DEVELOPMENT OF AN HYBRID K-ANONYMITY MODEL FOR DATA MINING PRIVACY PROTECTION

BY

ABUBAKAR MUHAMMAD HASHIMU
(PGS/02/15202611)

M.Sc. COMPUTER SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

JULY, 2018

KEBBI STATE UNIVERSITY OF SCIENCE AND TECHNOLOGY, ALIERO
(POSTGRADUATE SCHOOL)

DEVELOPMENT OF AN HYBRID K-ANONYMITY MODEL FOR DATA MINING PRIVACY PROTECTION

A Dissertation Submitted to the Postgraduate School

KEBBI STATE UNIVERSITY OF SCIENCE AND TECHNOLOGY
ALIERO, NIGERIA.
In Partial Fulfillment of the Requirements
For the Award of the Degree of
MASTER OF SCIENCE (COMPUTER SCIENCE)

BY

ABUBAKAR MUHAMMAD HASHIMU (PGS/02/15202611)
DEPARTMENT OF COMPUTER SCIENCE

JULY, 2018

DEDICATION

To my father, Alhaji Muhammad Abubakar, may Almighty Allah continue to protect and bless him and reward him with Al-Jannatul Al-Firdaus, Amin.

DECLARATION

I declare that the work in this Dissertation entitled DEVELOPMENT OF AN HYBRID K-ANONYMITY MODEL FOR DATA MINING PRIVACY PROTECTION has been carried out by me in the Department of Computer Science. The information derived from the literature has been duly acknowledged in the text and a list of references provided. No part of this dissertation was previously presented for another degree or diploma at this or any other institution.

Muhammad Hashimu Abubakar	__________________	_____________
Name of Student	Signature	Date

CERTIFICATION

This dissertation entitled DEVELOPMENT OF AN HYBRID K-ANONYMITY MODEL FOR DATA MINING PRIVACY PROTECTION by Muhammad Hashimu Abubakar meets the regulations governing the award of the degree of Master in Computer Science of the Kebbi State University of Science and Technology Aleiro, and is approved for its contribution to knowledge and literary presentation.

Dr. O. Alimi ______________________ ______________________
Major Supervisor Sign. Date

Dr. Muhammad Garba ______________________ ______________________
Co-Supervisor Sign. Date

Dr. I. Sulaiman ______________________ ______________________
Head of Department Sign. Date

______________________ ______________________ ______________________
External Examiner Sign. Date

ACKNOWLEDGEMENTS

In the Name of Allah, Most Beneficent, Most Merciful.

All praises are due to Allah (S.W.T.) who has given me the health, the courage and the ability to conduct this work successfully. May His peace and blessings be upon His beloved servant and Messenger, Muhammad (S.A.W.), his household, his Companions and all those who follow their footstep till the last day.

First of all, I am very much thankful to Allahu SubhanahuWaTa’ala for His blessings and showing me the right direction to reach here.

I am greatly indebted to many who have in one way or the other contributed to whatever is the success of this dissertation.

My profound gratitude goes to my parents, Alhaji Muhammad Abubakar, Malama Amina Abubakar, Malama Aishatu and Malama Hadiza Abubakar for their moral, material and financial support for which my mere expression of thanks does not suffice.

No words to express my gratitude and pleasure towards the guidance and help I have received from my supervisors, Dr. Alimi Olasunkanmi and Dr. Rahul Styakhm whose constant guidance and encouragement made the completion of my MSc. dissertation possible.

My appreciation also goes to all the Lecturers of the Department of Computer Science for instilling in me the basic knowledge about the field that greatly benefitted me while carrying out the project and achieving the goal.. They include: Dr. A. R. Yauri, the Deputy Dean, Faculty of Engineering and HOD ICT, KSUSTA, Dr S. Rahul, the PG Coordinator, Department of Computer Science, KSUSTA, Dr I. Sulaiman, the HOD, Department of Computer Science, KSUSTA, Dr. Muhammad Garba, Department of Computer Science, KSUSTA, Mal. Bashir Aliyu Yauri, Department of Computer Science, KSUSTA, Mr.Bakare A. Hakeem, Department of Computer Science, KSUSTA, Department of Computer Science, KSUSTA and Mal. M. S. Aliero of ICT Department, KSUSTA to mention but a few.

I have to express my sincere gratitude to Prof. Yakubu Aliyu, of the Department of physics with Electronics FUBK.

This dissertation would not have been possible without the support of best friend Abdulhakeem Ibrahim of the Department of Computer Science FUBK, I owe my sincerest gratitude to him for his endless support even with his tight schedules.

I remain thankful to my friends, Mohammad Saidu Aleiro, Faruk Musa Aliyu, Armayau Saadu, Sabiu Lawal, Aminu Aliyu, Shehu Abubakar for their friendly encouragement.

I feel sincerely appreciative to my lovely wife in person of Hussaina Zubairuand my little daughter Amina Muhammad Abubakar (Mama).I feel sincerely appreciative to my siblings in person of Aminu, Hassana, Maimuna, Hamza, Fatima, Ibrahim, Abdulmalik, Tijani, Quasim, Hafsatu, Shafatu, Nafisa, Summayya, Umar, Isah, Sadiq and Ummu-Salama.

Finally, to my fellow colleagues in Department of Computer Science Kebbi State University of Science and Technology, Aliero who are too numerous to mention for their love throughout the course of this program, I thank you all.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABBREVIATIONS AND SYMBOLS

PPDM: Privacy Preserving Data Mining

QI: Quasi-Identifier

K:The Anonymization Level

IL: Information Loss

GH: Generalization Hierarchy

HF: Hidden Failure

MD: Minimal Distortion

E-Health: Electronic Health

IDE: Integrated Development Environment

WSN: Wireless Sensor Network

LBS: Location-Based Service

NIST: National Institute of Standards and Technology

PT: Private Table

ABSTRACT

Data mining is a technique where massive amounts of both sensitive and non-sensitive data are collected and examined. While distributing such private data, privacy preserving becomes an important issue. Various methods and techniques have been introduced in privacy preserving data mining to undertake this problem. The main intention of privacy preserving data mining is to extract the knowledge without disclosing private data and it also concerns about the sequential release of data. Still privacy of individual in the existing system is exposed to background knowledge and homogeneity attack. Individual can be simply identified in a published data by simply linking of record, and this issue can be seen as main challenge of K-anonymity in privacy preserving data mining security domain. Proposed enhance K-anonymity will limit re-identification in a microdata set , which relaxes the indistinguishability requirement of k-anonymity and only requires that the probability of re-identification be the same as in k-anonymity. This method is resistant against homogeneity and background knowledge attack. To ripen this method of privacy preserving, L-Diversity was added to K-anonymity to make the identification level more secure and robust. The algorithm was evaluated based on Information loss and Privacy level. Our proposed system produced an anonymous dataset with information loss value lower to 10% than information loss of 50% of existing system. These enhancements are obtained at the expense of acceptable additional computation overheads.

CHAPTER ONE

INTRODUCTION

This chapter discusses the introductory part of this dissertation, which includes background of the study, research motivation, problem statement, research objectives, and finally the dissertation outline.

1.1 Background to the Study

As the information era continues to advance and diverse information technology architectures are designed and implemented, there has arisen the need to design educational, business and enterprise solutions that would take full advantage of these advancing technologies. At the center of this “information revolution” is the Internet, a global network of networks that has steadily become an integral part in the way people communicate, educate, do business and relate to one another. With rapid development of technology and large scale of network, information is stored, published and shared. Information such as consumption record, crime record, health services organization, credit card information can be used to help government, company and organization to do specific research or make certain decision which will benefit society.

As we spend more of our time online in information-rich and personalized environments, it becomes increasingly easier for details from our offline life to meld with their online presence. Through Facebook and other social networks, our preferences in friends, food, and games becomes visible to others; LinkedIn makes our employment history and professional colleagues public information; even public Pandora profiles reveal our detailed tastes in music. Records stored offline also contain just as rich a bank of data; public voter records show who is likely to vote, and purchase histories show what someone has bought (and is likely to buy again).

As this data becomes increasingly easy to access and collect, many web providers take advantage of the ability to analyze this data and use it to tailor a person's online experiences to their specific interests. The question becomes, how should one act on this data in a responsible manner? Not many companies have found a foolproof way to do this. Most recently, Facebook and Myspace have received public and media criticism for passing demographic and interest data to advertisers. An intuitive way to balance the need for privacy but provide customization is to use an anonymized data release to customize web experiences. It is reasonable to personalize a website based on a person's probable interests, as long as the website cannot determine the person's real identity (unless they intentionally log-in).

However, such access to information can lead to compromised individual privacy Since shared, published information contains certain attributes that can be linked with external knowledge to identify individual's record directly. Many organizations that involves in data sharing tend to overcome such problem by removing or changing individual information such as name, medical care, and card number before publishing data record containing aggregated data information about individual, with assumption, that record of individual will not be recognized. Thus privacy of such as medical status is preserved.

Privacy is the ability of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively. The boundaries and content of what is considered private differ among cultures and individuals, but share common themes. When something is private to a person, it usually means that something is inherently special or sensitive to them. The domain of privacy partially overlaps security (confidentiality), which can include the concepts of appropriate use, as well as protection of information. Privacy may also take the form of bodily integrity.

Unfortunately, releasing data and ensuring anonymity at the same time is very difficult. To keep a user's identity unknown, it is not enough to strip data of Personally Identifying Information (PII). Even non PII data can still be traced back to a single person if enough personal microdata is provided. For example, age, income, job, marital status, and movie interests can often identify a single individual. The field of privacy preserving data mining studies how to mine data in a way that simultaneously maintains privacy for the individuals whose records are being mined, while keeping the released dataset rich enough that it is useful for data-mining purposes. One popular strategy for maintaining privacy in a released dataset is simply to ensure that the dataset remains anonymous. K-anonymity was the first carefully studied model for data anonymity.The k-anonymity privacy assurance guarantees that a published record can be identified as one of no fewer than k individuals. The k-anonymity problem has traditionally been researched from the perspective of sensitive data disclosure a commonly cited domain is that of medical records released for data mining purposes, where it is important to be able to link diseases to demographic data, without revealing that an individual has a particular disease.

Most research on data exposure focuses on shielding individuals from the release of sensitive attributes which could be embarrassing or harmful, if released. conventional examples of sensitive data include medical records and criminal records. However, the objective here is somewhat different; the objective is to prevent a user from being identified by their available data, and none of the targeted information is considered to be sensitive data. The similarity of the data targeting problem described above to the k-anonymity problem however indicates that algorithms developed to ensure k-anonymity could be used to efficiently anonymize this targeting data. Unfortunately, the k-anonymity problem is a very difficult problem; most research has focused on developing efficient heuristic approaches. Since the original model of k-anonymity was developed, many other data anonymity models have been studied which ensure security under either more or less aggressive attack models for a variety of data models.

1.2 Research Motivation

Throughout the entire data mining operation (starting with collection of data through finding knowledge) the data used normally consist of sensitive personal information about many individuals which they do not want to disclose to anyone such as the owner of the dataset, collectors, users, and miners. There are lot of opportunities to mismanage the delicate information if the sensitive data about individuals are leaked (Benjamin, Ke, Rui and Philip, 2010). Availability of voluminous data assures the possibility of learning a great deal of information regarding individuals out of the public data. And this fact naturally leads to more responsibility with regard to privacy of person's individual data (Charu and Philip, 2008). Entire information regarding any person often contains certain private information. Careless distribution of such data means instant violation of the privacy with regard to the individual. Privacy has been defined as the condition or quality of being blocked or secluded from others view or presence. When data mining gets related with privacy, privacy suggests keeping an individual's information from freely being obtainable by other people. As long as one does not feel his or her personal information has been negatively used, privacy is not considered to be violated. When once personal sensitive information has been revealed, one is not in a position to prevent misuse of the same. Privacy preservation has been treated very crucial for evading information spillage for the most effective usage of voluminous data. It involves storing data in electronic format with no disturbance to the individual. Hence, privacy has to be preserved before, during and after the data mining procedure. Privacy Preserving is found to have emerged as one major concern with regard to the data mining process success. PPDM is used to protect the privacy of an individual's personal data or classified knowledge without having to sacrifice complete usage of the required data. Privacy intrusions on the part of personal data are common, people have become aware of this, and they are naturally very hesitant of sharing their classified information. Importance of the issues in PPDM have been realized more pronouncedly during the recent period. This is due to the fact that the ability to save users' personal data has increased and data mining algorithm leveraging these information's have become increasingly sophisticated. It is not possible to apply privacy constraints in a single step. One must remember that PPDM technique has to be applied throughout the data mining practice beginning with data collection through information/knowledge generation. The goal of PPDM consists of constructing procedures to transform the raw data in such a way as to maintain confidentiality of private knowledge and private data even subsequent to data mining process. Hence, as an ancient quote says it's true that "Old is Gold" those methods help in developing new ideas to sort out problems, thus leads to discovery of a new horizon. Little modifications or combinations of two or more algorithms make a lot of progress which leads to improvements.

1.3 Problem Statement

Much effort has been put into security to addressed breach of privacy in data mining, still privacy of individual is susceptible to background knowledge and homogeneity attack. Individual can be simply identified in a published data by simply linking of record, and this issue can be seen as main challenge in privacy preserving data mining security domain. Hence individual can be identified by linking attributes attack, since attribute such as name, date of birth, gender can potentially identify individual in population. Therefore, this dissertation focuses on study of different solution available to defend against homogeneity and background knowledge attack and designing and developing new approach or enhancing existing approach to avoid this problem.

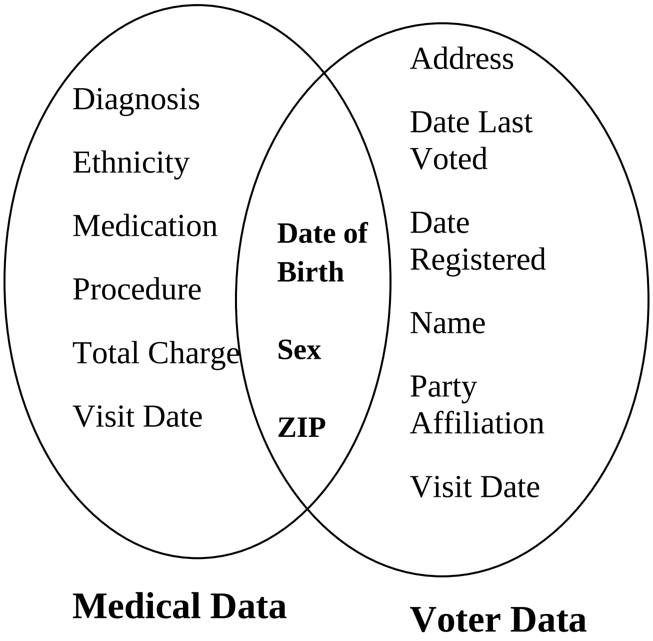


Figure 1.1 Linking to re-identify record owner

1.4 Research Aim and Objectives

The aim of this research work is to Design and implement algorithm, that guarantees privacy of sensitive data even when some aggregate data have been published, data owner cannot be identified. The specific objectives are to:

1. To design an algorithm that will allow data miner to exercise their activities without data about individual being disclosed.
2. Implement the designed algorithm.
3. Appraise performance of the algorithm with respect to anonymizing data, large data size, Running Time, Integrity and Preserving Privacy.

1.5 Research Contribution to Knowledge

The main contributions of this work are two folds:

Theoretical contribution, a new anonymizer is proposed and designed so as to produce a strong algorithm that cannot be compromised by background knowledge and homogeneity attack.

Practical contribution, the new system will be implemented and evaluated with data of different sizes. The algorithm is to treat the problem of background knowledge, homogeneity attack and also improve the robustness of data anonymization, unlike the existing algorithm which render the anonymzed data to attacks.

1.6 Research Methodology

- A critical review of the existing work on Privacy Preserving Data Mining.
- Design of the proposed algorithm that will diversify the anonymous dataset.
- Open source technologies (the Java Programming Language and Netbeans IDE) will be use in implementing the algorithm.
- Evaluate and compare the algorithm with the work of Srivastava, 2015 based on the following parameters: Information loss, privacy constraints and data utility.

1.7 Organization of the Dissertation

The rest of the dissertation is structured as follows:

Chapter Two: This chapter reviews the relevant literature that covers Privacy Preserving Data Mining. An overview of the K-anonymity is stated in detail. Categories of K-anonymity algorithms are described. Concepts related to the proposed research work and also review of related works/state-of-art.

Chapter Three: The methodology of the proposed system is demonstrated. The modification aspects of our proposed algorithm are presented in detail.

Chapter Four: The research implementation details of the system with screen shots showing the graphical user interface are displayed in this chapter. This chapter also presents all the experiments conducted to evaluate the proposed system and results of comparative analysis obtained from the research.

Chapter Five: This chapter states the summary of the study and presents the conclusion, recommendation and future work.

1.8 Summary

This chapter discussed the introductory part of this dissertation, which includes the background of the study, research motivations, problem statement, research aim and objectives, and finally the outline of the dissertation chapters. The next chapter looks into review of related works.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

This section reviews several important concepts that are related to the proposed research work and also reviews related works and technologies in our proposed area of study.

2.1 Privacy Preserving Data Mining

Our society has substantially enhanced the potential to spawn and gather data from diverse sources. The enormous amount of data is needed in our everyday aspect of live. There is an urgent need for tools and techniques for transforming this vast amount of data into useful information and knowledge. This has led to the generation of a promising and flourishing end called Data Mining. Data Mining is also referred to as Knowledge Discovery from Data (KDD) (Agrawal and Srikant, 2000).

A number of companies have inadvertently released insufficiently anonymized datasets in past years. Netflix, an online movie rental company, uses movie rental and ranking records to suggest movie rentals to users, and sponsored a data-mining competition where teams competed to beat Netflix's recommendation engine. Netflix has published a dataset consisting of more than 100 million movie ratings from over 480 thousand of its customers, and invited the research community to contend for improvements to its recommendation algorithm. To protect customer privacy, Netflix removed all personal information identifying individual customers and perturbed some of the movie ratings. Despite these precautions, researchers have shown that with relatively little auxiliary information anonymous customers can be re-identified (Arvind andVitaly, 2008).

The second phase of this competition was canceled out of privacy concerns when it was discovered that many individuals could be identified by a combination of movie recommendations (Neil, 2010). America Online, and stylized as (AOL) is a web portal and online service provider based in New York City, has previously released a data set of 20 million web search queries, without realizing that most of the queries could be traced back to a single individual via personalized queries (Masnick, 2006). As a result, many companies, governmental agencies and privacy researchers have been forced to look for ways to ensure the anonymity of their data, while ensuring that the important information is kept saved, but at the same time it lendsitself to statistical inference, data mining, and online personalization applications.

Rapleaf is a startup which specializes in web personalization (MacDonald, 2011). If a website is able to get clues about the interests or demographics of a user, it can tailor the experience to that person. To allow a website to learn about that individual, non-personally identifying microdata can be stored in a cookie on the user’s browser. The cookie contains basic non-sensitive information like age, gender, and interests. The fact that this targeting information is being served to websites and advertisers leads to concerns about user privacy; for privacy purposes, the microdata being served about an individual should not allow the receiver of the data to link a browser to an individual. There is no harm in associating a user with the fact ”Loves Football”; instead the fear is that if enough of these simple pieces of data are stored, the mixture will uniquely identify an individual. Unfortunately, the most specific information, and therefore, the most de-anonymizing, is often the most valuable and useful data (it’s somewhat useful to know if a person is male, but very interesting to know that he Rides a Power Bike.)

However according to Sweeney such de-identification procedure would not guarantee privacy since individual can be uniquely identified by mixture of some attribute. Hence individual can be identified by linking attributes attack, since attribute such as name, date of birth, gender can potentially identify individual in a population. Study shows that about 87% of the population of United States can be uniquely identified using the seemingly innocuous attributes such as gender, date of birth, social security number and zip code. Therefore, the existing methods provide solution for privacy to some extent however cannot prevent the identity disclosure attack (Mahesh and Meyyappan, 2013, Vijayarani, 2010). Many researchers propose various algorithm and technique to avoid attack in micro-data. Still privacy preserving method that enables protection of privacy of individual and at the same time keeps information accurate and available is the trend of development of information security. Achieving reasonable approach that guarantees data extraction and privacy protection is the ultimate goal of this project.

In data mining, the raw material is transactional data and data mining algorithm serves as filter which filters out valuable nuggets of information from huge amount of data(Malik, Ghazi and Ali, 2012). Data is collected from single or various organizations and stored at respective databases. For analytical purposes, the data is transformed into suitable format and then the modified data is stored into the data warehouse and various data mining procedures/algorithms are applied for the generation of useful information. Privacy cannot be applied at one step, but needs to be applied at all levels. At level 1, raw data is gathered from different sources and is changed into suitable appearance for systematic purposes and stored into data warehouse. Privacy techniques are applied at this stage also while collecting data(Lindell and Pinkas, 2000). At level two, in data warehouse, used for reporting and data analysis. They store current and historical data and are used for creating reports. Data from data warehouse is subjected to get through a number of processes. These processes are blocking, suppression, perturbation, modification, generalization, sampling etc. For the discovery of knowledge/information, data mining algorithms are applied to processed data. Even the data mining algorithms are modified for the purpose of protecting privacy without sacrificing the goals of data mining.

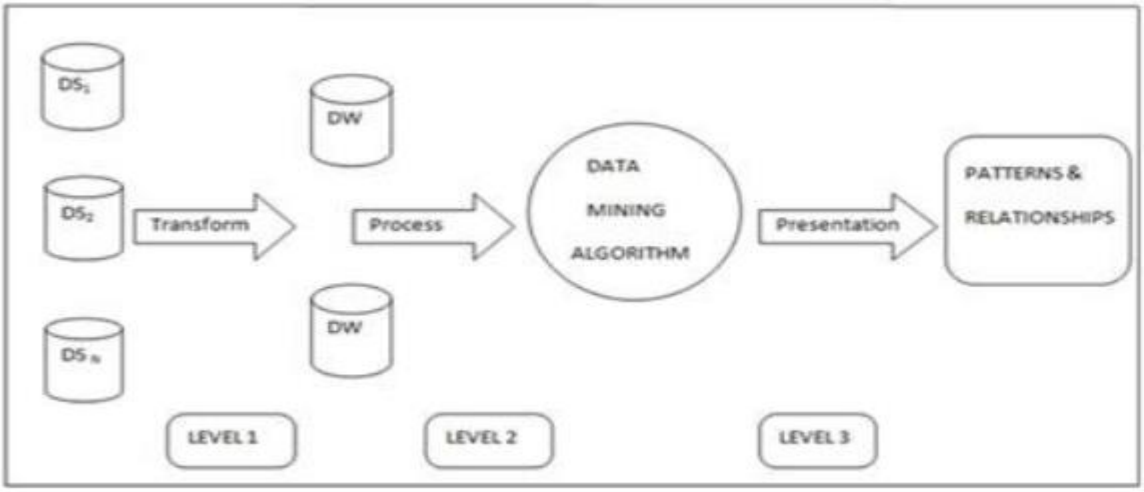


Figure 2.1: A Framework for PPD (Malik, Ghazi and Ali, 2012)

At level three, the knowledge revealed by data mining algorithms are checked for its sensitiveness towards disclosure risks. Privacy algorithms are applied at all three levels(Lindell and Pinkas, 2000).

PPDM has become an important issue in data mining research (Agrawal and Srikant, 2000). A set of new approaches are provided for mining of data and at the same time without allowing the privacy of data to be violated (Lindell and Pinkas, 2000). approaches can be classified into two main broad categories (2) Gkoulalas and Verikios, (1) (2010).

1. Procedures that save fragile information itself in the mining process.

2. Procedures that save fragile information mining results.

The first category refers to techniques that apply perturbation, sampling, modification, generalization etc to original datasets in order to generate their correlative that can be revealed to un-trusted parties. The second category refers to techniques that proscribe the disclosure of delicate data which is derived through the use of data mining algorithms. PPDM techniques can be classified into.

(1) Data distribution.

(2) Data modification.

(3) Data mining algorithms.

(4) Data or rule hiding.

(5) Privacy preservation.

The first dimension refers to distribution of data which can be either Centralized or Distributed. Distributed data can be further classified into Horizontal or Vertical distribution. Horizontal distribution refers to cases where different records reside in different places whereas vertical distribution refers to cases where all values of different attributes reside in different places. The second dimension refers to the data modification. In data modification, originalvalues are modified in the database and the altered values are released to the public.Methods of modification include:

1. Perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1- value to a 0-value, or adding noise).

2. Blocking, which is the replacement of an existing attribute value with a “?”.

3. Aggregation or merging which is the combination of several values into a coarser category,

4. Swapping that refers to interchanging values of individual records.

5. Sampling, which refers to releasing data for only a sample of a population.

The third dimension refers to the data mining algorithms which are applied on transformed data to get patterns and relationships which were not readily known to exist. The fourth dimension refers to whether the raw or aggregated data should be hidden. The final dimension refers to the techniques that are used for protecting privacy. Review of the Privacy Preserving Data Mining Techniques Based on different dimensions the PPDM techniques are classified into five categories:

(1) Randomized Response based PPDM

(2) Perturbation based PPDM

(3) Anonymization based PPDM

(4) Condensation approach based PPDM

(5) Cryptography based PPDM

We elaborate these in more detail in following subsections.

2.1.1. Randomized Response Based PPDM

In Randomized Response, the data is scatter such that original source cannot tell with the probabilities better than a pre-defined threshold, whether the data from user contains truthful information or false information. The information received from individual user is scrambled and if the numbers of users are significantly large, the result information of these users can be evaluated with good amount of accuracy. Randomized Response based PPDM is used in decision tree classification. Data collection inrandomized process is a two-step process (Nayak and Devi,2011).During first step; the data providers randomize their data and transmit the randomized data to the data receiver. In second step, the data receiver reconstructs the original distribution of the data by employing a distribution reconstruction algorithm. Randomization is very easy process and does not require the knowledge of location of records in data. It does not require the server to keep the original records in order to perform anonymization process (Aggarwal and Yu, 2008).

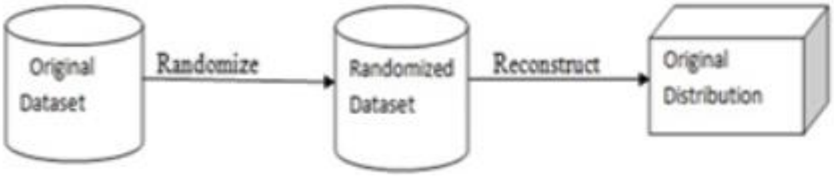


Figure 2.2.: Randomization response model (Malik, Ghazi and Ali, 2012)

2.1.2. Perturbation Based PPDM

Kargupta, Datta, Wang and Sivakumar, (2003) reported that in perturbation, the original values are replaced with some duplicate data values so that the statistical information computed from the perturbed data does not differ from the statistical information computed from the original data to a larger extent. The perturbed data records do not correspond to real-world record owners, so the attacker cannot violate the privacy of derived data or recover sensitive information from the modified data. In perturbation approach, records released is synthetic i.e. it does not correspond to real world entities represented by the original data. Therefore, the individual records in the perturbed data are meaningless to the human recipient as only statistical properties of the records are preserved. Perturbation can be done by using additive noise or data swapping or synthetic data generation (Jahan, Narsimha and Rao, 2012). Since the perturbation method does not reconstruct the original values but only the distributions, new algorithms are to be developed for mining of data. In perturbation approach, data mining algorithms treats each dimension independently.

2.1.3. Anonymization Based PPDM

Anonymization refers to hiding the sensitive information or identity of record owners. Explicit identifiers, are removed in this approach, set of attributes containing information that are reveling identifies of a record owner explicitly such as name, social security number etc. Explicit identifiers should be removed but still there is a danger of privacy intrusion when quasi identifiers, set of attributes that could Potentially identify a record owner when combined with publicly available data, are linked to publicly available data. Such attacks are called as linking attacks (Sweeney, 2002). For example, attributes such as DOB, Sex,Race, and Zipcode are available in public records such as voterlist. Such records are available in medical records also, whenlinked, can be used to infer the identity of the correspondingindividual with high probability. A value is replaced withless semantic consistent value called generalization and insuppression values are blocked. When such data is combinedwith publically available data, mining reduces risk ofidentification. Although the anonymization method ensuresthat the transformed data is true but suffers heavyinformation loss.

2.1.4. Condensation Approach

An additional approach used is Condensation approach. This approach uses a methodology which condenses the data into multiple groups of pre-defined size. It builds constrained clusters in the data set and after that produces pseudo-data. The basic concept of the method is to contract or condense the data into multiple groups of are maintained (Aggarwal and Yu, 2004). This approach is used in dynamic data update such as stream problems. Each group has a size of at least k“, which is referred to as the level of that privacy- preserving approach. The higher the level, the high is the amount of privacy. They use the statistics from each group in order to generate the corresponding pseudo-data. This is a simple privacy preservation approach but it is not efficient because it leads to loss of the information.

2.1.5. Fuzzy Algorithms Based PPDM

PPDM based on Fuzzy algorithms allow achievinganonymization without significant loss of information. The algorithms merge similar records into clusters. Each cluster formed is distinct from other clusters and the records of each cluster are not distinguishable from those of other clusters. A technique k-means clustering for anonymizing using Fuzzy logic is proposed in (Honda, Kawano, Notsu and Ichihashi, 2012). The record in cluster k is anonymized to make it indistinguishable from remaining k-1 clusters. A suggestion was made by Sathiya and Sadasivam, (2011) to modify apriori algorithm based on Fuzzy data in order to identify and then privatize sensitive rules in distributed scenarios. The method proposed by them for association rule hiding is efficient in terms of information hiding with fewer side effects. Cano and Torra, (2009) have used a fuzzy-based c-regression method to generate microdata (synthetic data). Trusted third party commodity servers are then entrusted with task of statistical computation with minimum risk of information loss.

2.1.6. Cryptography Based PPDM

In cryptographic technique private data can be encrypted safely. This technique is used where two or more than twoparties are involved in sharing their sensitive data but at the same time they are concerned for preserving privacy (Wang, 2010). Cryptographic techniques are ideally meant forsuch scenarios where multiple parties collaborate to computeresults or share non sensitive mining results and thereby avoiding disclosure of sensitive information (Aggarwal and Yu, 2008).Cryptographic techniques are used in such scenarios because it provides models for privacy and for implementing privacy preserving data mining algorithms. If the parties distributed across multiple sites are legally prohibited from sharing their datasets, a mining model to be built must be able to maintain the privacy of contributing parties. Previous categories of PPDM allow disclose of data beyond the control of the data collection. Lindell and Pinkas, (2010) have addressed the problem of reconstructing missing values by building a data model where the parties are distributed and data is horizontally partitioned. A cryptographic protocol based on decision-tree classification is described by them. A survey on cryptographic techniques for PPDM is studied by Shahn and Gulati, (2012). Distributed environment where the sharing is constrained either under legal or privacy policy issues use the cryptographic techniques. Oblivious transfer is used as building block for constructing an efficient PPDM model by Ding and Klein, (2010). The problem of distributed ID3 was addressed by Lindell and Pinkas, (2010). The implementations of these protocols consist of computationally intensive operations and generally consist of hard wired circuits.

Secure Multiparty Computation (SMC) is a technique in which computations are done beforehand on the basis of certain rules in statistical disclosure limitation. Basically there are three broad types of techniques under SMC: homomorphic encryption, circuit evaluation and secret sharing scheme. Bothsemi-honest and malicious adversaries are addressed by SMC protocols. A semi-honest adversary abides the protocol specification righteously but may try to learn facts by supplying incorrect information to the protocol. Most of the applications under SMC are built which address the semi honest adversaries. A SMC model was proposed for malicious adversaries (Jiang, Clifton and Kantarcıoğlu, 2008). The authors have proposed a framework that assigns liability for privacy to the responsible parties. Teo, Lee and Han, (2012) have made an analysis to support the accuracy and efficiency of SMC based protocols. Shashanka, (2010) provides a privacy preserving framework based on SMC using Gaussian mixture models. Also (Yi and Zhang, 2007) have devised a protocol based on encryption which will protect the privacy at each contributor end. Zhan, Matwin and Chang, (2007) have devised a method for privacy preservation based on homomorphic encryption for association rule mining.

Another form of cryptographic application is Pseudonomization. Here, the links between the personal and his medical information arebroke by anonymizing. Directly the information pertaining to personal identification is not removed from the dataset, but a pseudonym is generated and replaced. This information cannot be retrieved without compromising a secret shared previously. Ding and Klein, (2010) proposes encryption based technique for building pseudonyms. The pseudonyms are generated at the disseminated site by the contributor parties.

2.2 PPDM and Privacy Metrics

Since privacy has no single standard definition, measuring privacy level is quite challenging (Mendes and Vilela, 2017). In the context of PPDMs, some metrics have been proposed. Unfortunately, no single metric is enough, since multiple parameters may be evaluated (Mendes and Vilela, 2017). The existing metrics may be classified into three main categories, differing on what aspect of the PPDM is being measured: privacy level metrics measure how secure is the data from a disclosure point of view, data quality metrics quantify the loss of information/utility and complexity metrics, which measure efficiency and scalability of the different techniques. The level of Privacy and data quality metrics can be further categorized into two subsets (Bertino, Lin, and Jiang, 2008): data metrics and result metrics. Data metrics evaluate the privacy level/data quality by appraising the transformed data that resulted from applying a privacy-preserving method (e.g. randomisation or a privacy model). Result metrics make a similar evaluation, but the assessment is done to the results of the data mining (e.g. classifiers) that were developed with the transformed data.The following subsections present a survey on PPDM metrics concerning privacy level, data quality and complexity.

2.2.1. Privacy Level

As previously mentioned, the primal objective of PPDM methods is to preserve a certain level of privacy, while maximizing the utility of the data. The level of privacy metrics gives a sense of how secure is the data from possible privacy breaches. Recall from the aforementioned discussion that privacy level metrics can be categorised into data privacy metrics and result privacy metrics. In this context, data privacy metrics measure how the original sensitive information may be inferred from the transformed data that resulted from applying a privacy-preserving method, while result privacy metrics measure how the results of the data mining can disclose information about the original data. One of the first proposed metrics to measure data privacy is the **confidence level** (Agrawal and Srikant, 2000). This metric is used in additive-noise-based randomisation techniques, and measures how well the original values may be estimated from the randomised data. If an original value may be estimated to lie in an interval $[x1, x2]$ with c% confidence, then the interval $(x2-x1)$ is the amount of privacy at c% confidence. The problem with this metric is that it does not take into account the distribution of the original data, therefore making it possible to localise the original distribution in a smaller interval than $[x1, x2]$, with the same c% confidence. To address the issue of not taking into account the distribution of the original data, the average conditional entropy metric (Agrawal and Aggarwal, 2001) is proposed based on the concept of information entropy. Given two random variables X and Z, the average conditional privacy of X, given Z is $H(X|Z) = 2_{h(X|Z)}$, where $h(X|Z)$ is the conditional differential entropy of X, defined as:

$$h(X|Z) = -\int_{\Omega_{X,Z}} f_{X,Z}(x,z) \log_2 f_{X|Z=z}(x)\, dxdz$$

Figure 2.1 Equation to find the Confidence Level

where $f_X(\cdot)$ and $f_Z(\cdot)$ are the density functions of X and Z , respectively.

In multiplicative noise randomisation, privacy may be measured using the variance between the original and the perturbed data (Oliveira and Zaıane, 2010). Let x be a single original attribute value, and z the respective distorted value, $Var(x-z)Var(x)$ expresses how closely one can estimate the original values, using the perturbed data. Result privacy metrics, as opposed to data privacy metrics, are metrics that measure if sensitive data values may be inferred from the produced data mining outputs (a classifier, for example). These metrics are more application specific than the previously described. In fact, Fung et al. (Fung, Wang, Chen and Yu, 2010) defined these metrics as ‘‘special purpose metrics’’. One important result privacy metric is the hidden failure (HF), used to measure the balance between privacy and knowledge discovery (Bertino, Lin, and Jiang, 2008). The hidden failure may be defined as the ratio between the sensitive patterns that were hidden with the privacy-preserving method, and the sensitive patterns found in the original data (Oliveira and Zaiane, 2002). More formally:

$$HF = \frac{\#R_P(D')}{\#R_P(D)}$$

Figure 2.2 Equation to calculate the Hidden Failure

where HF is the hidden failure, D0 and D are the sanitized dataset and the original dataset, respectively, and #RP(·) is the number of sensitive patterns. If HF = 0, all sensitive patterns are successfully hidden, however, it is possible that more non sensitive information will be lost in the way. This metric maybe used in any pattern recognition data mining technique (e.g. classifier or an association rule algorithm). Note that this metric does not measure the amount of information lost.

2.2.2. Data Quality

Privacy-preserving techniques often degrade the quality of the data. Data quality metrics (also called functionality loss metrics)(Dua and Du, 2011) attempt to quantify this loss of utility. Generally, the measurements are made by comparing the results of a function over the original data, and over the privacy preserved transformed data. When evaluating data quality, three important parameters are often measured (Bertino and Fovino, 2005): the accuracy, which measures how close is the transformed data from the original data, the completeness, which evaluates the loss of individual data in the sanitized dataset, and consistency, which quantifies the loss of correlation in the sanitized data.

Furthermore, and similarly to the privacy level metrics, data quality measurements may be made from a data quality point of view, or from the quality of the results of a data mining application. Several metrics have been defined for both points of view, and for each of the parameters described above. Fletcher and Zahidul, (2015) surveyed a series of metrics used to measure information loss from the data quality perspective, for generalisation and suppression operations, and for equivalence classes algorithms (such as the k-anonymity). For the generalisation and suppression techniques, the authors described the Minimal Distortion (MD) (first proposed as generalisation height (Samarati, 2001)), the Loss Metric (LM) (Iyengar, 2002) and the Information Loss (ILoss) metric (Xiao and Tao, 2006). The MD metric is a simple counter that increments every time a value is generalized to the parent value. The higher the MD value, the more generalized is the data, and consequently, more information was lost. The LM and ILoss metrics measures the average information loss over all records, by taking into account the total number of original leaf nodes in the taxonomy tree. The ILoss differs from the LM metric by applying different weights to different attributes, for the average. The weight may be used to differentiate higher discriminating generalizations (Machanavajjhala, Kifer, Gehrke and Venkitasubramaniam, 2007). For the equivalence class algorithms, the Discernibility Metric (DM) was described (Bayardo and Agrawal, 2005). This metric measures how many records are identical to a given record, due to the generalizations.

The higher the value, the more information that is lost. For example, in the k-anonymity, at least k − 1 other records are identical to any given record, thus the discernibility value would be at least k − 1 for any record. Increasing k, will increase generalization and suppression, and consequently the discernibility value. For this reason, this metric is considered to be the opposite concept of the k-anonymity. A metric to measure the accuracy of any reconstruction algorithm (such as in randomisation) is defined (Agrawal and Aggarwal, 2001). The authors measure the information loss by comparing the reconstructed distribution and the original distribution. Let fX (x)be the original density function and fˆ X (x) the reconstructed density function. Then, the information loss is defined as:

$$I(f_X(x), \hat{f_X}(x)) = \frac{1}{2}E\left[\int_{\Omega_X} \left|f_X(x) - \hat{f_X}(x)\right| \, dx\right]$$

Figure 2.3: Equation to measure Information Loss

Where the expected value corresponds to the L1 distance between the original distribution fX (x) and the reconstructed estimation fˆ X (x). The metrics for evaluating the quality of the results are specific to the data mining technique that is used. These metrics are often based on the comparison between the results of the data mining with the perturbed data and with the original data. Two interesting metrics to measure data quality loss from the results of pattern recognition algorithms are the Misses Cost (MC) and the Artifactual Patterns (AP), presented in (Oliveira and Zaiane, 2002). The MC measures the number of patterns that were incorrectly hidden. That is non-sensitive patterns that were lost in the process of privacy preservation (recall the aforementioned discussion on association rule hiding). This metric is defined as follows. Let D be the original database and D' the sanitized database. The misses cost is given by:

$$MC = \frac{\# \sim R_P(D) - \# \sim R_P(D')}{\# \sim R_P(D)}$$

Figure 2.4: Equation to calculate the Misses Cost

where # ∼ RP(X) denotes the number of non-restrictive patterns discovered from database X. Ideally, an MC = 0%is desired, which means that all non-sensitive patterns are present in the transformed database. The AP metric measure satisfactory patterns, i.e. the number of patterns that did not exist in D, but were created in the process that led to D'. The following equation defines the AP metric.

$$AP = \frac{|P'| - |P \cap P'|}{|P'|}$$

Figure 2.5: Equation measures the Artifactual Patterns

where P and P' are the set of all patterns in D and D', respectively, and |·| represents the cardinality. In the best case scenario, AP should be equal to 0, indicating that no artificial pattern was introduced in the sanitisation process. For clustering techniques, the Misclassification Error (ME) metric proposed in (Oliveira and Za, 2010) measures the percentage of data points that ‘‘are not well classified in the distorted database’’. That is, the number of points that were not grouped within the same cluster with the original data and with the sanitised data. The misclassification is defined by the following equation:

$$M_E = \frac{1}{N} \times \sum_{i=1}^{k} \left(|Cluster_i(D)| - |Cluster_i(D')|\right)$$

Figure 2.6: The Equation is for misclassification

with N the number of total points in the database, k the number of clusters, and |Clusteri(X)| the number of legitimate data points of the ith cluster in database X. Additional metrics to evaluate the quality of results for classification and clustering are described in (Fletcher and Islam, 2015). These metrics include commonly used quantitative approaches to measure the quality of data mining results, such as the Randindex(Rand, 1971) and the F-measure (Rijsbergen, 1979). Finally note that cryptographic techniques implemented in distributed privacy preserve data quality since no sanitisation is applied to the data.

2.2.3. Complexity

The complexity of PPDM techniques mostly concern the efficiency and the scalability of the implemented algorithm (Bertino, Lin and Jiang, 2008). To measure the efficiency, one can use metrics for the usage of certain resources, such as time and space. Time may be measured by the CPU time or by the computational cost. Space metrics quantify the amount of memory required to execute the algorithm. In distributed computation, it may also be interesting to measure the communication cost, based either on the time, or the number of exchanged messages, and the bandwidth consumption. Both time and space are usually measured as a function of the input. Scalability refers to how well will a technique perform under increasing data. This is an extremely important aspect of any data mining technique since databases are ever increasing. In distributed computation, increasing the inputs may severely increase the amount of communications. Therefore, PPDM algorithms must be designed in a scalable way. Scalability may be evaluated empirically by subjecting the system to different loads (Bondi, 2000). For example, to test if a PPDM algorithm is scalable, one can make several experiments with increasing input data, and measure the loss of efficiency. The loss of efficiency over experiments can then be used to measure scalability, since a more scalable system will present lower efficiency losses when under the same ‘‘pressure’’ as a less scalable system.

2.3 PPDM Applications

A description of different privacy-preserving techniques, as well as a set of metrics to measure the privacy level, data quality and complexity was given. This section describes some existing PPDM applications, focusing on the employed privacy-preserving techniques and on the metrics used to measure the preservation of privacy. The following sections group the PPDM applications in the following fields: cloud computing, e-health, wireless sensor networks (WSN), and location-based services (LBS). Furthermore, in the Electronic-health applications, an emphasis on genome sequencing was given due to the rising privacy research interest in the area, and in the LBS applications, such as vehicular communications and mobile device location privacy was described. Note that this section does not extensively surveys existing PPDM applications. Nonetheless, it is sufficient to illustrate some of the described privacy-preserving methods described in this work, and relate the applicability with the assumptions and privacy requirements of the applications. For comprehensive reviews on privacy in genome sequencing, WSN and location privacy (Naveed et al, 2015, Zhang, Das, and Thuraisingham, 2009, Krumm, 2009), respectively.

2.3.1 Cloud PPDM

The U.S. National Institute of Standards and Technology (NIST) defined cloud computing (Mell and Grance, 2009) as ‘‘a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.’’ In other words, cloud is a distributed infrastructure with great storage and computation capabilities that is accessible through the network, anytime and anywhere. Therefore, applications (or services) that collect, store and analyze large data quantities often require the cloud. However, entities need to either trust cloud providers with data, or to apply techniques that protect data while stored and/or during distributed computation. The cloud may be also used to publish data, and in this case, query auditing and inference control may be required. Consequently, cloud-based services are one of the primary focus of privacy-preserving techniques (Mehmood, Natgunanathan, Xiang, Hua, and Guo, 2016). A scheme for classification over horizontally partitioned data under a semi-honest behaviour was proposed (Rong, Wang, Liu, and Xian, 2016).

This scheme allows owners to store encrypted data in the cloud, thus preserving privacy of data in communications and while stored. Furthermore, queries to the cloud are allowed to obtain the classes of a given set of inputs over encrypted data without the need for intermediate decryption. Using homomorphic encryption, the data, the query and the result are encrypted, and only the ‘‘querist’’ can decrypt the result, thus protecting the user from information leakage even against the cloud provider. The authors formally prove the security of the scheme and evaluate the computational and communicational complexity through simulations. Another approach that uses homomorphic encryption for cloud computing was presented in (Yi, Rao, Bertino, and Bouguettaya, 2015), for storing and mining association rules in a vertically distributed environment with multiple servers. This approach achieves privacy if at least one out of n servers is honest, and similarly to (Zhou, Cao, Dong and Lin, 2015), security is proven mathematically, based on cryptography. Additionally, the authors of the paper (Yi, Rao, Bertino, and Bouguettaya, 2015), also presented a series of efficient secure building blocks, which were required for their solution. Privacy in the cloud is not limited to the use of secure protocols. For instance, a technique to publish data to the cloud based on the concept of k-anonymity was presented (He, Wang, Li, and Hao, 2016). The authors described a novel approach where equivalence classes have less than k records, but still ensure the kanonymity principle by exploring overlaps in the definitions of the equivalence classes. That is, by creating classes with less than k records (a divisor of k records) such that each record could belong to multiple classes and, thus, provide k-anonymity. By having a lower number of records in each class, the number of required generalisations is lower, and thus, more utility is preserved. The authors show this result by measuring the information loss, and also show the good performance of the implementation.

2.3.2. E-Health PPDM

Health records are considered to be extremely private, as much of this data is considered sensitive. However, the increase in the amount of data, combined with the favourable properties of the cloud has led health services to store and exchange medical records through this infrastructure (Abbas and Khan, 2014). Thus, to protect from unwanted disclosures privacy-preserving approaches are considered. A survey on the state-of-the-art privacy-preserving approaches employed in the e-Health clouds was given (Abbas and Khan, 2014), where the authors divide PPDM techniques in either cryptographic and non-cryptographic. The cryptographic techniques are usually based on encryption, whereas non-cryptographic approaches are based on policies and/or some sort of restricted access. An example of a cryptographic technique was found in (Zhou, Cao, Dong and Lin, 2015), where the authors propose a privacy-preserving medical text mining and image feature extraction scheme based on fully homomorphic data aggregation under semi honest behaviour is presented. The authors formally prove that their encryption is secure from the data point of view and from the results point of view. They also evaluate the performance of the PPDM by measuring computation and communication costs over the amount of input data. An emerging field in e-health that is raising a growing privacy interest is genome sequencing. Genome sequencing is the process of studying genetic information about an individual through the study of sequences of DNA (Deoxyribonucleic acid). Genomic data sees applications in (Naveed et al, 2015) healthcare, forensics and even direct-to-consumer services. Due to the advances in genome sequencing technologies and the capabilities of the cloud for computation and communication of data, this area has experienced a recent boom in research, including in the privacy field. Genetic data is highly identifiable and can be extremely sensitive and personal, revealing health conditions and individual traits (Naveed, 2015). Furthermore, this type of data also reveals information about blood relatives, thus involving not only a single individual (Naveed, 2015). It is, therefore, critical to prevent unwanted disclosure of this type of data, while preserving maximum utility. For genome data publishing, (Uhlerop, 2013) proposed a solution for releasing aggregate data based on the SI-differential privacy. This approach was motivated where an attack to accurately identify the presence of an individual from a DNA mixture of numerous individuals was introduced. Thus, additive noise is added to the statistical data to be released (Uhlerop, 2013), in order to achieve SI-differential privacy. Simulations have shown that SI-differential privacy is achievable and good statistical utility was preserved. However, for big and sparse data, the release of simple summary statistics is problematic from both privacy and utility perspectives. Recently, (McLaren, 2016) proposed a framework for privacy-preserving genetic storing and testing. The depicted scenario involved the patients (P), a certified institution (CI), which has access to unprotected raw genetic data and therefore must be a trusted entity, a storage and processing unit (SPU) and medical units (MU). Both the SPU and the MU follow a semi-honest adversarial behaviour, i.e. they will follow the protocol, but may attempt to infer sensitive data about the patients. Essentially, the patient supplies the data to the CI, which stores such data encrypted in the SPU using a partially homomorphic encryption scheme. MUs can then use secure two-party protocols with the SPU to operate the data in encrypted form, to be decrypted only when the result is returned from the SPU to the MU. Their framework has proven to be efficient, although it was limited to some genetictests. Fully homomorphic encryption is suggested has a future solution to this limitation, however, the computational cost is currently prohibitive.

2.3.3 Wireless Sensor Networks PPDM

Wireless Sensor Networks (WSN), sometimes called Wireless Sensor and Actuator Networks (WSAN), are networks of sparsely distributed autonomous sensors (and actuators), that monitor (act upon changes in) the physical environment (Akyildiz and Kasimoglu, 2004) (e.g. light, temperature). Each sensor/actuator is referred to as node in the WSN and data is exchanged wirelessly between these devices. Since nodes have low battery capacity, one of the most important challenges in WSN networks is the efficiency in communication and processing of data at each node (Akyildiz and Kasimoglu, 2004). Thus, techniques to aggregate data from multiple sensors are often used to reduce network traffic and hence, improve battery life and consequently the sensor's lifetime. Data generated in WSNs may be considered sensitive in many different applications. For instance, sensed humidity of a room may determine room occupancy and house electrical usage over time may be used to track household behaviour (Taban and Gligor, 2009). Due to the aggregation of data and the WSNs' topology, attackers may try to control one or a few nodes to obtain access to all information. In this case, even if the communications are encrypted, the compromised nodes have the ability to decrypt the information, giving the adversary full access (He, Liu, Nguyen, Nahrstedt, and Abdelzaher, 2007). Therefore, privacy-preservation techniques may be required. An approach to leverage the advantage of data aggregation for efficiency and to preserve privacy on the collected data was proposed (Taban and Gligor, 2009). In this work, users can only query aggregator nodes to obtain aggregated data. Aggregator nodes query a set of nodes for the sensed values and proceed to compute the aggregation results over the received data, which is then forwarded to the inquirer.   However, users must be able to verify the integrity of the aggregated data, since malicious users may try to control aggregation nodes and send false data. The WSN owners, on the other hand, want to prevent disclosure of individual sensor data, thus restricting query results to aggregated data. The challenge here is how to verify the integrity of aggregated data, without access to the original data. To address this issue, a framework where the user has full access to encrypted sensor data, in addition to the access to the aggregated data is proposed. The user can verify the integrity of the aggregated data by making use of the encrypted sensed values, without decrypting such data. Four solutions were described, each providing a different privacy-functionality trade-off, where one of the solutions uses (partially) homomorphic encryption to achieve perfect privacy, that is, no individual sensed value is disclosed. The authors compare the four solutions in terms of the number of messages exchanged and the supported aggregation functions. Another approach that makes use of the aggregation of data to preserve privacy was proposed in (Groat, Hey, and Forrest, 2011). This approach is non-cryptographic and implements a similar concept to k-anonymity, referred to as k-indistinguishable, where instead of generalisations, synthetic data is added to camouflage the real values (obfuscation). Aside from using k to control the number of indistinguishable values, a discussion to decrease the probability of privacy breach under colluding nodes (combining information from multiple nodes) is given. The authors also compared the performance of their implementation against encryption approaches, where the results show that this method is more time and power efficient than such approaches. The above examples concern information leakage from within the WSN. However, large WSN may be queried by multiple entities (clients) that may not trust the network owners (Carbunar, Yu, Shi, Pearce, and Vasudevan, 2010). The network owners may infer clients' intentions through the respective queries and profiles. These queries may be specific to a given area, or a given event thus revealing the intention. As stated in (Carbunar, Yu, Shi, Pearce, and Vasudevan, 2010), one solution would be to query all sensors in the network and save only the readings of interest.

However, this would result in a significant load on the network, especially in large networks. To address this issue, (Carbunar, Yu, Shi, Pearce, and Vasudevan, 2010) proposed two approaches differing on the type of network models: querying server(s) that belong to a single owner (organization) and querying servers (at least two) belonging to different organizations. In both scenarios, servers are considered to be semi honest, i.e. they abide by the protocols, but attempt to learn more than allowed. In the single owner model, the idea is to create a trade-off between the area of sensors that is queried and the privacy that is achieved. If the client queries only the region of interest, then no privacy is achieved, but the cost is minimal, whereas if the query targets all the network, the cost is maximum, but the achieved privacy is also maximized. The solution is thus a function that transforms an original query sequence into a transformed sequence, in order to conceal the region(s) of interest. Two metrics were used to measure privacy: the spatial privacy level and the temporal privacy level. The spatial metric is the inverse of the probability of the server guessing the regions of interest from the transformed query. The temporal privacy level measures the distance between the distributions of frequency of access of the regions obtained with the transformed and original query. A higher distance value translates into a better obfuscation of the frequency of access to the regions of interest. In the multiple owners situation, cryptography is used to assign a virtual region to each sensor, that is only recognized by both the client and the sensor. A queried server then broadcasts the encrypted query, which is dropped by sensors that do not belong to the target virtual region. Sensors from the queried region encrypt the sensed values and return the results, which can only be decrypted by the client. This solution is fully private, as long as the servers used to create the virtual regions do not collude.

2.3.4 Location-Based Services PPDM

Pervasive technologies such as the Global Positioning System (GPS) allow to obtain highly accurate location information. LBSs use this spatiotemporal data to provide users with useful contextualized services (Jiang and Yao, 2006). However, this same information can be used to track users and consequently discover for example, their workplace, their houses' location and the places that they visit (Krumm, 2009). Furthermore, this information can also be used to identify users, since routes and behaviours often have characteristic patterns (Montjoye, Hidalgo, Verleysen, and Blondel, 2013). Therefore, the possibility of locating information leakage is a serious concern and a threat to one's privacy. This type of leakage occurs when attackers have access to the Location Based Services (LBS) data, or when LBS providers are not trustworthy. In computational location, privacy is achieved with anonymity, data obfuscation (perturbation), or through application-specific queries (Krumm, 2009). For location anonymity, users can be assigned IDs (pseudonyms) to prevent identity disclosure. However, these pseudonyms must be changed periodically so that users cannot be tracked over time and space, and consequently disclose identity. To prevent this type of disclosure, Beresford and Stajano, (2003) presented the concept of ‘‘mix zone’’. In this approach, IDs are changed every-time users enter a mix zone. In this type of zone, at least $k - 1$ other users are present, such that changing all pseudonyms prevents the linkage between the old and new pseudonyms. With this approach, and similarly to the k-anonymity privacy model, k may be used as a privacy metric.

In data obfuscation, the idea is to generate synthetic data or to add noise in order to degrade the quality of the spatial, and sometimes temporal, data. The assumption is that the LBS provider is untrustworthy. Simple examples include giving multiple locations and/or imprecise locations (Krumm, 2009). A solution to ‘‘cloak’’ users' locations using an intermediary anonymiser server (between the user and the LBS) was proposed (Meyerowitz and Choudhury, 2009). The user queries the intermediary server (named CacheCloak) and if this server has the correct data for the location in cache, the data is sent to the user without querying the LBS. If the location data is not cached, the CacheCloak server creates a prediction path from the queried point until reaching a point in another cached path, and then queries the LBS for all these points. The received data is then cached and the correct data is forwarded to the user. As the user moves through the predicted path, the CacheCloak will return the cached information. When the user changes from the predicted path, and if the new position is not yet cached, then the same process is repeated. Since the predicted path is queried at the same time to the LBS, the service provider has no way to know the exact user location nor the movement direction. The authors presented a metric based on the concept of (location) entropy to measure the achieved privacy leveland how their solution to location privacy can work in realtime LBS services.

Furthermore, an implementation to workunder the assumption of an untrusted CacheCloak server is also discussed.Another type of technique to achieve location privacy isto implement private queries, that is, location queries that do not disclose user location to the LBS provider. An approach using a secure protocol was presented by Ghinita, Kalnis, Khoshgozaran, Shahabi, and Tan, (2008), that allows users to query the LBS server through an encrypted query that does not reveal user's location. The protocol used is the private information retrieval (PIR), that has many similarities with the oblivious transfer protocol. With the encrypted query, the server computes the nearest neighbor, to retrieve the closest point of interest from the user location. The authors implemented data mining techniques to optimize the performance of their solution, to identify redundant partial products, and show through simulation that the final cost in server time and the cost of communications is reasonable for location-based applications.

This solution achieves full privacy in the sense that it is computationally infeasible for the server to decipher the encrypted query. Vehicular communication privacy may be seen as a particular case of location privacy. These location-based systems are essentially networks, where cars and roadside units are nodes that communicate wirelessly to exchange spatiotemporal information (Lim, Yu, Kim, Kim, and Lee, 2017). Location-based services (LBS) make use of this data to provide drivers with useful content, such as traffic conditions, targeted advertising, and others. In this scenario, the highly accurate spatiotemporal information provided by the GPS is transferred to a third party server, that accumulates routes information that can be used to track drivers (Lim, Yu, Kim, Kim, and Lee, 2017). Privacy preservation is thus required, to protect drivers from being tracked. A privacy preservation approach under the assumption of untrusted collector was presented (Lim, Yu, Kim, Kim, and Lee, 2017). This technique uses synthetic data generation to obfuscate the real trajectory of the car, by providing consistent fake locations.

The authors present three measures of privacy: the tracking process, which is measured by the attacker's belief (probability) that a given location-time sample corresponds to the real location of the car; the location entropy, to measure the location uncertainty of an attacker; and the tracking success ratio, which measures the chance that the attacker's belief is correct when targeting a driver over some time. In this section we provide an overview of a set of relevant applications of PPDM methods, yet several other applications for the aforementioned domains and others exist.

2.4.0 Review of Related Works

To give more specific about K-anonymity Privacy Preserving model, this part describes and examines previous work done in field of data Mining.

Friedman, (2011), presented a privacy preserving data mining using Anonymization and Decision Tree (ID3 and C45) perspective techniques which are currently used for data mining purposes. The paper focused on the problem of guaranteeing privacy of data mining output.

Tiancheng and Ninghui, (2008) suggested a novel approach which uses a bottom-up method to group and then anonymize quasi-identifiers. Another work by Poovammal and Ponnavaikko, (2009) suggested a task-based technique which satisfactorily balances both the privacy and utility trade-offs in data mining. Mining is done after the algorithm in Poovammal and Ponnavaikko, (2009) was applied which hides the sensitive data effectively. Anonymizing quasi-identifiers and sensitiveattributes in datasets pose an information loss which is notdesirable for mining. Zhu and Peng, (2007) focus on medical datasets and try to address the issues related to privacy requirements. Anonymization methods are also useful for addressing specific problems. Matatov, Rokach and Maimon, (2010) used k-anonymity based method for optimal feature set partitioning. (Zhu, Li and Wu, 2009) proposed data reconstruction approach which achieves k-anonymity protection in predictive data mining. The potentially identifiable attributes are first mapped using aggregation for numeric data and swapping is done for nominal data. A technique based on genetic algorithm is applied to the masked data for finding a better subset from it. The subset is replicated to generate published dataset which satisfies the k-anonymity constraint.

Condensation is a statistical approach which constructs constrained clusters in a dataset and then generates pseudo data from statistics of these clusters (Liu, Kantarcioglu and Thuraisingham, 2009). Clusters of non-homogeneous size are constructed from whole data, such that, each record lay in a group whose size is at least equal to its anonymity level. After this pseudo data is generated from each group, and synthetic dataset is created with similar aggregate distribution as that of the original dataset. Condensation is effectively used for solving the classification problem. An additional layer of protection is provided with pseudo data making it difficult for adversaries. Also, aggregate behaviour of data is preserved with condensation, making it useful for data mining tasks.

Srivastava, (2015), proposed a Privacy Preserving Data Mining (PPDM) approach for medical research called “k-anonymity with decision tree”. The main goal of this research was to provide tradeoffs between privacy and utility. The author create hierarchy of each attributes which define the privacy level, which means how much range we want to generalize in record table for each attribute and applied k-anonymity on raw input dataset, by using k-anonymity algorithm and get privacy preserve dataset which is useful for make identity private and also prevent for linking attack in medical research. The anonymous dataset is sent to a classification decision tree (C4.5) module for mining purpose. Anonymization is a PPDM approach that hides the identity and sensitive data of record owners, assuming that sensitive data must be retained for data analysis. The first step he did was to remove explicit identifiers, even with all explicit identifiers being removed, the dataset showed a real-life privacy threat in which an individual was identified uniquely using his name in a public voter list linked with his record in a published medical database through the combination of zipcode, date of birth, and sex. Each of these attributes does not uniquely identify a record owner, but their combination, called the quasi identifier identifies a unique or a small number of record owners.

2.4.1 Attacks On k-Anonymity

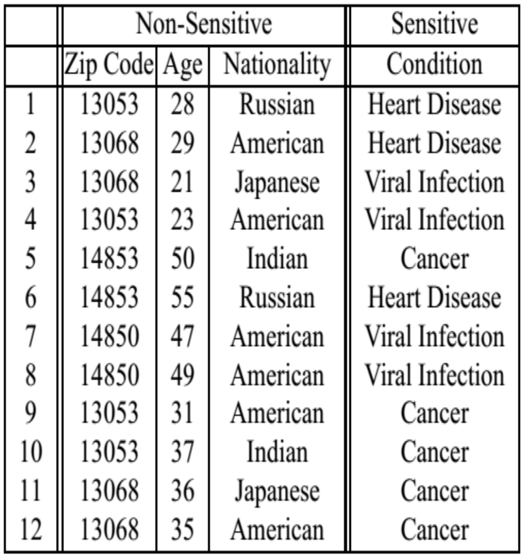
Here is an illustration of the two attacks that Srivastava, (2015) work is venerable to, the homogeneity attack and the background knowledge attack. This is how they can be used to compromise a k-anonymous dataset.

2.4.1.1 Homogeneity Attack:

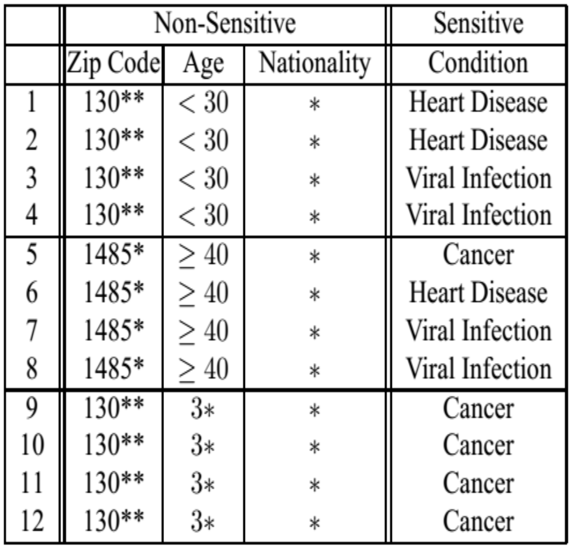
Take for example Alice and Bob are antagonisticneighbors. One day Bob falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Alice sets outto discover what disease

Bob is suffering from. Alice discovers the 4-anonymous table of current inpatient records published by the hospital (Table 2.1), and so she knows that one of the records in this table contains Bob’s data. Since Alice is Bob’s neighbor, she knows that Bob is a 31-year-oldAmerican male who lives in the zip code 13053. Therefore, Alice knows that Bob’s record number is 9,10,11, or 12.Now, all of those patients have the same medical condition(cancer), and so Alice concludes that Bob has cancer. (Machanavajjhala, Gehrke and Kifer, 2012)

Table 2.1.: Inpatient MicroData (**Machanavajjhala, Gehrke and Kifer, 2012**) Table 2.2.: 4 anonymous Inpatient Microdata



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer



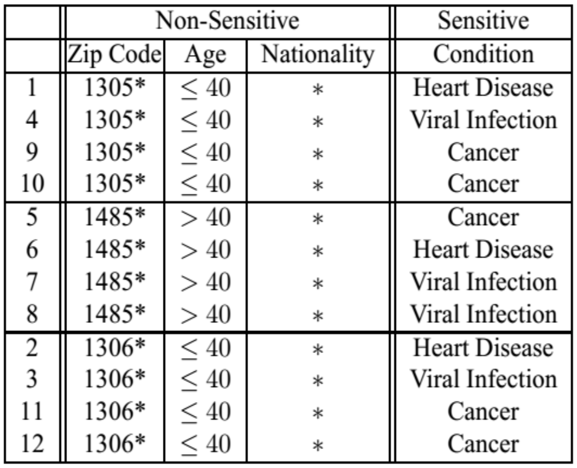
	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	∗	Heart Disease
2	130**	< 30	∗	Heart Disease
3	130**	< 30	∗	Viral Infection
4	130**	< 30	∗	Viral Infection
5	1485*	≥ 40	∗	Cancer
6	1485*	≥ 40	∗	Heart Disease
7	1485*	≥ 40	∗	Viral Infection
8	1485*	≥ 40	∗	Viral Infection
9	130**	3∗	∗	Cancer
10	130**	3∗	∗	Cancer
11	130**	3∗	∗	Cancer
12	130**	3∗	∗	Cancer

Note that such a situation is not uncommon. As a back of-the-envelope calculation, suppose we have a dataset containing 60,000 distinct tuples where the sensitive attribute can take 3 distinct values and is not correlated with the non-sensitive attributes. A 5-anonymization of this table will have around 12,000 groups and, on average, 1 out of every 81 groups will have no diversity (the values for the sensitive attribute will all be the same). Thus, about 148 groups with no diversity is to expected. Therefore, information about 740 people would be compromised by a homogeneity attack. This suggests that in addition to k-anonymity, the sanitized table should also ensure “diversity” – all tuples that share the same values of their quasi-identifiers should have diverse values for their sensitive attributes. Our next observation is that an adversary could use “background” knowledge to discover sensitive information.(Machanavajjhala, Gehrke and Kifer, 2012).

2.4.1.2 Background Knowledge Attack:

Alice has a pen friend named Umeko who is admitted to the same hospital as Bob, and whose patient records also appear in the table shown in (Table 2.3), Alice knows that Umeko is a 21 yearold Japanese female who currently lives in zip code 13068. Based on this information, Alice learns that Umeko’s information is contained in record number 1,2,3, or 4. Without additional information, Alice is not sure whether Umeko caught a virus or has heart disease. However, it is well known that Japanese have an extremely low incidence of heart disease. Therefore Alice concludes with near certainty that Umeko has a viral infection.(Machanavajjhala, Gehrke and Kifer, 2012). The privacy level of the system can however be further enhanced with introduction of L-diversity. Table 2.3 shows the encryption and decryption process of the work.

Table 2.3: 3-diverse Inpatient Microdata.(Machanavajjhala, Gehrke and Kifer, 2012)



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Table 2.3 is a 3-diverse version of the Table 2.2. Comparing it with the 4-anonymous table the attacks against the 4-anonymous table are prevented by the 3-diverse table. For example, Alice cannot infer from the 3-diverse table that Bob (a 31 year old American from zip code 13053) has cancer. Even though Umeko (a 21 year old Japanese from zip code 13068) is extremely unlikely to have heart disease, Alice is still unsure whether Umeko has a viral infection or cancer.

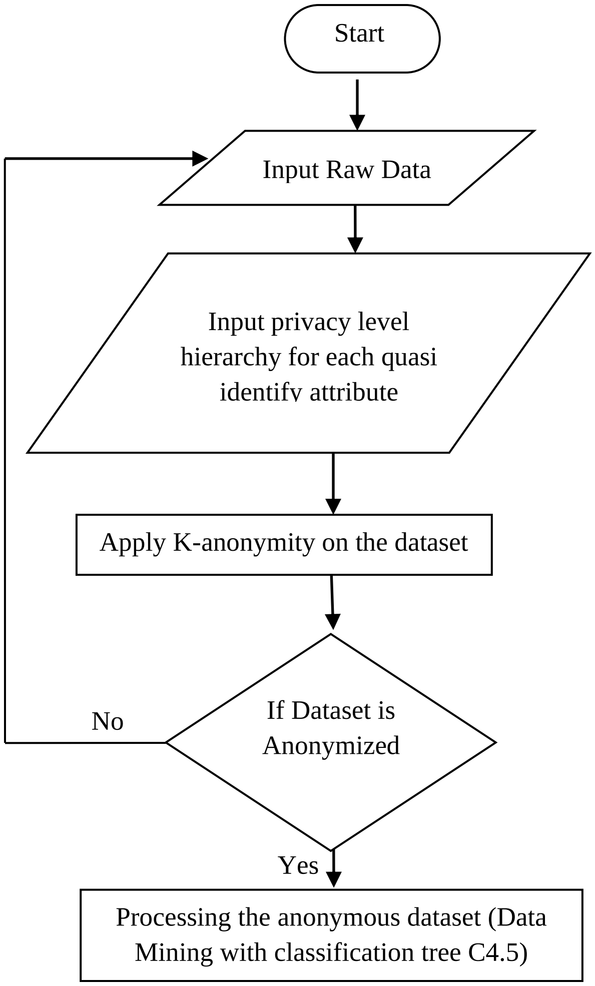

Start
Input Raw Data
Input privacy level hierarchy for each quasi identifv attribute
Apply K-anonymity on the dataset
If Dataset is Anonymized
No
Yes
Processing the anonymous dataset (Data Mining with classification tree C4.5)

Figure 2.3: Flowchart of Anonymization Module (Exiting System)

2.4.2 Gap in Literature

Though some of the literature reviewed modified the K-anonymity, it is not sufficient to say the modification is not correct but the modification could not handle some of the problems like homogeneity attack associated with the privacy and data utility.

Hence, this research work extends Srivastava (2015) by applying L-diversity to anonymous dataset sensitive attributes so as to prevent homogeneity and background knowledge attack which as a result of that improves the algorithm.

2.4.3 Summary

This chapter presented the overview of Privacy Preserving Data Mining, Privacy Metrics, PPDM Applications and also reviewed related works. The next chapter describes the methodology followed to accomplish the proposed research.

CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter describes the methodology of the proposed development of an Hybrid k-anonymity model for data mining privacy protection. The chapter presents the step-by-step procedures taken to develop the underlying approach that accepts datasets and anonymize the datasets that to present privacy of individuals being disclosed. This chapter presents the architecture of the proposed system, based on the existing system of Srivastava (2015), the k-anonymization enhancement is also presented.

3.1 Proposed System

The main weakness of K-anonymity is being venerable to homogeneity and background knowledge attacks; thus a an improvement to K-anonymity is need to strengthen the definition of privacy. To curb this problem, the proposed is system builds on the work of Srivastava, (2015) by improving on the K-anonymity module using L-diversity so as to prevent homogeneity and background knowledge attacks, thereby making the algorithm more secure and robust.

3.2 Description of the Proposed System

To L-diversify a dataset, the system will have to anonymize the dataset. For the purpose of privacy preserving of individual record, explicit identifier such as P-id, Name, Social Security Numbers(SSN) are removed from the micro data, but de-identifying of data generally not provide the guarantee of anonymization. The anonymization is done with any K-anonymity algorithm and the anonymous dataset is further diversify using L-diversity principle that divides and group the anonymised dataset based on the privacy attribute level specify.

3.3 Architecture of the proposed System

The pseudo code for Hybrid K-anonymity process of the proposed system as shown in figure 3.1is as follows:

3.3.1 Proposed System Algorithm

Stage 1

Input: Private table PT; quasi-identifier QI=(A1,....,An), k constraint; hierarchies DGHAi, where i=1,...n.

```
Step1   Consider a table MGT = PT[QI]
Step2   While k-anonymity is not achieved and the count of the remaining rows that donot
        comply to k-anonymity is more than k:
        •   Get the number of distinct values of each attribute in MGT
        •   Generalize the attribute with the most distinct values
Step3   Suppress the remaining rows and return MGT.
```

Stage 2

Diversify the sensitive attributes anonymous dataset using L-diversity, (c, ℓ)-diversity

```
Step1   Enter a constant number (integer type).
Step2   Enter the level of diversity ie. integer.
Step3   Get the diversified dataset in a file.
```

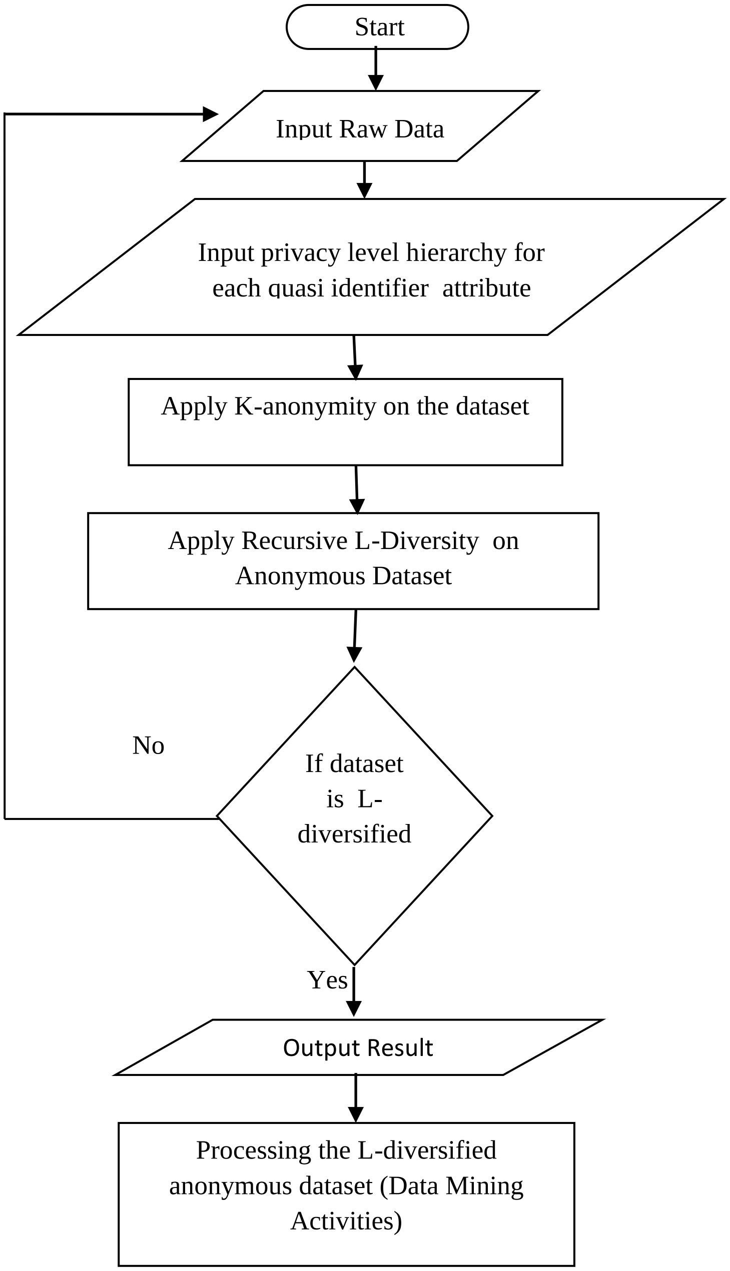

Start
Input Raw Data
Input privacy level hierarchy for each quasi identifier attribute
Apply K-anonymity on the dataset
Apply Recursive L-Diversity on Anonymous Dataset
No
If dataset is L-diversified
Yes
Output Result
Processing the L-diversified anonymous dataset (Data Mining Activities)

Figure 3.1:Flowchartof Proposed System

3.6 Summary

This chapter presented the architecture, flow diagrams, encryption and decryption algorithms for the proposed system. The next chapter discusses the implementation and result analysis of the proposed system.

CHAPTER FOUR

SYSTEM IMPLEMENTATION

4.0 Introduction

This chapter will provide a detailed explanation of how the Hybrid K-anonymity privacy preserving model is implemented. Development of anHybrid K-anonymity Model for Data Mining Privacy Protection comprises utilisation of the K-anonymity and L-diversity algorithmss described in Chapter 3. The dataset is a .CSV format raw medical data and is being K-anonymized. The anonymise dataset is the further anonymised by using L-diversity technique and the final dataset can then be mined thereby protecting individual privacy and also maintaining the usability of the dataset. This chapter discusses the implementation details of the system with screen shot showing the graphical user interface. In addition, the section also presents all the experiments conducted to evaluate the proposed system and results of comparative analysis obtained from the research.

4.1 System Requirements

The software requirement for the effective development and implementation of this system are as follows:

1. Java Programming
2. Netbeans IDE

The hardware requirements are:

1. A Pentium Dual Core CPU, 2.1 GHz processor with 6GB memory
2. A 500 GB hard drive capacity
3. A graphic adaptor with screen resolution of 800 * 600 pixels and 64 bit quality

4. A CD/DVD rewritable drive

5. Laser jet printer for draft copies

4.2 Implementation Details

This section discusses the requirement needed for the implementation of the system. The system "Development of An Hybrid K-Anonymity Model For Data Mining Privacy Protection" is considered to be a standalone application where it does not require using neither a database nor web server. It is expected to perform two main operations: Anonymizing dataset and writing the result into a file as CSV format. The system was implemented on Windows 7 Ultimate Operating System. It was implemented on a machine that has processor of 2.27 GHz Intel Core i3 with memory 2 GB 350 MHz DDR3. The system was written in java language and implemented in Netbeans environment.

4.3 Dataset

This sub-section describes the Medical dataset that were used for the experiment. The medical dataset were generated from www.generatedata.com website.

4.4 Dataset Load Description

The user first authenticate into the system by providing the login credentials, if successful then proceed to the Dataset Loading stage. During the dataset loading stage, the user clicks on the load data button, the system locate the dataset in the data folder directory which is sub-folder in the system project library. In few seconds depending on the data size, the data is loaded in the dataset area in tabular form, showing the original data including the attributes and their values.

4.5 Prepare Values (Quasi-Identifiers)

During the prepare values stage, the user selects list of Quasi-identifier (Age, Zipcode,) from datasets, then proceeds to specify the K-value which can only take integer values and finally anonymize the dataset.

4.6 Dataset Anonymization Process

In this section, based on user's requirements (users may only require one or more constraintsi.e the quasi-identifiers), the algorithm first applies k-anonymity by generalizing or suppression based on the data type andlater apply L-diversity to the anonymized datasets.Attributes with unique identifiers such as Name (if we consider it to be a unique identifier) will suppressed to a value ∗ in each of the tuples.

4.7 Measures of Loss of Information

The Entropy Measure, we suggest to use the standard measure of information, namely entropy, in order to assess more accurately the amount of information that is lost by anonymization. The public database D={R_1, . . . , R_n}induces a probability distribution for each of the public attributes. Let X_j, 1 ≤ j ≤ r, denote hereinafter the value of the attribute A_j in a randomly selected record from D. Then

$$\Pr(X_j = a) = \frac{\#\{1 \leq i \leq n : R_i(j) = a\}}{n} .$$

Figure 4.1 Information Loss Equation

4.8 Graphical User Interface of the Proposed System

The Development of a Hybrid K-Anonymity Model for Data Mining Privacy Protection has the following functions. Login Button, Load Dataset Button, Prepare Values Button,Quasi-identifier Text Area,Anonymize Buttons, Dataset Tex Area, Anonymized Dataset Output File and the Quit Button. Figure 4.1 shows the graphical user interface of the designed system.

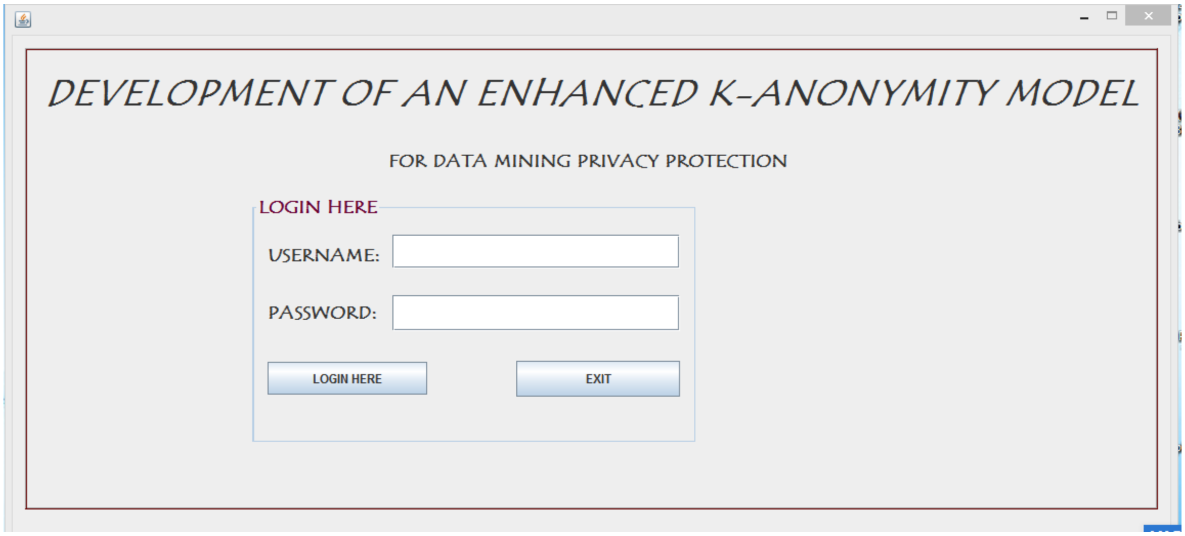
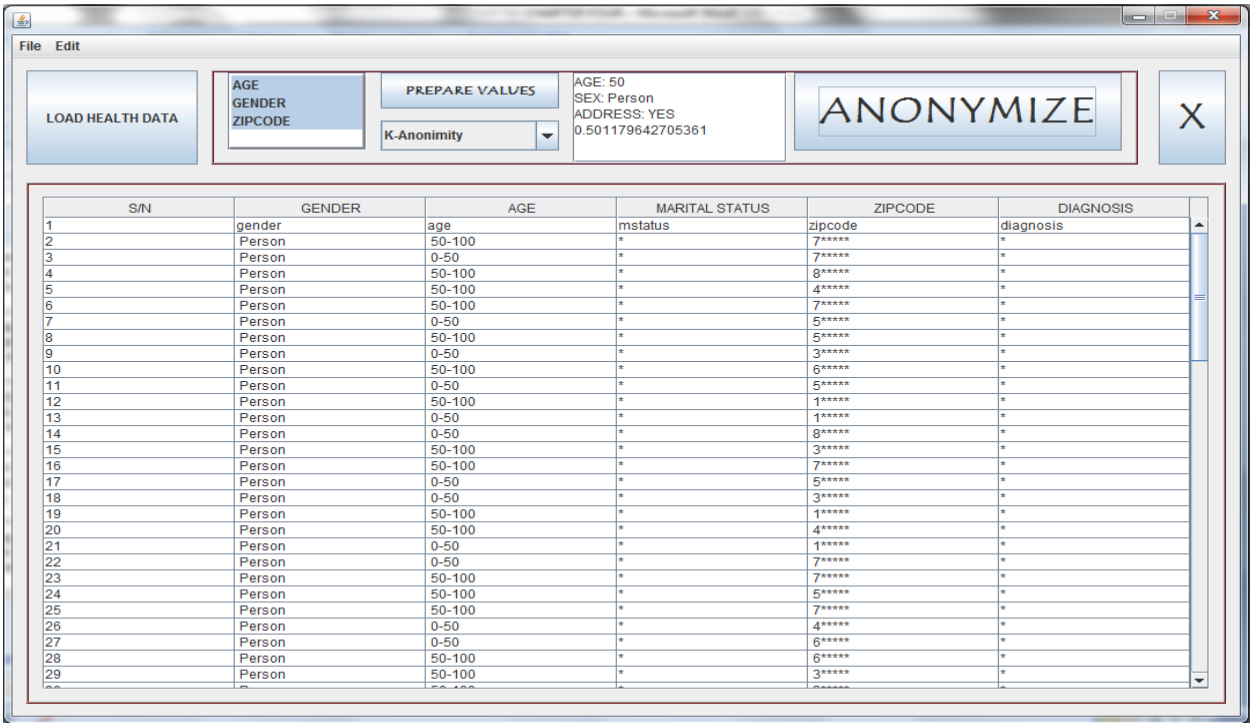

DEVELOPMENT OF AN ENHANCED K-ANONYMITY MODEL
FOR DATA MINING PRIVACY PROTECTION
LOGIN HERE
USERNAME:
PASSWORD:
LOGIN HERE
EXIT

Figure 4.2: Graphical User interface of Designed System

This is the first interface of the system that authenticates the user of the system. Then user enters the login details, if correct then have access to the home page of the system.

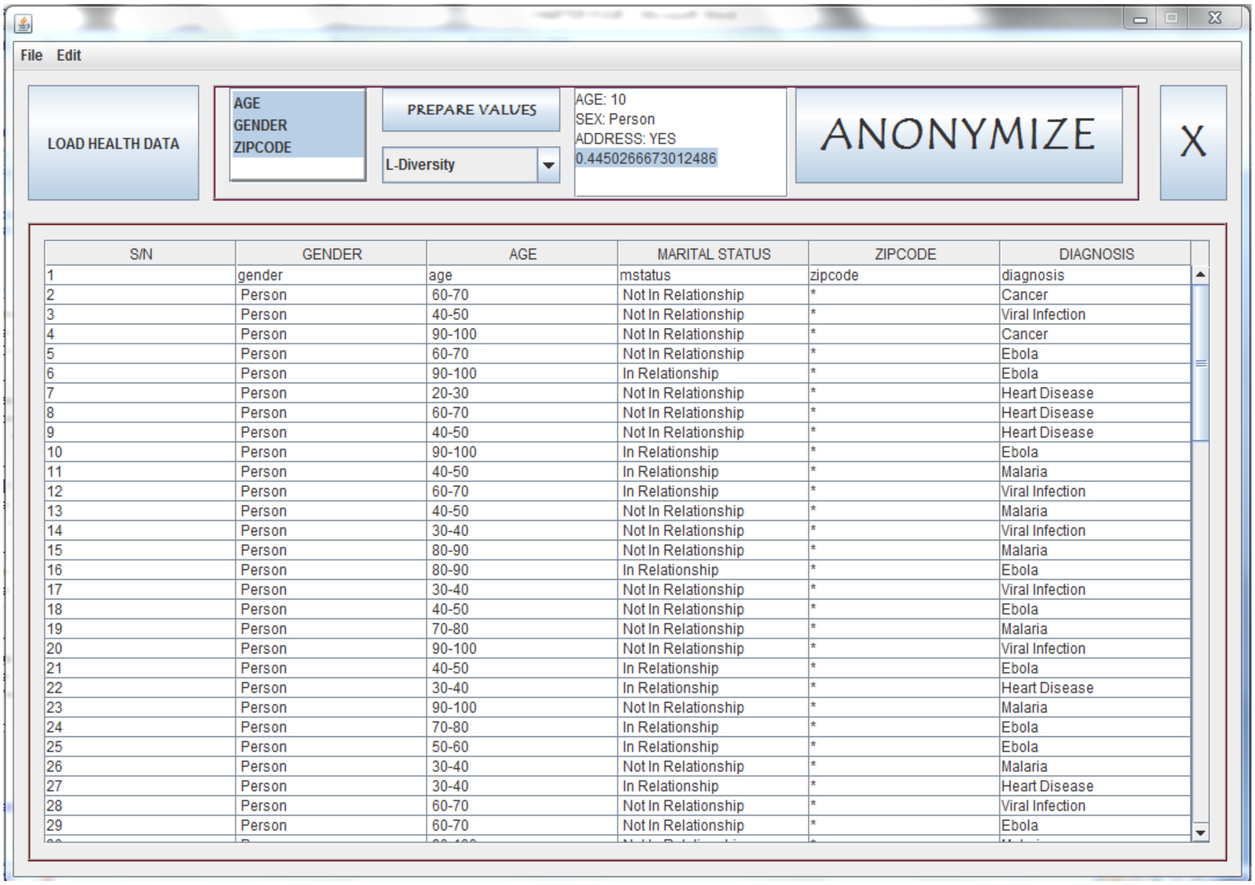


AGE: 50
SEX: Person
ADDRESS: YES
0.501179642705361

S/N	GENDER	AGE	MARITAL STATUS	ZIPCODE	DIAGNOSIS
1	gender	age	mstatus	zipcode	diagnosis
2	Person	50-100	*	7*****	*
3	Person	0-50	*	7*****	*
4	Person	50-100	*	8*****	*
5	Person	50-100	*	4*****	*
6	Person	50-100	*	7*****	*
7	Person	0-50	*	5*****	*
8	Person	50-100	*	5*****	*
9	Person	0-50	*	3*****	*
10	Person	50-100	*	6*****	*
11	Person	0-50	*	5*****	*
12	Person	50-100	*	1*****	*
13	Person	0-50	*	1*****	*
14	Person	0-50	*	8*****	*
15	Person	50-100	*	3*****	*
16	Person	50-100	*	7*****	*
17	Person	0-50	*	5*****	*
18	Person	0-50	*	3*****	*
19	Person	50-100	*	1*****	*
20	Person	50-100	*	4*****	*
21	Person	0-50	*	1*****	*
22	Person	0-50	*	7*****	*
23	Person	50-100	*	7*****	*
24	Person	50-100	*	5*****	*
25	Person	50-100	*	7*****	*
26	Person	0-50	*	4*****	*
27	Person	0-50	*	6*****	*
28	Person	50-100	*	6*****	*
29	Person	50-100	*	3*****	*

Figure 4.3: Graphical User interface of Designed System (K-Anonymity)

This is the main home page of the system, by default remains blank on the data load area, below the menu bars. But on shows the first step of anonymised dataset by K-anonymity.

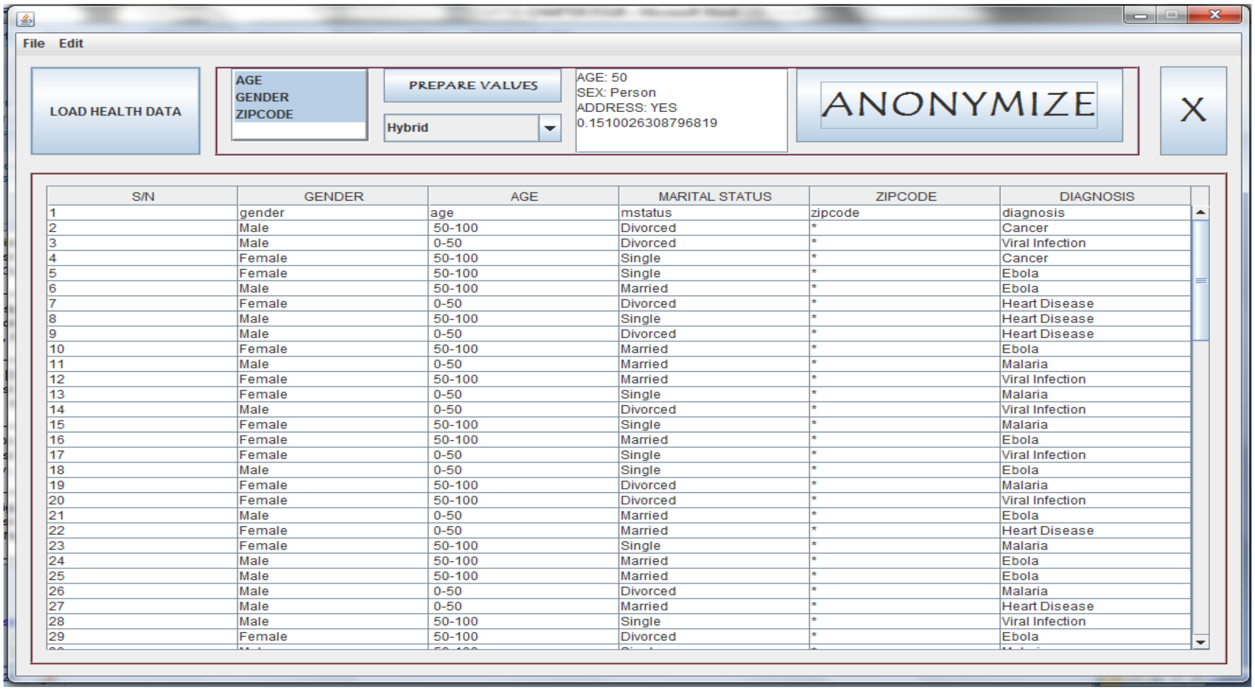


AGE: 10
SEX: Person
ADDRESS: YES
0.4450266673012486

S/N	GENDER	AGE	MARITAL STATUS	ZIPCODE	DIAGNOSIS
1	gender	age	mstatus	zipcode	diagnosis
2	Person	60-70	Not In Relationship	*	Cancer
3	Person	40-50	Not In Relationship	*	Viral Infection
4	Person	90-100	Not In Relationship	*	Cancer
5	Person	60-70	Not In Relationship	*	Ebola
6	Person	90-100	In Relationship	*	Ebola
7	Person	20-30	Not In Relationship	*	Heart Disease
8	Person	60-70	Not In Relationship	*	Heart Disease
9	Person	40-50	Not In Relationship	*	Heart Disease
10	Person	90-100	In Relationship	*	Ebola
11	Person	40-50	In Relationship	*	Malaria
12	Person	60-70	In Relationship	*	Viral Infection
13	Person	40-50	Not In Relationship	*	Malaria
14	Person	30-40	Not In Relationship	*	Viral Infection
15	Person	80-90	Not In Relationship	*	Malaria
16	Person	80-90	In Relationship	*	Ebola
17	Person	30-40	Not In Relationship	*	Viral Infection
18	Person	40-50	Not In Relationship	*	Ebola
19	Person	70-80	Not In Relationship	*	Malaria
20	Person	90-100	Not In Relationship	*	Viral Infection
21	Person	40-50	In Relationship	*	Ebola
22	Person	30-40	In Relationship	*	Heart Disease
23	Person	90-100	Not In Relationship	*	Malaria
24	Person	70-80	In Relationship	*	Ebola
25	Person	50-60	In Relationship	*	Ebola
26	Person	30-40	Not In Relationship	*	Malaria
27	Person	30-40	In Relationship	*	Heart Disease
28	Person	60-70	Not In Relationship	*	Viral Infection
29	Person	60-70	Not In Relationship	*	Ebola

Figure 4.3: Graphical User interface of Designed System (L-Diversity)

This figure shows the anonymised dataset that only L-diversity is applied no to diversify the anonymised dataset by K-anonymity.



Hybrid

AGE: 50
SEX: Person
ADDRESS: YES
0.1510026308796819

S/N	GENDER	AGE	MARITAL STATUS	ZIPCODE	DIAGNOSIS
1	gender	age	mstatus	zipcode	diagnosis
2	Male	50-100	Divorced	*	Cancer
3	Male	0-50	Divorced	*	Viral Infection
4	Female	50-100	Single	*	Cancer
5	Female	50-100	Single	*	Ebola
6	Male	50-100	Married	*	Ebola
7	Female	0-50	Divorced	*	Heart Disease
8	Male	50-100	Single	*	Heart Disease
9	Male	0-50	Divorced	*	Heart Disease
10	Female	50-100	Married	*	Ebola
11	Male	0-50	Married	*	Malaria
12	Female	50-100	Married	*	Viral Infection
13	Female	0-50	Single	*	Malaria
14	Male	0-50	Divorced	*	Viral Infection
15	Female	50-100	Single	*	Malaria
16	Female	50-100	Married	*	Ebola
17	Female	0-50	Single	*	Viral Infection
18	Male	0-50	Single	*	Ebola
19	Female	50-100	Divorced	*	Malaria
20	Female	50-100	Divorced	*	Viral Infection
21	Male	0-50	Married	*	Ebola
22	Female	0-50	Married	*	Heart Disease
23	Female	50-100	Single	*	Malaria
24	Male	50-100	Married	*	Ebola
25	Male	50-100	Married	*	Ebola
26	Male	0-50	Divorced	*	Malaria
27	Male	0-50	Married	*	Heart Disease
28	Male	50-100	Single	*	Viral Infection
29	Female	50-100	Divorced	*	Ebola

Figure 4.4: Graphical User interface of Designed System (Hybrid)

This shows the dataset after the Hybrid option from the select bar on the menu. Once the Hybrid option is select and the anonymised button is clicked on , the dataset is anonymised using K-anonymity and Diversify with L-diversity all at once.

4.9 Summary

This chapter described in detail the proposed Development of a Hybrid K-Anonymity Model for Data Mining Privacy Protection system. It showed the step-by-step implementation of the proposed approach from the dataset input to retrieval of anonymise result. In the next chapter a comprehensive results analysis of the proposed system will be discussed.

CHAPTER FIVE

RESULT ANALYSIS AND DISCUSSION

5.1 Introduction

In this chapter, an evaluation of the proposed Hybrid K-Anonymity Model for Data Mining Privacy Protection is conducted. The Hybrid K-Anonymity Model approach accepts dataset and apply K-anonymity by generalization is performed on the dataset the L-diversity is applied on the anonymized datasets transforms into a result ready for data mining activities. The method proposed by this research took into consideration problems of homogeneity attack and background knowledge attack current Hybrid K-Anonymity Model for Data Mining Privacy Protection approach solve the identified problems.

5.2 Performance Discussion

Data Mining indicates mining or deriving wisdom from voluminous data. Data Mining has been defined as the procedure of finding intriguing knowledge from huge volume of data that has been saved in databases, or any data archives. Through data mining, it becomes possible to extract consistencies, interesting facts, or advanced information out of the database checked or searched from various angles. Such extracted knowledge can then be used for query processing, decision making, data management, and process control. Considered as the most crucial benchmarks in database systems, Data mining certainly is one among the best reliable interdisciplinary advancements in the industry of Information. Data mining investigation involves extracting possibly fruitful information from voluminous selection of data that includes a wide range of application domains like client relationship management and market basket research. There are lot of opportunities to mishandle the delicate information if the sensitive data about persons are leaked. Availability of voluminous data assures the possibility of learning a great deal of information regarding individuals out of the public data. This is due to the fact that the ability to save users' personal data has increased and data mining algorithm leveraging this information have become increasingly sophisticated. It is not possible to apply privacy constraints in a single step. One must remember that PPDM technique has to be applied throughout the data mining practice beginning with data collection through information/knowledge generation. The goal of PPDM consists of constructing procedures to transform the raw data in such a way as to maintain confidentiality of private knowledge and private data even subsequent to data mining process.

In order to experiment using the proposed Hybrid K-Anonymity Model for Data Mining Privacy Protection in this thesis, a medical data was generated containing medical records of different individuals. The dataset was cleaned and change to .csv format and stored in the Netbeans project data folder using IDE editor. The proposed Hybrid K-Anonymity Model for Data Mining Privacy Protection approach involves application of k-anonymity and L-diversity techniques to anonymize dataset and made ready for data mining activities. The proposed approach goes beyond anonymization of the dataset; it also diversifies the dataset to prevent linking attack.

5.3 Results and Analysis

During this stage, the performance of the system was tested with medical dataset as shown in figure 4.4 and compared with the work of Srivastava, (2015) in line with the stated objectives of our research. The comparative performance and privacy level of the proposed system was based on the following parameters: Information Loss, data size, Running Time privacy level.

Table 5.1: Result obtained from calculating Information Loss

	Existing System		**Proposed System**
	K-Anonymity	L-diversity	K-Anonymity + L-diversity
Information Loss	0.501179642705361	0.4450266673012486	0.1510026308796819

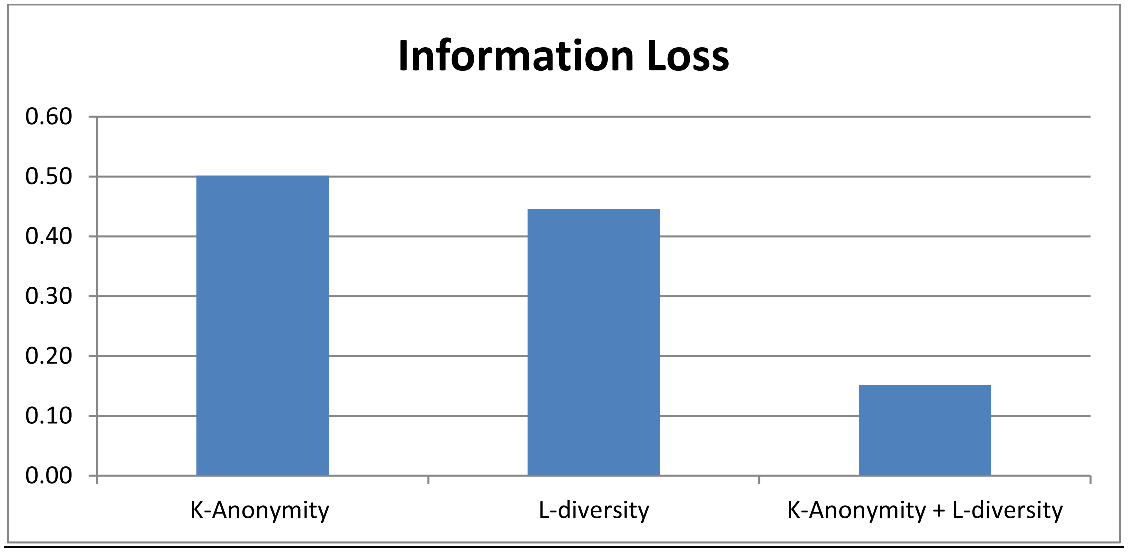


Figure 5.1: Bar chart of result obtained from calculating information loss

From the analysis it is seen that the proposed system (K-Anonymity + L-Diversity) has a lower Information Loss of 0.1510when compared with the existing system which has 0.5011 if used with K-Anonymity alone and 0.4450 if used with L-Diversity alone . Thus, making the new system less prone to Homogeneity and Background knowledge attacks.

5.4 Summary

This chapter has presented the evaluation of the proposed Hybrid K-Anonymity Model for Data Mining Privacy Protection approach of this thesis, which shows the performance of the proposed approach in comparison with the existing Data Mining Privacy Protection approach in the work of Srivastava, (2015). Several evaluation processes have been presented to show the strength of the approach proposed in this thesis in terms of protecting privacy of individuals and also making the data available for use. Chapter 6 will discuss on the summary and provide a conclusive analysis of what the results in this chapter mean.

CHAPTER SIX

SUMMARY, CONCLUSION

6.1 Summary

This dissertation introduced a new approach for Privacy Preserving Data Mining. Although there have been many researches on K-anonymity, which shown that a k-anonymized dataset permits strong attacks due to lack of diversity in the sensitive attributes. This dissertation introduced ℓ-diversity, a framework that gives stronger privacy guarantees. most of the existing algorithms have several weaknesses either caused by information loss or data utility due to the designs of the algorithms themselves. Our proposed algorithm was implemented and tested against homogeneity and background knowledge attacks and proved more privacy than the base algorithm we extended. Therefore, it can be considered as a good alternative to some applications such as electronic commerce.

6.2 Conclusion

In line with the objectives of the research, the following have been achieved:

i. An improved diversity has been merged with K-anonymity which produced a strong algorithm that is devoid of being attacked by homogeneity and background knowledge attacks.

ii. An algorithm where the anonymized dataset cannot be used to identify individuals privacy in any other published dataset, unlike the existing system where the anonymized dataset could be used to identify individuals privacy in a published dataset have been produced.

iii. The algorithm was successfully implemented in java and Netbeans IDE.

iv. A comparison analysis carried out between the existing system and the proposed system showed that the proposed system gives a better Privacy Preservation against Homogeneity and background attack.

REFERENCES

A. Abbas and S. U. Khan, (2014) ‘‘A review on the state-of-the-art privacy preserving approaches in the e-health clouds,’’ IEEE J. Biomed. Health Inform., vol. 18, no. 4, pp. 1431–1441.

A. B. Bondi, (2000)"Characteristics of scalability and their impact on performance", Proc. 2nd Int. Workshop Softw. Perform., pp. 195-203.

A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, (2007) " ell -diversity: Privacy beyond k-anonymity ", ACM Trans. Knowl. Discovery Data, vol. 1, no. 1, pp. 3.

A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, (2016)‘‘Protection of big data privacy,’’ IEEE Access, vol. 4, pp. 1821–1834.

A. R. Beresford and F. Stajano, (2003)‘‘Location privacy in pervasive computing,’’ IEEE Pervasive Comput., vol. 2, no. 1, pp. 46–55.

Aggarwal C, Philip S Yu, (2004) "A condensation approach to privacy preserving data mining", EDBT, 183-199.

Aggarwal C, Philip S Yu, (2008) "A General Survey of Privacy- Preserving Data Mining Models and Algorithms", Springer Magazine, XXII, 11-52.

Agrawal, R., and Srikant, R. (2000). Privacy-preserving data mining. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data - SIGMOD ’00, 29(2), 439–450. https://doi.org/10.1145/342009.335438

Alpa K. Shah, Ravi Gulati, (2012) " A Survey on Cryptographic Techniques for Privacy Preserving Data Mining", IIJDWM, Mining Vol 2 Issue1 pp: 8-12

ArisGkoulalas-Divanis and Vassilios S. Verikios, (2010)"An Overview of Privacy Preserving Data Mining", Published by The ACM Student Magazine.

ArunaKumari D, Rajasekhara Rao K, Suman M (2014) PrivacyPreserving Data Mining. Springer International Publishing. 517–24.

ArunaKumari D, Rajasekhara Rao K, Suman M. (2014) Privacy Preserving Data Mining. Springer International Publishing.; 517–24.

Arvind Narayanan and VitalyShmatikov, (2008) Robust de-anonymization of large sparse datasets. In IEEE Symposium on Security and Privacy, pages 111-125.

Arvind Narayanan and VitalyShmatikov. (2006) How to break anonymity of the netflix prize dataset. CoRR, abs/cs/0610105.

B. C. M. Fung, K. Wang, R. Chen, P. S. Yu, (2010) "Privacy-preserving data publishing: A survey of recent developments", ACM Comput. Surveys., vol. 42, no. 4, pp. 14:1-14:53.

B. Carbunar, Y. Yu, W. Shi, M. Pearce, and V. Vasudevan, ‘‘Query privacy in wireless sensor networks,’’ ACM Trans. Sensor Netw., vol. 6, no. 2, p. 14, 2010.

B. Jiang and X. Yao, ‘‘Location-based services and GIS in perspective,’’ Comput., Environ. Urban Syst., vol. 30, no. 6, pp. 712–725, 2006.

Benjamin CMF, Ke W, Rui C, Philip SY (2010) Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Computing Surveys. Jun; 42(4).

Benjamin CMF, Ke W, Rui C, Philip SY (2010). Privacy-PreservingData Publishing: A Survey of Recent Developments. ACM Computing Surveys.42(4).

C. J. van Rijsbergen, (1979) Information Retrieval, Newton, MA, USA:Butterworth-Heinemann.

Cano I., Torra V, (2009) “Generation of synthetic data by means of fuzzy c-Regression” . IEEE International Conference on Fuzzy Systems, FUZZ-IEEE, pp: 1145 – 1150

Charu CA, Philip SY. (2008) A General Survey of PrivacyPreserving Data Mining Models and Algorithms, Springer US.; 11–52.

D. Agrawal and C. C. Aggarwal, (2001) ‘‘On the design and quantification of privacy preserving data mining algorithms,’’ in Proc. 20th ACM SIGMODSIGACT-SIGART Symp. Principles Database Syst.pp. 247–255.

D. Agrawal, C. C. Aggarwal, (, 2001) "On the design and quantification of privacy preserving data mining algorithms", Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst., pp. 247-255.

Dan Zhu, Xiao-Bai Li, Shuning Wu, (2009)“Identity disclosure protection: A data reconstruction approach for privacypreserving data mining”, Decision Support Systems 48 133– 140.

E. Bertino, D. Lin, and W. Jiang, (2008) ‘‘A survey of quantification of privacy preserving data mining algorithms,’’ in Privacy-Preserving Data Mining. New York, NY, USA: Springer, pp. 183–205.

E. Bertino, I. N. Fovino, (2005)"Information driven evaluation of data hiding algorithms", Proc. Int. Conf. Data Warehousing Knowl. Discovery, pp. 418-427.

E. Poovammal and M. Ponnavaikko, (2009) “Task Independent Privacy Preserving Data Mining on Medical Dataset”, International Conference on Advances in Computing, Control and Telecommunication Technologies.

G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, (2008) ‘‘Private queries in location based services: Anonymizers are not necessary,’’ in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 121–132.

G. Taban and V. D. Gligor, ‘‘Privacy-preserving integrity-assured data aggregation in sensor networks,’’ in Proc. IEEE Int. Conf. Comput. Sci. Eng. (CSE), vol. 3. Aug. 2009, pp. 168–175.

Ghalehsefidi, Narges J, Mohammad ND (2016). "A HybridAlgorithm based on Heuristic Method to Preserve Privacy in Association Rule Mining" Indian Journal of Science and Technology. https://doi.org/10.17485/ijst/2016/v9i27/97476.

H. Rong, H.-M. Wang, J. Liu, and M. Xian, (2016)‘‘Privacy-preserving k-nearest neighbor computation in multiple cloud environments,’’ IEEE Access, vol. 4, pp. 9589–9603.

Honda, K. ; Kawano, A. ; Notsu, A. ; Ichihashi, H., (2012) “A fuzzy variant of k-member clustering for collaborative filtering with data anonymization”, Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on, pp: 1-6

I. F. Akyildiz and I. H. Kasimoglu, ‘‘Wireless sensor and actor networks: Research challenges,’’ Ad Hoc Netw., vol. 2, no. 4, pp. 351–367, Oct. 2004.

J. Krumm, (2009)‘‘A survey of computational location privacy,’’ Pers. Ubiquitous Comput., vol. 13, no. 6, pp. 391–399.

J. Krumm, ‘‘A survey of computational location privacy,’’ (2009) Pers. Ubiquitous Comput., vol. 13, no. 6, pp. 391–399.

J. Lim, H. Yu, K. Kim, M. Kim, and S.-B. Lee, (2017) ‘‘Preserving location privacy of connected vehicles with highly accurate location updates,’’ IEEE Commun. Lett., vol. 21, no. 3, pp. 540–543.

J. Meyerowitz and R. R. Choudhury, (2009)‘‘Hiding stars with fireworks: Location privacy through camouflage,’’ in Proc. 15th Annu. Int. Conf. Mobile Comput. Netw, pp. 345–356.

J. Zhou, Z. Cao, X. Dong, and X. Lin, (2015)‘‘PPDM: A privacy-preserving protocol for cloud-assisted e-healthcare systems,’’ IEEE J. Sel. Topics Signal Process., vol. 9, no. 7, pp. 1332–1344.

Jiang, Clifton and Kantarcıoğlu, (2008) “Transforming SemiHonest Protocols to Ensure Accountability”, Data and Knowledge Engineering, Elsevier.

Justin Zhan, Stan Matwin, Li Wu Chang, (2007) “Privacypreserving collaborative association rule mining”, Journal of Network and Computer Applications 30 1216–1227.

L. Sweeney. (2002) k-Anonymity: A Model for Protecting Privacy. International journal of uncertainty, fuzziness, and knowledge-based systems.

Li Liu, Murat Kantarcioglu and BhavaniThuraisingham, (2009) “Privacy Preserving Decision Tree Mining from Perturbed Data”, Proceedings of the 42nd Hawaii International Conference on System Sciences.

Lindell, Y., and Pinkas, B. (2000). Privacy Preserving Data Mining. Crypto, 1880(3), 36–54. https://doi.org/10.1007/3-540-44598-6_3

M. M. Groat, W. Hey, and S. Forrest, (2011) ‘‘KIPDA: k-indistinguishable privacy-preserving data aggregation in wireless sensor networks,’’ in Proc. IEEE INFOCOM, pp. 2024–2032.

M. Naveed et al.(2015.) ‘‘Privacy in the genomic era,’’ ACM Comput. Surv., vol. 48, no. 1, p. 6.

MadhusudanaShashanka, (2010) “A Privacy–Preserving Framework for Gaussian Mixture Models”, IEEE International Conference on Data Mining Workshops.

Majid Bashir Malik, M. Asger Ghazi, Rashid Ali (2012). ” Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”, Third International Conference on Computer and Communication Technology.

Mike Masnick. (2006) Forget the government, aol exposes search queries to everyone. http://www.techdirt.com/articles/20060807/0219238.shtml.

N. Li, N. Zhang, S. K. Das, and B. Thuraisingham, (2009)‘‘Privacy preservationin wireless sensor networks: A state-of-the-art survey,’’ Ad Hoc Netw.,vol. 7, no. 8, pp. 1501–1514.

Neil Hunt. (2010) Netflix prize update. http://blog.netflix.com/2010/03/this-is-neil-hunt-chief-product-officer.html.

NissimMatatov, LiorRokach, OdedMaimon, (2010) “Privacypreserving data mining: A feature set partitioning approach”, Information Sciences 180 2696–2720.

P. J. McLaren et al., (2016)‘‘Privacy-preserving genomic testing in the clinic:A model using HIV treatment,’’ Genet. Med., vol. 18, no. 8, pp. 814–822.

P. Mell and T. Grance, (2009)‘‘The NIST definition of cloud computing,’’ Nat.Inst. Standards Technol., vol. 53, no. 6, p. 50.

P. Samarati, (2001) "Protecting respondents identities in microdata release", IEEE Trans. Knowl. Data Eng., vol. 13, no. 6, pp. 1010-1027.

R. Agrawal and R. Srikant, (2000) ‘‘Privacy-preserving data mining,’’ ACMSIGMOD Rec., vol. 29, no. 2, pp. 439–450.

R. J. Bayardo, R. Agrawal, (2005) "Data privacy through optimal k-anonymization", Proc. IEEE 21st Int. Conf. Data Eng. (ICDE), pp. 217-228.

S. Dua, X. Du, (2011) Data Mining and Machine Learning in Cybersecurity, Boca Raton, FL, USA:CRC Press.

S. Fletcher, M. Zahidul. Islam, (2015) "Measuring information quality for privacy preserving data mining", Int. J. Comput. Theory Eng., vol. 7, no. 1, pp. 21-28.

S. R. M. Oliveira and O. R. Zaıane, (2010)‘‘Privacy preserving clustering by data transformation,’’ J. Inf. Data Manage., vol. 1, no. 1, p. 37.

S. R. M. Oliveira, O. R. Za, (2010)"Privacy preserving clustering by data transformation", J. Inf. Data Manage., vol. 1, no. 1, pp. 37-52.

S. R. Oliveira, O. R. Zaiane, (2002) "Privacy preserving frequent itemset mining", Proc. IEEE Int. Conf. Privacy Secur. Data Mining, vol. 14, pp. 43-54.

S. R. Oliveira, O. R. Zaiane, (2002)"Privacy preserving frequent itemset mining", Proc. IEEE Int. Conf. Privacy Secur. Data Mining, vol. 14, pp. 43-54.

S. Vijayarani, A Tamilarasi and Sampoorna M. (2010) Analysis of privacy preserving k-anonymity methods and techniques. Proceedings of IEEE international conference on communication and computational intelligence (INCOCCI). pp. 540-545.

SathiyaPriya, K.; Sadasivam, G.S.;Celin, (2011) “A new method for preserving privacy in quantitative association rules using DSR approach with automated generation of membership function”, World Congress on Information and Communication Technologies (WICT), pp:148-153

Sin G Teo, Vincent Lee, Shuguo Han, (2012) “A Study of Efficiency and Accuracy of Secure Multiparty Protocol in Privacy-Preserving Data Mining”, 26th International Conference on Advanced Information Networking and Applications Workshops.

Sweeney L, (2002) "Achieving k-Anonymity privacy protection using generalization and suppression" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571-588.

Tiancheng Li, Ninghui Li, (2008) “Towards Optimal kanonymization”, Data and Knowledge Engineering, Elsevier. 303

V. S. Iyengar, (2002) "Transforming data to satisfy privacy constraints", Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 279-288.

W. He, X. Liu, H. Nguyen, K. Nahrstedt, and T. Abdelzaher, (2007) ‘‘PDA: Privacy-preserving data aggregation in wireless sensor networks,’’ in Proc. 26th IEEE Int. Conf. Comput. Commun. (INFOCOM), pp. 2045–2053.

W. M. Rand, (1971)"Objective criteria for the evaluation of clustering methods", J. Amer. Statist. Assoc., vol. 66, no. 336, pp. 846-850.

Wang P, (2010) "Survey on Privacy preserving data mining", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9.

Wang PS (2010). Survey on Privacy Preserving Data Mining.International Journal of Digital Content Technology and itsApplications. 2010; 4(9):1–7. https://doi.org/10.4156/jdcta.vol4.issue9.1

X. Xiao, Y. Tao (2006) "Personalized privacy preservation", Proc. VLDB, pp. 139-150.

X. Yi, F.-Y. Rao, E. Bertino, and A. Bouguettaya, (2015) ‘‘Privacy-preserving association rule mining in cloud computing,’’ in Proc. 10th ACM Symp. Inf., Comput. Commun. Secur, pp. 439–450.

X.-M. He, X. S. Wang, D. Li, and Y.-N. Hao, (2016.)‘‘Semi-homogenous generalization: Improving homogenous generalization for privacy preservation in cloud computing,’’ J. Comput. Sci. Technol., vol. 31, no. 6,pp. 1124–1135.

Xun Yi, Yanchun Zhang, (2007) “Privacy-preserving distributed association rule mining via semi- trusted mixer”, Data and Knowledge Engineering 63 550–567.

Y. Lindell and B. Pinkas (2000) “Privacy Preserving Data Mining”, Journal of Cryptology, 15(3), pp.36-54.

Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, (2013) ‘‘Unique in the crowd: The privacy bounds of human mobility,’’ Sci. Rep., vol. 3, p. 1376.

Yan ZHU and Lin PENG, (2007) “Study on K-anonymity Models of Sharing Medical Information”, 1-4244-0885- 7/07/$20.00 ©2007 IEEE.

Yehuda Lindell, Benny Pinkas, “Privacy Preserving Data Mining”, http://www.pinkas. net /PAPERS/id3-final.pdf.

Yun Ding and Karsten Klein, (2010)“Model-Driven Application-Level Encryption for the Privacy of EHealth Data”, International Conference on Availability, Reliability and Security.

Yun Ding and Karsten Klein, “Model-Driven Application-Level Encryption for the Privacy of EHealth Data”, International Conference on Availability, Reliability and Security, 2010.

APPENDIX 1: SAMPLE PROGRAM CODE

```
package panels;


import abutech.Utility;
import static data.DBCSV.executeQuery;
import java.io.File;
import java.io.IOException;
import java.nio.charset.Charset;
import java.text.AttributedCharacterIterator;
import java.util.ArrayList;
import java.util.Arrays;
import java.util.Iterator;
import java.util.Scanner;
import java.util.logging.Level;
import java.util.logging.Logger;
import javax.swing.JOptionPane;
import javax.swing.table.DefaultTableModel;
import org.deidentifier.arx.ARXAnonymizer;
import org.deidentifier.arx.ARXConfiguration;
import org.deidentifier.arx.ARXResult;
import org.deidentifier.arx.AttributeType;
import org.deidentifier.arx.AttributeType.Hierarchy;
import org.deidentifier.arx.AttributeType.Hierarchy.DefaultHierarchy;
import org.deidentifier.arx.Data.DefaultData;
import org.deidentifier.arx.criteria.KAnonymity;
import org.deidentifier.arx.criteria.LDiversity;
import org.deidentifier.arx.criteria.RecursiveCLDiversity;
import org.deidentifier.arx.metric.Metric;
import static panels.Example.printResult;


/**
 *
 * @author Caps-Lap
 */
public class Home extends javax.swing.JFrame {


    /**
     * Creates new form Home
     */
    public static String name_range = null;
    public static String age_range = null;
    public static String gender_range = null;
    public static String address_range = null;


    public Home() {
initComponents();
setLocationRelativeTo(null);
setResizable(false);
operation_panel.setVisible(false);
viz_panel.setVisible(false);
        //selectedchoices.setVisible(false);
        //executeButton.setVisible(false);
    }


    /**
```

```
     * This method is called from within the constructor to initialize the form.
     * WARNING: Do NOT modify this code. The content of this method is always
     * regenerated by the Form Editor.
     */
    @SuppressWarnings("unchecked")
    // <editor-fold defaultstate="collapsed" desc="Generated Code">
    private void initComponents() {


        jMenu1 = new javax.swing.JMenu();
viz_panel = new javax.swing.JPanel();
        jScrollPane1 = new javax.swing.JScrollPane();
        jTable1 = new javax.swing.JTable();
operation_panel = new javax.swing.JPanel();
        jButton3 = new javax.swing.JButton();
        jScrollPane2 = new javax.swing.JScrollPane();
anonymize_option = new javax.swing.JList<>();
executeButton = new javax.swing.JButton();
        jScrollPane3 = new javax.swing.JScrollPane();
        jTextArea1 = new javax.swing.JTextArea();
        algorithm = new javax.swing.JComboBox<>();
        jButton2 = new javax.swing.JButton();
        jButton1 = new javax.swing.JButton();
        jMenuBar1 = new javax.swing.JMenuBar();
        jMenu2 = new javax.swing.JMenu();
        jMenuItem2 = new javax.swing.JMenuItem();
        jMenuItem1 = new javax.swing.JMenuItem();
        jMenu3 = new javax.swing.JMenu();
        jMenu1.setText("jMenu1");

        setDefaultCloseOperation(javax.swing.WindowConstants.EXIT_ON_CLOSE);

        viz_panel.setBorder(javax.swing.BorderFactory.createEtchedBorder(new java.awt.Color(102, 0, 0),
null));

        jTable1.setModel(new javax.swing.table.DefaultTableModel(
          new Object [][] {

          },
          new String [] {
            "S/N", "GENDER", "AGE", "MARITAL STATUS", "ZIPCODE", "DIAGNOSIS"
          }
        ) {
boolean[] canEdit = new boolean [] {
            false, false, false, false, false, true
          };

          public booleanisCellEditable(introwIndex, intcolumnIndex) {
            return canEdit [columnIndex];
          }
        });
        jScrollPane1.setViewportView(jTable1);

javax.swing.GroupLayoutviz_panelLayout = new javax.swing.GroupLayout(viz_panel);
viz_panel.setLayout(viz_panelLayout);
viz_panelLayout.setHorizontalGroup(
      viz_panelLayout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(javax.swing.GroupLayout.Alignment.TRAILING, viz_panelLayout.createSequentialGroup()
.addContainerGap()
.addComponent(jScrollPane1)
.addContainerGap())
    );
viz_panelLayout.setVerticalGroup(
      viz_panelLayout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(javax.swing.GroupLayout.Alignment.TRAILING, viz_panelLayout.createSequentialGroup()
.addContainerGap()
.addComponent(jScrollPane1, javax.swing.GroupLayout.DEFAULT_SIZE, 482, Short.MAX_VALUE)
.addContainerGap())
    );

    operation_panel.setBorder(javax.swing.BorderFactory.createEtchedBorder(new java.awt.Color(102,
0, 51), null));

    jButton3.setFont(new java.awt.Font("Tempus Sans ITC", 1, 12)); // NOI18N
    jButton3.setText("PREPARE VALUES");
    jButton3.addActionListener(new java.awt.event.ActionListener() {
      public void actionPerformed(java.awt.event.ActionEventevt) {
        jButton3ActionPerformed(evt);
      }
    });

anonymize_option.setBorder(javax.swing.BorderFactory.createBevelBorder(javax.swing.border.BevelBo
rder.RAISED));
anonymize_option.setModel(new javax.swing.AbstractListModel<String>() {
```

```
String[] strings = { "AGE", "GENDER", "ZIPCODE" };
            public intgetSize() { return strings.length; }
            public String getElementAt(inti) { return strings[i]; }
        });
anonymize_option.setToolTipText("Choose Values to Anonymize");
        jScrollPane2.setViewportView(anonymize_option);


executeButton.setFont(new java.awt.Font("Tempus Sans ITC", 1, 36)); // NOI18N
executeButton.setText("ANONYMIZE");
executeButton.addActionListener(new java.awt.event.ActionListener() {
            public void actionPerformed(java.awt.event.ActionEventevt) {
executeButtonActionPerformed(evt);
            }
        });


        jTextArea1.setEditable(false);
        jTextArea1.setColumns(20);
        jTextArea1.setLineWrap(true);
        jTextArea1.setRows(5);
        jTextArea1.setWrapStyleWord(true);
        jScrollPane3.setViewportView(jTextArea1);


algorithm.setModel(new javax.swing.DefaultComboBoxModel<>(new String[] { "K-Anonimity", "L-
Diversity", "Hybrid" }));


javax.swing.GroupLayoutoperation_panelLayout = new javax.swing.GroupLayout(operation_panel);
operation_panel.setLayout(operation_panelLayout);
operation_panelLayout.setHorizontalGroup(
          operation_panelLayout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(operation_panelLayout.createSequentialGroup()
.addContainerGap()
.addComponent(jScrollPane2, javax.swing.GroupLayout.PREFERRED_SIZE, 120,
javax.swing.GroupLayout.PREFERRED_SIZE)
.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.UNRELATED)
.addGroup(operation_panelLayout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addComponent(jButton3, javax.swing.GroupLayout.PREFERRED_SIZE, 144, Short.MAX_VALUE)
.addComponent(algorithm, 0, javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE))
.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.UNRELATED)
.addComponent(jScrollPane3, javax.swing.GroupLayout.PREFERRED_SIZE, 185,
javax.swing.GroupLayout.PREFERRED_SIZE)
.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)
.addComponent(executeButton, javax.swing.GroupLayout.PREFERRED_SIZE, 283,
javax.swing.GroupLayout.PREFERRED_SIZE)
.addContainerGap())
       );
operation_panelLayout.setVerticalGroup(
          operation_panelLayout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addComponent(jScrollPane3, javax.swing.GroupLayout.PREFERRED_SIZE, 0, Short.MAX_VALUE)
.addGroup(operation_panelLayout.createSequentialGroup()
.addGroup(operation_panelLayout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addComponent(jScrollPane2, javax.swing.GroupLayout.PREFERRED_SIZE, 0, Short.MAX_VALUE)
.addComponent(executeButton, javax.swing.GroupLayout.DEFAULT_SIZE, 78, Short.MAX_VALUE)
.addGroup(operation_panelLayout.createSequentialGroup()
.addComponent(jButton3, javax.swing.GroupLayout.PREFERRED_SIZE, 36,
javax.swing.GroupLayout.PREFERRED_SIZE)
.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.UNRELATED)
.addComponent(algorithm)))
.addContainerGap())
       );

       jButton2.setFont(new java.awt.Font("Tempus Sans ITC", 1, 36)); // NOI18N
       jButton2.setText("X");

       jButton1.setText("LOAD HEALTH DATA");
       jButton1.addActionListener(new java.awt.event.ActionListener() {
          public void actionPerformed(java.awt.event.ActionEventevt) {
              jButton1ActionPerformed(evt);
          }
       });

       jMenu2.setText("File");

       jMenuItem2.setText("Home");
       jMenu2.add(jMenuItem2);

       jMenuItem1.setText("Exit");
       jMenuItem1.addActionListener(new java.awt.event.ActionListener() {
          public void actionPerformed(java.awt.event.ActionEventevt) {
              jMenuItem1ActionPerformed(evt);
          }
       });
       jMenu2.add(jMenuItem1);

       jMenuBar1.add(jMenu2);
        jMenu3.setText("Edit");
        jMenuBar1.add(jMenu3);

setJMenuBar(jMenuBar1);

javax.swing.GroupLayout layout = new javax.swing.GroupLayout(getContentPane());
getContentPane().setLayout(layout);
layout.setHorizontalGroup(
layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(layout.createSequentialGroup()
.addContainerGap()
.addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addComponent(viz_panel, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
.addGroup(layout.createSequentialGroup()
.addComponent(jButton1)
.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.UNRELATED)
.addComponent(operation_panel, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
.addGap(18, 18, 18)
.addComponent(jButton2)))
.addContainerGap())
        );
layout.setVerticalGroup(
layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
.addGroup(javax.swing.GroupLayout.Alignment.TRAILING, layout.createSequentialGroup()
.addContainerGap()
.addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING, false)
.addComponent(operation_panel, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)

.addComponent(jButton2, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)

.addComponent(jButton1, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE))

.addGap(18, 18, 18)

.addComponent(viz_panel, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)

.addContainerGap())

        );


pack();

    }// </editor-fold>


    private void jMenuItem1ActionPerformed(java.awt.event.ActionEventevt) {

        // TODO add your handling code here:

    }


    private void jButton1ActionPerformed(java.awt.event.ActionEventevt) {

        try {

            // TODO add your handling code here:

operation_panel.setVisible(true);

ArrayList<String>executeQuery = executeQuery("select * from MedicalData");

System.out.println("Total Query: " + executeQuery.size());

DefaultTableModel model = (DefaultTableModel) jTable1.getModel();

model.setRowCount(0);

inti = 1;

            for (String s :executeQuery) {
```

```
String[] split = s.split("¬");
Object[] obj = new Object[6];
obj[0] = i++;
obj[1] = split[1];
obj[2] = split[2];
obj[3] = split[3];
obj[4] = split[4];
obj[5] = split[5];
model.addRow(obj);
            }

viz_panel.setVisible(true);
        } catch (ClassNotFoundException ex) {
Logger.getLogger(Home.class.getName()).log(Level.SEVERE, null, ex);
        }
    }

    private void jButton3ActionPerformed(java.awt.event.ActionEventevt) {
        // TODO add your handling code here:
PrepareDataprepareData = new PrepareData(anonymize_option);
prepareData.setVisible(true);

    }

    private void executeButtonActionPerformed(java.awt.event.ActionEventevt) {
        try {
            // TODO add your handling code here:
            String param = "";
```

```
          if (PrepareData.age_range != null) {
param += "AGE: " + PrepareData.age_range;
age_range = PrepareData.age_range;
          }
          if (PrepareData.gender_range != null) {
param += "\nSEX: " + PrepareData.gender_range;
gender_range = PrepareData.gender_range;
          }
          if (PrepareData.address_range != null) {
param += "\nADDRESS: " + PrepareData.address_range;
address_range = PrepareData.address_range;
          }
System.out.println("Parameters: " + param);
          jTextArea1.setText(param);

          //Get data and Heirachies
DefaultData data = Utility.loadData("MedicalData");

          Scanner scan = new Scanner(new File("src/data/zipdata"));
DefaultHierarchyziphierarchy = Hierarchy.create();
          while (scan.hasNext()) {
String[] vals = scan.nextLine().split(",");
ziphierarchy.add(vals[0], vals[1], vals[2], vals[3], vals[4], vals[5]);
          }
          scan = new Scanner(new File("src/data/agedata"));
DefaultHierarchyagehierarchy = Hierarchy.create();
          while (scan.hasNext()) {
String[] vals = scan.nextLine().split(",");
```

```
agehierarchy.add(vals[0], vals[1], vals[2]);
          }

          scan = new Scanner(new File("src/data/genderdata"));
DefaultHierarchygenderhierarchy = Hierarchy.create();
          while (scan.hasNext()) {
String[] vals = scan.nextLine().split(",");
genderhierarchy.add(vals[0], vals[1], vals[2]);
          }
          scan = new Scanner(new File("src/data/maritaldata"));
DefaultHierarchymaritalhierarchy = Hierarchy.create();
          while (scan.hasNext()) {
String[] vals = scan.nextLine().split(",");
maritalhierarchy.add(vals[0], vals[1], vals[2]);
          }

          final ARXAnonymizer anonymizer = new ARXAnonymizer();
          final ARXConfigurationconfig = ARXConfiguration.create();

          if (algorithm.getSelectedItem().equals("K-Anonimity")) {
data.getDefinition().setAttributeType("age", agehierarchy);
data.getDefinition().setAttributeType("gender", genderhierarchy);
data.getDefinition().setAttributeType("zipcode", ziphierarchy);
config.addPrivacyModel(new KAnonymity(2));

          } else if (algorithm.getSelectedItem().equals("L-Diversity")) {
data.getDefinition().setAttributeType("age", agehierarchy);
data.getDefinition().setAttributeType("gender", genderhierarchy);
data.getDefinition().setAttributeType("zipcode", AttributeType.IDENTIFYING_ATTRIBUTE);
data.getDefinition().setAttributeType("mstatus", maritalhierarchy);
data.getDefinition().setAttributeType("diagnosis", AttributeType.SENSITIVE_ATTRIBUTE);
config.addPrivacyModel(new RecursiveCLDiversity("diagnosis", 3, 2));

      } else if (algorithm.getSelectedItem().equals("Hybrid")) {
data.getDefinition().setAttributeType("age", agehierarchy);
data.getDefinition().setAttributeType("gender", genderhierarchy);
data.getDefinition().setAttributeType("zipcode", AttributeType.IDENTIFYING_ATTRIBUTE);
data.getDefinition().setAttributeType("mstatus", maritalhierarchy);
data.getDefinition().setAttributeType("diagnosis", AttributeType.SENSITIVE_ATTRIBUTE);
config.addPrivacyModel(new KAnonymity(2));
config.addPrivacyModel(new RecursiveCLDiversity("diagnosis", 3, 2));
      }

      // Create an instance of the anonymizer
      try {
ARXResult result = anonymizer.anonymize(data, config);// Now anonymize
        String printResult = printResult(result, data); // Print info
        if (result.getOutput() != null) {
System.out.println(" - Transformed data:");// Process results
          final Iterator<String[]> transformed = result.getOutput(false).iterator();
DefaultTableModel model = (DefaultTableModel) jTable1.getModel();
model.setRowCount(0);
          for (inti = 0; i<model.getRowCount(); i++) {
model.removeRow(i);
          }
inti = 1;
```

```
                    while (transformed.hasNext()) {
String[] row = transformed.next();
Object[] obj = new Object[row.length + 1];//creates an array of length 7 as the table has 7 rows
obj[0] = i++;
System.arraycopy(row, 0, obj, 1, row.length);
model.addRow(obj);
                    }
                    jTextArea1.setText(jTextArea1.getText()+"\n"+printResult);
                } else {
JOptionPane.showMessageDialog(this, " Criteria cannot be enforced!");
                }

            } catch (Exception e) {
e.printStackTrace();
            }
        } catch (IOException ex) {
Logger.getLogger(Home.class.getName()).log(Level.SEVERE, null, ex);
        }
    }

    /**
     * @paramargs the command line arguments
     */
    public static void main(String args[]) {
        /* Set the Nimbus look and feel */
        //<editor-fold defaultstate="collapsed" desc=" Look and feel setting code (optional) ">
        /* If Nimbus (introduced in Java SE 6) is not available, stay with the default look and feel.
         * For details see http://download.oracle.com/javase/tutorial/uiswing/lookandfeel/plaf.html
     */
    try {
      for (javax.swing.UIManager.LookAndFeelInfo info :
javax.swing.UIManager.getInstalledLookAndFeels()) {
        if ("Nimbus".equals(info.getName())) {
javax.swing.UIManager.setLookAndFeel(info.getClassName());
          break;
        }
      }
    } catch (ClassNotFoundException ex) {
java.util.logging.Logger.getLogger(Home.class.getName()).log(java.util.logging.Level.SEVERE, null, ex);
    } catch (InstantiationException ex) {
java.util.logging.Logger.getLogger(Home.class.getName()).log(java.util.logging.Level.SEVERE, null, ex);
    } catch (IllegalAccessException ex) {
java.util.logging.Logger.getLogger(Home.class.getName()).log(java.util.logging.Level.SEVERE, null, ex);
    } catch (javax.swing.UnsupportedLookAndFeelException ex) {
java.util.logging.Logger.getLogger(Home.class.getName()).log(java.util.logging.Level.SEVERE, null, ex);
    }
    //</editor-fold>

    /* Create and display the form */
java.awt.EventQueue.invokeLater(new Runnable() {
      public void run() {
        new Home().setVisible(true);
      }
    });
  }
```

```
    // Variables declaration - do not modify
    private javax.swing.JComboBox<String> algorithm;
    private javax.swing.JList<String>anonymize_option;
    private javax.swing.JButtonexecuteButton;
    private javax.swing.JButton jButton1;
    private javax.swing.JButton jButton2;
    private javax.swing.JButton jButton3;
    private javax.swing.JMenu jMenu1;
    private javax.swing.JMenu jMenu2;
    private javax.swing.JMenu jMenu3;
    private javax.swing.JMenuBar jMenuBar1;
    private javax.swing.JMenuItem jMenuItem1;
    private javax.swing.JMenuItem jMenuItem2;
    private javax.swing.JScrollPane jScrollPane1;
    private javax.swing.JScrollPane jScrollPane2;
    private javax.swing.JScrollPane jScrollPane3;
    private javax.swing.JTable jTable1;
    private javax.swing.JTextArea jTextArea1;
    private javax.swing.JPaneloperation_panel;
    private javax.swing.JPanelviz_panel;
    // End of variables declaration

}
```