating mechanism based on micro aggregation is suggested. 2. In order to achieve privacy protection while the data is altered, the technique also suggests a dynamic updating strategy. 3.Additionally, a Laplace noise technique is used to safeguard the delicate properties of the result set. 4.With its dynamic updating function, this method effectively lowers information loss and ensures information availability after data anonymization.	1. Clustering time is increased because of the greater number of tuples.
2019	Yuichi Sei, Hiroshi Okumura, Takao Takenouchi, Akihiko Ohsuga. Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness.	1.In this paper, Numerous studies have been conducted on the models of l-diversity and t-closeness for privacy protection. 2.They put forth two brand-new privacy models, (l1,..., lq)-diversity and (t1,..., tq)-closeness, as well as reconstruction and anonymization algorithms that can handle sensitive QIDs.	1. It is still possible for disclosure to occur even if we remove all explicit identifiers from a database.

2014	Salvatore Ruggieri. Using t-closeness anonymity to control for non-discrimination	1.This paper made two contributions in total. 2.The analytical methods of t-closeness in privacy data anonymization and of α-protection in non-discrimination data analysis have first been associated. 3.Second, we have methodically developed dMondrian, a multidimensional generalisation algorithm, and dSabre, a bucketization and redistribution algorithm, as adaptations of well-known algorithms for k-anonymity and tcloseness, by taking advantage of the observed implication. This is a methodological advancement that connects non-discrimination research with data anonymization research.	1.For low-dimensionality datasets, experiments have shown that dSabre outperforms dMondrian, however it also has a dimensionality problem.
2013	Liang, H., Yuan, H. On the Complexity of *t*-Closeness Anonymization and Related Problems.	1.Authors of the paper started the first comprehensive theoretical investigation of the t-closeness principle in accordance with the widely-used attribute suppression concept. We show that finding an ideal t-closeness is NP-hard for every constant t such that 0 ≤ t < 1, expansion of a specific table. 2.Additionally, they gave the first polynomial time precise algorithm for 2-Diversity and a conditionally improved approximation approach for k-Anonymity.	1.No theoretical method for finding the most appropriate t value. 1. 2. NP – hardness of the t-closeness.
May 2016	S.Sarswathi and K,ThiruKumar. Enhancing utility and privacy using t-closeness for multiple sensitive attributes.	1.For enhancing the utility of any organization the data they have must be analyzed to provide the better user experience. 2. The SLOMS method which helps to tackle the linkage attacks have been introduced. 3. In SLOMS method the sensitive attributes are divided in m parts based on the	1.The more the utility for a dataset the more there will be the disclosure risk that means the less privacy. 2.Just generalizing the data cannot be protected from the linkage attacks. 3. Even though the usage of SLOMs method there is a chance for the probabilistic inference attack.

		principle that is highly co related. 4.Introduced about the different types of attacks on a dataset such as probabilistic inference attack, slicing attack and many techniques such as t-closeness, discretization and MSB KACA algorithm which generalizes the quasi identifiers to implement the k-anonymization.	4. So the t-closeness is introduced over the MSB-KACA method to protect the privacy.
May 2013	Debaditya Roy, Determining t in t-closeness using Multiple Sensitive Attributes	1.As the k-anonymization failed to protect the data over linkage attacks and l-diversity failure over the skewness attack the t-closeness method is introduced. 2.There are two methods till now to find the t-value for the single attribute data those are Earth Movers DItance and Hellinger Distance method. 3. Decompose+ a framework which was introduced for the attacks on the multiple sensitive attributes. 4.The lower the value of*t* i.e. $t \rightarrow 0$; the more diverse the original data is and the equivalence class is required to be as close to original data as possible to give the Required anonymization. Secondly, the higher the value of $t \rightarrow 1$, the less diverse the original data and the equivalence class is required to be as different as possible from the original data to give the appropriate anonymization.	1. There is no mention of any method for determining *t*. 2.If the optimum value of *t*has to be determined using the utility vs. privacy curve it is not possible do so because of the inherent nature of the curve i.e. diverging. 3.The Inference method cannot be applied for finding the value of t for the single valued attributes.

December 2012	Sergio Martínez, David Sánchez, Aida Valls, A semantic framework to protect the privacy of electronic health records with non-numerical attributes	1.Removal of external identifiers cannot solve the problem related to the privacy protection, so to protect this the statistical disclosure control methods have been proposed. 2.The general framework is introduced that enables the anonymization of structured non-numerical medical data such as names of symptoms or diagnosis from a semantic perspective. 3. The framework uses the three main methods those are comparison, aggregation and sorting to protect the privacy. 4. The framework is used to adapt three well differenced SDC methods, so that structured non-numerical data could be k-anonymised while retaining their semantics as much as possible. 5.After the proposed framework is implemented on the dataset the utility and privacy of release data is improved to the highest possibility.	1.Implementation of the different methods on a dataset can be hard because of the multi attributes in the data. 2.The major limitation for any privacy protection technique is the proportion to the utility. 3.Even after the implementation of the framework the dataset released is vulnerable to the probabilistic inference attack.
June 2018	Gordon Sande, METHODS FOR DATA DIRECTED MICROAGGREGATON IN ONE OR MORE DIMENSIONS	1.Microaggregation is a perturbated method that release the average clustered groups of the data instead of the anonymized data such that no attribute is over dominated. 2. The univariate micro aggregation technique can be extended to allow for varying group size. This permits the groups to be chosen for greater within group homogeneity. 3.The data at first is sorted into fixed groups based on the homogentisic attributed and then shuffled to vary the size. 4.There are different approaches of micro aggregation to protect the privacy such as clustering approach and optimal approach and reference approach. 5.Using clustering and optimal approaches can give most privacy protected data with much utility possible.	1.Without the usage of the computational geometry the determination of the variable size in the micro aggregation is difficult. 2.The robustness and quality of the data after optimization and approximation techniques are not manageable for most of the companies the release data can have more diverse data in it. 3.For two or more-dimensional data the Adjacency can make quite normal but the simplicity of the code implementation is not well defined because of the no proper sorting techniques.

2008	Domingo-Ferrer, Josep & Sebé, Francesc & Solanas, Agusti. An Anonymity Model Achievable Via Microaggregation	1.They have developed (k, p, q, r)-anonymity computational method to attain this new model that relies on microaggregation as a brand new security model which outperforms most current security models within the literature. 2.The model behaves in a very pragmatic way to scale back information loss.	
2012	Domingo-Ferrer, Josep & Trujillo-Rasua, Rolando. (2012). Microaggregation- and permutation-based anonymization of movement data.	1.They have proposed two heuristics for trajectory anonymization which yield anonymized directions shaped by completely exact genuine unique areas. 2.The principal heuristic depends on direction microaggregation utilizing the above distance and on the spot change; it successfully accomplishes trajectory k-anonymity. 2.The subsequent heuristic depends just on the spot stage; it surrenders direction k-secrecy and focuses on the spot k-diversity. The solid point of the subsequent heuristic is that it considers reachability limitations when registering anonymized directions.	
2017	Sonu Khapekar, Prof. Lomesh Ahire, Privacy Protection of Sensitive Microdata in Healthcare System using t-Closeness through Microaggregation	1. By using microaggregation, the suggested t-closeness model effectively and securely preserves the privacy of sensitive attributes in the healthcare system. 2. The microaggregation causes data to be perturbed, and this additional masking freedom enables boosting data utility in a number of ways, including enhancing data granularity, minimising the influence of outliers, and avoiding discretization of numerical data.	1. Other privacy models like k-anonymity and l-diversity do not offer attribute disclosure protection. 2. On the other hand, attribute disclosure is not protected by k-Anonymity and happens when there is insufficient variation in a set of k individuals' secret values.

2015	Josep Domingo-Ferrer, Jordi Soria-Comas, From t-Closeness to Differential Privacy and Vice Versa in Data Anonymization.	1.Several linkages between k-anonymity, t-closeness, and ε-differential privacy have been identified and taken advantage of in this paper. 2. They have demonstrated that stochastic t-closeness is produced via k-anonymity for the quasiidentifiers and -differential privacy for the private attributes, with t being a function of, the size of the data set, and the size of the equivalence classes.	1. Basic k-anonymity only protects against identity disclosure. 2. The effects of the distance between distributions suggested in the article and the earth mover's distance could also be compared in terms of privacy and utility.
2015	Jordi Soria-Comas, Josep Domingo-Ferrer, David Sanchez and Sergio Mart’inez T-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation	1.To achieve k-anonymous t-closeness, they suggested and tested the usage of microaggregation. 2. To produce kanonymous t-close data sets, they have proposed and assessed three distinct microaggregation-based techniques. The first is a straightforward merging process that may be applied following any microaggregation procedure. the two additional algorithms, t-closeness-first and k-anonymity-first.	1. The attribute disclosure that happens if the variability of the secret values in a group of k participants is too small is not protected by k-Anonymity. 2. Generalization-based methods have some limitations.
2013	Jordi Soria-Comas, Josep Domingo-Ferrer Differential Privacy via t-Closeness in Data Publishing	1. They demonstrated the validity of the k-anonymity family of models. strong enough to achieve context-differentiated privacy of publishing data. 2. They have demonstrated that exp(ε)-closeness implies approximate -differential privacy for informed intruders and ε-differential privacy for ignorant intruders using a suitable approach.	1.The proposed approach for nominal confidential attributes cannot be ordered.
2012	Jordi Soria-Comas, Josep Domingo-Ferrer Probabilistic k-Anonymity through Microaggregation and Data Swapping	1. They have provided two computational strategies to attain probabilistic k-anonymity, primarily based totally on microaggregation and swapping. 2. This is in particular applicable whilst handling an information set that includes many quasi-identifier attributes.	1. Like general k-anonymity, probabilistic k-anonymity ensures that the chance of accurate re-identity is at maximum 1/k, however with out explicitly requiring that the quasi-idetifier attributes take equal values inside every organization of k records.

2009	Jun-Lin Lin , Tsung-Hsien Wen, Jui-Chien Hsieh, Pei-Chann Chang, Density-based microaggregation for statistical disclosure control.	1. This paper demonstrates how microaggregation problem of minimizing information loss has been shown to be NPhard for multivariate data. 2. None of the methods based on heuristics performs the best for every microdata set and various k values. 3. This work presents a density based algorithm (DBA) for microaggregation. The performance of the DBA is compared against the latest microaggregation methods.	
2007	Josep Domingo-Ferrer_, Francesc Seb´e, Agusti Solanas, A polynomial-time approximation to optimal multivariate microaggregation.	1. As optimal microaggregation can only be computed in polynomial time for univariate data. 2. For multivariate data, it has been shown to be NP-hard. 3. In this paper a polynomial-time approximation to microaggregate multivariate numerical data for which bounds to optimal microaggregation can be derived at least for two different optimality criteria: minimum within-groups.	
2017	Prof. Sarita Lalchand Tanay, Prof. Vivek Jaysing Nagargoje, Sayali Avinash Inamdar, Security for Personal Credentials in Big Data: Through Microaggregation and TCloseness.	1. In microaggregation, the data perturbs and masking allows improving data in many ways. 2. K-Anonymity, alone cannot provide the protection for the data, as it provides protection against identity disclosure but prone to attribute disclosure. 3. To solve this problem, many refinements of k-anonymity is being proposed, in which t-closeness is one providing the solution for personal privacy for information of the subjects.	
2017	A. PRAVALLIKA, I. SAPTHAMI, T-Closeness Through Microaggregation: Strict Privacy With Enhanced	1. As K-Anonymity, alone cannot provide protection against identity disclosure but is prone to attribute disclosure. 2. Hence t-closeness is one providing the solution for personal privacy for information of the subjects.	

	Utility Preservation.		
2018	Wang, R., Zhu, Y., Chen, T.S. and Chang, C.C., 2018. Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness.	1.Extend the definition of t-closeness of a single attribute to a new definition of multiple attributes. 2.The values of SAs in an equivalence class must be spread to the maximum extent possible over all of the data to make the class satisfy t-closeness. 3.The main aim of proposed algorithms is to heterogenize the values of the SAs in different equivalence classes. And the more similar the QI attribute values of all records in an equivalence class are, the lower the information loss caused by anonymization should be. 4.t-closeness of multiple sensitive attributes is based on considering each attribute separately, namely, if an equivalence class satisfies t-closeness, all sensitive attributes of it should satisfy t-closeness, respectively. 5.Proposed Algorithm - Cluster-Based Algorithm for Multiple Sensitive Attributes Satisfying t-Closeness, PCA-Based Algorithm for Multiple Sensitive Attributes Satisfying t-Closeness. 6.The first algorithm partitions all records into different clusters and generates equivalence classes by selecting records from these clusters separately. The second algorithm processes the multiple sensitive attributes by analyzing the principal components, sorts the original records according to the results of the projection, and partitions these records into different subsets by the sorting order.	

1998	Masking Microdata Using Micro-Aggregation D. Defays and M.N. Anwar	1.In order to minimise data losses, it is proposed that the different unidimensional variables be aggregated separately, by sorting the values according to their ranks, and by an aggregation in size groups of contiguous values. 2.Creation of classes of three individuals of minimum variance, using the average as a replacement value. 3.Micro-aggregation is also one way to recode data or to replace them by missing values. 4.When applied to numerical variables, micro-aggregation can be seen as a disturbance method. 5.Micro aggregation is simple and flexible in its approach, it offers a compromise between data protection and utility.	1.It is an empirical approach based on empirical rules which have proved useful, rather than on pure statistical theory. 2.This method is not much powerful and still can be edeveloped a lot in theoretical front.
2021	Sarah Zouinina*, Younès Bennani, Nicoleta Rogovschi, and Abdelouahid Lyhyaoui Data Anonymization through Collaborative Multi-view Microaggregation	1.They have proposed two techniques to achieve k-anonymity through microaggregation: k-CMVM and Constrained-CMVM. The first one determines the k levels automatically and the second defines it by exploration. 2.Multi-view collaborative Self Organizing Maps to achieve data anonymization. 3.Multi-view clustering is an efficient way to deal with multisources data and high dimensional elements. 4.Constrained collaborative Self Organizing Maps to attain a predetermined k anonymity level. 5.The introduction of the discriminative information and the use of the pLVQ2 to achive highest anonymity levels with a good utility trade-off. 6.pLVQ2 is used which gives weights to each of the features what results in better preservation of the utility of the anonymized dataset.	1.Other better ways to anonymize data. 2.They are experiencing 1D clustering as a way to anonymize data without loosing the information it is containing and we want to explore new methods to anonymize unbalanced datasets.

2018	Wang, M., Jiang, Z., Zhang, Y. and Yang, H.T-closeness slicing: A new privacy-preserving approach for transactional data publishing	1.This study develops a novel method named t-closeness slicing (TCS) to better protect transactional data against various attacks. The time complexity of TCS is O(nlogn), where n is the number of records in the dataset, hence the algorithm scales well with large data. 2.Connections among the three types of disclosures - Membership Disclosure , Identity Disclosure , Attribute Disclosure. Identity disclosure can lead to membership disclosure and attribute disclosure, However membership disclosure may not cause identity disclosure, Attribute disclosure may occur even without identity disclosure. 3.Vertical partition divides the item set into columns based on the correlations between the sensitive item and non-sensitive items. I c the time complexity of vertical partition is O(〖(log_2 n)〗^2). 5.Horizontal partition divides the transactional dataset D into b buckets based on the correlations between the values of their QIs. the time complexity of horizontal partition algorithm is O(nlog_2 n). TCS offers a high level of privacy protection, reduces the risks of multiple types of privacy disclosures including membership disclosure, identity disclosure, and attribute disclosure, and makes a better tradeoff between privacy protection and data utility.	1.Similarity attack and skewness attacks are more subtle in nature and have not been as well analyzed and protected against as the other types of attacks. 2.Lacks to preserve more correlations between items. 3.Don’t have a more in-depth study with a clear focus on protecting membership privacy. They have considered only single attribute to protect in their study.
2015	NAUSHEEN FATHIMA1, MISBAH KOUSER, Privacy Preserving with Utility Preservation through	1. Author describes how k-anonymity does not secure against trait divulgence, which happens if the changeability of the private values in a gathering of k subjects is too little. 2.To address this issue, a few refinements of k-secrecy have been proposed, among which t-	1.Other better ways to anonymize data. 2.They are experiencing 1D clustering as a way to anonymize data without loosing the information it is containing and we want to explore new methods to anonymize unbalanced datasets.

	Microaggregation.	closeness emerges as giving one of the strictest security ensures.	
2012	Dangi, A.P. and Mogili, R., 2012. Privacy preservation measure using t-closeness with combined l-diversity and k-anonymity	(n, t)- closeness model better protects the data while improving the utility of the released data.(n, t)- closeness allows us to take advantage of anonymization techniques other than generalization of quasi-identifier and suppression of records.Extended the closeness with Anticloseness or diversity. After removing closeness from data, again data rows are re-grouped such that no similar data even after reduction appears together.combining both diversity and anonymity based methods – Entorpy measures,closeness of columns and their aggregation, rows reduction by group aggregate.	1.(n,t) – closeness technique does not affect quasiidentifiers, it does not help achieve k-anonymity. Removing a sensitive value in a group reduces diversity and therefore, it does not help in achieving l-diversity. 2.Technique proposed is not enough if more rows are added which increases the redundancy in data.
2008	Solanas, A. and Pietro, R.D. A linear-time multivariate micro-aggregation for privacy protection in uniform very large data sets. In International Conference on Modeling Decisions for Artificial Intelligence (pp. 203-214). Springer, Berlin, Heidelberg..	1.The microaggregation - Given a data set D with n records in a characteristic space Rd, the problem consists in obtaining a k-partition2 P of D, so that the SSE of P is minimised. Once P is obtained, each record of every part of P is replaced by the average record of the part. 2. Micro-aggregating the group which belongs to the same hyper-space(Hpercube). 3.This method saves Time and Information loss does not have significant different from the other solutions. The difference between MDAV and proposed method tends to decrease when the number of records increases 4.complete characterisation for model, is dependant on three	1. Even though there is not significant loss in the information, it is still present and is more if the size of the data is less.

		parameters only: n, k(security parameter), and ε. The error probability (Pr[Bad]) decreases exponentially fast as free parameters increase. 5. Current micro-aggregation algorithms are very costly (i.e. at least O(n2)) but proposed algorithm is able to micro-aggregate very large data sets in linear time O(n).	

GAPS IDENTIFIED:

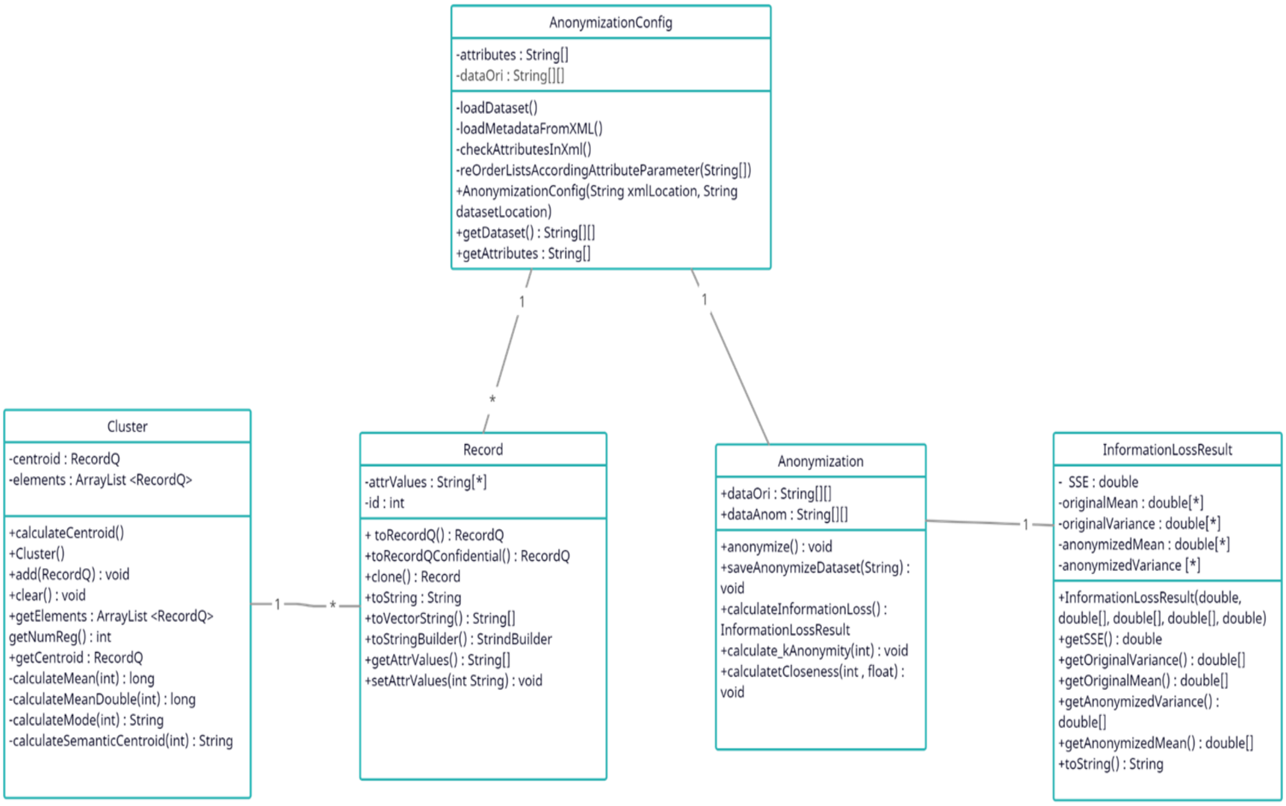
An order preservation guarantee is not provided by the micro aggregation algorithm for stream k-anonymity. Also there are several better algorithms such as L-diversity and T-closeness etc.. which are used in protecting the data more securely than K-anonymization but they alone cannot provide a balance between utility and privacy of the data.

PROBLEM STATEMENT:

K-anonymous data sets are prone to attribute disclosure even though k-anonymity protects against identity disclosure. The l-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure but it can undergo skewness attack. It is difficult to achieve and may not provide sufficient privacy protection against attribute disclosure. So, we are using t-closeness through micro aggregation which can withstand more than k-anonymization and l-diversity in protecting the data and to have a better utility of the data.

PROPOSED WORK:

This project's work flow begins with classifying the data according to its data kinds, attributes, and identities. Suppression is first applied on Explicit Identifiers, followed by MDAV clustering and K anonymization and finally T-closeness on Quasi Identifiers. The data is clustered using the MDAV Clustering method. The mean and variance of the equivalence classes are determined using MDAV clustering prior to K anonymization, and this information is then used to perturb the equivalence class using Microaggregation based K anonymization. By clustering up to the point where the t-closeness criteria must be met, micro aggregation is carried out. Then the anonymized data is subsequently subjected to T-closeness. We must mix all of these to achieve privacy that fulfils the required t-closeness. UML Class Diagram of main classes in Anonymization package:



METHODOLOGY:

Our project will involve a few key steps, including:
1. Data Classification
2. MDAV
3. Cluster construction and Anonymization through micro-aggregation

1.Data Classification: Here, we classify the data by dividing it into categories by watching the data types and values to numeric, categoric etc... Later, we categorize them by observing the attributes and the identities of the data into groups of Explicit Identifiers, Quasi Identifiers, Confidential and Non-Confidential. After the classification, we apply suppression on the Explicit Identifiers of the data to protect the privacy of the users.

2.MDAV: A popular micro aggregation-based method called MDAV (Maximum Distance to Average Vector) which divides the dataset into homogeneous clusters.

In order to run this algorithm, we need two inputs:
(i) a dataset centroid
(ii) a distance measure that calculates the distance between records in order to determine how similar they are.

In light of this, the records are grouped into clusters according to their distance scores.

3.Cluster construction and Anonymization through micro-aggregation: The original data set's records are divided up into a number of clusters, by the help of MDAV algorithm and then we apply k anonymization on the clustered data each of which has at least k records. And we could see the anonymized data by the help of K-Anonymization. Using the distance between the quasi-identifiers of the micro-aggregated clusters as the quality criterion, we choose which groups are to be merged after micro-aggregating and merging groups of records in the micro-aggregated data set. The original data set's quasi-identifier properties are first used to execute the micro aggregation method. Then, until t-closeness is satisfied, clusters of micro-aggregated records are combined. By first choosing the cluster that is the furthest from satisfying t-closeness (i.e., the one that has the most different confidential attribute distribution from the confidential attribute distribution across the board), and then by merging it with the cluster that is closest to it in terms of quasi-identifiers, we incrementally increase the level of t-closeness.

<u>RESULTS AND DISCUSSION:</u>

<u>SNAP SHOTS OF CODES:</u>

Here we attached few snapshots of different functions of our code:

```
public void calculate_kAnonymity(int k) throws InvalidValueException{
    ArrayList<Record>data;
    ArrayList<Record>dataAnomRecords;

    data = createRecords(dataOri);
    dataAnomRecords = kAnonymize(data, k);
    dataAnom = createMatrixStringFromRecords(dataAnomRecords);

}
public void calculate_tCloseness(int k, float t) throws InvalidValueException, InvalidConfidentialAttributeException{
    ArrayList<Record>data;
    ArrayList<Record>dataAnomRecords;

    data = createRecords(dataOri);
    dataAnomRecords = kAnonymize_tCloseness(data, k, t);
    dataAnom = createMatrixStringFromRecords(dataAnomRecords);
}
public void saveAnonymizedDataset(String locationAnonymized){
    File file;
    String s;

    file = new File(locationAnonymized);
    FileWriter fw = null;
    try {
        fw = new FileWriter(file);
        BufferedWriter bw = new BufferedWriter(fw);

        s = "";
        for(String attr:Record.getListNames()){
            s += attr + ",";
        }
        s = s.substring(beginIndex: 0, s.length()-1);
        bw.write(s);
        bw.newLine();

        for(String[] rec:dataAnom){
```

```java
                s = "";
                for(String attr:rec){
                    s += attr + ",";
                }
                s = s.substring(beginIndex: 0, s.length()-1);
                bw.write(s);
                bw.newLine();
            }

            bw.close();
        } catch (IOException e) {
            e.printStackTrace();
        }
        System.out.println("Protected file saved: " + locationAnonymized);
    }
    public InformationLossResult calculateInformationLoss() throws InvalidValueException{
        InformationLossResult informationLossResult;
        double SSE, recordDist;
        double originalVariance[], originalMean[], anonymizedVariance[], anonymizedMean[];
        int numAttr, numRecords;
        String values[];

        numAttr = Record.getNumAttr();
        numRecords = dataOri.length;
        originalVariance = new double[numAttr];
        originalMean = new double[numAttr];
        anonymizedVariance = new double[numAttr];
        anonymizedMean = new double[numAttr];
        Distances.calculateTypicalDeviations(dataOri);
        SSE = 0;
        for(int i=0; i<numRecords; i++){
            recordDist = Distances.euclideanDistNorm(dataOri[i], dataAnom[i]);
            SSE += (recordDist*recordDist);
        }
        SSE /= numRecords;
```

```java
        for(int i=0; i<numAttr; i++){
            if(Record.getListDataTypes().get(i).equalsIgnoreCase(Constants.categoric) ||
               Record.getListDataTypes().get(i).equalsIgnoreCase(Constants.categoricOrdinal) ||
               Record.getListDataTypes().get(i).equalsIgnoreCase(Constants.semantic)){
                originalMean[i] = 0;
                originalVariance[i] = 0;
            }
            else{
                values = new String[numRecords];
                for(int j=0; j<numRecords; j++){
                    if(Record.getListDataTypes().get(i).equalsIgnoreCase(Constants.date)){
                        values[j] = String.valueOf(Distances.getLongFromStringDate(dataOri[j][i]));
                    }
                    else{
                        values[j] = dataOri[j][i];
                    }
                }
                originalMean[i] = Statistics.calculateMean(values);
                originalVariance[i] = Statistics.calculateVariance(values, originalMean[i]);
            }
        }
        for(int i=0; i<numAttr; i++){
            if(Record.getListDataTypes().get(i).equalsIgnoreCase(Constants.categoric) ||
               Record.getListDataTypes().get(i).equalsIgnoreCase(Constants.categoricOrdinal) ||
               Record.getListDataTypes().get(i).equalsIgnoreCase(Constants.semantic)){
                originalMean[i] = 0;
                originalVariance[i] = 0;
            }
            else{
                values = new String[numRecords];
                for(int j=0; j<numRecords; j++){
                    if(Record.getListDataTypes().get(i).equalsIgnoreCase(Constants.date)){
                        values[j] = String.valueOf(Distances.getLongFromStringDate(dataAnom[j][i]));
                    }
                    else{
                        values[j] = dataAnom[j][i];
```

```java
                anonymizedVariance[i] = Statistics.calculateVariance(values, anonymizedMean[i]);
            }
        }

        informationLossResult = new InformationLossResult(SSE, originalVariance, originalMean,
                anonymizedVariance, anonymizedMean);

        return informationLossResult;
    }

    private static ArrayList<Record> createRecords(String[][] data){
        ArrayList<Record> records = new ArrayList<Record>();
        Record record = null;
        int id;

        id = 0;
        for(int i=0; i<data.length; i++){
            record = new Record(id);
            id++;
            for(int j=0; j<data[i].length; j++){
                record.setAttrValues(j, data[i][j]);
            }
            records.add(record);
        }

        System.out.println("Records loaded: " + records.size());
        return records;
    }
```

```java
public class AnonymizationConfig {
    String[] attributes;
    String[][] dataOri;
    public AnonymizationConfig(String xmlLocation, String datasetLocation)
            throws DatasetNotFoundException, XmlNotFoundException,AttributeNameNotFoundException,
                InvalidCSVFormatFoundException, NullValueException, InvalidAttributeTypeException,
                InvalidDataTypeException, InvalidProtectionException, InvalidConfidentialAttributeException,
                NoOntologyInSemanticDataTypeException, OntologyNotFoundException{
        loadDataset(datasetLocation);
        loadMetadataFromXML(xmlLocation);
        checkAttributesInXml();
        reOrderListsAccordingAttributeParameter(attributes);
    }
    private void loadDataset(String datasetLocation)
            throws DatasetNotFoundException, InvalidCSVFormatFoundException, NullValueException{
        DatasetParser datasetParser;
        File file;

        file = new File(datasetLocation);
        datasetParser = new DatasetParser(file , ",");
        attributes = datasetParser.parseHeaders();
        datasetParser.setNumAttr(attributes.length);
        dataOri = datasetParser.parseDataset();
    }
    private void loadMetadataFromXML(String xmlLocation)
            throws XmlNotFoundException, InvalidAttributeTypeException,
                InvalidDataTypeException, InvalidProtectionException, InvalidConfidentialAttributeException,
                NoOntologyInSemanticDataTypeException, OntologyNotFoundException{

        XmlReader.loadXmlFile(xmlLocation);
    }

    public String[][] getDataset(){
        return dataOri;
    }
```

```java
public String[] getAttributes() {
    return attributes;
}
private void checkAttributesInXml() throws AttributeNameNotFoundException{
    boolean ok;
    String nameInXml;

    for(int j=0; j<Record.getListNames().size(); j++){
        nameInXml = Record.getListNames().get(j);
        ok = false;
        for(int i=0; i<attributes.length; i++){
            if(nameInXml.equalsIgnoreCase(attributes[i])){
                ok = true;
                break;
            }
        }
        if(!ok){
            throw new AttributeNameNotFoundException(nameInXml);
        }
    }

}
private static void reOrderListsAccordingAttributeParameter(String attributes[]) {
    ArrayList<String>newListNames = new ArrayList<String>();
    ArrayList<String>newListAttrTypes = new ArrayList<String>();
    ArrayList<String>newListDataTypes = new ArrayList<String>();
    String attr, name;
    boolean ok;

    for(int i=0; i<attributes.length; i++){
        attr = attributes[i];
        ok = false;
        for(int j=0; j<Record.getListNames().size(); j++){
            name = Record.getListNames().get(j);
            if(attr.equals(name)){
                newListNames.add(name);
                newListAttrTypes.add(Record.getListAttrTypes().get(j));
                newListDataTypes.add(Record.getListDataTypes().get(j));
                ok = true;
                break;
            }
        }
        if(!ok){
            newListNames.add(attr);
            newListAttrTypes.add(Constants.non_confidential);
            newListDataTypes.add(Constants.categoric);
        }
    }
    Record.setListNames(newListNames);
    Record.setListAttrTypes(newListAttrTypes);
    Record.setListDataTypes(newListDataTypes);
    Record.setNumAttr(newListNames.size());

}

}
```

```java
private long calculateMean(int attr) throws InvalidValueException{
    long mean;
    String value = null;

    try {
        mean = 0;
        for(RecordQ reg:elements){
            value = reg.attrValues[attr];
            mean += Long.parseLong(value);
        }
        mean /= elements.size();
    } catch (NumberFormatException e) {
        throw new InvalidValueException(value);
    }

    return mean;
}

private long calculateMeanDouble(int attr) throws InvalidValueException{
    double mean;
    String value = null;

    try {
        mean = 0;
        for(RecordQ reg:elements){
            value = reg.attrValues[attr];
            mean += Long.parseLong(value);
        }
        mean /= elements.size();
    } catch (NumberFormatException e) {
        throw new InvalidValueException(value);
    }

    return (long)mean;
}
```

```java
public class InformationLossResult {
    private double SSE;
    private double originalVariance[];
    private double originalMean[];
    private double anonymizedVariance[];
    private double anonymizedMean[];
    public InformationLossResult(double SSE, double[] originalVariance, double[] originalMean,
        double[] anonymizedVariance, double[] anonymizedMean){
        this.SSE = SSE;
        this.originalVariance = originalVariance;
        this.originalMean = originalMean;
        this.anonymizedVariance = anonymizedVariance;
        this.anonymizedMean = anonymizedMean;
    }
    public double getSSE() {
        return SSE;
    }
    public double[] getOriginalVariance() {
        return originalVariance;
    }
    public double[] getOriginalMean() {
        return originalMean;
    }
    public double[] getAnonymizedVariance() {
        return anonymizedVariance;
    }
    public double[] getAnonymizedMean() {
        return anonymizedMean;
    }
    public String toString(){
        String s;
        s = "\n";
        s += "SSE: " + SSE + "\n";
        for(int i=0; i<originalVariance.length; i++){
            s += "Mean original dataset attribute " + i + ": " + originalMean[i] + "\n";
            s += "Variance original dataset attribute " + i + ": " + originalVariance[i] + "\n";
            s += "Mean anonymized dataset attribute " + i + ": " + anonymizedMean[i] + "\n";
            s += "Variance anonymized dataset attribute " + i + ": " + anonymizedVariance[i] + "\n";
        }
        return s;
    }
}
```

```java
ckage cat.urv.anonymization;
public class MdavClusteringAlgo {
    public void computeClusters(MdavData dataSet, int k) {
        int clustersDone = 0;
        long remainingRows = dataSet.rowCount();
        double lastCalc = System.currentTimeMillis();
        while ( remainingRows >= 3 * k) {
            Point avgPoint = dataSet.computeAverages();
            Point mostDistantFromAvg = dataSet.findMostDistantFrom(avgPoint);
            Point mostDistandFromRow = dataSet.findMostDistantFrom(mostDistantFromAvg);
            dataSet.computeClusterNearPoint(mostDistantFromAvg, k);
            dataSet.computeClusterNearPoint(mostDistandFromRow, k);
            clustersDone += 2;
            remainingRows -= 2 * k;
            if (clustersDone%100 == 0 ) {
                long newCalc = System.currentTimeMillis();
                System.out.println("Cluster done :" + clustersDone +  " in " + (newCalc - lastCalc) + " ms remaining " + rema
                lastCalc = newCalc;
            }
        }
        if ( remainingRows > 2 * k ) {
            Point avgEnd = dataSet.computeAverages();
            dataSet.computeClusterNearPoint(avgEnd, k);

        }
        dataSet.computeLastGroup();
    }
}
```

<u>OUTPUTS:</u>

Original Dataset text file:

data_example_snomed - Notepad

File Edit View

```
Patient_ID,First_Name,Last_name_1,Last_name_2,Sex,Age,PinCode,Serial_ID,Discharge_date,Admission_date,Mob_no,Systolic_number
00000946,Raquel,Manzano,Gallego,F,46,BCNCI,762656009,2015/08/21,2015/08/30,762656009,0000000001
00005923,Mireia,Calvo,Roman,F,40,BCNCI,722372005,2015/07/03,2015/07/17,722372005,0000000012
00020750,Teresa,Garcia,Guerrero,F,45,BCNPR,202821008,2015/11/08,2015/12/07,202821008,0000000021
00017868,Loida,Gomez,Vazquez,F,63,BCNCI,734009000,2015/11/19,2015/12/06,734009000,0000000033
00015553,Maria,Garcia,Padilla,F,73,BCNCI,716324008,2015/12/19,2015/01/13,716324008,0000000045
00022957,Sandra,Lopez,Jimenez,F,31,BCNPR,365445003,2015/07/21,2015/08/06,365445003,0000000057
00000364,Josefa,Gomez,Martinez,F,23,BCNCI,249413006,2015/02/26,2015/03/06,249413006,0000000067
00015556,Francisco,Diaz,Garcia,M,69,BCNCI,168627008,2015/03/02,2015/03/29,168627008,0000000076
00000996,Juana,Romero,Dominguez,F,40,BCNCI,299367000,2015/03/05,2015/03/05,299367000,0000000089
00018043,Francesc,Conde,Gonzalez,M,54,BCNPR,249411008,2015/09/07,2015/09/21,249411008,0000000091
00012929,Maria,Lopez,Hernandez,F,24,BCNCI,423316001,2015/07/19,2015/07/20,423316001,0000000010
00003253,Josep,Ortega,Sanchez,M,36,BCNCI,422840005,2015/05/12,2015/06/05,422840005,0000000011
00012184,Luis,Fernandez,Rueda,M,57,BCNCI,64314006,2015/06/02,2015/06/21,64314006,0000000012
00022546,Paula,Soriano,Lozano,F,66,BCNPR,711374009,2015/04/04,2015/04/21,711374009,0000000013
00013200,Sandra,Galvez,Calvo,F,76,BCNCI,299016006,2015/03/04,2015/03/24,299016006,0000000014
```

2 - Anonymized Dataset

data_example_snomed_anom - Notepad

File Edit View

```
Patient_ID,First_Name,Last_name_1,Last_name_2,Sex,Age,PinCode,Serial_ID,Discharge_date,Admission_date,Mob_no,Systolic_number
*,*,*,*,F,35,BCNCI,722372005,2015/07/03,2015/07/27,722372005,0000000012
*,*,*,*,F,35,BCNCI,365445003,2015/07/21,2015/07/27,365445003,0000000057
*,*,*,*,F,39,BCNCI,249411008,2015/09/07,2015/08/20,249411008,0000000091
*,*,*,*,F,39,BCNCI,423316001,2015/07/19,2015/08/20,423316001,0000000010
*,*,*,*,F,47,BCNCI,762656009,2015/08/21,2015/04/16,762656009,0000000001
*,*,*,*,F,47,BCNCI,716324008,2015/12/19,2015/04/16,716324008,0000000045
*,*,*,*,F,53,BCNCI,299367000,2015/03/05,2015/03/28,299367000,0000000089
*,*,*,*,F,53,BCNCI,711374009,2015/04/04,2015/03/28,711374009,0000000013
*,*,*,*,F,54,BCNCI,202821008,2015/11/08,2015/12/06,202821008,0000000021
*,*,*,*,F,54,BCNCI,734009000,2015/11/19,2015/12/06,734009000,0000000033
*,*,*,*,F,56,BCNCI,299016006,2015/03/04,2015/04/29,299016006,0000000014
*,*,*,*,F,56,BCNCI,422840005,2015/05/12,2015/04/29,422840005,0000000011
*,*,*,*,M,63,BCNCI,168627008,2015/03/02,2015/05/10,168627008,0000000076
*,*,*,*,M,63,BCNCI,64314006,2015/06/02,2015/05/10,64314006,0000000012
```

3 - Anonymized Dataset

data_example_snomed_anom - Notepad

File Edit View

```
Patient_ID,First_Name,Last_name_1,Last_name_2,Sex,Age,PinCode,Serial_ID,Discharge_date,Admission_date,Mob_no,Systolic_number
*,*,*,*,F,41,BCNCI,762656009,2015/08/21,2015/08/23,762656009,0000000001
*,*,*,*,F,41,BCNCI,249411008,2015/09/07,2015/08/23,249411008,0000000091
*,*,*,*,F,41,BCNCI,423316001,2015/07/19,2015/08/23,423316001,0000000010
*,*,*,*,F,45,BCNCI,716324008,2015/12/19,2015/02/16,716324008,0000000045
*,*,*,*,F,45,BCNCI,249413006,2015/02/26,2015/02/16,249413006,0000000067
*,*,*,*,F,45,BCNCI,299367000,2015/03/05,2015/02/16,299367000,0000000089
*,*,*,*,F,46,BCNPR,202821008,2015/11/08,2015/10/26,202821008,0000000021
*,*,*,*,F,46,BCNPR,734009000,2015/11/19,2015/10/26,734009000,0000000033
*,*,*,*,F,46,BCNPR,365445003,2015/07/21,2015/10/26,365445003,0000000057
*,*,*,*,M,55,BCNCI,722372005,2015/07/03,2015/06/01,722372005,0000000012
*,*,*,*,M,55,BCNCI,64314006,2015/06/02,2015/06/01,64314006,0000000012
*,*,*,*,M,55,BCNCI,168627008,2015/03/02,2015/06/01,168627008,0000000076
*,*,*,*,F,59,BCNCI,711374009,2015/04/04,2015/04/26,711374009,0000000013
*,*,*,*,F,59,BCNCI,299016006,2015/03/04,2015/04/26,299016006,0000000014
*,*,*,*,F,59,BCNCI,422840005,2015/05/12,2015/04/26,422840005,0000000011
```

4 - Anonymized Dataset

data_example_snomed_anom - Notepad

File Edit View

```
Patient_ID,First_Name,Last_name_1,Last_name_2,Sex,Age,PinCode,Serial_ID,Discharge_date,Admission_date,Mob_no,Systolic_number
*,*,*,*,F,46,BCNCI,762656009,2015/08/21,2015/06/25,762656009,0000000001
*,*,*,*,F,46,BCNCI,299367000,2015/03/05,2015/06/25,299367000,0000000089
*,*,*,*,F,46,BCNCI,249411008,2015/09/07,2015/06/25,249411008,0000000091
*,*,*,*,F,46,BCNCI,711374009,2015/04/04,2015/06/25,711374009,0000000013
*,*,*,*,F,48,BCNCI,722372005,2015/07/03,2015/04/14,722372005,0000000012
*,*,*,*,F,48,BCNCI,716324008,2015/12/19,2015/04/14,716324008,0000000045
*,*,*,*,F,48,BCNCI,249413006,2015/02/26,2015/04/14,249413006,0000000067
*,*,*,*,F,48,BCNCI,168627008,2015/03/02,2015/04/14,168627008,0000000076
*,*,*,*,F,54,BCNCI,202821008,2015/11/08,2015/08/18,202821008,0000000021
*,*,*,*,F,54,BCNCI,734009000,2015/11/19,2015/08/18,734009000,0000000033
*,*,*,*,F,54,BCNCI,64314006,2015/06/02,2015/08/18,64314006,0000000012
*,*,*,*,F,54,BCNCI,299016006,2015/03/04,2015/08/18,299016006,0000000014
```

Information Loss Result for 2- Anonymized Dataset

Command Prompt

```
SSE: 132.09578464919207
Mean original dataset attribute 0: 0.0
Variance original dataset attribute 0: 0.0
Mean anonymized dataset attribute 0: 0.0
Variance anonymized dataset attribute 0: 0.0
Mean original dataset attribute 1: 0.0
Variance original dataset attribute 1: 0.0
Mean anonymized dataset attribute 1: 0.0
Variance anonymized dataset attribute 1: 0.0
Mean original dataset attribute 2: 0.0
Variance original dataset attribute 2: 0.0
Mean anonymized dataset attribute 2: 0.0
Variance anonymized dataset attribute 2: 0.0
Mean original dataset attribute 3: 0.0
Variance original dataset attribute 3: 0.0
Mean anonymized dataset attribute 3: 0.0
Variance anonymized dataset attribute 3: 0.0
Mean original dataset attribute 4: 0.0
Variance original dataset attribute 4: 0.0
Mean anonymized dataset attribute 4: 0.0
Variance anonymized dataset attribute 4: 0.0
Mean original dataset attribute 5: 49.53333333333333
Variance original dataset attribute 5: 283.71555555555557
Mean anonymized dataset attribute 5: 49.4
Variance anonymized dataset attribute 5: 78.24
Mean original dataset attribute 6: 0.0
Variance original dataset attribute 6: 0.0
Mean anonymized dataset attribute 6: 0.0
Variance anonymized dataset attribute 6: 0.0
Mean original dataset attribute 7: 0.0
Variance original dataset attribute 7: 0.0
Mean anonymized dataset attribute 7: 0.0
Variance anonymized dataset attribute 7: 0.0
Mean original dataset attribute 8: 1.435406 76E12
Variance original dataset attribute 8: 7.12908324864E19
Mean anonymized dataset attribute 8: 1.435406 76E12
Variance anonymized dataset attribute 8: 7.12908324864E19
Mean original dataset attribute 9: 1.43468676E12
Variance original dataset attribute 9: 7.02162100224E19
Mean anonymized dataset attribute 9: 1.4346522E12
```

Information Loss Result for 3 - Anonymized Dataset

```
SSE: 136.36105644443438
Mean original dataset attribute 0: 0.0
Variance original dataset attribute 0: 0.0
Mean anonymized dataset attribute 0: 0.0
Variance anonymized dataset attribute 0: 0.0
Mean original dataset attribute 1: 0.0
Variance original dataset attribute 1: 0.0
Mean anonymized dataset attribute 1: 0.0
Variance anonymized dataset attribute 1: 0.0
Mean original dataset attribute 2: 0.0
Variance original dataset attribute 2: 0.0
Mean anonymized dataset attribute 2: 0.0
Variance anonymized dataset attribute 2: 0.0
Mean original dataset attribute 3: 0.0
Variance original dataset attribute 3: 0.0
Mean anonymized dataset attribute 3: 0.0
Variance anonymized dataset attribute 3: 0.0
Mean original dataset attribute 4: 0.0
Variance original dataset attribute 4: 0.0
Mean anonymized dataset attribute 4: 0.0
Variance anonymized dataset attribute 4: 0.0
Mean original dataset attribute 5: 49.53333333333333
Variance original dataset attribute 5: 283.71555555555557
Mean anonymized dataset attribute 5: 49.2
Variance anonymized dataset attribute 5: 44.96000000000001
Mean original dataset attribute 6: 0.0
Variance original dataset attribute 6: 0.0
Mean anonymized dataset attribute 6: 0.0
Variance anonymized dataset attribute 6: 0.0
Mean original dataset attribute 7: 0.0
Variance original dataset attribute 7: 0.0
Mean anonymized dataset attribute 7: 0.0
Variance anonymized dataset attribute 7: 0.0
Mean original dataset attribute 8: 1.43540676E12
Variance original dataset attribute 8: 7.12908324864E19
Mean anonymized dataset attribute 8: 1.43540676E12
Variance anonymized dataset attribute 8: 7.12908324864E19
Mean original dataset attribute 9: 1.43468676E12
Variance original dataset attribute 9: 7.02162100224E19
```

Information Loss Result for 4 - Anonymized Dataset

```
SSE: 288.1887437371673
Mean original dataset attribute 0: 0.0
Variance original dataset attribute 0: 0.0
Mean anonymized dataset attribute 0: 0.0
Variance anonymized dataset attribute 0: 0.0
Mean original dataset attribute 1: 0.0
Variance original dataset attribute 1: 0.0
Mean anonymized dataset attribute 1: 0.0
Variance anonymized dataset attribute 1: 0.0
Mean original dataset attribute 2: 0.0
Variance original dataset attribute 2: 0.0
Mean anonymized dataset attribute 2: 0.0
Variance anonymized dataset attribute 2: 0.0
Mean original dataset attribute 3: 0.0
Variance original dataset attribute 3: 0.0
Mean anonymized dataset attribute 3: 0.0
Variance anonymized dataset attribute 3: 0.0
Mean original dataset attribute 4: 0.0
Variance original dataset attribute 4: 0.0
Mean anonymized dataset attribute 4: 0.0
Variance anonymized dataset attribute 4: 0.0
Mean original dataset attribute 5: 49.53333333333333
Variance original dataset attribute 5: 283.71555555555557
Mean anonymized dataset attribute 5: 49.333333333333336
Variance anonymized dataset attribute 5: 11.555555555555554
Mean original dataset attribute 6: 0.0
Variance original dataset attribute 6: 0.0
Mean anonymized dataset attribute 6: 0.0
Variance anonymized dataset attribute 6: 0.0
Mean original dataset attribute 7: 0.0
Variance original dataset attribute 7: 0.0
Mean anonymized dataset attribute 7: 0.0
Variance anonymized dataset attribute 7: 0.0
Mean original dataset attribute 8: 1.43540676E12
Variance original dataset attribute 8: 7.12908324864E19
Mean anonymized dataset attribute 8: 1.43540676E12
Variance anonymized dataset attribute 8: 7.12908324864E19
Mean original dataset attribute 9: 1.43468676E12
Variance original dataset attribute 9: 7.02162100224E19
Mean anonymized dataset attribute 9: 1.4346522E12
```

EFFICIENCY:

Thus, from the results generated, the calculation of SSE which is a parameter of Euclidean distance is done using the metrics such as Mean and Variance of the data in the clusters. Therefore, higher the value of SSE the higher the privacy of the data.

CONCLUSIONS AND FUTURE ENHANCEMENTS:

The suggested t-closeness model uses micro aggregation to maintain the privacy of sensitive characteristics safely and effectively in any system. Other privacy models like k-anonymity and l-diversity do not offer attribute disclosure protection. The micro aggregation disturbs the data, and the additional masking freedom enables enhancing the usability of the data in several ways, including enhancing data granularity, minimizing the influence of outliers, and avoiding discretization of numerical data. One of the tightest privacy assurances is provided by the suggested micro aggregation technique to produce t-close data sets in microdata.Thus, this study demonstrates the use of microaggregation to provide k-anonymous t-closeness. The microaggregation-based t-closeness algorithm is described and analyzed using three different K values. The algorithm we have used is based on executing micro aggregation in the typical manner, followed by clustering to the extent required to meet the t-closeness criteria. Although it is easy to use and can be used with any method, it could not perform well in terms of utility since clusters could become quite large. This can be accepted as a challenge and included in our future work to potentially enhance this value for huge data clusters. In an effort to increase the usefulness of the anonymized data, one potential enhancement could involve changing the Microaggregation method to take t-closeness into consideration.

VIDEO LINK:

Presentation Link

REFERENCES:

1.Josep Domingo-Ferrer and Jordi Soria-Comas. “Steered Microaggregation: A Unified Primitive for

Anonymization of Data Sets and Data Streams”.IEEE Transactions on Information Forensics and Security (Volume: 14, Issue: 12, December 2019)

2. David Sánchez, Sergio Martínez , Josep Domingo-Ferrer, Jordi Soria-Comas and Montserrat Batet. “μ-ANT: Semantic Microaggregation-based Anonymization Tool”. Bioinformatics, Volume 36, Issue 5, March 2020, Pages 1652–1653

3. J.M. MATEO-SANZ J and DOMINGO-FERRER. “A COMPARATIVE STUDY OF MICROAGGREGATION METHODS”. Institut d'Estadística de Catalunya.

4. D. Rebollo-Monedero, J. Forné and J. Domingo-Ferrer, "From t-Closeness-Like Privacy to Postrandomization via Information Theory," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 11, pp. 1623-1636, Nov. 2010, doi: 10.1109/TKDE.2009.190.

5. Shi, Yancheng & Zhang, Zhenjiang & Shen, Bo. “Data Privacy Protection Based on Micro Aggregation with Dynamic Sensitive Attribute Updating”. Sensors. 18. 2307. 10.3390/s18072307. (2018).

6. Y. Sei, H. Okumura, T. Takenouchi and A. Ohsuga, "Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness," in IEEE Transactions on Dependable and Secure Computing, vol. 16, no. 4, pp. 580-593, 1 July-Aug. 2019, doi: 10.1109/TDSC.2017.2698472.

7. Salvatore Ruggieri. “Using t-closeness anonymity to control for non-discrimination”.

Transactions on Data PrivacyVolume 7,Issue 2-August 2014 pp 99–129

8. Liang, H., Yuan, H. (2013). On the Complexity of t-Closeness Anonymization and Related Problems. In: Meng, W., Feng, L., Bressan, S., Winiwarter, W., Song, W. (eds) Database Systems for Advanced Applications. DASFAA 2013. Lecture Notes in Computer Science, vol 7825. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37487-6_26

9.Saraswathi, S., and K. Thirukumar. "Enhancing utility and privacy using t-closeness for multiple sensitive attributes." Advances in Natural and Applied Sciences, vol. 10, no. 5, May 2016, pp. 6+. Gale Academic OneFile, link.gale.com/apps/doc/A465808911/AONE?u=anon~eb9b9c21&sid=googleScholar&xid=edbf2d02. Accessed 12 Sept. 2022.

10.Roy, Debaditya & Jena, Sanjay. (2013). Determining t in t-closeness using Multiple Sensitive Attributes. International Journal of Computer Applications. 70. 47-51. 10.5120/12179-8291.

11.Sergio Martínez, David Sánchez, Aida Valls,A semantic framework to protect the privacy of electronic health records with non-numerical attributes,Journal of Biomedical Informatics,Volume 46, Issue 2,2013,Pages 294-303.

12.Sande, Gordon. "Methods for Data Directed Microaggregation in One Dimension." Proceedings of the NTTS&ETK 2001 (2001).

13.Domingo-Ferrer, Josep & Sebé, Francesc & Solanas, Agusti. (2008). An Anonymity Model Achievable Via Microaggregation. 209-218. 10.1007/978-3-540-85259-9_14.

14.Domingo-Ferrer, Josep & Trujillo-Rasua, Rolando. (2012). Microaggregation- and permutation-based anonymization of movement data. Information Sciences. 208. 55–80. 10.1016/j.ins.2012.04.015.

15.Khapekar, Sonu V., and Lomesh Ahire. "Privacy Protection of Sensitive Microdata in Healthcare System using t-Closeness through Microaggregation." (2017).

16.Domingo-Ferrer, Josep, and Jordi Soria-Comas. "From t-closeness to differential privacy and vice versa in data anonymization." Knowledge-Based Systems 74 (2015): 151-158.

17. Soria-Comas, Jordi, et al. "t-closeness through microaggregation: Strict privacy with enhanced utility preservation." IEEE Transactions on Knowledge and Data Engineering 27.11 (2015): 3098-3110.

18. Soria-Comas, Jordi, and Josep Domingo-Ferrert. "Differential privacy via t-closeness in data publishing." 2013 Eleventh Annual Conference on Privacy, Security and Trust. IEEE, 2013.

19.Soria-Comas, Jordi, and Josep Domingo-Ferrer. "Probabilistic k-anonymity through microaggregation and data swapping." 2012 IEEE International Conference on Fuzzy Systems. IEEE, 2012.

20. Solanas, A. and Pietro, R.D., 2008, October. A linear-time multivariate micro-aggregation for privacy protection in uniform very large data sets. In International Conference on Modeling Decisions for Artificial Intelligence (pp. 203-214). Springer, Berlin, Heidelberg.

21. Dangi, A.P. and Mogili, R., 2012. Privacy preservation measure using t-closeness with combined l-diversity and k-anonymity. International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), 1(8), pp.28-33.

22. Wang, R., Zhu, Y., Chen, T.S. and Chang, C.C., 2018. Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness. Journal of Computer Science and Technology, 33(6), pp.1231-1242.

23. Wang, M., Jiang, Z., Zhang, Y. and Yang, H., 2018. T-closeness slicing: A new privacy-preserving approach for transactional data publishing. INFORMS Journal on Computing, 30(3), pp.438-453.

24. Zouinina, S., Bennani, Y., Rogovschi, N. and Lyhyaoui, A., 2021. Data anonymization through collaborative multi-view microaggregation. Journal of Intelligent Systems, 30(1), pp.327-345.

25. Defays, D. and Anwar, M.N., 1998. Masking microdata using micro-aggregation. Journal of Official Statistics, 14(4), p.449.

26. FATHIMA, N. and KOUSAR, M., 2017. Privacy Preserving with Utility Preservation through Microaggregation.

27. PRAVALLIKA, A. and SAPTHAMI, I., 2017. T-Closeness Through Microaggregation: Strict Privacy With Enhanced Utility Preservation.

28. Tanay, I.S.L., Nagargoje, V.J. and Inamdar, A., 2013. Security for Personal Credentials in Big Data: Through Microaggregation and TCloseness.

29. Domingo-Ferrer, J., Sebé, F. and Solanas, A., 2008. A polynomial-time approximation to optimal multivariate microaggregation. Computers & Mathematics with Applications, 55(4), pp.714-732.

30. Lin, J.L., Wen, T.H., Hsieh, J.C. and Chang, P.C., 2010. Density-based microaggregation for statistical disclosure control. Expert Systems with Applications, 37(4), pp.3256-3263.