Achieving High Data Utility K-Anonymization Using Similarity-Based Clustering Model

MOHAMMAD RASOOL SARRAFI AGHDAM

DOCTOR OF
PHILOSOPHY

Department of Informatics,
School of Multidisciplinary Sciences,
SOKENDAI
(The Graduate University for Advanced Studies)

2016 (School Year)

March 2017

A dissertation submitted to
Department of Informatics,
School of Multidisciplinary Sciences,
SOKENDAI (The Graduate University for Advanced Studies),
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Advisory Committee:

Advisor	Prof. Noboru Sonehara
	National Institute of Informatics (NII), SOKENDAI
Sub-advisor	Prof. Isao Echizen
	National Institute of Informatics (NII), SOKENDAI
Committee Members	
	Prof. Kento Aida
	National Institute of Informatics (NII), SOKENDAI
	Prof. Akihisa Kodate
	Tsuda College
	Prof. Shigeki Yamada
	National Institute of Informatics (NII)

(Alphabet order of last name except advisor and sub-advisor)

Table of Contents

List of Figures

List of Tables

List of Equations

Acknowledgments

I owe my deepest gratitude to my supervisor Professor Noboru Sonehara, for welcoming me in his laboratory and facilitating me with opportunity to work in a unique research and educational environment with elite researchers during my years in National Institute of Informatics (NII). I will always be grateful for all his constructive advices he taught me to continue and succeed in my research. Also I am thankful for the precious time he dedicated to me during my studies at The Graduate University for Advanced Studies (Sokendai) for discussions regarding my research future and correcting my articles to be acceptable for submissions at international conferences and journals. Professor Noboru Sonehara also introduced me to respectful researchers and business owners, with whom I had the opportunity to discuss, cooperate and work on successful projects and learn the value of research and innovation in society.

I also would like to thank all the members of my PhD committee, Professor Isao Echizen, Professor Kento Aida, Professor Shigeki Yamada at Sokendai and Professor Akihisa Kodate at Tsuda College, for their constructive comments and invaluable advice during my intermediate presentations. I am indebted to them for investing and dedicating time and effort for discussions and suggestions to guide me through my PhD.

I owe special thanks to Professor Kenzo Takahashi who has introduced me to Professor Noboru Sonehara. He has been particularly helpful and always encouraged me to continue my studies at higher level.

Moreover, I would like to thank all my lab mates for sharing their experiences and reporting their feedback on my work. Also I would like to thank all the staff in the

International Affairs and Education Support Team at The Graduate University for Advanced Studies (Sokendai) for their assistance, which made my life comfortable.

Last but not least, I would like to thank my family, my father Saeid, my mother Afkham and my brother Ali, who have been always supporting me emotionally and financially so I could continue and proceed with my education up to this PhD degree. Certainly none, and I mean NONE, of this work would exist without their love and support.

Abstract

Nowadays, in the communication, smart devices, social networks and Big Data era, privacy has become one of the main concerns of all socially active individuals. Mainly because of advanced information processing, storage capacity, data mining technologies and the necessity of information sharing in social life. There are lots of government agencies, organizations and service providers that collect and store huge amount of information containing personal information of individuals as their common procedure. The collected micro-data which is a combination of categorical and numerical attributes, contains identifying attributes (e.g., Date of Birth, Sex and Zip code) and private sensitive attributes (e.g., Salary, Credit Card Records and Disease). Sharing the collected micro-data for research and education purposes would be very helpful for researchers and data miners to investigate the correlation between different attributes and get some useful outcomes. However, individual's privacy is one of the main concerns in data publishing especially when releasing datasets involving human subjects contain private sensitive information. Even though information such as name and social security number are discarded in the shared dataset re-identification of individuals is still very much possible due to the existence of other identifying attributes. Therefore to protect the privacy of individuals, a model that is widely used for privacy preservation in publishing micro-data, k-anonymity model was proposed by Samarati and Sweeney. It suggests “For every record in a released dataset there should be at least k-1 other records identical to it along the quasi-identifier attributes”. K-anonymity from clustering aspect is defined as “clustering with constrain of minimum k tuples in each group”.

K-anonymity protects the privacy of individuals by modifying the values of identifying attributes through generalization and suppression. Through this modification some information loss occurs. Information loss in k-anonymity model is an unfortunate and inevitable consequence. This information loss reduces the utility of anonymized-data and makes the anonymized-data to be less accurate and accordingly less useful for further analysis. In addition, there is a trade-off relationship between the privacy and data utility. Due to this trade-off, performing anonymization with maximum privacy and attaining maximum data utility is not possible. Moreover, the problem of optimal k-anonymization and computational complexity of finding an optimal solution for the k-anonymity problem has been proven to be NP-hard. Furthermore, real world and census datasets contain both numerical and categorical type data. As a matter of fact most of the QID attributes in micro-data are assume to be categorical. The combination of numerical and categorical attributes makes anonymization process rather complicated and very often results in an inefficient anonymization with very high information loss. Most of the previous approaches and techniques to achieve k-anonymity suffer from huge information loss and very low data utility. Also most of the approaches are mainly designed for continuous numerical attributes and in case of considering categorical attributes, they depend on hierarchical taxonomies or require some additional information, which more often than not, are not defined or available in real life applications.

Therefore, in order to maximize the utility of anonymized-data in real life applications in this work a new approach is proposed. The proposed model is called Similarity-Based Clustering (SBC) Anonymization. It is based on clustering and local recoding anonymization method. SBC model concentrates on clustering the original dataset containing both numerical and categorical attributes efficiently based on given k value so after anonymization the information loss kept as minimum as possible. SBC model

suggests a new similarity measurement and distance calculation based on the measured similarity for categorical attributes so the total distance between tuples can be calculated for clustering. This approach does not depend on hierarchical taxonomies regarding categorical attributes. Based on the proposed model a bottom-up greedy algorithm for k-anonymization is proposed and evaluated on two different real datasets. Our extensive study on information loss and data utility show that the proposed algorithm based on SBC model in comparison with existing well-known algorithms offers data utility above 80% and reduces the information loss to less than 20% within the wide range of various k values.

Keywords: Privacy Preserving Data Mining, Anonymization, Privacy and Algorithm

Thesis Overview

At present, in the communication, smart devices (smartphones, tablets, Google glass), social networks and Big Data era, privacy has become one of the main concerns of all socially active individuals. Mainly because of advanced information processing, storage capacity, data mining technologies and the necessity of information sharing in social life. Also, there are lots of government agencies, organizations and service providers in real and cyber world (hospitals, universities, banks, social network and etc.) that collect and store huge amount of information containing personal information of individuals as their common procedure. The collected data at individual level is called micro-data. The attributes in micro-data, which is a combination of categorical (e.g., Gender, Nationality) and numerical (e.g., Age) attributes, can be divided into two main categories. The attributes which are used to identify an individual, called quasi-identifier (QID) attributes (e.g., Date of Birth, Sex and Zip code) and the attributes which are not normally shared with public or strangers, called private sensitive attributes (e.g., Salary, Credit Card Records and Disease). Information sharing and data publishing has a long history in information technology and due to the regulations, mutual benefits or for some other reasons such as business, marketing, research and education purposes there is always a demand for sharing the collected information among various parties. Publishing or sharing the collected micro-data for research and education purposes would be very helpful for researchers and data miners to investigate the correlation between different attributes and get some useful outcomes. For instance, the relation between a particular disease and location or Sex could be very helpful to prevent and possibly find a cure for that specific disease.

However, individual's privacy is one of the main concerns in data publishing especially when releasing datasets involving human subjects contain private sensitive information such as financial information or medical history. Even though information such as name and other identifiers (such as driver license number and social security number) are discarded in the shared dataset (anonymous shared dataset) identifying information about specific individuals is still very much possible since a particular record can often be uniquely identified from the combination of other QID attributes. There are lots of data and information available on Internet, which is accessible to everyone, and the QID attributes of the released dataset could be linked with the QID attributes in the external datasets. This linking, which technically known as “linking attack”, could lead to re-identifying individuals uniquely and consequently lead to releasing the sensitive information which were not meant to be released by individuals.

Therefore given the thread of re-identification in our growing digital society, guaranteeing privacy of individuals while providing accurate and high utility data for data mining and knowledge discovery has become rather difficult issue. Therefore in order to protect the privacy of individuals, a model that is widely used for privacy preservation in publishing micro-data, k-anonymity model was proposed. Samarati and Sweeney proposed K-anonymity in 2002 which suggests “For every record in a released dataset there should be at least k-1 other records identical to it along the quasi-identifier attributes”. Generally k-anonymity can be also defined from clustering point of view. Clustering is the process of arranging similar records in groups so that the records belonging to the same cluster have high similarity, while records belonging to different clusters have high dissimilarity. K-anonymity from clustering aspect is defined as “clustering with constrain of minimum k tuples in each group”.

The K value in k-anonymity model is the minimum number of data records with identical QID attributes in k-anonymous dataset. K value basically is the anonymization degree representing the level of desired privacy. Obviously by having larger k value the privacy protection is higher and having low k value (k=2) is providing minimum privacy for the dataset. K-anonymity protects the privacy of individuals by modifying the values of QID attributes through generalization and suppression so that each record in the released dataset is indistinguishable from at least k-1 other records within the same dataset. Therefore the linking confidence between the k-anonymous released dataset and the external dataset will reduce by 1/k ratio. Thus it can be concluded that the privacy of individuals is protected to some extent.

By modifying QID attributes using generalizing or suppression in original dataset to form k-anonymous dataset some information loss occurs. Information loss in k-anonymity model is an unfortunate and inevitable consequence. This information loss reduces the utility of anonymized-data and makes the anonymized-data to be less accurate and accordingly less useful for data mining, knowledge discovery and research purposes.

In addition, there is a trade-off relationship between the privacy level and the quality of anonymized-data. Choosing larger k value means providing higher privacy and consequently obtaining less utility k-anonymous dataset. Due to this trade-off, performing anonymization with maximum privacy and attaining maximum utility for anonymized-data is not possible. Moreover the problem of optimal k-anonymization and computational complexity of finding an optimal solution for the k-anonymity problem has been proven to be NP-hard.

Furthermore, real world and census datasets contain both numerical and categorical type data. As a matter of fact most of the QID attributes in micro-data are assume to be categorical with no hierarchical taxonomies. The combination of numerical and categorical attributes makes anonymization process rather complicated and very often results in an inefficient anonymization with very high information loss. Most of the previous approaches and techniques to achieve k-anonymity suffer from huge information loss and very low data utility (anonymized-data). Also most of the approaches are mainly designed for continuous numerical attributes and in case of considering categorical attributes, they depend on hierarchical taxonomies or require some additional information, which more often than not, are not defined or available in real life applications.

Therefore, in order to maximize the utility of anonymized-data in real life applications (datasets containing both numerical and categorical attributes with NO hierarchical taxonomies), in this work a new approach is proposed. The proposed model is called Similarity-Based Clustering (SBC) Anonymization. It is based on clustering and local recoding anonymization method. As it was mentioned earlier k-anonymity is actually clustering with constrain of minimum k tuples in each group, thus SBC model concentrates on clustering the original dataset containing both numerical and categorical attributes efficiently based on given k value so after anonymization the information loss kept as minimum as possible. The key point to reduce the information loss is to retain the records in a cluster (equivalent class) as similar to each other as possible. Therefore when all the records in the same cluster are modified through anonymization process to have identical QID values, the anonymized-data will be less distorted. However, clustering the datasets containing both numerical and categorical attributes is quite challenging mainly due to the presence of categorical attributes. Therefore in SBC model a new similarity measurement function and distance calculation based on the measured similarity for categorical attributes is introduced so the total distance between tuples (data records) with numerical and categorical attributes can be calculated for clustering. This approach does not depend on hierarchical taxonomies regarding categorical attributes. Based on the proposed model a bottom-up greedy algorithm for k-anonymization is suggested. In order to evaluate our Similarity-Based Clustering model the proposed algorithm is simulated on two different real datasets with around 5000 data records of individuals (Adult dataset and N. Corporation ISP Dataset). Our extensive study on information loss and data utility show that the proposed algorithm based on SBC model in comparison with existing well-known algorithms offers much higher data utility and reduces the information loss significantly within the wide range of various k values.

1. Chapter 1: Introduction

1.1. Summary of Chapter

In this chapter an introduction of main topics which are going to be discussed in this work has been given. Regarding privacy of individuals, Act on Protection of Personal Information (APPI) is introduced. The privacy issues and concerns in data mining applications, data publishing and data sharing are explained and elaborated in details.

Privacy Preserving Data Mining (PPDM) concept is introduced. In addition the relevant reasons of why such a concept is necessary is explained as well. K-anonymity model as a general solution to protect privacy of individuals is introduced and the current challenges in this model is indicated and elaborated.

Finally the motivation and objectives on this work is stated and explained in details. The organization of this thesis is also explained briefly at the end of this chapter.

1.2. Introduction to Data Mining and Privacy Issues

Data mining is relatively new and interdisciplinary field of computer science and it is regarded as the process of discovering new and insightful patterns from large datasets [1] [2] [52] . In recent years by growing the amount of data in databases, cyber world (e.g., social network, online shopping, online banking and advertising) data mining has become a significant technology for getting and extracting useful and handy information from huge quantities of data [1] [2] [5] [52] .

Although data mining is still at the early stage of growth and development, it has been using in various fields such as science, engineering, education, healthcare [53] , medicine, genetics, bioinformatics and business [1] [5] [52] .

Even though the goal of most data mining approaches is to develop generalized knowledge rather than identify information about specific individuals, but the existence of comprehensive and accurate datasets brings up privacy issues regardless of their intended use. An example of such datasets and privacy issues will be discussed later in this chapter.

Nowadays in our digitalized social life, individuals leave lots of electronical trails through their daily activities such as using electronic cards to use public transportation, credit cards for shopping and booking hotels, using different applications and services on their smart devices or even emails for communication [5] .

Based on Act on Protection of Personal Information (APPI) Personal information should be collected with the consent and permission of the individuals who actually are providing the information and they are in fact the data subjects [54] . The data collectors (e.g., credit card companies, hospitals and service providers) should provide some assurance that the individual privacy will be protected based on Act on Protection of Personal Information (APPI).

However, in real life the data collectors use the collected data for some secondary purposes which means using the collected data in any other ways or for any other purposes that the data were collected initially. Moreover it is also very common practice that data collectors sell the collected data to other organizations and entities which utilize the purchased data for their own purposes. These kind of personal information utilizations increases the privacy concerns of individuals [5] .

Therefore it can be concluded that data mining and data privacy are in disagreement with each other and in fact by having more accurate and complete data the data mining results will become better [45] . In order to exercise data mining while protecting the privacy of individuals, Privacy-Preserving Data Mining (PPDM) was proposed [1] [5] [31] [32] [41] [42] [43] .

1.3. Privacy Preserving Data Mining (PPDM)

It has been proven that data mining is so crucial and beneficial for organizations, yet high public concerns regarding individual privacy has actually made the implementation of privacy preserving data mining techniques to become a demand at the moment. A privacy preserving data mining provides individual privacy while allowing extraction of useful knowledge from data. In another word, privacy preserving data mining techniques allow researcher and data miner to extract the useful information while protecting the privacy of individuals [5] [31] [32] [41] [42] .

There are various different methods, techniques and models that is employed to enable privacy preserving data mining. One particular way of such techniques modifies the collected dataset before it is released to protect individual records from being re-identified. When a dataset has been modified and then released an intruder or third party user cannot be very sure and certain about the correctness of re-identification even by having additional knowledge. This way of privacy preserving data mining techniques relies on the fact that the datasets which are used for data mining purposes do not necessarily need to contain 100% accurate data. In the context of data mining it is very important to maintain the patterns in the dataset. Additionally, maintenance of statistical parameters such as means, variances and covariance of attributes is important in the context of statistical databases [1-7].

There are two main factors that a useful privacy preserving technique requires to satisfy, one is high data quality and the other one is high privacy and security. Therefore, we need to evaluate the data quality and the degree of privacy of a modified dataset. The information loss or data quality of a modified dataset can be evaluated through a few quality indicators such as extent to which the original patterns are preserved, and maintenance of statistical parameters. There is no single agreed upon definition of privacy. Therefore, measuring privacy and security is a challenging task [1-7].

1.4. Motivation

As it was mentioned earlier nowadays data mining is widely used in most of organizations and they are extremely dependent on data mining in their daily routine activities. During the whole process of data mining, from collection of data to discovery of knowledge, these data, which usually contain private sensitive individual information, such as medical background and detailed financial information, often shared and get exposed to several parties including data collectors, data owners, data users and finally researcher and data miners. Disclosure of such sensitive information among various parties raises privacy concerns and also can cause a breach of individual privacy [1-7]. For instance, the detailed credit card record of an individual can expose the private life style with sufficient precision. Private sensitive information can also be disclosed by linking multiple databases and accessing web log data. A malicious data miner can learn sensitive attribute values such as income and disease type of a certain individual, through re-identification of the record from an exposed dataset [1-7].

Simply removing the names and other identifiers (such as social security number, driver license number and passport number) does not guarantee the confidentiality of individual records. Because of particular individual record can often be uniquely identified from the combination of other attributes in datasets [3] .

An example on combination of attributes, which are shared between two different datasets, is shown in Figure 1-1. In Medical dataset there are plenty of sensitive information about individuals from the date and time of doctor visits to medication, disease and insurance coverage data. On the other hand in voter list dataset there are information about the political party, name and address.

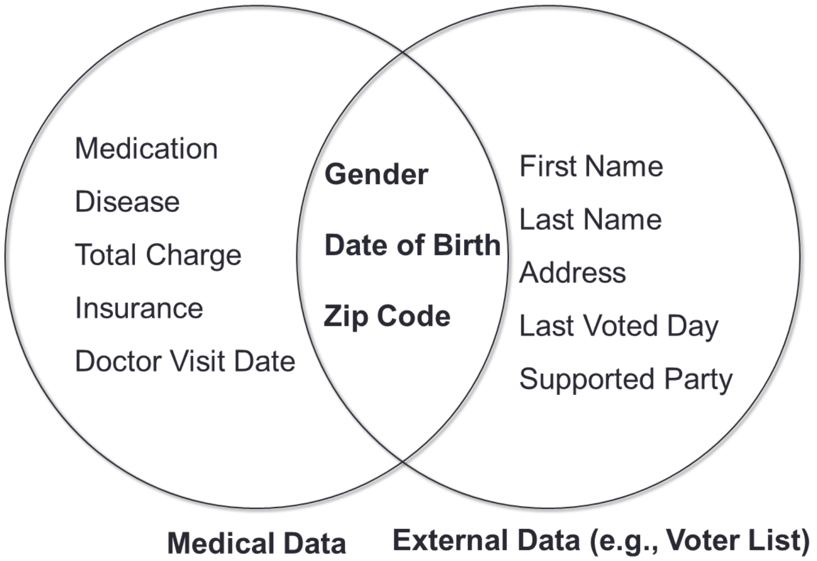


Figure 1-1: Common Attributes between Medical Dataset and Voter Registration List Linking Datasets to Re-Identify [3]

These two datasets are linked together using three attributes, which are in common between them. Attributes are {Gender, Date of Birth and Zip Code}. Therefore it is not really difficult for a malicious data miner or an intruder to re-identify a record from a dataset if sufficient additional knowledge about an individual is provided [3] [4] [5] [6] .

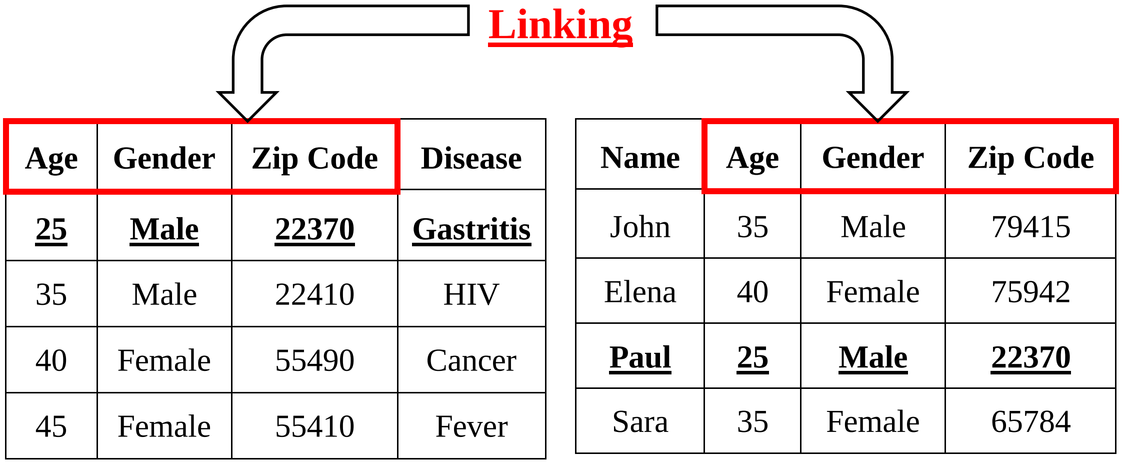
1.5. Privacy Concerns in Data Publishing

In real world, there are lots of government agencies and organizations such as hospitals that collect and store huge amount of information containing personal information of individuals. The individual level collected data, which is called micro-data, contains quasi-identifier attributes and private sensitive attributes. Quasi-identifier attributes such as Age, Zip code and Gender are type of attributes, which are used to identify an individual. On the other hand private sensitive attributes, for example Disease name, are type of attributes, which are not shared with public or strangers by individuals. Information sharing and data publishing has a long history in information technology and due to the regulations, mutual benefits or for some other reasons such as research and education purposes there is a demand for sharing the collected information among various parties [35] [48] [49] [50] .

Publishing or sharing the collected micro-data by the hospital for research and education purposes would be very helpful and interesting for researchers and data miners to investigate the correlation between different attributes such as the relation between a certain disease and gender, the relation between the area of living and certain type of disease and so on. However, publishing the collected data containing private sensitive information would bring up some privacy concerns. Even though the identifying information such as name and social security number (ID number) are discarded before releasing the data, disclosing the private sensitive information of individuals and re-identifying them uniquely is still very much possible due to the existence of quasi-identifier attributes in the released dataset [3] [30] [48] [49] [50] .

Based on the previous research and study on US population in [4] , disclosing one’s gender, Zip code and full date of birth allows for unique identification of 63% of the US population. This study clearly shows the high possibility of re-identification, the importance of quasi-identifier attributes in data sharing and eventually the main reason of privacy concerns in data publishing.

The common quasi-identifier attributes between the released dataset and other existing datasets like Voter Registration dataset, which are accessible to everyone through Internet, can be used to establish a link by matching records from the released dataset to other records in the Voter Registration dataset, which have the same values. This established link between these two datasets could result in identifying the individuals uniquely and disclosure of their private sensitive information. Technically this is known as “linking attack” [3] [33] [48] [49] [50] .



Age	Gender	Zip Code	Disease
25	**Male**	**22370**	**Gastritis**
35	Male	22410	HIV
40	Female	55490	Cancer
45	Female	55410	Fever

(a)Patient Dataset

Name	Age	Gender	Zip Code
John	35	Male	79415
Elena	40	Female	75942
Paul	**25**	**Male**	**22370**
Sara	35	Female	65784

(b)Voter Registration Dataset

Figure 1-2: Sample of Linking Attack between the Released Dataset (a) and External Dataset (b) through shared attributes between two datasets

A sample of linking attack between the Patient Dataset released by a hospital and Voter Registration Dataset as an external dataset is shown in Figure 1-2. In this linking attack, the privacy of Paul is violated because his disease is disclosed and we know Paul has Gastritis [48] [49] [50] .

1.6. K-anonymity Model as The General Solution

In order to protect the privacy of individuals against the possible re-identification through linking attacks explained in previous section, k-anonymity model was proposed by Samarati and Sweeney as an approach towards privacy preserving data mining [1] [4] [6] . K-anonymity definition stated as follows:

K-anonymity suggests “For every record in a released dataset there should be at least k-1 other records identical to it along the quasi-identifier attributes”.

K-anonymity model is widely used for privacy preservation in data publishing and information sharing. A method of k-anonymization suggests to modify the values of quasi-identifier attributes through generalization so that each record in the dataset is indistinguishable from at least k-1 other records within the same dataset [34] [68] .

By applying this method on the released dataset, the linking confidence between the released dataset and the external dataset will reduce by $^1/_k$ ratio, which means the privacy of individuals is protected to some extent. Clearly by having larger k value the privacy protection is higher. The effect of applying k-anonymity model on the Patient dataset in Figure 1-2 is illustrated in Table 1-1.

Table 1-1: The Effect of K-anonymity Model

Equivalent Class	**Age**	**Sex**	**Zip code**	**Disease**
E1	25 ~ 35	Male	22***	Gastritis
	25 ~ 35	Male	22***	HIV
E2	40 ~ 45	Female	554**	Cancer
	40 ~ 45	Female	554**	Fever

(a) 2-Anonymous Patient Dataset

Name	**Age**	**Sex**	**Zip code**
John	35	Male	79415
Elena	40	Female	75942
Paul	25	Male	22370
Sara	35	Female	65784

(b) Voter Registration Dataset

As it is shown in Table 1-1, having the Patient Dataset anonymized with privacy degree of k=2, the linking confidence between 2-Anonymous Patient Dataset and Voter Registration Dataset is reduced by the ration of $^1/_2$. Therefore the exact identification of Paul as an individual and his specific disease is not simply possible. It can be concluded that the linking confidence is reduced the ration of $^1/_2$ and the privacy of individuals is somewhat protected. At the same time it is realized that the values are distorted and the anonymized data is less accurate than original dataset. This matter, data privacy and data utility is discussed in the next subsection [48] [49] [50] .

1.7. Data Privacy and Data Utility Basic Definition

In anonymization there are two terms, which will be used a lot in this thesis, the first one is data privacy. Privacy itself is very difficult to be defined, as privacy meaning is actually different from person to person. The second term is data utility, which actually defines the usefulness of data and its originality. Considering the Patient Dataset in the previous section it can be said that the original dataset has 100% utility and 0% privacy. On the other hand 2-anonymous dataset has privacy to some extent (more than 0%) but the utility of dataset decreased due to the anonymization (less than 100%) [44] .

Considering the original dataset T which contains the information on each individual in n attributes {A1... An} the main terminologies are defined as below.

Quasi-Identifier attributes: set of attributes in dataset T that can potentially join with external datasets to reveal private information of individuals. For example Age, Gender and Zip Code attributes in Figure 1-2 are quasi- identifiers, which can establish a link between Patient dataset and Voter Registration dataset.

Equivalent class: An equivalent class E of dataset T is a set of all tuples in T containing identical values with respect to QID attributes. For instant T1 (tuple 1) and T2 (tuple 2) in Table 1-1 form an equivalent class (E1) with respect to attributes Age, Gender and Zip Code.

K-anonymity: A dataset T is said to be k- anonymous with respect to the QID attributes if the size of every equivalent class is greater or equal to pre-defined k value.

1.8. Challenges in K-anonymity Model

There are several different methods for anonymizing a dataset which will be mentioned in chapter two. However regarding k-anonymity model there are two main methods which are very well used, Generalization and Suppression on quasi-identifier attributes (QIDs) [1] [3] [6] .

Generally K-anonymity is achieved through 1) Generalization and 2) Suppression methods. The generalization method itself can be further subcategorized as follows.

1. Global Recoding
2. Local Recoding

The Generalization method essentially is modifying the value of the data record into more generalized form and suppression is basically removing the data record from the published dataset [1] [3] [6] [16] [17] [22] .

By generalizing the original data record to the generalized form in k-anonymous dataset or suppressing the original data record, some information loss occurs. Information loss in k-anonymity model is an unfortunate and inevitable consequence. This information loss reduces the utility of anonymized-data and makes the anonymized-data to be less accurate and accordingly less useful for data mining and research purposes.

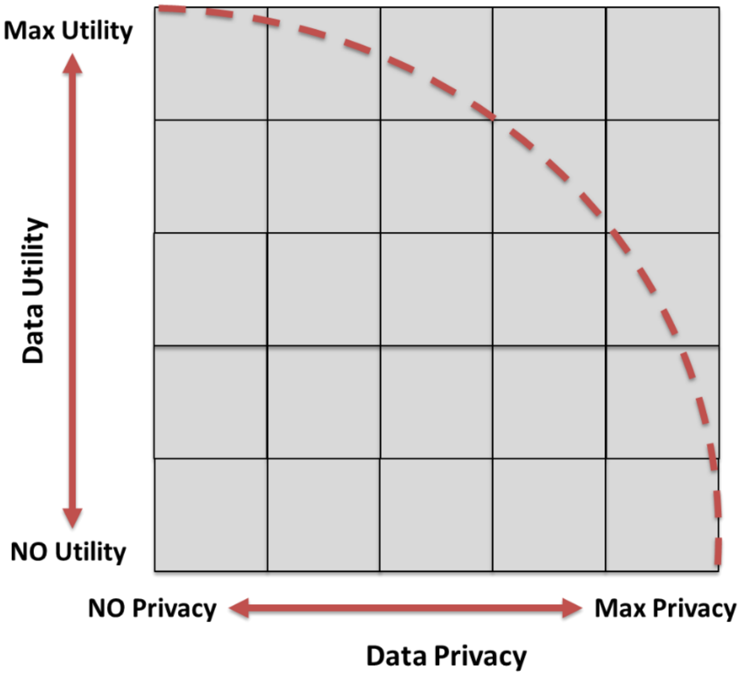


Figure 1-3: Tradeoff Relation between Data Privacy and Data Utility

Therefore one of the main challenges in k-anonymization is to minimize the information loss while the privacy of individuals is protected, so high utility anonymized-data can be obtained.

Privacy protection level in k-anonymity could be measured by “k” value. As it was mentioned in section 1.6, minimum value for “k” value is k=2. If k=1 it means the dataset is original dataset and it is not anonymized. Regarding maximum “k” value, it depends on number of records in dataset. If k is equal to number of records in dataset (k >= n/2 and k <= n) that means there is only one group and all values across all quasi-identifiers are identical to each other. The data publisher usually chooses the anonymization degree or the k value in k-anonymity model based on the desired level of privacy.

In addition, as shown in Figure 1-3, there is a tradeoff relationship between the privacy level and the quality of anonymized-data and due to this tradeoff, performing anonymization with maximum privacy and attaining maximum utility for anonymized-data is not possible [55] .

Furthermore on challenges in k-anonymization, real world and census datasets contain both numerical and categorical type data. As a matter of fact according to Leon Willenborg and Ton de Waal, one of the author of Elements of Statistical Disclosure Control, most of the quasi-identifier attributes in micro-data are assumed to be categorical (nominal) [23] , which by itself does not have hierarchical taxonomy or generalization hierarchy. Therefore it can be concluded that categorical attributes play a very important role in actual real life datasets.

The combination of numerical and categorical attributes in dataset makes anonymization process rather complicated and very often results in an inefficient anonymization with very high information loss and low data utility. Most of the previous approaches and techniques to achieve k-anonymity, which will be reviewed carefully in the next chapter, suffer from huge information loss and very low data utility (quality of anonymized-data). Also most of the approaches are mainly designed for continuous numerical attributes and in case of considering categorical attributes, they depend on hierarchical taxonomies or require some additional information, which are mostly not defined or available in real life applications [20] [48] [49] [50] .

Moreover, the optimal anonymization and optimal selection of k value is shown to be NP hard problem [7] [8] [9] . Therefore, one of the possible approaches to solve high information loss problem in k-anonymization could be through heuristic algorithms [10] [11] [12] [71] .

1.9. Main Objectives

In the previous sections the importance of privacy protection of individuals in data publishing and information sharing was pointed out. Also privacy concerns due to the linking attack were explained and elaborated with an example. In order to protect the privacy of individual a model, which is widely used and called k-anonymity, was introduced. Furthermore the challenges in k-anonymity model were elaborated from our point of view.

Therefore, the main objectives of this research can be summarized and stated as here under:

I. To study and understand the concept of Privacy Preserving Data Mining (PPDM) and k-anonymity model as the widely used model for privacy preservation in data publishing and information sharing.

II. To investigate and study the current challenges in k-anonymization and possible solution methods.

III. To propose and define clustering approach in k-anonymity model.

IV. To propose a new model based on clustering which implements k-anonymity and at the same time minimizes the information loss and maximizes the utility of anonymized data.

V. To evaluate the proposed model using real dataset in practice.

1.10. Contributions

In this work the privacy issues in data publishing and information sharing is studied. K-anonymity model as a general solution to overcome the privacy violation of individuals issue in data publishing and data sharing is introduced. Particularly the current challenges in k-anonymization were identified and studied in depth are stated as follows.

1) Huge information loss issue in k-anonymization due to generalization and suppression as an unfortunate and inevitable consequence
2) Tradeoff relationship between data privacy and data utility
3) Real life datasets most likely consist of a combination of numerical and categorical attributes.

Some of the main information quality metrics in this field are studied. Most importantly, main previous works and existing well-known models and algorithms, which are implementing k-anonymization, are reviewed in detail.

By considering the drawbacks of the previous works and current challenges in k-anonymity model, a new model, which is based on clustering, to achieve k-anonymity with high data utility is proposed. In the proposed model k-anonymity is defined and viewed from clustering point of view. Since datasets containing numerical and categorical attributes is the core and heart of this work attention, a new similarity and distance measurement between the variables in categorical attributes for clustering purposes, which will be employed in anonymization, is introduced.

Also a bottom-up greedy algorithm is introduced based on the proposed clustering model and finally the proposed model and algorithm is evaluated regarding the information loss and data utility. The results are compared with other existing well-known algorithms which proves the proposed model increases the utility of anonymized data and reduces the information loss.

1.11. Organization of Thesis

This thesis consists of 8 chapters in total. In chapter 1 of this thesis, data mining and its relation to privacy of individuals is explained. The benefits of data mining and the privacy concerns which data mining and data sharing can cause is elaborated and explained in details. Privacy preserving data mining and its benefits are introduced. Also the technical term “Linking Attack” is introduced and explained with an example which shows exactly how one’s privacy could get violated. K-anonymity model as a general solution and its effect is introduced. Moreover the current challenges in k-anonymization are explained. Also motivation, objectives and contributions of this work are clearly stated in this chapter.

Chapter 2 of this thesis is more focused on related works and background study. In this chapter, we have defined the terms related to anonymization. Different data and attribute types are explained. As a background study, various anonymization and disclosure protection techniques are mentioned. K-anonymity model and related techniques are fully explained and elaborated. We have also reviewed Mondrian, Datafly and Incognito, which are the main models implementing k-anonymity, and explained them in details with example. Their drawbacks are also mentioned in this chapter. At the end some of the data quality metrics and information loss measurements are introduced. In this work for measuring information loss Normalized Certainty Penalty (NCP) methods has been chosen.

In chapter 3, we introduce our model, which is called Similarity-Based Clustering model. The basics of clustering, clustering in k-anonymization and how it is employed in anonymization to reduce information loss and increase data utility is explained. The proposed model clusters the dataset based on measured similarity over all quasi identifier attributes. The similarity is measured based on calculated distances over categorical and numerical attributes. Particularly distances over categorical attributes is defined based on the context and observation probability of values in each categorical attribute. Based on the Similarity-Based Clustering model a bottom-up heuristic algorithm is presented.

Chapter 4 is dedicated to empirical evaluation and results. The proposed model which was explained in chapter 3 is simulated on two different real datasets. The proposed model and other well-known algorithms were compared to each other with respect to information loss and data utility.

Finally in chapter 5 this work is concluded and the future works is discussed. Chapter 6 is on main publications of the author and chapters 7 and 8 are appendix and reference materials which have been used to write this thesis.

2. Chapter 2: Background Study and Related Works

2.1. Summary of Chapter

This chapter is dedicated to literature review and background study. At first an introduction is given regarding different types of data and its classifications. Particularly the difference between numerical and categorical attributes is elaborated.

Attributes classification from anonymization point of view is also explained with an example. It is stated that from this point of view the attributes are mainly categorized in 2 groups. First category is called identifying attributes and second type is private sensitive attributes.

Moreover, different anonymization techniques are explained. Then k-anonymity model and its implementation through generalization and suppression is explained. After the fundamental definitions, very well-known algorithms which implement k-anonymity model are explained with examples and results are analyzed. The drawbacks on each model is also discussed briefly in this section.

Finally, the main two terms, data utility and information loss in k-anonymity model is discussed. Data utility and basically calculation and measurement of information loss and the quality of anonymization in addition to data utility after anonymization is very crucial. Different calculation and measurement methods are explained in the last section. At the end, the information loss measurement and data utility which is used in this work is explained in details with an example.

2.2. Data Types and Classification

Typically variables and attributes in real datasets contain different type of data. In this section an overview of various attributes and data types is presented. Two of the most common type of data can be classified as follows [23] [52] .

1) Numerical data (continuous)

2) Categorical data (nominal)

Boolean data are a special case of categorical data, which can take only two possible values, 0 or 1 (true or false). Gender attribute could be a very good example of Boolean data type. Because an individual can be either “Male” or “Female”.

Numerical data are formed by continuous digits, which basically means numbers represents them and different kind of math and calculations can be performed on them. For example the Age attribute is a numerical variable that takes numbers. Income, Profit and Turnover are other example of numeric attributes [23] [52] .

However categorical values lack natural ordering in them. For example, an attribute “Education” can have values such as “High School”, “Bachelor Degree”, “Master Degree”, and “PhD Degree”. There is no straightforward way of ordering these values. In addition mathematical operation could not be performed on this type of attributes.

The collected data at individual level is called microdata. Microdata is a series of data records in which each data record containing information on an individual [23] [52] . Microdata also can be defined as individual level data which consists of a series of records and each record contains information on an individual as a person, or a firm or an institution. Microdata in their simplest form maybe represented as a single data matrix where the rows correspond to the units and the columns to the variables and attributes. An example of individual level data, microdata, is shown in Table 2-1 as follows.

Real datasets are combination of numerical and categorical attributes. At the same time, the datasets are not perfect. Meaning that there could be unknown values for different attributes. For instance in the data shown in Table 2-1 the sex of the individual with ID number “322” is not known. Same applies for Marital Status for ID number “324”.

Table 2-1: Sample of Collected Microdata

ID	Age	Sex	Marital Status	Education	Income
321	45	Male	Married	Bachelor Degree	100,000
322	65	-?-	Divorced	Bachelor Degree	65,000
323	57	Female	Married	Master Degree	50,000
324	41	Male	-?-	Master Degree	120,000
325	34	Male	Not Married	PhD Degree	100,000

The fundamental difference between categorical and numerical attributes forces the privacy protection techniques to take different approaches. This topic will be discussed deeper in chapter 3 when the proposed model is presented.

In categorical attributes, sometimes it is possible to define a tree-like structure that defines the relation between various values in that categorical attributes. The tree-like structure is called hierarchical structure. The importance of such structure is when applying anonymization. It is also important to note that the hierarchical structure for categorical attributes do not always exist. More often than not the hierarchical taxonomies are not defined for all categorical attributes in a dataset [23] .

A sample of the dataset including categorical and numerical attributes along with a sample of hierarchical taxonomy of categorical attribute (e.g., Education) is shown in the Figure 2-1.

As it was briefly explained in the first chapter, looking at attribute and variable types from another aspect, which is privacy and disclosure risk perspective, the attributes could be divided into two types as follows.

1) Identifying Attributes (Quasi-identifier Attributes)
2) Sensitive Attributes (Private Sensitive Attributes)

Age	Zip Code	Gender	Education	Salary (USD/Year)
30	1430020	Male	PhD	$ 80,000
21	1570012	Female	Bachelor Degree	$ 50,000
32	1430025	Female	PhD	$ 96,000
25	1570121	Female	Master Degree	$ 78,500
51	1570001	Male	High School	$ 44,000
33	1440120	Male	Master Degree	$ 65,000

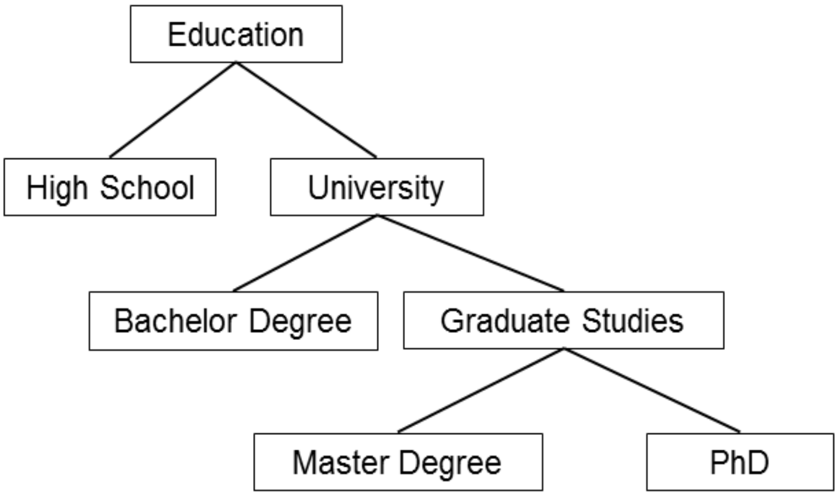


Figure 2-1: Sample of a Dataset with Numerical and Categorical Attributes along with Education Attribute Hierarchical Taxonomy

First type of attributes is called identifying attributes. This type of variables are used to identify an individual (quasi-identifier attributes). For instance Gender, Date of Birth, Zip Code, address and Phone Number which are very common attributes and probably in any registration form for a survey or service subscription this information are necessary and required.

The second type of attributes is called sensitive attributes. The private sensitive attributes are type of information such as personal income, credit card history or medical background and disease which are not usually shared with public or strangers. Obviously there can be some situations, which are exception, such as when a disaster strikes or when ones disease is very rare therefore the information are shared with various parties for further investigation or assistance.

In anonymization models, having a combination of numerical and categorical attributes in the real datasets bring up some difficulties in order to have an efficient anonymization process. Mainly because the numerical data are the type of data that the arithmetic operations are defined for them however regarding the categorical attributes the arithmetic operations are not defined and they are not applicable in the same way.

In fact most of the identifying data and microdata are assumed to be categorical. Therefore as the datasets in real life are consist of both types of data then having a model which can operate efficiently for both types of data is necessary and going to be very useful in real life applications [23] .

2.3. Common Anonymization Techniques

Generally in anonymization in order to protect privacy of individuals in microdata there are several methods which all anonymize the data though data modification [43] [23] . Privacy preserving techniques can be classified based on the protection methods used by them. The classification is shown in Figure 2-2 and explained briefly with an example in this section. Noise addition usually adds a random number (random number and noise are same) to numerical attributes. This random number is generally drawn from a normal distribution with zero mean and a small standard deviation. Noise is added in a controlled way so as to maintain means, variances and co-variances of the attributes of a data set. However, noise addition to categorical attributes is not as straightforward as the noise addition to numerical attributes, due to the absence of natural ordering in categorical values [5] [23] [56] [57] [58] .

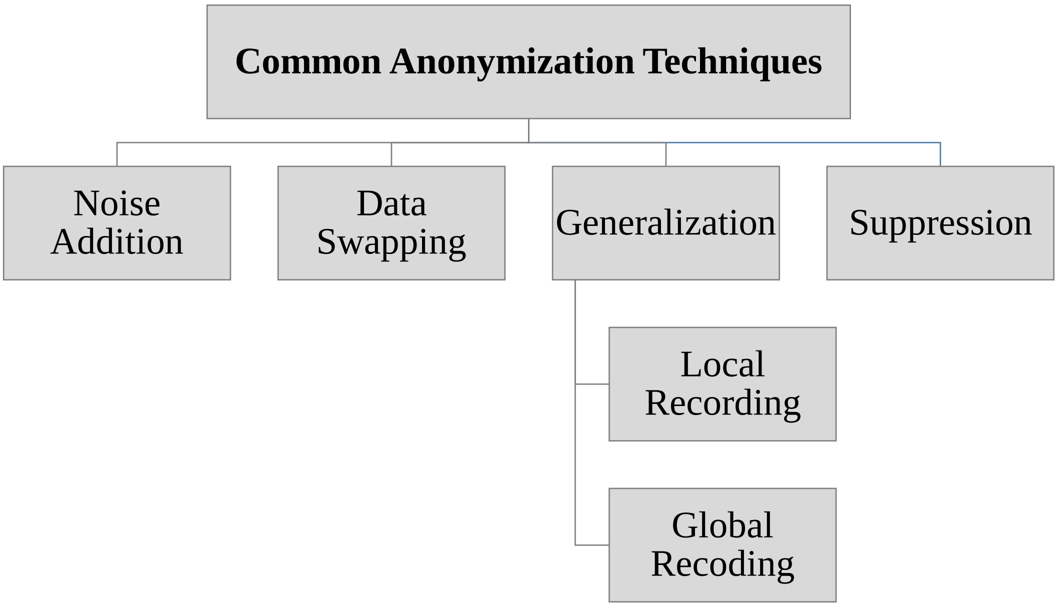


Figure 2-2: Common Anonymization Techniques

Data swapping interchanges the attribute values among different records. Similar attribute values are interchanged with higher probability. All original values are kept within the dataset and just the positions are swapped. Data swapping is often explained as a special case of noise addition. The reason is because swapping two numerical values can be seen as the addition of a number (the difference between the values) to the smaller value, and subtraction of the same number from the larger value. Therefore, data swapping results in the addition of noise having zero mean. Similar explanation can be given for swapping categorical values [5] [23] [56] [57] [58] .

Generalization refers to both combining a few attribute values into one, or grouping a few records together and replacing them with a group representative for numerical and categorical attributes. Regarding categorical attributes depending on the generalization type, which will be explained in details in later section, generalization hierarchy is necessary (e.g., generalization hierarchy for Education attribute shown in Figure 2-1) [22] [23] [48] [49] [50] [66] .

Finally, Suppression means replacing an attribute value in one or more records by a missing value. Many such techniques for different scenarios have already been proposed [1] [3] [4] . It is unlikely to have a single privacy preserving technique that outperforms all other existing techniques in all aspects. Each technique has its strength and weakness. Hence, a comprehensive evaluation of a privacy preserving technique is crucial. It is important to determine the evaluation criteria and related benchmarks.

2.4. K-anonymity Model through Generalization and Suppression

K-anonymity model was defined in Chapter 1 and explained briefly. In this section the details on k-anonymity model will be elaborated. Specifically the type of anonymization methods, which are utilized in k-anonymity model, will be explained in details with the help of some examples.

Technically from our point of view the operation in k-anonymity model can be divided into two different steps. The first operation is to cluster the data records into groups with a minimum group size of k. This will be explained more in details in the third chapter when k-anonymity is defined from clustering point of view. Afterwards, the second operation is the process of anonymization using anonymization methods mentioned earlier and shown in Figure 2-2.

The two main methods which are utilized in k-anonymity to anonymize the quasi-identifier attributes are as follows [1] [3] [43] .

1) Generalization
2) Suppression

Both methods are technically recoding the values of quasi-identifier attributes in original dataset. Suppression can be defined as specific type of recoding in which the values of data record in original dataset is recoded to null values [3] [4] [16] [17] .

In generalization the original values of quasi-identifier attributes are replaced by intervals for numerical attributes. Regarding categorical attributes if generalization hierarchy (taxonomy tree) is provided the original values are replaced by the more general value according to the provided generalization taxonomy. If generalization taxonomies for categorical attributes are not defined the original values are replaced by set of distinct values.

For instance in attribute {Age}, the value 23 could be replaced by [20~25] and for attribute {Gender} with its corresponding generalization hierarchy shown in Figure 2-3, [“Male”] could be replaced by [“Person”].

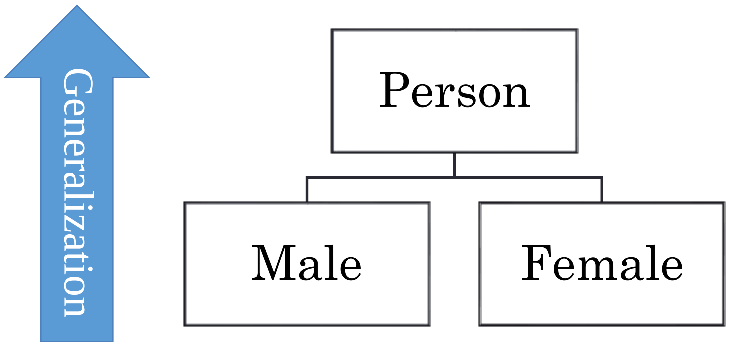


Figure 2-3: Gender (Sex) Attribute Generalization Hierarchy

The generalization method in anonymization can be further divided into three different types as follows.

1) Global recoding generalization
2) Multidimensional recoding generalization
3) Local recoding generalization

The differences between these three types will be explained using proper examples. In global recoding generalization, the dataset is generalized at the domain level. There are many works, which are based on global recoding generalization such as [3] [13] [14] [17] [19] [35] . In global recoding generalization if a lower level domain needs to be generalized to the higher domain, all the values in the lower level domain are generalized to the higher domain. An example of original dataset with the total of 104 data records with 2 attributes (Att1 and Att2) and its 2-dimentional representation is shown in Table 2-2.

Table 2-2: Original Dataset with Its 2-dimensional Representation

Att1	Att2	Frequency
A	X	1
A	Y	11
B	X	7
B	Y	4
B	Z	5
C	X	20
C	Y	35
D	Y	21

Original Dataset

	X	**Y**	**Z**
A	**1**	11	0
B	7	**4**	5
C	20	35	0
D	0	21	0

2-Dimentional representation of Original Dataset

If the dataset which is shown in Table 2-2 is anonymized through global recoding with k=5 as the k-anonymity condition, all the data records with value “X” and value “Y” will be replaced by a generalized value as “(X, Y)”. Because k value condition is k=5 and data records with values (A, X) and (B, Y) do not satisfy the k value condition. The result of anonymization through global recoding is illustrated in Table 2-3.

Table 2-3: Global Recoding Generalization of Original Dataset Shown in Table 2-2

Att1	Att2	Frequency
A	(X, Y)	12
B	(X, Y)	11
B	Z	5
C	(X, Y)	55
D	(X, Y)	21

5-Anonymous Dataset

	(X, Y)	**Z**
A	12	0
B	11	5
C	55	0
D	21	0

2-Dimentional representation of 5-Anonymous Dataset

This particular effect is called over generalization, which will cause more information loss and reduces the anonymized data utility. Over generalization will be investigated in chapter 4 where the proposed model is discussed and compared with other well-known models which implement k-anonymity.

One of the global generalization methods is Incognito [15] . Incognito produces minimal full domain generalizations with an optional tuple suppression threshold. This model will be reviewed with an example in the next session.

In multidimensional and local recoding generalization, the generalization is taking place at cell levels [8] [9] [11] [12] . They do not cause overgeneralization or reduce the effect of over generalization, which lead to more flexible generalization and have the potential of less information loss.

Table 2-4: Multidimensional Recoding Generalization of Original Dataset Shown in Table 2-2

Att1	Att2	Frequency
A	(X, Y)	12
B	(X, Y)	11
B	Z	5
C	X	20
C	Y	35
D	Y	21

5-Anonymous Dataset

<table>
<tr><th></th><th>X</th><th>Y</th><th>Z</th></tr>
<tr><td><b>A</b></td><td colspan="2">12</td><td>0</td></tr>
<tr><td><b>B</b></td><td colspan="2">11</td><td>5</td></tr>
<tr><td><b>C</b></td><td>20</td><td>35</td><td>0</td></tr>
<tr><td><b>D</b></td><td>0</td><td>21</td><td>0</td></tr>
</table>

2-Dimentional representation of 5-Anonymous Dataset

As it is shown in Table 2-4 the unnecessary generalization regarding data records with value of (C, X) or (D, Y) is not taking place in multidimensional generalization as both Att1 and Att2 dimensions are considered for anonymization. One of the best performing algorithms is Mondrian heuristic algorithm [18] . It studies the single dimension partitioning and suggests an efficient partitioning method for multidimensional recoding anonymization. Mondrian algorithm and its model on multidimensional generalization will be explained in detailed in the next section.

Table 2-5: Local Recoding Generalization of Original Dataset Shown in Table 2-2

Att1	Att2	Frequency
A	(X, Y)	5
A	Y	7
B	X	6
B	(X, Y)	5
B	Z	5
C	X	20
C	Y	35
D	Y	21

5-Anonymous Dataset

	X	Y	Z
A	5	7	0
B	6	5	5
C	20	35	0
D	0	21	0

2-Dimentional representation of 5-Anonymous Dataset

Eventually, Table 2-5 is representing the generalization through Local recoding generalization. As it is shown only the necessary data records, which do not satisfy the k value condition, will be generalized. In local recoding generalization, attributes are generalized at the cell level. Therefore over generalization does not take place. In local recoding the numerical attributes are generalized into intervals from minimum to maximum (e.g., [20~25]) and categorical attributes are generalized into set of distinct values (e.g., {Malaysia, Japan, China}) or in case generalization hierarchy is defined a single value that represents such a set (e.g., Asia).

The work “Utility-Based Anonymization Using Local Recoding” [20] is based on utility anonymization through local recoding. It introduces the new quality metric for both numerical and categorical attributes. However regarding the categorical data it assumes that the hierarchical structure for each categorical attribute is defined and provided. In most of the real life applications the hierarchical structures often do not exist which makes this approach not so practical.

In our proposed model we consider local recoding generalization, as it is more flexible and efficient with the possibility of lower information loss.

2.5. Related Works

There are several models and strategies, which have been proposed on k-anonymity model. In this section some of the well-known models on k-anonymity, which were mentioned in previous section briefly, will be carefully reviewed.

The first k-anonymization model that will be reviewed is called Datafly, which is presented and reviewed previously [16] . Datafly is one of the very first and famous algorithms in k-anonymization. Datafly utilizes generalization and suppression to achieve k-anonymity. It use heuristics in order to make approximations and it has been shown not to be efficient always [16] [64] .

Datafly algorithm requires original dataset, quasi identifier attributes and the corresponding generalization hierarchies (for all quasi identifier attributes) in addition to k value constraint as an input to function and operate. An example of the original dataset with Date of Birth, Gender, Zip Code and Race as its quasi identifier attributes with their generalization hierarchies is shown in Table 2-6 and Figure 2-4 respectively.

Table 2-6: Original Dataset from Hospital

Tuple	Date of Birth	Gender	Zip Code	Race	Disease
T1	9/20/1995	Female	1141	Asian	Fever
T2	2/14/1995	Female	1141	Asian	Back Pain
T3	10/23/1995	Male	1138	Asian	Chest Pain
T4	8/24/1995	Male	1138	Asian	HIV
T5	11/7/1994	Male	1138	Asian	Painful Eye
T6	12/1/1994	Male	1138	Asian	Headache
T7	10/23/1994	Female	1139	Black	Stomachache
T8	3/15/1965	Male	1139	Black	Brocken Leg
T9	8/13/1964	Female	1139	Black	HIV
T10	5/5/1964	Female	1139	Black	Brocken Hand
T11	2/13/1967	Female	1138	Black	Asthma
T12	3/21/1967	Female	1138	Black	Heart Attack

Table 2-6 presents a sample dataset with twelve tuples on five different attributes.

Date of Birth, Gender, Zip Code and Race are considered to be the quasi identifier attributes and Disease is the only private sensitive attributes in this example.

Among quasi identifier attributes Date of Birth, and Zip Code is assumed to be numeric and for categorical attributes there are Gender and Race attributes. Regardless of the type of quasi identifier attributes, value generalization hierarchies must be prepared and given to the algorithm for anonymization operation.

The generalization hierarchies are defined for every quasi identifier attribute individually. Moreover, generalization hierarchies tend to vary depending on the dataset and the type of quasi identifier attribute.

For instance for "Race" quasi identifier attribute the generalization hierarchy is defined as shown in Figure 2-4. "Asian", "Black" or "White" will be generalized to "Person" in the first step and for the second and last step it will be replaced by "*". The star means that the value is completely removed from the dataset which is known as suppression technique in anonymization.

G2 = {*****}

⬆

G1 = {Person}

⬆

G0 = {Male, Female}

Values Generalization Hierarchies for Gender Attribute

G2 = {*****}

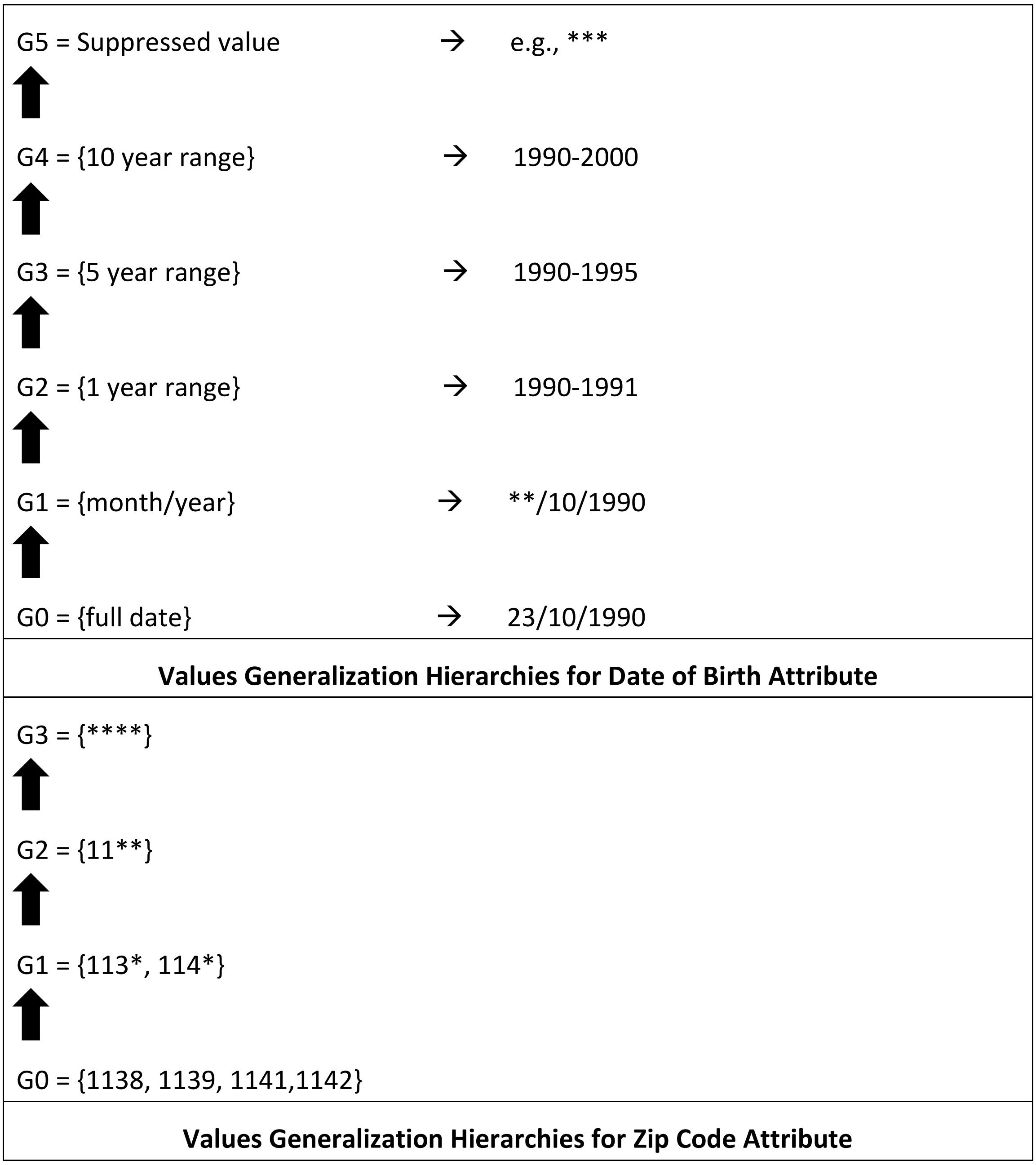
⬆

G1 = {Person}

⬆

G0 = {Asian, Black, White}

Values Generalization Hierarchies for Race Attribute



G5 = Suppressed value → e.g., ***

↑

G4 = {10 year range} → 1990-2000

↑

G3 = {5 year range} → 1990-1995

↑

G2 = {1 year range} → 1990-1991

↑

G1 = {month/year} → **/10/1990

↑

G0 = {full date} → 23/10/1990

Values Generalization Hierarchies for Date of Birth Attribute

G3 = {****}

↑

G2 = {11**}

↑

G1 = {113*, 114*}

↑

G0 = {1138, 1139, 1141,1142}

Values Generalization Hierarchies for Zip Code Attribute

Figure 2-4: Generalization Hierarchies for Gender, Race, Date of Birth and Zip Code Attributes

On the other hand, the generalization hierarchy for “Date of Birth” quasi identifier attribute, which is a numerical type of attribute, is a bit more complicated as the height of the generalization tree is longer. The lowest part of the hierarchy is a full “Date of Birth” and by going higher in generalization hierarchy the “Date of Birth” gets less detailed. The “Day” part is removed at the second level of the hierarchy and the “Year” is going into 1 year, 5 years and 10 years range in the next levels to finally replaced by “*”.

Regarding Datafly anonymization algorithm, the core of algorithm is summarized in steps as follows. In the first step Datafly will create a frequency list based on the original dataset which is going to be anonymized. The frequency list is actually containing distinct sequence of values from the original dataset (cardinality of each quasi-identifier attribute) along with the number of occurrence of each sequence. Each sequence in frequency list represents one or more tuples in the original dataset [16] [65] .

The corresponding frequency list of original dataset which is shown in Table 2-6 is constructed and illustrated in Table 2-7 as follows.

Table 2-7: Constructed Frequency List of Original Dataset

Date of Birth	**Gender**	**Zip Code**	**Race**	**Frequency**	**Tuple**
9/20/1995	Female	1141	Asian	1	T1
2/14/1995	Female	1141	Asian	1	T2
10/23/1995	Male	1138	Asian	1	T3
8/24/1995	Male	1138	Asian	1	T4
11/7/1994	Male	1138	Asian	1	T5
12/1/1994	Male	1138	Asian	1	T6
10/23/1994	Female	1138	Black	1	T7
3/15/1995	Male	1139	Black	1	T8
8/13/1994	Female	1139	Black	1	T9
5/5/1994	Female	1139	Black	1	T10
2/13/1997	Female	1138	Black	1	T11
3/21/1997	Female	1138	Black	1	T12
12	2	3	2		

Frequency List (No.1) of Original Dataset

In the Figure 2-4 the numbers in the last row of the table indicate the cardinality (number of distinct values) of each quasi identifier attributes. Based on the original dataset shown in Figure 12 the cardinality of quasi identifier attributes is {Date of Birth: 12, Gender: 2, Zip Code: 3 and Race: 2}.

In next step the Datafly algorithm uses some heuristics to perform generalization. The quasi identifier attribute which has the highest number of distinct values, quasi identifier attribute with the highest cardinality, in the constructed frequency list is generalized based on the defined generalization hierarchies. The generalization will be continued until there remains “k” or fewer tuples having distinct sequences in frequency list. Regarding frequency list example shown in Table 2-7 in the first constructed frequency list all the tuples frequency is equal to one however by going through the second step and generalizing {Date of Birth} attribute to year of birthday only and updating the frequency list, shown in Table 2-8, some of the tuples are actually merged together and they have identical values along all quasi-identifier attributes.

Table 2-8: Constructed Frequency List of Original Dataset

Date of Birth	Gender	Zip Code	Race	Frequency	Tuple
1995	Female	1141	Asian	2	T1, T2
1995	Male	1138	Asian	2	T3, T4
1994	Male	1138	Asian	2	T5, T6
1994	Female	1138	Black	1	T7
1995	Male	1139	Black	1	T8
1994	Female	1139	Black	2	T9, T10
1997	Female	1138	Black	2	T11, T12
3	2	3	2		

Frequency List (No.2) of Original Dataset

For instance T1 and T2 data records were different from each other on the month and date in {Date of Birth} attribute. By generalizing “Date of Birth” attribute to year of birthday these two tuples are exactly identical along all quasi identifier attributes with values equal to {Date of Birth (Generalized): 1995, Gender: Female, Zip Code: 1141 and Race: Asian}.

The next step which is the 3rd step in Datafly anonymization algorithm, any tuple with frequency less than “k” will be suppressed from the table (frequency list). Complimentary suppression is performed in step 4 so that the number of suppressed tuples satisfies “k” requirement in k-anonymity model. “K” value is assumed to be equal to two (k=2) for this example. So as shown in Table 2-8, T7 and T8 with frequency equal to one will be removed from the dataset, as they do not satisfy “k” value condition. In the final step, k-anonymous table is produced based on frequency list such that the values stored as a sequence in frequency list appear as tuple or tuples in k-anonymous table replicated in accordance to the stored frequency.

Table 2-9: 2-anonymous Dataset Using Datafly

Date of Birth	Gender	Zip Code	Race	Disease
1995	Female	1141	Asian	Fever
1995	Female	1141	Asian	Back Pain
1995	Male	1138	Asian	Chest Pain
1995	Male	1138	Asian	HIV
1994	Male	1138	Asian	Painful Eye
1994	Male	1138	Asian	Headache
1994	Female	1139	Black	HIV
1994	Female	1139	Black	Brocken Hand
1997	Female	1138	Black	Asthma
1997	Female	1138	Black	Heart Attack

The k-anonymous table is also called Minimal Generalization of Table (MGT) [16] . Regarding this example, the 2-anonymous table is presented in Table 2-9 as follows.

The Datafly algorithm always satisfies k-anonymity however it does not necessarily provide k-minimal generalization or distortion. Datafly utilizes heuristics to make approximation therefore it does not always yield optimal results. One of the problems in Datafly is that Datafly makes rough decisions on generalizing all values associated with an attribute and suppressing all values within a tuple. There is also another issue, which is related to the attribute selection for generalization. Datafly select the highest cardinality attribute for generalization, which causes unnecessary and over generalization. This will result in loss of data utility and high information loss in the k-anonymous table which lead to having low quality data with accuracy below expectation.

In addition to Datafly algorithm, another famous model and algorithm that implements k-anonymity model is called Incognito [15] . Incognito word means “with your true identity kept secret” or “with one's identity concealed” and the model produces minimal full-domain generalization [15] .

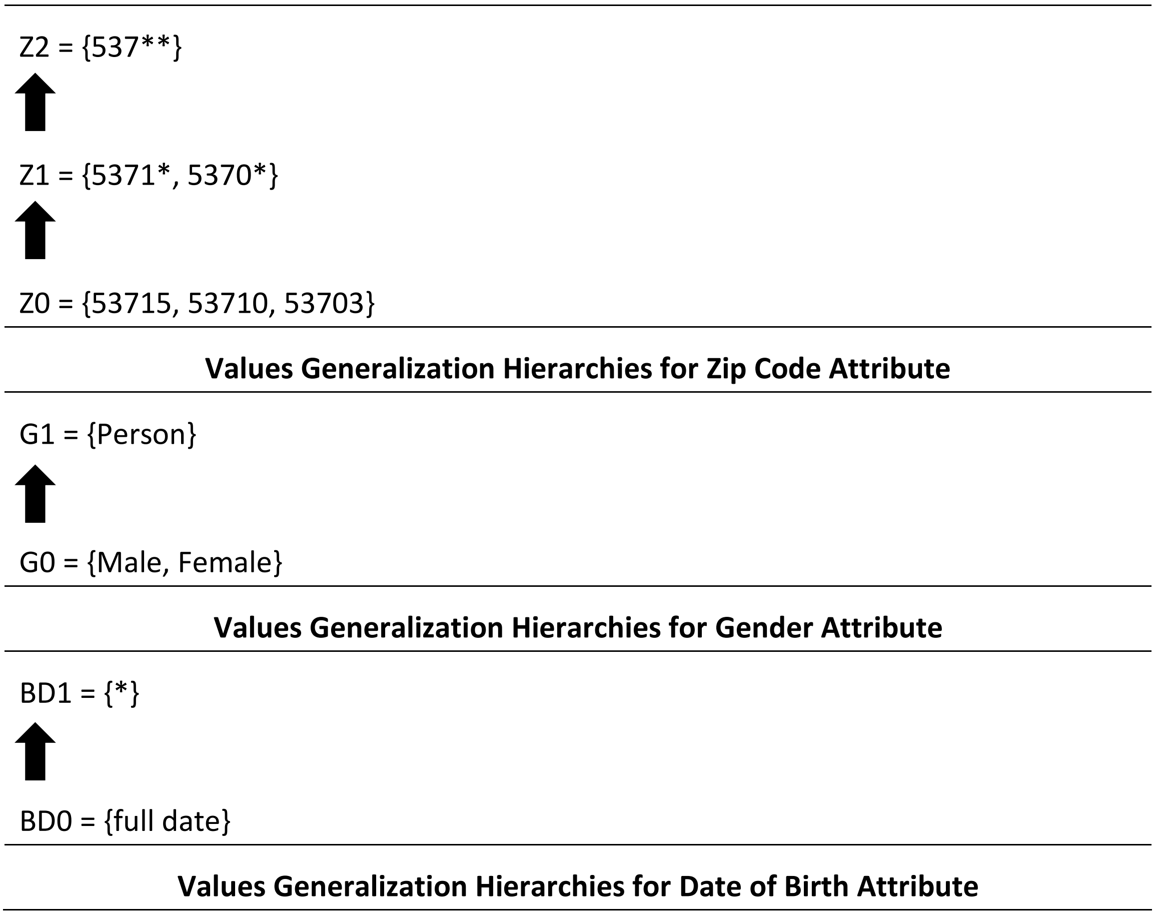
Table 2-10: Original Patient Dataset

Date of Birth	Gender	Zip Code	Disease
1/21/76	Male	53715	HIV
4/13/86	Female	53715	Painful Eye
2/28/76	Male	53703	Headache
1/21/76	Male	53703	HIV
4/13/86	Female	53706	Brocken Hand
2/28/76	Female	53706	Asthma

Incognito generates the set of all possible k-anonymous full-domain generalizations. Full-domain generalization is a specific global recoding model as it was mentioned earlier and it means if a lower level domain needs to be generalized to higher domain, all the values within the lower level domain are going to be generalized to higher domain. For instance if Date of Birth attribute is going to be generalized to year of birthday, all values in Date of Birth attribute or domain are going to be replaced by year of birthday.

The core of Incognito algorithm will be explained through the following example step by step. Consider the original patient dataset shown in Table 2-10 with quasi-identifier attributes of {Gender, Date of Birth and Zip Code} in addition to {Disease} as the private sensitive attribute. The generalization hierarchies of all quasi-identifier attributes are required in Incognito and for this example all generalization hierarchies for quasi identifiers are provided as illustrated in Figure 2-5.

In the first iteration, Incognito discovers that the patient dataset is k-anonymous with respect to every quasi-identifier attribute which in this case are {Gender, Date of Birth and Zip Code} considering “k” value is “k=2”. This actually means that in original dataset, as shown in Figure 2-5, there is at least two or more values which are same in each quasi-identifier attributes.



Z2 = {537**}

Z1 = {5371*, 5370*}

Z0 = {53715, 53710, 53703}

Values Generalization Hierarchies for Zip Code Attribute

G1 = {Person}

G0 = {Male, Female}

Values Generalization Hierarchies for Gender Attribute

BD1 = {*}

BD0 = {full date}

Values Generalization Hierarchies for Date of Birth Attribute

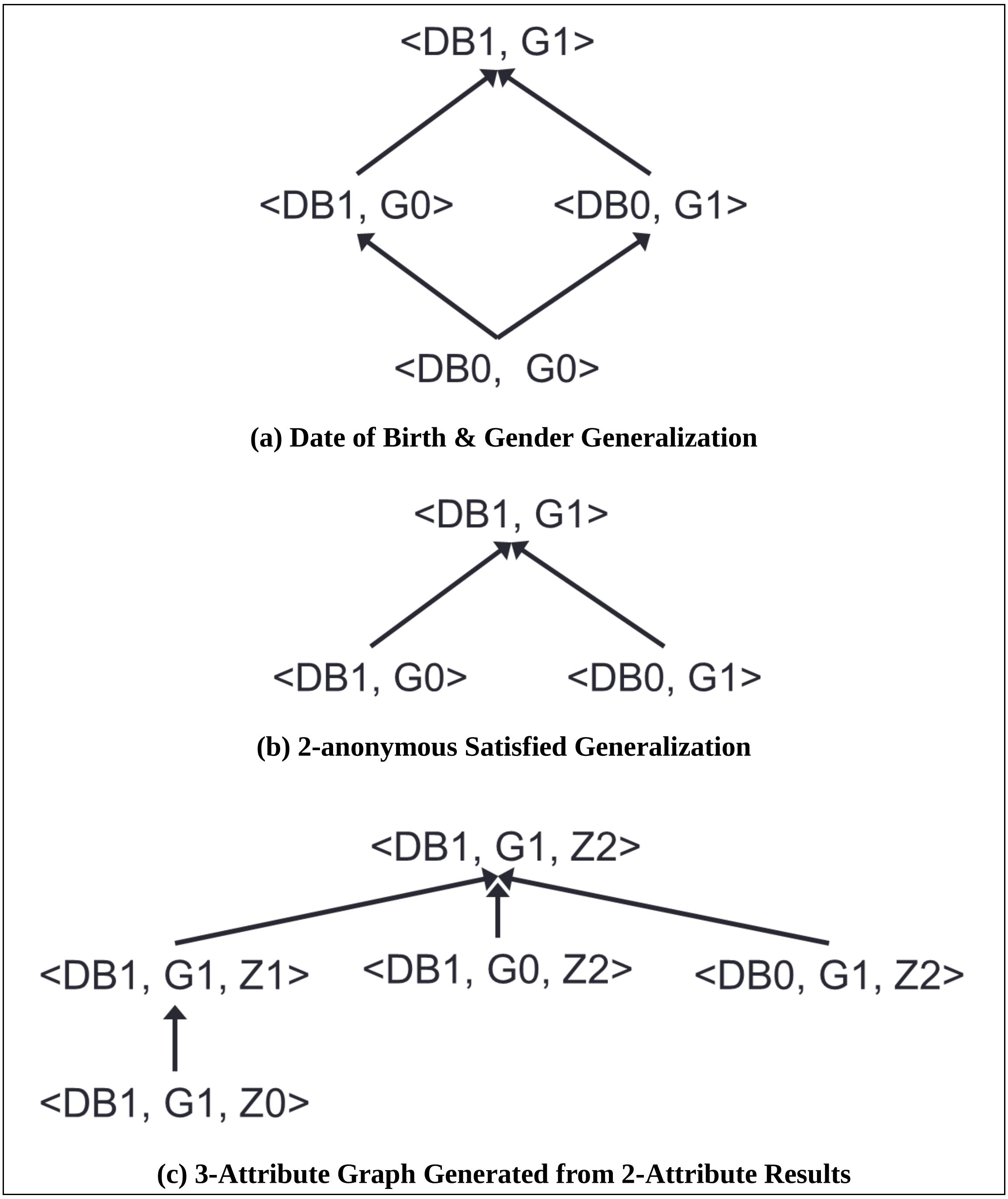
Figure 2-5: Quasi-Identifier Attributes Generalization Hierarchy

Afterwards the generalization lattice for multi-attribute such as <Date of Birth, Gender> are constructed and the second iteration performs Breadth-First searches to the multi-attribute generalization of <Date of Birth, Gender >, < Date of Birth, Zip Code> and <Gender, Zip Code> in order to check whether k-anonymity condition is satisfied or not.

For instance, multi-attribute generalization of <Date of Birth, Gender> is constructed as shown in Figure 2-6(a), Incognito generates query on <DB0, G0>. In Figure 2-6(b) after Breadth-First search, the generalization between Date of Birth and Gender which satisfy "k" value condition when "k=2" is presented.

In the next step the 3-attribute graph is constructed using the second iteration result. The 3-attribute graph generated from 2-attribute results which is shown in Figure 2-6(c). The generalization lattice for all 3 quasi-identifier attributes is constructed and checked against k-anonymity condition considering "k=2".

Among those generalization in 3-attribute generalization lattice the algorithm chooses the generalization with minimum anonymization height which in this case is <DB1, G1, Z0> generalization. The 2-anonymous Patient Dataset is shown in Table 2-11.



(a) Date of Birth & Gender Generalization

(b) 2-anonymous Satisfied Generalization

(c) 3-Attribute Graph Generated from 2-Attribute Results

Figure 2-6: Multi-attribute Generalization of <Date of Birth, Gender>

Table 2-11: 2-Anonymous Patient Dataset Using Incognito

Date of Birth	Gender	Zip Code	Disease
*	Person	53715	HIV
*	Person	53715	Painful Eye
*	Person	53703	Headache
*	Person	53703	HIV
*	Person	53706	Brocken Hand
*	Person	53706	Asthma

Previously explained approaches employ full domain generalization to reach k-anonymity, which tend to over generalize the dataset. It is mainly because the generalization is performed at the domain level. Mondrian algorithm is another well-known algorithm in k-anonymity proposed by Kristen LeFevre and David J.DeWitt [18] , which tries to resolve this particular over generalization issue by performing generalization at equivalence level. Mondrian suggests a partitioning model, which is based on multi-dimensional approach to achieve k-anonymity.

Table 2-12: Original Patient Dataset

Age	Gender	Zip Code	Disease
35	Female	23111	Broken Leg
35	Male	23112	Flu
36	Female	23111	HIV
37	Female	23110	Ulcer
37	Male	23112	Gastric
38	Female	23111	Pneumonia

Its solution provides additional degree of flexibility, which has not been considered in the previously reviewed approaches. The core of Mondrian multi-dimensional anonymization consists of two main steps. The first step is to perform partitioning of the d-dimensional space where d is the number if quasi identifier attributes in a way that every partition contains at least k number of data records. The second step is to generalize all the records in each partition so that they are all share the same quasi identifier values. Considering a medical dataset from a hospital to be anonymized using Mondrian algorithm, the original Patient dataset with quasi-identifier attributes of {Age, Gender, Zip Code} and private sensitive attribute of {Disease} is shown in Table 2-12.

As it was mentioned earlier Mondrian suggests a partitioning model. In suggested partitioning model Mondrian partitions the dataset starting with one dimension and then moving to the next dimension looking for allowable cuts. An allowable cut is defined as the cut that results in region with number of records equal or more than k value. Therefore if a cut results in a region with number of records less than k value it is not allowed and the partition will not be formed.

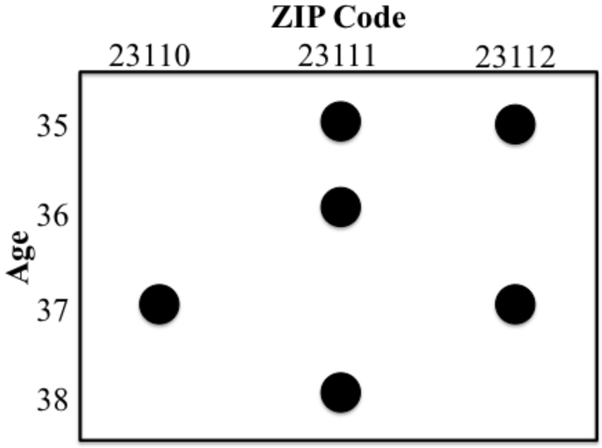


Figure 2-7: Spatial Representation of Age and Zip Code Attributes in Original Patient Dataset

In order to explain portioning model spatial representation of dataset is very helpful. The spatial representation for quasi identifier attributes {Age, Zip Code} in Patient dataset is illustrated in Figure 2-7.

Considering the spatial representation in Figure 2-7, starting with {Zip Code} attribute the first allowable cut is between “23111” and “23112”, which divides the dataset into two groups. This cut is allowable as the number of records in both groups is greater than or equal to k value. K value is assumed to be k=2 in this example. However to give an example of a not allowable cut, the cut between “23110” and “23111” is not allowable.

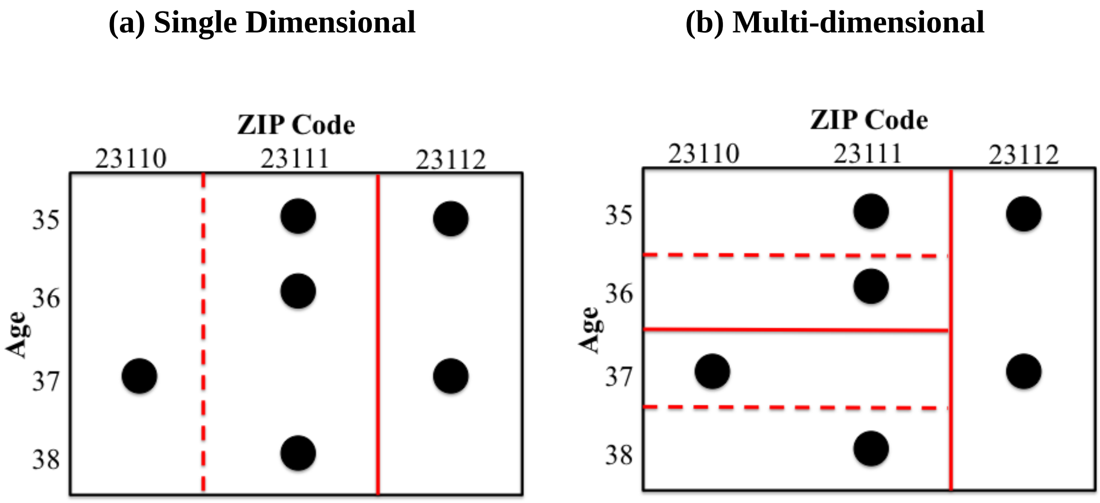


Figure 2-8: Spatial Representation with Single and Multi-dimensional Allowable Cuts

Because this cut, which is represented by dashed line in Figure 2-8(a) divides the dataset into two groups and the group on the left side of the dashed line has only one data record, which is less than the specified k value therefore k value condition is not satisfied and this cut is not allowable. The cut on {Zip Code} attribute dimension is a single dimension cut. However as it was mentioned earlier, Mondrian suggests a multi-dimensional solution to reduce or possibly remove the over generalization effect on anonymized data.

Therefore in multi-dimensional case the dataset will be cut again on the {Age} attribute dimension if any allowable cut exists. The allowable cut on {Age} attribute dimension is also shown in Figure 2-8(b).

In single dimensional case, after the first cut is made on Zip Code dimension there are no more allowable single dimension cuts because any cut perpendicular to Age dimension would violate the k value condition by resulting in a region with points (data records) fewer than k value. Using Mondrian multi-dimensional anonymization on Patient dataset shown in and considering k value is defined as k=2, anonymization process will result in 2-anonymous dataset which is shown in Figure 2-9.

In order to present the Mondrian multi-dimensional anonymization effect on reducing the over generalization cost a simple comparison is performed between Mondrian multi-dimensional anonymization and single dimensional anonymization.

As it is shown in Figure 2-9(b), 2-anonymous dataset using multi-dimensional anonymization is less distorted and generalized, comparing to single dimensional anonymization.

Age	Gender	Zip Code	Disease
[35-38]	Female	[23110-23111]	Broken Leg
[35-38]	Male	23112	Flu
[35-38]	Female	[23110-23111]	HIV
[35-38]	Female	[23110-23111]	Ulcer
[35-38]	Male	23112	Gastric
[35-38]	Female	[23110-23111]	Pneumonia

(a) 2-anonymous Patient dataset Using Single Dimensional Anonymization

Age	Gender	Zip Code	Disease
[35-36]	Female	23111	Broken Leg
[35-37]	Male	23112	Flu
[35-36]	Female	23111	HIV
[37-38]	Female	[23110-23111]	Ulcer
[35-37]	Male	23112	Gastric
[37-38]	Female	[23110-23111]	Pneumonia

(b) 2-anonymous Patient dataset Using Multi-Dimensional Anonymization

Figure 2-9: 2-Anonymous Dataset Using Mondrian

Particularly, {Age} attribute in all data records of Patient dataset is generalized from 35 to 38 when single dimensional anonymization is utilized. However {Age} attribute is generalized in smaller ranges depending on which equivalent class the data record belongs to in multi-dimensional anonymization approach. Same pattern and difference can be observed in {Zip Code} attribute when single dimensional anonymization is used comparing to multi-dimensional anonymization.

Therefore as it was mentioned earlier the 2-anonymous dataset anonymized using multi-dimensional anonymization is not over generalized comparing to single dimensional anonymization model. However as much as the multi-dimensional anonymization is efficient comparing to single dimensional anonymization, this approach has some disadvantages especially when categorical data are involved in the dataset. Mondrian requires the total order for each quasi identifier attribute in the dataset. As categorical data do not have meaningful orders this requirement makes Mondrian impractical in most cases involving categorical data [18] [72] .

2.6. Information Loss and Data Utility Metrics

Various models have been proposed to measure the quality of anonymized-data and information loss. In this section we review some of those metrics [37] [44] .

Minimal Distortion (MD) [16] is a single attribute measure and it defines the information loss as number of instances, which are made indistinguishable. For example if ten records are generalized in Sex attributes from “Male” or “Female” to “People”, the information loss is equal to ten. The Discernibility Metric (DM) [14] assigns penalty to each record based on the number of records indistinguishable from that record in anonymized table. The DM metric defines information loss for generalization and suppression, which can be expressed mathematically as follows.

Equation 2-1: DM Metric Information Loss for Generalization and Suppression

$$C_{DM}(g,k) = \sum_{\forall E s.t.|E \geq k|} |E|^2 + \sum_{\forall E s.t.|E < k|} |D||E|$$

In this expression E is the equivalent class and |D| is the size of the original dataset. The first sum calculates the information loss for generalized tuples and the second sum computes the information loss due to suppression. The information loss in both MD and DM is defined based the size of the group that the record is generalized and even though the DM is more accurate than MD, in k-anonymization methods which are near optimum the size of the groups are close to k value which makes these metrics less practicable. The ILoss metric proposed in [21] calculates the information loss of a specific value of a record, which is generalized. ILoss metric is expressed in Equation 2-2 as follows.

Equation 2-2: ILoss Metric

$$ILoss\ (v_g) = \frac{|v_g| - 1}{|D_A|}$$

In this expression $|v_g|$ is the number of domain values that are descendants of v_g and $|v_g|$ is the number of domain values in the attribute A of v_g and this metric requires all original data values to be at the leaves in the taxonomy. The Classification Metric CM [13] , charges a penalty for a record if its private value differs from the majority of the private values in its group or if the record is totally suppressed.

The more exact metric is the Normalized Certainty Penalty (NCP) [20] , which defines information loss due to generalization for both numerical and categorical attributes. For numerical attributes the NCP of a cell on numerical attribute A_i which joins in equivalent class G is defined as below.

Equation 2-3: Normalized Certainty Penalty Information Loss Numerical Attributes

$$NCP_{A_i}(G) = \frac{Max_{A_i}^G - Min_{A_i}^G}{Max_{A_i} - Min_{A_i}}$$

In case of categorical attributes, the NCP of the equivalent class G in A_i attribute is defined as follows.

Equation 2-4: Normalized Certainty Penalty Information Loss Categorical Attributes

$$NCP_{A_i}(G) = \begin{cases} 0 & Card(u) = 1 \\ \dfrac{Card(u)}{Card(D_i)}, & Otherwise \end{cases}$$

Where, $Card(u)$ is the number of distinct values of A_i in G and $Card(D_i)$ is the total number of distinct values of attribute A_i. For information loss measurement, it is very crucial to choose the right measurement metric. In work [20] the information loss and data utility is measured using NCP, however NCP only measures the information loss due to generalization and in [22] both suppression and generalization have been used for anonymization. Therefore the evaluation results may not be reliable and precise because the information loss due to suppression is not calculated [22] [66] .

Since in the evaluation section of this work Normalized Certainty Penalty (NCP) [20] [67] is employed in order to calculate the total information loss (total NCP) and data utility of anonymized dataset an example of NCP calculation on 3-Anonymous dataset shown in Figure 2-10 will be explained as follows.

Considering the Patient dataset and its corresponding 3-anonymous dataset, Normalized Certainty Penalty is defined for Age as numerical and Sex as categorical attributes. In this example Zip Code attribute has no modification therefore the total NCP for Zip Code is equal to null. On the other hand the tuple 1, 2 and 3 are grouped together and formed an equivalent class while tuple 4 is suppressed. For suppression the total NCP is considered to be maximum as 1 for each attribute. The range of Age attribute is assumed to be 90 (10 ~ 100) and Gender attribute has two distinct values of Male and Female.

All the calculation of NCP for each tuple with respect to each attribute can be summarized as shown in Table 2-13.

Tuple	Age	Gender	Zip Code	Disease
T1	25	Male	2370	Gastritis
T2	35	Male	2370	HIV
T3	40	Female	2370	Cancer
T4	65	Female	5300	Fever

(a) Original Patient Dataset

Equivalent Class	Tuple	Age	Gender	Zip Code	Disease
E1	T1	25~40	Male, Female	2370	Gastritis
	T2	25~40	Male, Female	2370	HIV
	T3	25~40	Male, Female	2370	Cancer
Suppressed	~~T4~~	~~65~~	~~Female~~	~~5300~~	~~Fever~~

(b) 3-Anonymous Patient Dataset

Figure 2-10: Original Patient Dataset and its Corresponding 3-Anonymous Dataset

Of course after the total NCP calculation for each tuple the sum of all tuples' NCP would make the total information loss of the k-anonymous dataset, which in this case is 3-anonymous Patient dataset. The total NCP of k-anonymous dataset must be normalized between 0 and 1 so it can be compared with the k-anonymous result of other models with respect to information loss.

We have introduced the metrics to calculate information loss and an example is provided in this section particularly on how Normalized Certainty Penalty is actually calculating the information loss of k-anonymous dataset for each tuple and with respect to each attribute (numerical and categorical) for generalization and suppression. However regarding utility of data nothing is mentioned yet.

Table 2-13: Normalized Certainty Penalty (NCP) Calculation for Each Tuple with Respect to Each Attribute

$NCP_{Age}(t1) = \frac{40-25}{100-10}$	$NCP_{Gender}(t1) = \frac{2}{2}$	$NCP_{ZipCode}(t1) = 0$	$NCP_{Total}(t1) = 1.167$
$NCP_{Age}(t2) = \frac{40-25}{100-10}$	$NCP_{Gender}(t2) = \frac{2}{2}$	$NCP_{ZipCode}(t2) = 0$	$NCP_{Total}(t2) = 1.167$
$NCP_{Age}(t3) = \frac{40-25}{100-10}$	$NCP_{Gender}(t3) = \frac{2}{2}$	$NCP_{ZipCode}(t3) = 0$	$NCP_{Total}(t3) = 1.167$
$NCP_{Age}(t4) = 1$	$NCP_{Gender}(t4) = 1$	$NCP_{ZipCode}(t4) = 1$	$NCP_{Total}(t4) = 3$

As it was mentioned data utility is opposite of information loss. Therefore by having the information loss that is caused in k-anonymous dataset due to generalization and suppression, data utility can also be calculated. By normalizing the total NCP between zero and one, the utility of anonymized-data could be defined as follows.

Equation 2-5: Data Utility

$$\text{Utility} = 1 - \text{NCP}_{\text{Total}} \qquad \text{where, } 0 \leq \text{NCP}_{\text{Total}} \leq 1$$

Using the above formula the data utility of k-anonymous dataset can be calculated. If the total NCP of dataset is equal to 1, which means maximum information loss, the utility of that dataset is practically zero. The utility of tuple for in the given example at Table 2-13 is equal to zero since the tuple 4 is removed from the 3-anonymous dataset. On the other hand if the total NCP (Information loss) of dataset is equal to zero then the utility of dataset is maximum and equal to 1.

3. Chapter 3: Similarity Based Clustering Model for K-anonymity

3.1. Summary of Chapter

In this chapter, the proposed Similarity Based Clustering Model is explained in details. This chapter starts with an introduction on k-anonymity and its definition from clustering point of view. Later in this chapter an introduction of clustering is given. The basic clustering is defined and different types of clustering is explained to some extent.

In section 3.3 clustering application in k-anonymity is mentioned which talks about clustering application in k-anonymity model. Then in section 3.4 the proposed model on defining similarity between values in categorical attributes is explained. It is mentioned that construction of contingency tables are necessary for measuring similarity between different values in categorical attributes. Moreover, the measured similarities are solely based on probability of observation. It does not depend on any additional information on dataset such as hierarchical taxonomies of categorical attributes.

After similarity measurement, in next section it is explained how to measure distance between values in categorical attributes using the measured similarities. In addition it is explained how the total distance between tuples including numerical and categorical attributes is calculated. This will allow us to measure total distance between tuples considering all attributes over all dimensions.

At the end of this chapter, finally a bottom-up algorithm based on the proposed Similarity Based Clustering Model is suggested and explained.

3.2. Proposed Similarity-Based Clustering Model for K-Anonymization

As it was mentioned briefly in introduction part and also in chapter 2 the background study on k-anonymity and different models and algorithms which implement k-anonymity, generally performing any kind of anonymization of a dataset causes distortion on original dataset. This distortion, which is known as information loss, is inevitable and it reduces the utility of data. As data utility is very important to researchers and data miners regarding the quality of the result they are going to get by working on a particular dataset, one of the primary objectives in this work is to minimize the information loss caused by anonymization process through generalization or suppression to be able to get high utility anonymized data. So the data is still useful for data mining and research purposes while preserving the privacy of individuals at the same time.

The other issue that was mentioned which is rather a more practical problem is about the datasets and the type of data that is considered for anonymization in real life applications. In most of the datasets there are various categorical attributes and as a matter of fact most of the quasi-identifier attributes are considered to be categorical [23] . Therefore having a model that can actually perform efficiently and independently with datasets including both numerical and categorical data type attributes is considered.

In order to rectify the above stated issues in k-anonymization, clustering approach is considered. Technically a dataset is called k-anonymous dataset when for every record in the dataset there are at least k-1 other records identical to it along the quasi-identifier attributes. As it was mentioned, k-anonymity can also be defined and explained from clustering point of view. Normally in clustering approaches finding number of clusters are more important than the number of records in each cluster.

However in k-anonymization number of records in each cluster is defined through the k value condition. Therefore k-anonymity could be defined as clustering with constrain of minimum of k tuples (data records) in each cluster. Considering the example shown in Figure 3-1, the dataset is assumed to contain Age and Zip Code attributes as quasi identifier attributes. The dataset is clustered into 3 separated groups and each group contains 3 tuple (data records).

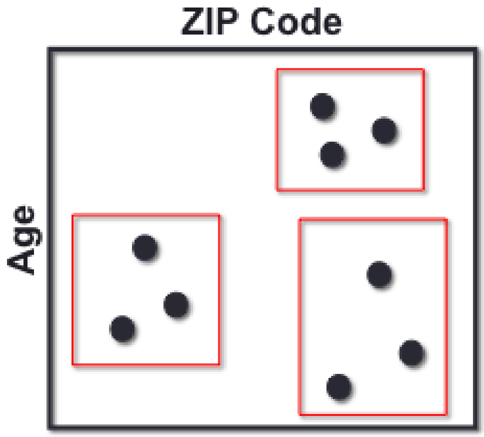


Figure 3-1:2-Dimensional Representation of Dataset Clustered with Constrain of k=3

The number of groups is not important in anonymization application however every cluster containing at least 3 tuples is the k-anonymity condition, which is full filled in this example. The clustering approach in k-anonymization and proposed model to achieve k-anonymity will be explained in the following sections.

3.3. Clustering Approach in K-anonymization

3.3.1. Clustering

Clustering is the process of grouping data objects. In clustering process the data is arranged in a way that the most similar records, along all attributes, belong to the same cluster or group, while records with high dissimilarity put in different clusters. Clustering is a main task of exploratory data mining, and it is used commonly for statistical data analysis and various fields including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. In data mining, clustering is a type of unsupervised classification. Some common applications of clustering as a stand-alone tool is to get insight into data distribution for instance, discovery of distinct customer groups, categorization of genes with similar functionality and identification of areas of similar land use [5] [24] . There exists various number of clustering methods. In this section some of the methods, such as partitioning, hierarchical, density based, grid based and model based methods, are mentioned and briefly discussed [5] [24] .

A partitioning method in clustering generally divides the records of a dataset into k non-empty and mutually exclusive partitions. In this method k, number of portions, is defined by the user. The method then uses an iterative relocation in order to improve the quality of the partitions or clusters by grouping similar records in a cluster and dissimilar records in different clusters. The two common heuristics used in this method are k-means and k-medoids [5] [24] [59] [62] .

A hierarchical method in clustering creates a hierarchical decomposition of data objects using some criterion. This method can be further divided into two types as follows.

1) Agglomerative

2) Divisive

An agglomerative hierarchical method first considers each single data object or data record as a separate cluster (cluster with single record). Based on some defined similarity criteria, it then merges the two most similar records or groups of records in consecutive iteration until the termination condition is fulfilled or all records are merged into one single cluster [5] [61] [63] .

On the other hand, the divisive hierarchical clustering method starts with all records in a single cluster which is exactly opposite to agglomerative hierarchical method. In iteration it splits the initially formed cluster into two clusters in order to improve the criteria that measure the quality of the overall clustering. Finally, the method stops when a termination condition is met or each record is separated into a cluster [5] [61] [63] .

A density based clustering forms clusters of dense regions where a high number of records are located. This method initially selects a core record that has large number of neighbor records. The core record and all its neighbor records are included in a cluster. If a record “R” among these neighbors is itself a core, then all neighbors of “R” are also added in the cluster. The process terminates when there is no more record left that can be added to a cluster. This clustering technique is also used to filter out noise from a dataset [5] [59] [60] [61] [62] [63] .

A grid-based method performs all clustering operations on a grid like structure obtained by quantizing the data space into a finite number of cells. The main advantage is a faster processing speed, which mainly depends on the number of cells. Unlike conventional clustering, a model based clustering attempts to find a characteristic description of each cluster, in addition to just clustering the unlabeled [5] [59] [60] [61] [62] [63] .

3.3.2. Clustering in K-anonymity

Clustering approach in k-anonymization is considered to be one of the ultimate solutions to solve the information loss issue in k-anonymization. Moreover the main challenge in clustering approach in k-anonymization application is slightly different than common clustering problem, which is the number of clusters in the dataset [38] [39] .

The main goal and target in clustering approach in anonymization application is to find the k closest tuple in dataset and group them all together. So in each cluster there are at least k tuples, which satisfies k value condition in k-anonymization, and all the tuples in the same cluster have minimum possible distance from each other thus the information loss in each cluster is minimized.

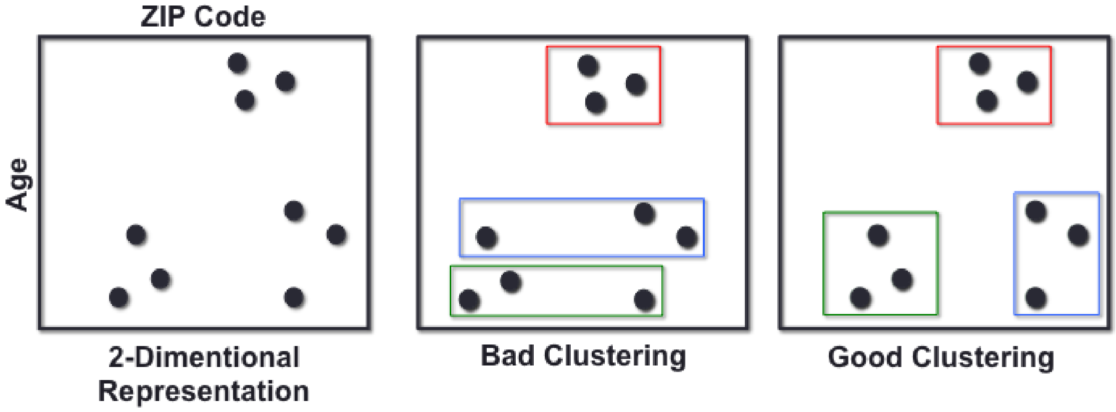


Figure 3-2: Good versus Bad Clustering

By keeping the distance, which is technically representative of information loss and a measurement on how similar or different data tuples are from each other, minimized in every cluster eventually the information loss after the generalization could be minimized as well.

K-anonymization as a clustering problem can be defined as follows. Clustering problem in k-anonymization is to find a set of groups or clusters in the provided dataset with n number of tuples so that each group consists of at least k tuples (data records) and all the tuples in the same group or cluster have the minimum distance from each other [38] [39] .

However performing efficient and good clustering as it is shown in Figure 3-2 on datasets including numerical and categorical attributes for anonymization is challenging especially due to the existence of categorical attributes. In order to be able to cluster the whole dataset the distance between every tuples should be calculated. However because of the nature of categorical attributes and the fact that the values are actually numeric and continuous it is not simply possible.

There is a need of a model that actually defines the distances between all distinct values in categorical attributes. Regarding categorical attributes, in some of the previous works on k-anonymization such as Mondrian it is assumed that a total order exists on all values in categorical attributes however in many applications such an order may not exist. Zip code attribute could be a good example on this issue. Sorting all Zip Codes in their numeric order may not represent their distance and not reflect the utility property as two regions may be next to each other but their Zip Codes are not consecutive.

In some other works it is suggested to use hierarchical taxonomies and generalization trees in order to define the distances and measure utility between all the distinct values in categorical attributes [20] [67] . Obviously there is a need to make an assumption that the dedicated attribute's hierarchical taxonomies always provided as shown in Figure 3-3 [40] .

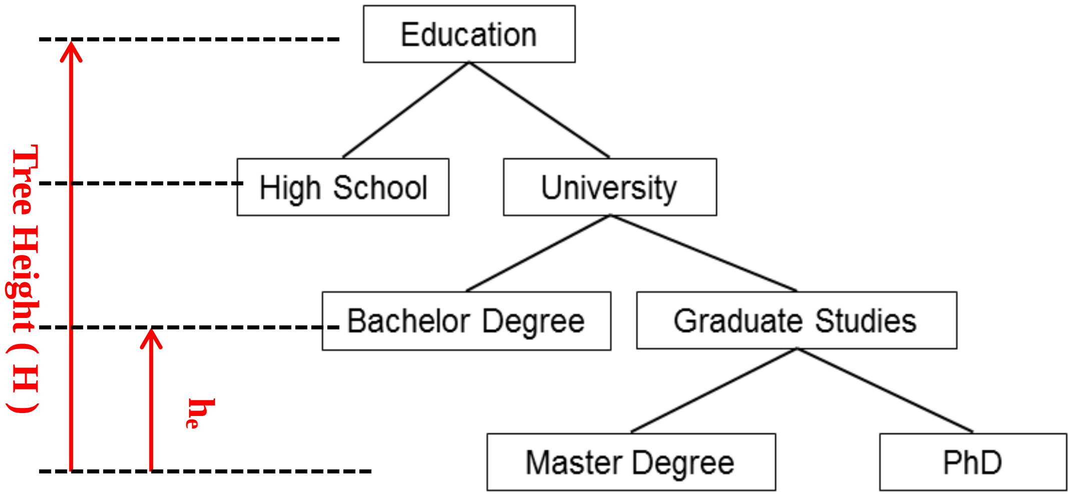


Figure 3-3: Education Attribute and Its Generalization Tree

The problem of these approaches is that all the process is actually depends on the hierarchical taxonomies, which mostly is not provided a long with the dataset itself in real life applications. Also hierarchical taxonomies tend to vary based on the datasets so for every dataset there should be additional information telling about the hierarchical taxonomies for every categorical attribute in the dataset so it can be anonymized. Moreover, as the hierarchical taxonomies are designed and fixed, once the distances are defined between distinct values in a particular datasets as the given example during the process the distances will not be changed [48] [49] [50] .

Therefore having another model to define the distance between all distinct values for categorical attribute could actually solve the entire above-mentioned problems. In the next section the approach to rectify these problems are explained in detail.

3.4. Similarity Measurement in Categorical Attributes

In order to calculate the distance between tuples in datasets including numerical and categorical attributes the distance between distinct values in categorical attribute must be defined.

Regarding categorical attributes, distance is not well defined between the values mainly due to the nature of categorical attributes and the problem of representing the value of categorical attribute in numerical form. In some previous works the distance in categorical attributes is defined with the help of the hierarchical structure [13] [20] [38] [39] [40] [67] .

However, the hierarchies may not exist or defined in real life applications. In our model the distance between the values in categorical attributes is defined based on the context and the observation probability of values in each attribute [46] . It is efficient and easily adjustable depending on the number of categorical attributes and more importantly it is derived from the dataset itself therefore there is no need of any extra information along with the dataset such as hierarchical taxonomy.

This method and idea is generated from another work which is on context based distance learning for categorical data clustering [46] . The key intuition of work [46] is actually defining distance between two values of categorical attribute A_i is determined by the method in which the values of other attributes A_j are distributed in the dataset [46] . In this approach, the first step is to construct the contingency table. The contingency table helps us to measure the observation probability for each value of categorical attribute A_j and assess the similarity between y_1, the value of the first tuple (t_1) in A_j, and the rest of the values in A_j. By knowing which values in A_j has the most and least similarity to y_1 the distances between y_1 and the rest of the values in A_j could be defined.

For instant, let's consider dataset T with total twenty tuples as shown in Figure 3-4. There are two categorical attributes, Sex = {Male, Female} and Nationality = {Japan, US, Iran} in which the attributes are arranged with respect to the cardinality order, lower to higher cardinality from left to right. The contingency table for categorical attributes in dataset T is constructed and shown in Table 3-1.

In this example since there are only two categorical attributes having one contingency table is adequate. However, if there are more categorical attributes in dataset T, in order to find the similarities between the values in the next higher cardinality attribute, we add the next higher cardinality attribute to the existing contingency table.

In the contingency table the attribute which has higher cardinality and the similarity measurement between its values are going to take place is placed horizontally and the attributes with lower cardinality are placed in the left side of the table vertically with respect to cardinality order.

Tuple	**Gender**	**Nationality**
T1	Female	Japan
⋮	⋮	⋮
T5	Male	USA
⋮	⋮	⋮
T10	Female	Iran
⋮	⋮	⋮
T15	Male	USA
⋮	⋮	⋮
T20	Female	Japan

Figure 3-4: Categorical Attributes in Dataset T

As shown in Figure 3-4 the values of Gender and Nationality attributes in t_1 are {Male} and {Japan}. By indicating that we start the similarity measurement from the attribute with cardinality more than two, which in this example is Nationality. For the attribute with cardinality less or equal to two there is no need to measure the similarity, because there is only one distance to be defined, in case the cardinality of attribute is equal to two, and the distance for such cases is already defined as maximum which is equal to one. The minimum distance is zero, which is the distance between the identical values. Also, we need to calculate the total number of tuples in each row of the contingency table as shown in Table 3-1. This calculation is for the purpose of confirmation on k value condition in k-anonymity.

Table 3-1: Contingency Table of Dataset T Shown in Figure 3-4

	Japan	**US**	**Iran**
Male	4	4	1
Female	4	1	6

If the number of Male tuples (in this example since the first tuple value for Gender attribute is Male) is greater than or equal to the pre-defined k value then the similarities are measured with respect to the total number of tuples in that row only, else other rows in that specific attribute needs to be considered regarding similarity measurement. In this example k value is considered to be k=3 and the total number of tuples in Male row in Table 3-2 is greater than 3.

Table 3-2: Contingency Table (1) of Dataset T and the Total Number of Tuples in Each Row

	Japan	**US**	**Iran**	**Total No. of Tuples**
Male	**4**	4	1	4+4+1 = 9 $\geq$ k=3
Female	4	1	6	4+1+6 = 11 $\geq$ k=3

However if the total number of tuples with value equal to Male were not greater than k value if k=10 for example, the contingency table would be modified for the similarity measurement between values of Nationality attribute. As shown in Table 3-3, the Female row also would be considered, merged with Male row, for similarity measurement between the values in Nationality attributes.

Table 3-3: Contingency Table (2) of Dataset T if Total Number of Male is less then k value k = 10

	Japan	US	Iran	Total No. of Tuples
Male, Female	**8**	5	7	8+5+7 = 20 $\geq$ k=10

The main reason for such confirmation is if k value k=10 is considered no matter how we try, the tuples which are grouped and clustered together are going to be a mixture of Male and Female regarding to Gender attribute as there are not enough tuples (more than or equal to 10) with only Male value in their Gender attribute.

Considering a dataset T with two categorical attributes $M=\{m_1,\cdots,m_i\}$ and $N=\{n_1,\cdots,n_j\}$, the probability of observation for each value in attribute N when $i<j$, $1\leq K\leq i$, $1\leq L\leq j$ and the total number of tuples in m_K is more than k value, is defined as follows.

Equation 3-1: Probability of Observation for Each Value in Attribute N

$$P(n_L)_{m_k}=\frac{(|n_L|)_{m_K}}{(|n_1+\cdots+n_j|)_{m_K}}$$

The notation $(|n_L|)_{m_K}$ indicates the number of tuples with value of n_L in N and value of m_K in M attribute and $(|n_1+\cdots+n_j|)_{m_K}$ means the total number of tuples in attribute N which have the value of m_K. The Equation 3-1 can be expanded for multiple categorical attributes with multiple values.

By calculating all the observation probabilities for each value in attribute $N=\{n_1,\cdots,n_j\}$ and obtaining $P(n_1)_{m_K},\cdots,P(n_j)_{m_K}$, the similarity between the value of t_1 in attribute N and rest of other values in N could be defined. The closer the $P(n_L)_{m_k}$ is to $P(n_1)_{m_k}$, the more similar n_L is to n_1. To apply this calculation on dataset provided earlier in Figure 3-4, the similarity between the values in Nationality attribute in in Table 3-1 is calculated. The calculation is presented in Figure 3-5.

By looking at the result of observation probability of values in Nationality attribute when first tuple has value of Male and Japan, since $P(\text{Japan})_{\text{Male}}$ is closer to $P(\text{US})_{\text{Male}}$ than $P(\text{Iran})_{\text{Male}}$ then it can be concluded that Japan is more similar to US and less similar to Iran considering k value is k=3 (Japan, USA, Iran is the similarity order).

$$P(\text{Japan})_{\text{Male}}= \frac{(|\text{Japan}|)_{\text{Male}}}{(|\text{Japan}|+|\text{US}|+|\text{Iran}|)_{\text{Male}}}=\frac{4}{9}$$

$$P(\text{USA})_{\text{Male}}= \frac{(|\text{USA}|)_{\text{Male}}}{(|\text{Japan}|+|\text{US}|+|\text{Iran}|)_{\text{Male}}}=\frac{4}{9}$$

$$P(\text{Iran})_{\text{Male}}= \frac{(|\text{Iran}|)_{\text{Male}}}{(|\text{Japan}|+|\text{US}|+|\text{Iran}|)_{\text{Male}}}=\frac{1}{9}$$

Figure 3-5: Similarity Calculation based on Contingency Table Shown in Table 3-1

If there were other attributes with cardinality greater or equal to Nationality, we could add the next attribute to the contingency table and investigate the similarity between its values likewise.

3.5. Distance Measurement

After measuring the similarity between the values of all categorical attributes in dataset, (in dataset T presented in section 3.4 the similarities in Gender and Nationality attributes are measured) the distances between the values can be defined.

We start with attribute with the lowest cardinality to the highest and the distances are defined with respect to the measured similarities from the least similarity to the most. In this example Gender attribute with cardinality two is the lowest and since there is only one distance to be defined (distance between Male, Female) it is defined as maximum distance, D(Male, Female) = 1.

For the second minimum cardinality attribute, which is Nationality in this example, the least similarity is between Japan and Iran and Japan and USA are the most similar values. Distances are defined with respect to the similarity order {Japan, USA, Iran} as follows:

$$\mathrm{D(Japan, Japan)} = 0$$

$$\mathrm{D(Japan, USA)} = \frac{\text{Index of USA in Similarity Order}}{|\text{Card (Nationality)}|\text{-1}} = \frac{1}{2}$$

$$\mathrm{D(Japan, Iran)} = \frac{\text{Index of Iran in Similarity Order}}{|\text{Card (Nationality)}|\text{-1}} = \frac{2}{2}$$

As shown above all the distance between values in Nationality attribute is calculated. The numerator is the index of the value in the similarity order that was measured and the denominator is the cardinality of attribute minus one, which basically indicates the number of distances, which need to be defined. Therefore all the distances between values are defined between 0 and 1. The most similar values have smaller distance and the most dissimilar values have the highest possible distance, which is equal to 1. By defining the distances using this method, the most similar values in different categorical attribute will have smallest distances to values at t_1.

In our example in this section, finally by having $\mathrm{D(Male, Female)} = 1$, $\mathrm{D(Japan, US)} = 1/2$ and $\mathrm{D(Japan, Iran)} = 1$ defined, the total distance between t_1 and other tuples in dataset T can be calculated as the sum of the $\mathrm{D(Male, Female)}$ and $\mathrm{D(Japan, US\ or\ Iran)}$. If a dataset is a combination of numerical and categorical attributes there is a separated process necessary for numerical attributes only for distance calculation and normalization. The distance between numerical values is calculated using the equation below.

Equation 3-2: Distance between Values Numerical Attributes

$$\text{Distance}(t_1,\ t_2)_{\text{Att}_i} = \frac{|x_1 - x_2|}{R(\text{Att}_i)}$$

For numerical attributes the distance measurement is rather conventional. The distance between two tuples t_1 and t_2 with respect to attribute A_i with values of x_1 and x_2 is defined using Equation 3-2 where, $R(\text{Att}_i)$ is the range of attribute A_i. The range of A_i attribute is defined as $R(A_i) = \text{Max}(A_i) - \text{Min}(A_i)$. Based on this, the total distance between t_1 and t_2 for numerical attributes in dataset T is the sum of the $D(t_1,t_2)_{A_i}$ for every A_i, where A_i is the numerical quasi-identifier attribute in dataset T.

After calculation of all distance in numerical attributes and categorical attributes and normalizing both separately, the total distance between tuples can be calculated. Considering the original dataset T with the numerical attributes $\{X_1,\cdots,X_m\}$ and categorical attributes $\{Y_1,\cdots,Y_m\}$, the total distance between two tuples t_1 and t_2 is defined as a sum of the distances in numerical and categorical attributes. Obviously after the addition the total distance will be normalized between 0 and 1.

Equation 3-3: Total Distance between Tuples

$$D_T(t_1,t_2) = \sum_{i=1,\ldots,m} \left(D(t_1[X_i],t_2[X_i])\right) + \sum_{j=1,\ldots,n} \left(D\left(t_1\left[Y_j\right],t_2\left[Y_j\right]\right)\right)$$

3.6. Similarity-Based Clustering (SBC) Model

As it was mentioned before, in k-anonymity model selecting the tuples and placing them into equivalent classes for anonymization is one of the essential parts that directly affect the performance of anonymization and the utility of anonymized-data. In addition, real world datasets contain numerical and categorical attributes. This combination of different type of attributes makes the anonymization process rather complicated and very often results in inefficient anonymization.

By going through section 3.4 and 3.5, the total distances between the first tuple and the rest of tuples in the given dataset are calculated successfully without a need of any additional information about the dataset or hierarchical taxonomies of all categorical attributes. Having all the distances clustering the dataset is made possible.

As the distances were all defined based on the similarity measurement, which introduced in section 3.4, and our approach toward k-anonymity is from clustering point of view this model is called similarity-based clustering.

In the next section we will introduce a bottom-up greedy algorithm based on similarity measurement and distance calculation that were just explained for clustering the tuples for anonymization through local recoding which does not require any additional information such as the total order for each attribute domain or the hierarchical structure of attributes.

3.7. Extension of Similarity-Based Clustering (SBC) Model for Multiple Categorical Attributes

In the example given in previous section there are only two categorical attributes {Gender and Nationality}. The similarity of values for {Nationality} attributes are actually measured using observational probability of those values with respect to {Gender} values and the first tuple in the dataset. In this section we would like to discuss the possibility of extending this model for cases and datasets which have more than two categorical attributes.

In previous example, since there are only two categorical attributes having one contingency table was enough. However, if there are more categorical attributes in dataset T, in order to find the similarities between the values in the next higher cardinality attribute, there is a necessity to create a new contingency table in order to measure similarity between the values in the higher cardinality attribute. Considering a new dataset T having

{Gender, Nationality and Education} attributes the contingency for this particular dataset could be constructed as shown in Table 3-4 .

In the contingency table the attribute which has higher cardinality and the similarity measurement between its values are going to take place is placed horizontally and the attributes with lower cardinality are placed in the left side of the table vertically and the similarity between values in higher cardinality attribute is measured based on lower cardinality attribute and how the values are distributed considering the first tuple values in the dataset.

Table 3-4: Contingency Table for More than Two Categorical Attributes

	Japan	**US**	**Iran**
Male	4	4	1
Female	4	1	6

(a) First Contingency Table for Similarity Measurement for Values in Nationality Attribute

	High School	**Bachelor Degree**	**Master Degree**	**PhD Degree**
Japan	0	4	3	1
US	4	1	0	0
Iran	1	3	1	2

(b) Second Contingency Table for Similarity Measurement for Values in Education Attribute

As it is shown, similar to first example, having more than two categorical attributes does not make a lot of changes in the proposed model itself. Just additional contingency table is required in order to measure the similarity between values in higher cardinality attribute. Same set of rules and definition is applicable in the second contingency table for similarity measurements and distance definition between values based on the measured similarities.

Therefore it can be concluded that this model and similarity measurement for categorical attribute could be extended and adjusted depending on the datasets and how many categorical attributes actually exist in the dataset.

Moreover there are some limitations regarding Similarity-Based Clustering (SBC) Model. These limitations are mainly related to datasets and data types are realized and will be briefly discussed in the next section.

3.8. Limitations of Similarity-Based Clustering (SBC) Model

In this section the limitation of the proposed Similarity-Based Clustering (SBC) model are discussed. One of the main limitations for defining similarity and distances between categorical attributes is related to the data types. It was mentioned that there are various type of data. The type of data which are considered in this model are based on real applications and data types which are considered by k-anonymity model. In most of applications the datasets are actually exported from data base systems. Therefore the data types are tabular data which we can apply this model efficiently and define similarity and calculate distances based on the measured similarities.

However if the data is not tabular data such as log datasets which are mainly text based this model is not going to function. For anonymizing those kind of data, there is a need of developing new models which could be text mining based models to anonymize texts.

Moreover, there is a point regarding {Gender} attribute as the lowest cardinality attribute in dataset. In all datasets which I personally worked with or datasets mentioned in other works related to k-anonymity there always exist {Gender} attribute. However if this attribute does not exist or removed from dataset for whatever reason, Similarity-Based Clustering (SBC) Model can still function without any problem. However there is a need to define distances for the lowest cardinality attribute. We actually have an idea which measures similarity and distances solely based on the number of values for that attribute only. For example if in previous example, which is shown in Table 3-4, {Nationality} attribute is the lowest cardinality attribute, based on the number of values it can be conclude that Japan is more similar to Iran comparing to Japan because the number of Japanese nationals in dataset is closer to Iranian nationals.

Since this situation is rare it is not considered for experimental evaluation, however the possibility of such situation is realized and solutions are considered as it was explained using an example earlier.

3.9. Proposed Similarity-Based Clustering (SBC) Algorithm

Based on the proposed model for similarity measurement, distance calculation and clustering in k-anonymization we introduce a greedy algorithm with bottom-up approach.

In the proposed algorithm, every single tuple is considered as a point in the Euclidean space and the dimension of the space is actually the number of attributes. K value is given to the model as the k-anonymity condition. Then the original dataset is sorted and the numerical quasi-identifiers are separated from the categorical quasi-identifiers for similarity measurement and distance calculation. The contingency table for categorical attributes is constructed and after the similarity measurement all the distances between the values in first tuple (t_1) and other values in categorical attributes are defined. By having all the distances for categorical attributes and using the formula for distance calculation in numerical attributes the total distances between t_1 and other tuples are calculated and normalized.

Then in order to find the k-1 closest tuples to t_1 to be placed in the same equivalent class, the total distance between t_1 and the rest of the tuples in dataset T is calculated and t_1 and the k-1 tuple with minimum distances are moved to merge clause and deleted from T. Considering k value the number of tuples in merge clause must be greater or equal to k, therefore if the group size in merge clause is less than k then more tuples need to be added to merge clause. Once the number of tuples in merge clause is equal or greater than k value the tuples in merge clause considered as an equivalent class and they anonymized through local recoding anonymization. Which means, a range from minimum to maximum will replace the numeric values for numerical attributes and values in categorical attributes will be replaced by a set of distinct values in that equivalent class.

After each equivalent class is made the contingency table will be updated. Therefore similarity is measured again between the values in categorical attributes and the new distances are calculated. This operation repeated until the total tuples in dataset T is none or less than k value.

```
Input: Original dataset T & K-Value
Output: K - anonymous table T'
1:  Sort “T” Dataset
2:  WHILE $|dataset\ T| \geq$ K-Value DO    {
3:  Obtain First Tuple in Sorted “T”
4:  FOR Categorical_ Att:
        4.1: Contingency table constructed
        4.2: K-Value check
        4.3: Similarity measurement
        4.4: Calculate distances
5:  FOR Numerical_ Att:
        5.1: Numerical_ Att Distance calculation
6:  Calculate Total Distance between “first tuple” and the rest of the tuples in T
7:  Cluster k-1 closest tuples to First Tuple into equivalent class
8:  Anonymize the equivalent class through local recoding & DELETE from T & SAVE T’
9:  IF $|dataset\ T| <$ K Value DO Suppression or Add to last Equivalent class
10: Publish K-anonymous table T'
```

Figure 3-6: Pseudo code for Similarity-Based Clustering Algorithm

The remaining tuples could be suppressed (removed from the dataset) or could join the already existing equivalent classes with minimum distance. However in most of the cases that last equivalent class has the highest information loss and the remaining tuples could be also added to the lastly created equivalent class. After this process this process there will be no more tuple left in original dataset T and the k-anonymous dataset can be published. The pseudo code for the algorithm is shown in Figure 3-6.

4. Chapter 4: Empirical Evaluation

4.1. Summary of Chapter

In previous chapter the concept of clustering in anonymization and specifically in k-anonymity model was introduced and elaborated. In addition the proposed model, Similarity-Based Clustering Model (SBCM), was introduced and explained in details. In this chapter the proposed model is evaluated and compared to three other well-known models which implement k-anonymity model. As it was mentioned high information loss and low data utility is one of the main issues in k-anonymity, so in order to evaluate models and compare different models to each other, measuring information loss is one of the popular and logical methods. In order to evaluate proposed Similarity-Based Clustering Model (SBCM) and compare it to other three well-known models, information loss measurement is used. All models are simulated on two different real datasets and information loss is measured using Normalized Certainty Penalty (NCP) method which was explained earlier in chapter two. Moreover detailed information on datasets such as frequency distribution of each attribute are illustrated and explained. The result of simulations are shown, compared and analyzed. Finally at the end the performance and outcome of evaluation is concluded.

4.2. Experimental Evaluation

In this section the proposed Similarity-Based Clustering Model is evaluated and the result of simulations are compared to other well-known models in k-anonymization. In k-anonymization generally the evaluation is done regarding information loss and data utility.

As it was stated earlier information loss occurs due to the generalization and suppression in k-anonymity and high information loss and low data utility is one of the main challenges in this domain. Therefore the models are compared to each other with respect to information loss and data utility of k-anonymous dataset (anonymized-data).

In addition regarding information loss and data utility measurements, in chapter two section 2.5 various measurements and calculations of information loss and data utility are introduced. In this work, Normalized Certainty Penalty (NCP) is selected to measure the information loss and data utility of anonymized data [20] [67] . The main reason for selecting Normalized Certainty Penalty (NCP) to measure information loss is because different methods are using different techniques to achieve k-anonymity and the two datasets which were utilized for simulation in this work consist of both numerical and categorical attributes. Therefore for information loss a calculation metric that calculates information loss for both generalization and suppression anonymization in numerical and categorical attribute is necessary [20] [67] .

Regarding the other three well-known models, which are Mondrian, Incognito and Datafly models and algorithms UTD anonymization toolbox, which is an open source software and developed by UT (The University Of Texas At Dallas) Dallas Data Security and Privacy Laboratory [47] , is employed in order to anonymize data through those models and get the anonymization results on our real datasets [26] . Mondrian, Incognito and Datafly models and algorithms which implement them are carefully reviewed and analyzed in chapter two section 2.4 [15] [16] [18] .

The UTD anonymization toolbox requires to be configured separately for each model and because Mondrian, Incognito and Datafly models are different than each other the configuration file is different than each other as well. In addition, Incognito and Datafly models require hierarchical taxonomies for each attribute to be defined in the configuration file in order for anonymization process can be started. Obviously the hierarchical taxonomies are depend on the dataset. Therefore it is defined separately for each dataset used for simulations [26] [47] . The configuration XML file for each algorithm in the toolbox is attached in Appendix section for more detailed information.

Unlike Mondrian, Incognito and Datafly models, the proposed Similarity-Based Clustering Model (SBCM), do not depend on attribute hierarchical taxonomies therefore it is not necessary to define the attribute hierarchical taxonomies. Similarity-Based Clustering Model (SBCM) requires the dataset itself and desired privacy level, k value, only to start the anonymization process [48] [49] [50] .

The proposed Similarity-Based Clustering Model (SBCM) along with the three well-known models are simulated on two different real datasets. The datasets and the simulated results will be explained in details in the following sections.

4.3. Simulation Result on Adult Dataset

For experimental evaluation, one of the datasets which are used is called Adult dataset. Adult dataset, also known as "Census Income" dataset, is provided by University of California, School of Information and Computer Science (UCI Machine Learning Repository) [25] [51] , which contains census data. Its primary purpose was the prediction task to determine whether a person makes over 50K a year. Adult Dataset has become a benchmark for k-anonymity as well and most of the related works on anonymization are evaluating their model by simulating it on Adult Dataset [25] [51] .

In this particular experiment, the dataset for simulation is taken from Adult dataset and it contains 5000 tuples (data records). Dataset has four different attributes. The quasi-identifier attributes are {Age, Gender and Native-Country}. The private sensitive attribute is {Salary}. The statistical information and frequency distribution of data over quasi identifier attributes is shown as follows in Figure 4-1.

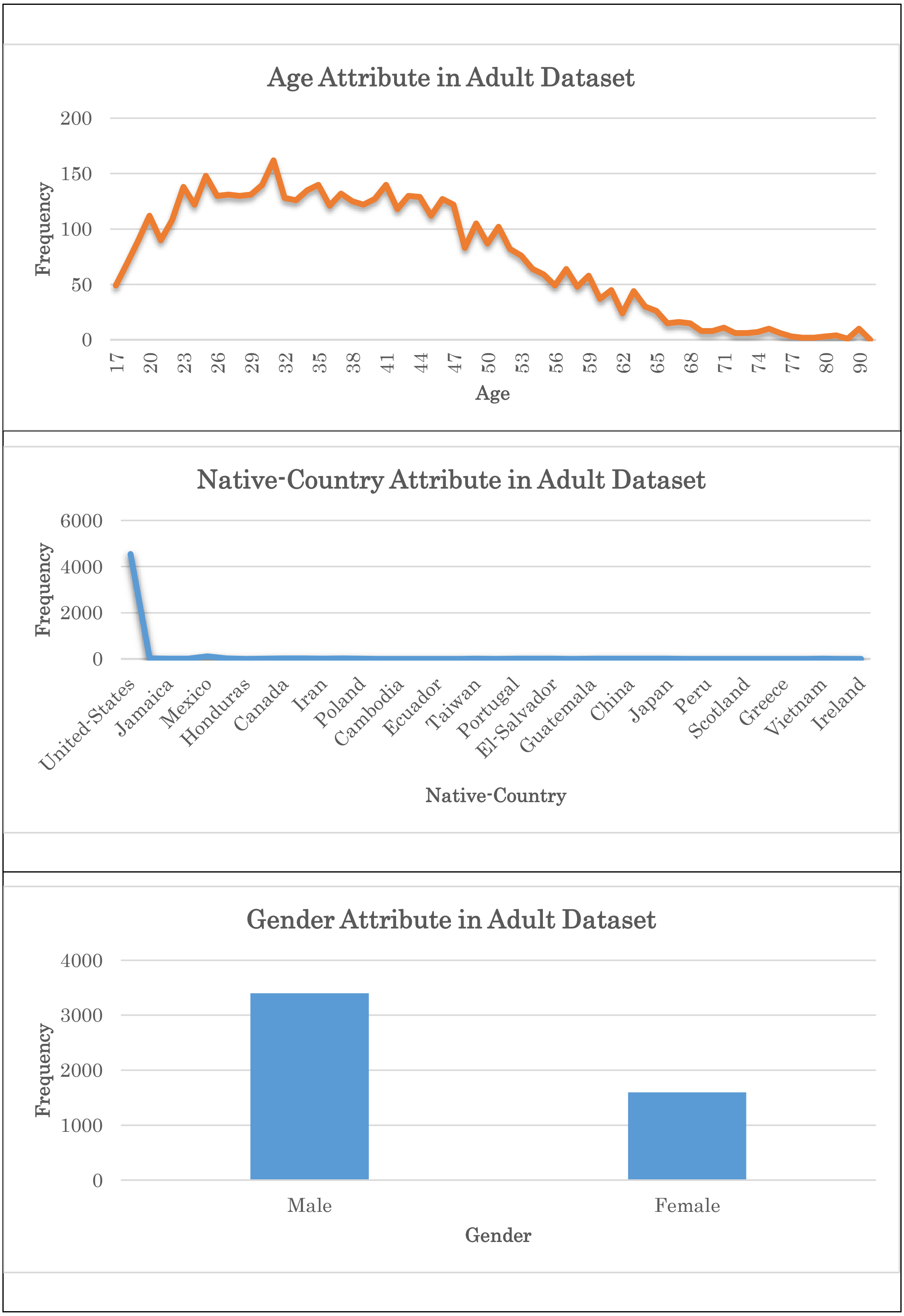


Figure 4-1: Frequency Distribution of Adult Dataset

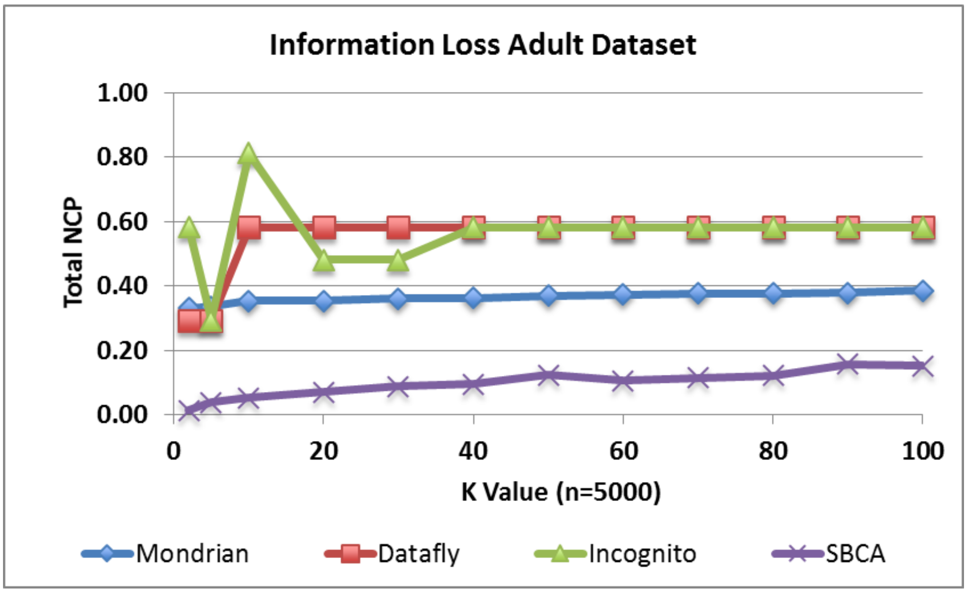
As it is presented, the frequency distribution of the 5000 tuples in Adult dataset over {Age and Native-Country} attributes have a very long tail. In addition in {Sex} attribute frequency of “Male” values is almost double of “Female” values. As a result of the long tail distribution in dataset there is a high possibility of outlier existence within the dataset [69] [70] . Having dataset with long tail distribution could be a very good opportunity to evaluate how different anonymization models are handling such situation. Regarding the values of attributes in dataset, {Age} attribute is a continuous attribute (numerical) and it ranges between 17 years old to 90 years old which is quite a wide range. The cardinality of {Gender} attribute is 2, with values of {Male, Female}. Regarding {Native-Country} the cardinality is 39. There are 39 different countries in this attribute. The highest population is for United-states and the rest of the countries have considerably smaller population as it is shown in Figure 4-1. The generalization hierarchy for {Gender and Native-Country} attributes is necessary for Mondrian, Incognito and Datafly models and provided in configuration file which is shown in Appendix section.

The {Salary} attribute is the private sensitive attribute. The values of {Salary} attribute are {‘<=50K’ and ‘>50K’} and they will not be modified during the anonymization process since {Salary} attribute is not considered a quasi-identifier attribute.

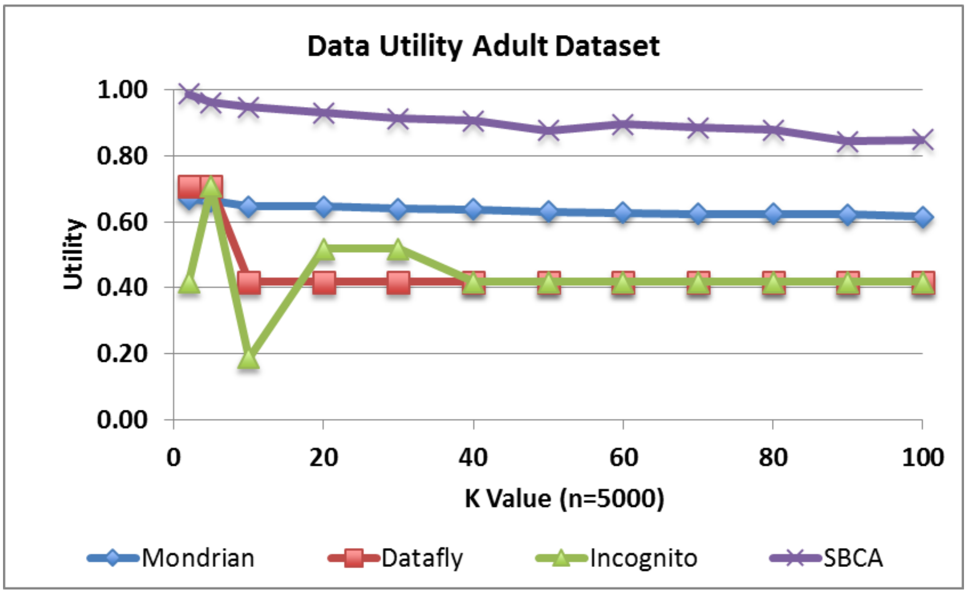
After providing the dataset and related information such as hierarchical structure for Mondrian, Incognito and Datafly algorithms [15] [16] [18] the privacy level or k value must be selected.

As “k” value represents the privacy level and due to the trade-off relationship between data privacy and data utility which was explained earlier all the algorithms are simulated over a range of “k” values starting with minimum value of k=2 all the way to k=100. From the theory and explanation which was given earlier regarding the trade-off relationship between data privacy and data utility, the information loss (total NCP) is expected to be minimum for k=2 and till k=100 the information loss should be having an increasing trend.

As it is illustrated in Figure 4-2, the total NCP of Similarity-Based Clustering Algorithm (SBCA) for the range of k values (k = 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100) is lower than other algorithms.



(a) Information Loss Measured by NCP



(b) Data Utility

Figure 4-2: Information Loss and Data Utility Comparison between Mondrian, Datafly, Incognito and SBCA on Adult Dataset

It is clear that Similarity-Based Clustering algorithm offers anonymization with much lower information loss while maintaining the same privacy level (k value) as other algorithms.

This advantage in reducing the information loss significantly while maintaining the anonymity level would result in a very high utility for anonymized-data. The comparison on utility of anonymized-data between SBCA, Mondrian, Incognito and Datafly algorithms are also shown in Figure 4-2. As it was expected by having the result on total NCP measurement, the utility of anonymized-data in Similarity-Based Clustering algorithm is much higher than other algorithms in the range of “k” values.

By looking at the results in Figure 4-2, the trade-off relationship between the privacy and data utility is clear that by increasing the “k” value, which is the privacy level, the utility is decreasing. However, in other algorithms the trade-off relationship is not very clear due to the overgeneralization or inefficient clustering that caused high information loss even in small “k” values. For instance in Incognito the Information loss in k=20 is much lower than k=10 even though the privacy level is increased the information loss dropped. This represents that the models such as Incognito do not work efficiently specially for small “k” values. In order to confirm the results on Adult dataset we will evaluate our algorithm and other algorithms on N. Corporation ISP dataset in the following section.

4.4. Simulation Result on N. Corporation ISP Dataset

N. Corporation is an Internet Service Provider (ISP) in Japan. As an ISP N. Corporation offers more than 2000 services online and obviously it collect and store a lot of different kinds of data such as search log data, registered customers log data and their credit card information. In a real situation as a case study N. Corporation needed to publish some of the data they have collected. This data actually tells about their customers and how much monthly they are paying to use the subscribed services provided by N. Corporation. Because there is some private sensitive information in the dataset this dataset obviously could not be published as it is an original dataset. It had to be anonymized for publication.

The dataset was given through a Non-Disclosure Agreement (NDA) with N. Corporation, which helped to evaluate our model using another real dataset in addition to Adult dataset.

The N. Corporation ISP dataset has 5750 tuples (data records) with four attributes. The quasi-identifier attributes are {Age, Gender and Location}. The monthly {Charge} of the service is the private sensitive attribute in this dataset. The frequency distribution of quasi-identifier attributes regarding the dataset is shown in Figure 4-3.

As it is shown in Figure 4-3 this dataset also has a long tail distribution and it is not normally distributed over {Age, Gender and Location}. By having the distribution of data over {Age} attribute apparently there are some flaws in original dataset, as there are some data records with Age value of 4 years old. The {Location} attribute is actually all in Japan and it is shown in coded values. N. Corporation provides the real Location values and its hierarchical taxonomy, which is shown in Appendix section. Regarding the {Gender} attribute, as it is shown most of the service subscriber seems to be “Male” than “Female”.

The range of {Age} attribute is from 4 years old to 99 years old. The cardinality of {Gender} attribute is 2 with values of {Male and Female}. {Location} attribute’s cardinality is 47 which includes are the provinces in Japan. The population in big cities is way more than the countryside locations therefore in Figure 4-3 the picks in {Location} attribute are definitely represents the big cities such as Tokyo and Osaka.

Same as previous section in Adult Dataset, the private sensitive attribute which is the monthly {charge} attribute will not be modified during the anonymization process, since it is not considered as a quasi-identifier attribute.

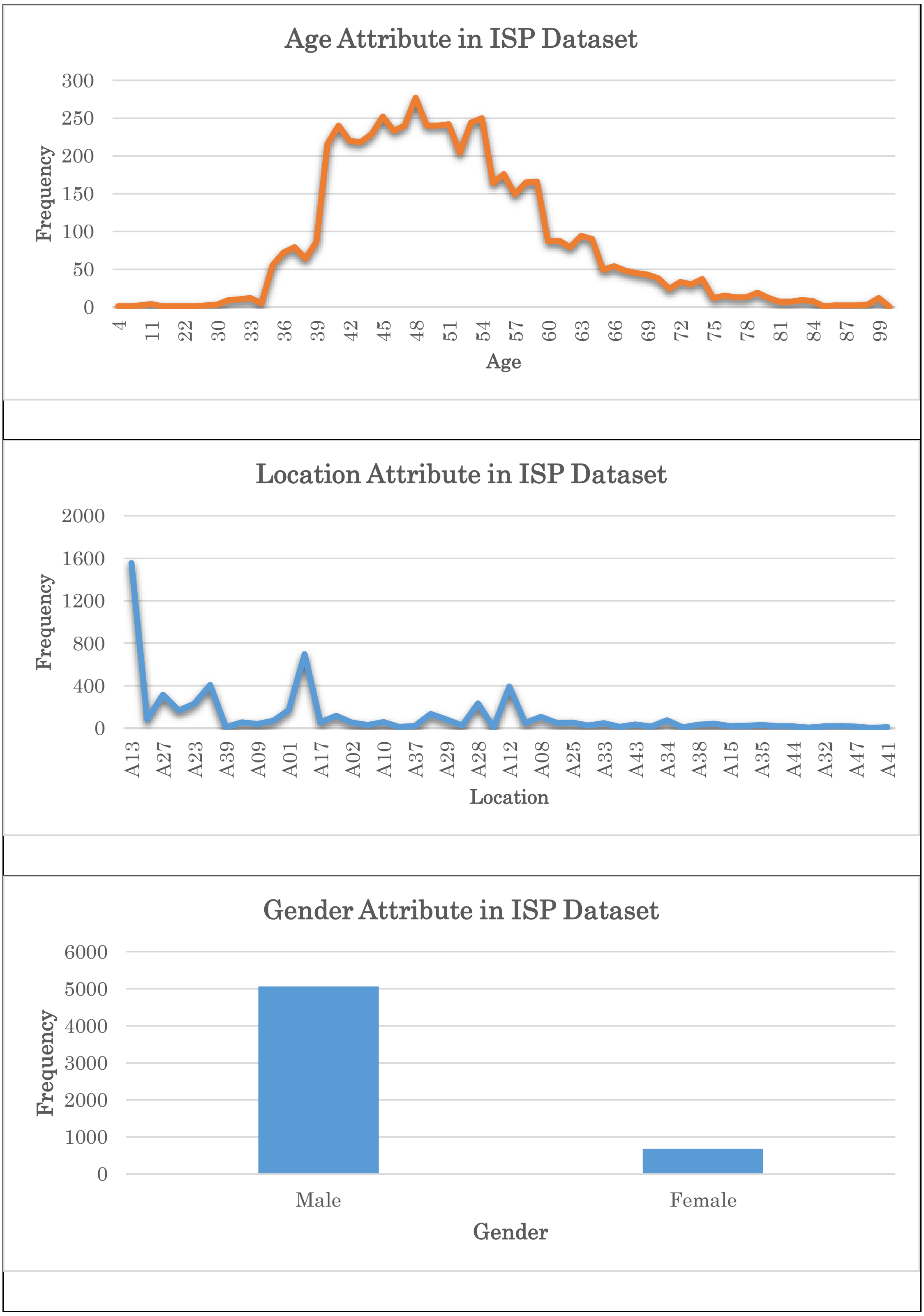
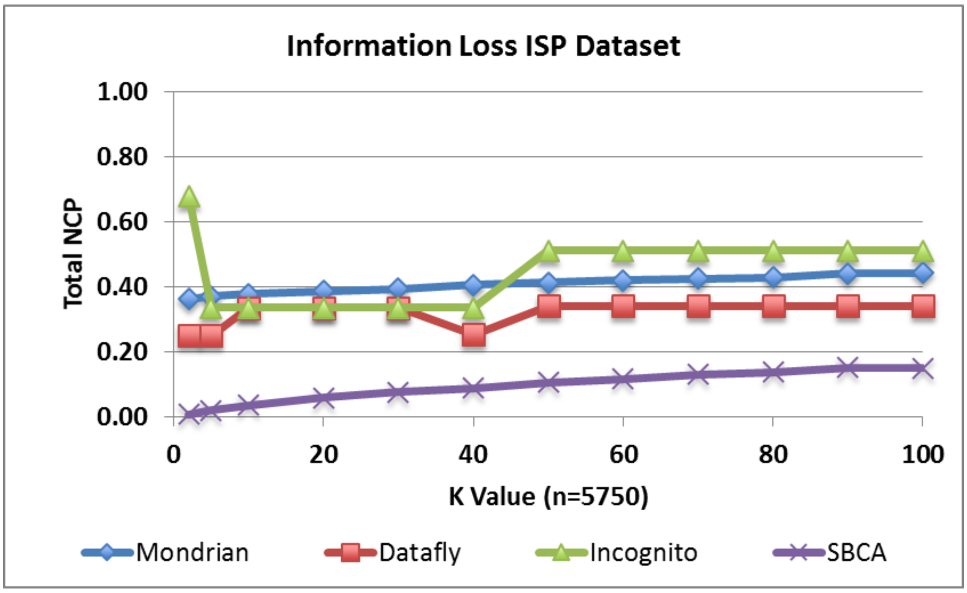
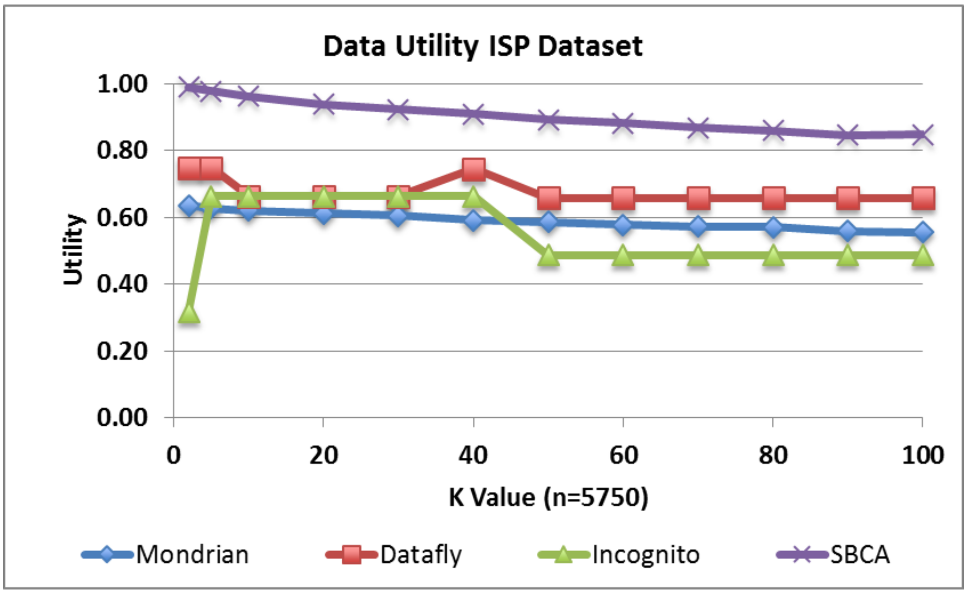


Figure 4-3: Frequency Distribution of N. Corporation ISP Dataset

After providing the dataset and related information such as hierarchical structure for Mondrian, Incognito and Datafly algorithms [15] [16] [18] the privacy level or k value must be selected. The information loss and Data utility is calculated for different anonymity degree (k = 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100) and the simulation result is shown in Figure 4-4.



(a) Information Loss Measured by NCP



(b) Data Utility

Figure 4-4: Information Loss and Data Utility Comparison between Mondrian, Datafly, Incognito and SBCA on N. Corporation ISP Dataset

As the results show in Figure 4-4 the total NCP of Similarity-Based Clustering Algorithm (SBCA) for the range of “k” values (k = 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100) is much less than other algorithms. Same as the result in Adult Dataset, it is clear that Similarity-Based Clustering Algorithm (SBCA) offers anonymization with much lower information loss by keeping the same privacy level (k value) as other algorithms. As the information loss is reduced significantly the utility of anonymized data is increased comparing to other algorithms as it is also shown in Figure 4-4.

By looking at the results in Figure 4-4 the trade-off relationship between the privacy and data utility is clear that by increasing the k value, which is the privacy level, the utility is decreasing. However, in other algorithms the trade-off relationship is not very clear. In algorithm like Mondrian the information loss is not really changing in the k value range. Changing k value from k=2 to k=100 does not really affect the information loss or data utility which could be due to the overgeneralization or inefficient clustering that caused high information loss even in small k values.

As it is shown in both dataset the information loss in the highest privacy level k=100 is less than 20% and the obtained anonymized data utility is 80% which is quite reasonable comparing to the results of other algorithms.

4.5. Analysis on Simulation Results

In this section the simulation results on Adult and ISP datasets are analyzed. As it was mentioned in previous section the information loss over various “k” values (k = 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100) for Similarity-Based Clustering Algorithm (SBCA) is maintained less than 20% which is considerably less than other algorithms. As a result of 0low information loss the measured data utility is maintained over 80% which is significantly higher than other algorithms as shown in Figure 4-2 and Figure 4-4.

One of the reasons that total Normalized Certainty Penalty (NCP) is lower in Similarity-Based Clustering Algorithm (SBCA) is efficient clustering and generalization that taking place in Similarity-Based Clustering Algorithm (SBCA). As it was mentioned before overgeneralization is one of the main reason of having huge information loss and low data utility after anonymizing dataset through k-anonymity model. In addition, how close the tuples are together in each cluster is also a very crucial point in minimizing information loss.

One way to investigate overgeneralization is by counting number of tuples in each group. K-anonymity definition is strictly states that each cluster must have at least “k” number of tuples. As for “k” value it is range between 2<= k <=n/2. It means the minimum “k” value is 2 and the maximum is half of the number of records which depends on the length of original dataset. Therefore it is desired to keep number of tuples in each group (cluster) equal to “k” value. For instance if “k” value is k=2 then it is desirable to keep all clusters a group of only two tuples if possible.

Table 4-1: Anonymized-data Analysis On k = 2 In Adult and ISP Datasets

Algorithm	Dataset	K-value	NO. of Unique Set of Results
Mondrian	Adult	2	209
	ISP		509
Incognito	Adult	2	22
	ISP		49
Datafly	Adult	2	16
	ISP		14
SBCA	Adult	2	582
	ISP		2016

For simulating other models, Mondrian, Incognito and Datafly algorithms [15] [16] [18] , UTD anonymization toolbox was used and unfortunately the information regarding clusters and number of tuples in each cluster could not be extracted [47] [26] . However in order to investigate possibility of over generalization from k-anonymous datasets number of unique combination of results could be obtained. Number of unique set of results is a proper indication of number of groups and tuple in each cluster. Having higher number of unique set of results means that there are more number of clusters and accordingly number of tuples in each cluster is closer (equal) to given "k" value.

Table 4-2: Anonymized-data Analysis On k = 50 and 100 In Adult and ISP Datasets

Algorithm	**Dataset**	**K-value**	**NO. of Unique Set of Results**
Mondrian	Adult	50	42
	ISP		66
Incognito	Adult	50	4
	ISP		4
Datafly	Adult	50	4
	ISP		8
SBCA	Adult	50	100
	ISP		115

(a) K value k=50

Algorithm	**Dataset**	**K-value**	**NO. of Unique Set of Results**
Mondrian	Adult	100	35
	ISP		36
Incognito	Adult	100	4
	ISP		4
Datafly	Adult	100	4
	ISP		8
SBCA	Adult	100	50
	ISP		57

(b) K value k=100

As it is shown in Table 4-1, number of unique set of results are counted for "k" value k=2 for Mondrian, Incognito, Datafly and SBCA algorithms for both Adult and ISP datasets. SBCA has the highest number of unique set of results for both Adult and ISP datasets which indicates there are more number of clusters created through SBCA and number of tuples in each cluster closer to given “k” value. Therefore the effect of overgeneralization is minimized and information loss due to over generalization is reduced.

In Table 4-1, “k” value is k=2 which is the minimum. Number of unique set of results for “k” value k=50 and k=100 on Adult and ISP dataset is investigated and the result is concluded and shown in Table 4-2 as follows.

As it is shown, SBCA has the highest number of unique set of results not only in small “k” value k=2, but also in k=50 and k=100. Therefore SBCA is more flexible in cluster and anonymizing the datasets which result in reducing information loss and producing high utility k-anonymous datasets.

5. Chapter 5: Conclusion

5.1. Summary of Main Results

In this research we have specifically mentioned the privacy issues in data publishing due to the existence of quasi-identifiers such as Zip Code and Gender attributes, which can link datasets together or in case of databases can link tables together. We have shown an example of privacy violation due to linking attacks [32] . Also we have mentioned about the benefits of data mining and knowledge discovery techniques and their applications in business, research and education in our fast growing digital world. It was mentioned that data mining and data privacy are in disagreement with each other and having more accurate data will result in better data mining analysis [44] . Therefore privacy preserving data mining concept was proposed [1] [3] . Also k-anonymity the general solution for protecting privacy of individuals, which is widely, is introduced. We have introduced the current main challenges in k-anonymity model from our point of view, which can be summarized as follow.

1) Information loss issue: By modifying the original data record to the more generalized form in k-anonymous dataset or suppressing the original data record, some information loss occurs. Information loss in k-anonymity model is an unfortunate and inevitable consequence. This information loss reduces the utility of anonymized-data and makes the anonymized-data to be less accurate and accordingly less useful for data mining and research purposes.

2) Real world dataset consist of categorical and numerical attributes: Real world and census datasets contain both numerical and categorical type data. As a matter of fact most of the QID attributes in micro-data are assume to be categorical with no hierarchical taxonomies [22] [66] . The combination of numerical and categorical attributes makes anonymization process rather complicated and very often results in an inefficient anonymization with very high information loss. Most of the previous approaches and techniques to achieve k-anonymity suffer from huge information loss and very low data utility (anonymized-data). Also most of the approaches are mainly designed for continuous numerical attributes and in case of considering categorical attributes, they depend on hierarchical taxonomies or require some additional information, which often are not defined or available in real life applications.

3) Trade-off relationship between data privacy and data utility: There is a trade-off relationship between the privacy level and the quality of anonymized-data. Choosing larger k value means providing higher privacy and consequently obtaining less utility k-anonymous dataset. Due to this trade-off, performing anonymization with maximum privacy and attaining maximum utility for anonymized-data is not possible. Also the problem of optimal k-anonymization and computational complexity of finding an optimal solution for the K-anonymity problem has been proven to be NP-hard.

We have mentioned that k-anonymity can be defined from clustering point of view as “clustering with constrain of minimum k tuples in each group”. In order to overcome these main challenges we have introduced a new model based on clustering. We introduced Similarity-Based Clustering Model (SBC). It is based on clustering and local recoding anonymization method. Similarity-Based Clustering Model concentrates on clustering the original dataset containing both numerical and categorical attributes efficiently based on given k value so after anonymization the information loss kept as small as possible. In this model specifically for categorical attributes new similarity measurement is introduced. Having the similarity between values in categorical attributes assists us to calculate the distance between different values in categorical attributes. Then along numerical attributes we can calculate the distances between tuples in dataset. By having the distances we can choose the closest tuple to be in the same group with each other. Having clusters that have very close (distance wise) or similar tuples to each other will reduce the information loss significantly when the clusters are anonymized through local recoding anonymization. Local recoding generalization for numerical attribute is the range of minimum value to maximum value in equivalent class and for categorical attribute is the collection of all distinct values in equivalent class.

The bottom-up greedy algorithm, which was introduced and evaluated on Adult and N. Corporation ISP dataset, showed that the information loss in Similarity-Based Clustering algorithm (SBCA) is significantly reduced comparing to other well-known algorithms. Therefore the resolved issues, which were resolved in this work, can be summarized as follows.

1) The information loss is significantly reduced which results in enhancing the data utility in anonymized data especially for small k value range (k value between 2 to 100).

2) The anonymization through Similarity-Based Clustering functions for datasets containing both numerical and categorical attributes. Unlike other models [14] [15] [17] [19] it does not depend on attribute hierarchical taxonomies. It can function independently as the distances are calculated based on the measured similarity based on the context and the observation probability of values in each attribute.

3) Regarding the trade-off relationship, as the result regarding Similarity-Based Clustering algorithm (SBC) shows by increasing the k value the information loss increases. K value represents the privacy level of the anonymized dataset and the dataset owner

selects it. As it was mentioned optimal k-anonymity problem is proven to be NP-hard problem therefore it cannot be stated that the information loss for the selected k value (for instance k=10) is the minimal information loss.

Regarding trade-off relation between data utility and data privacy, it is important to consider that defining those two terms are challenging. At the same time, because of the trade-off relationship and the responsibility of data publisher to protect the privacy of individual the k value must be decided and full field.

One suggestion which could be a help to resolve this particular issue is giving the opportunity to individuals to choose the desired privacy level and not the data publisher. The idea that we are suggesting is giving the opportunity to individuals to choose the privacy level and protection level they desire when they are actually sharing their data with data publisher. At the same time some method could be considered to encourage individuals to pick lower privacy levels and k values. Therefore the quality of anonymized data expected to be higher and individuals are actually chosen their desire privacy protection level.

5.2. Applications and Future Works

In privacy preserving data mining field there are research topics are ongoing. As our digital world is growing very fast people are getting more connected to each other through smart phones and tablets and at the same time the individuals are becoming more concern about their privacy. Having a highly accurate and efficient model for k-anonymization certainly increase the reliability on anonymization and the result, which is driven by working on anonymized data.

Generally one of the most used applications of k-anonymity is in data mining and data analysis field. For example in classification applications such a decision tree working on real data and anonymized data may result in totally different outcomes. The work [27] mentions about the application of anonymization in data mining applications. It states the fact that the range of applications is quite wide. Therefore knowing what kind of application the dataset is going to be used is helpful for anonymization process [40] [41] [42] [43] .

The examples we have given are mostly on medical data in order to show the real problems in real life applications and also present the meaning the private sensitive information. However now a days in the communication, smart devices (smartphones, tablets, Google glass), social networks [29] and Big Data era there are some other information which can be shared instantly. Location information is one of those, which were not possible to be shared right away, but nowadays with the help of the smart devices it is easily done. Therefore location privacy has become one of the main concerns of individuals specially the people who are very active on social networks [29] [27] and using smart devices as most of the applications trying to collect data as much as they can.

K-anonymity could be used to anonymize the location of individuals. So it can provide not very accurate locating information for service providers and at the same time protect the privacy of individuals who are using the application on their smart phone for example. Obviously there are challenges in specific application of k-anonymity, which need to be resolved, and it still requires more research. However real time anonymization could be very helpful for location privacy application.

We have mentioned about data mining earlier, however we did not cover so much on Big Data topic, which is a very hot topic at the moment. As the companies are storing more data on their customers and businesses dealing with Big Data in different applications (from social network to finance) has become challenging. Considering k-anonymization application on Big Datasets, considering Map-Reduce application on k-anonymity would be very interesting, as it will widen the range of applications of k-anonymization model.

Online advertising is another topic, which brings up some privacy concerns as the advertising agencies are tracking the users on each page and trying to bring them back on the shopping pages. Finally, in [36] k-anonymity model is utilized in Search Engine Marketing (SEM) and finally another application of k-anonymity could be on Cloud to increase the security of the Cloud. The process of obscuring published data to prevent the identification of key information. Data anonymization makes data worthless to others, while still allowing the Cloud owner to process it in a useful way [28] .

Data distribution plays an important role in applications such as clustering. Datasets with long tail distribution tend to have more outliers and managing the efficient clustering could be challenging. It was mentioned in previous chapter that two real datasets that we have evaluated out model on have long tail dataset, however it is not studied in details. This topic is interesting and we believe it could be studied more in the future.

6. Related Publications

Transactions and Journals:

[1] Mohammad Rasool Sarrafi Aghdam, and Noboru Sonehara, “Achieving High Data Utility K-Anonymization Using Similarity-Based Clustering Model”, IEICE Transactions on Information and Systems, Special Section on Security, Privacy and Anonymity of Internet of Things, VOL.E99-D, NO.8, PP. 2069-2078, August 2016. (Peer reviewed , Full paper)

[2] Mohammad Rasool Sarrafi Aghdam, and Noboru Sonehara, “ON ENHANCING DATA UTILITY IN K-ANONYMIZATION FOR DATA WITHOUT HIERARCHICAL TAXONOMIES”, International Journal of Cyber-Security and Digital Forensics (IJCSDF), Vol. 2, No. 2, PP. 12-22, July 2013. (Peer reviewed, Full paper)

International Conferences:

[1] Mohammad Rasool Sarrafi Aghdam, and Noboru Sonehara, “EFFICIENT LOCAL RECODING ANONYMIZATION FOR DATASETS WITHOUT ATTRIBUTE HIERARCHICAL STRUCTURE”, The Second International Conference on Cyber Security, Cyber Peacefare and Digital Forensic (CyberSec2013), PP. 130-140, March 2013. (Peer reviewed, Full paper)

Co-authored Conferences and Journals:

[1] Hidenobu Oguri, Noboru Sonehara, Kunio Matsui, Mohammad Rasool Sarrafi Aghdam, “An Efficient k-anonymization Algorithm by Predictive Model of the Power Approximation”, Journal of Information Processing Society of Japan, Volume 57, Issue 9, PP. 2034-2044, September 2016. (Peer reviewed, Full paper）

[2] Hidenobu Oguri, Noboru Sonehara, Kunio Matsui, Atsushi Kuromasa, Mohammad Rasool Sarrafi Aghdam, "A Proposal on Data Prediction and Evaluation Method, Using the Relationship between Division Number of Dataset and k-Anonymity", Computer Security Symposium 2015 Proceedings, Volume 2015, Issue 3, PP. 387-394, October 2015. (Not Peer reviewed, Research paper)

[3] Hidenobu Oguri, Noboru Sonehara, Kunio Matsui, Mohammad Rasool Sarrafi Aghdam, "The prediction of limit number of classification that satisfies k-anonymity, And suggestion of efficient k-anonymizing process", IPSJ SIG Technical Report, Vol.2015-SPT-13 No.3, PP. 1-8, May 2015.(Not Peer reviewed, Research paper)

Presentations and Technical Reports：

[1] Mohammad Rasool Sarrafi Aghdam, and Noboru Sonehara, “Anonymization Cost: Utility of Anonymized Data in Data Mining Applications 匿名化コスト：データ・マイニングアプリケーションにおける匿名化データの効用”, Poster exhibition, NII open house 2014. (Not Peer reviewed, Poster)

[2] Mohammad Rasool Sarrafi Aghdam, “Enhancing Data Utility in K-anonymization by Similarity-Based Clustering Model”, The 7th International Workshop on Information Systems for Social Innovation, Poster session February 2014. (Not Peer reviewed, Poster)

[3] Mohammad Rasool Sarrafi Aghdam, and Noboru Sonehara, “Privacy Preserving in data publishing, balancing Privacy Protection with Data Utilization An Efficient Local Recoding Anonymization in Data without Hierarchical Taxonomies”, Poster Exhibition, NII open house 2013. (Not Peer reviewed, Poster)

[4] Mohammad Rasool Sarrafi Aghdam, “Efficient Local Recoding Anonymization for Datasets without Attribute Hierarchical Structure”, International Workshop on Information Systems for Social Innovation, Session B privacy and life Log February 2013. (Not Peer reviewed, Presentation)

7. Appendix

7.1. Adult Dataset Configuration File for ToolBox

```
<?xml version="1.0"?>
<config method = " k = 100 ">
  <input filename='dataset/adult.data' separator=','/>
  <output filename='adult02_Anon_k.data' format ='genVals'/>

  <qid>
    <att index='0' name ='age'>
      <vgh value='[0:100)'>
        <node value='[0:50)'>
          <node value='[0:20)'>
            <node value='[0:10)'/>
            <node value='[10:20)'/>
          </node>
          <node value='[20:50)'>
            <node value='[20:30)'/>
            <node value='[30:40)'/>
            <node value='[40:50)'/>
          </node>
        </node>
        <node value='[50:100)'>
          <node value='[50:70)'>
            <node value='[50:60)'/>
            <node value='[60:70)'/>
          </node>
          <node value='[70:100)'>
            <node value='[70:80)'/>
            <node value='[80:90)'/>
            <node value='[90:100)'/>
          </node>
        </node>
      </vgh>
    </att>
    <att index='1' name='sex'>
      <map>
        <entry cat='Female' int='0' />
        <entry cat='Male' int='1' />
      </map>

      <vgh value='[0:1]'>
      </vgh>
    </att>


    <att index='2' name ='native-country'>
      <map>
        <entry cat='United-States' int='0' />
        <entry cat='Outlying-US(Guam-USVI-etc)' int='1' />
        <entry cat='Canada' int='2' />
        <entry cat='Mexico' int='3' />
        <entry cat='Honduras' int='4' />
        <entry cat='Guatemala' int='5' />
        <entry cat='Nicaragua' int='6' />
        <entry cat='El-Salvador' int='7' />
        <entry cat='Ecuador' int='8' />
        <entry cat='Peru' int='9' />
        <entry cat='Columbia' int='10' />
        <entry cat='Caribbean' int='11' />
        <entry cat='Puerto-Rico' int='12' />
        <entry cat='Dominican-Republic' int='13' />
        <entry cat='Jamaica' int='14' />
        <entry cat='Cuba' int='15' />
        <entry cat='Haiti' int='16' />
        <entry cat='TrinadadTobago' int='17' />
        <entry cat='France' int='18' />
        <entry cat='England' int='19' />
        <entry cat='Ireland' int='20' />
        <entry cat='Scotland' int='21' />
        <entry cat='Holand-Netherlands' int='22' />
        <entry cat='Italy' int='23' />
        <entry cat='Greece' int='24' />
        <entry cat='Portugal' int='25' />
        <entry cat='Yugoslavia' int='26' />
        <entry cat='Hungary' int='27' />
        <entry cat='Germany' int='28' />
        <entry cat='Poland' int='29' />
        <entry cat='Philippines' int='30' />
        <entry cat='Thailand' int='31' />
        <entry cat='Cambodia' int='32' />
        <entry cat='Vietnam' int='33' />
        <entry cat='Laos' int='34' />
        <entry cat='India' int='35' />
        <entry cat='Japan' int='36' />
        <entry cat='China' int='37' />
        <entry cat='South' int='38' />
        <entry cat='Hong' int='39' />
        <entry cat='Taiwan' int='40' />
        <entry cat='Iran' int='41' />
      </map>

      <vgh value='[0:41]'>
        <node value='[0:7]'>
          <node value='[0:2]'/>
          <node value='[3:7]'/>
        </node>
        <node value='[8:17]'>
          <node value='[8:10]'/>
          <node value='[11:17]'/>
        </node>
        <node value='[18:29]'>
          <node value='[18:22]'/>
          <node value='[23:25]'/>
          <node value='[26:29]'/>
        </node>
        <node value='[30:41]'>
          <node value='[30:34]'/>
```

```
          <node value='[35:35]'/>
          <node value='[36:40]'/>
          <node value='[41:41]'/>
        </node>
      </vgh>
    </att>
  </qid>
</config>
```

7.2. N. Corporation ISP Dataset Configuration File for ToolBox

```
<?xml version="1.0"?>
<config method = 'Mondrian' k = '100'>
<input filename='dataset/ N. Corporation ISP.data' separator=','/>
<output filename=' N. Corporation ISP_Anon_k100.data' format ='genVals'/>
<qid>

  <att index='2' name ='age'>
    <vgh value='[0:100)'>
      <node value='[0:50)'>
        <node value='[0:20)'>
          <node value='[0:10)'/>
          <node value='[10:20)'/>
        </node>
        <node value='[20:50)'>
          <node value='[20:30)'/>
          <node value='[30:40)'/>
          <node value='[40:50)'/>
        </node>
      </node>
      <node value='[50:100)'>
        <node value='[50:70)'>
          <node value='[50:60)'/>
          <node value='[60:70)'/>
        </node>
        <node value='[70:100)'>
          <node value='[70:80)'/>
          <node value='[80:90)'/>
          <node value='[90:100)'/>
        </node>
      </node>
    </vgh>
  </att>

  <att index='0' name='sex'>
    <map>
      <entry cat='Female' int='0' />
    <entry cat='Male' int='1' />
  </map>
  <vgh value='[0:1]'>
  </vgh>
</att>

<att index='1' name ='location'>
  <map>
    <entry cat='A01' int='0' />
    <entry cat='A02' int='1' />
    <entry cat='A03' int='2' />
    <entry cat='A04' int='3' />
    <entry cat='A05' int='4' />
    <entry cat='A06' int='5' />
    <entry cat='A07' int='6' />
    <entry cat='A08' int='7' />
    <entry cat='A09' int='8' />
    <entry cat='A10' int='9' />
    <entry cat='A11' int='10' />
    <entry cat='A12' int='11' />
    <entry cat='A13' int='12' />
    <entry cat='A14' int='13' />
    <entry cat='A15' int='14' />
    <entry cat='A16' int='15' />
    <entry cat='A17' int='16' />
    <entry cat='A18' int='17' />
    <entry cat='A19' int='18' />
    <entry cat='A20' int='19' />
    <entry cat='A21' int='20' />
    <entry cat='A22' int='21' />
    <entry cat='A23' int='22' />
    <entry cat='A24' int='23' />
    <entry cat='A25' int='24' />
    <entry cat='A26' int='25' />
    <entry cat='A27' int='26' />
    <entry cat='A28' int='27' />
    <entry cat='A29' int='28' />
    <entry cat='A30' int='29' />
    <entry cat='A31' int='30' />
    <entry cat='A32' int='31' />
    <entry cat='A33' int='32' />
    <entry cat='A34' int='33' />
    <entry cat='A35' int='34' />
    <entry cat='A36' int='35' />
    <entry cat='A37' int='36' />
    <entry cat='A38' int='37' />
    <entry cat='A39' int='38' />
    <entry cat='A40' int='39' />
    <entry cat='A41' int='40' />
    <entry cat='A42' int='41' />
    <entry cat='A43' int='42' />
    <entry cat='A44' int='43' />
    <entry cat='A45' int='44' />
    <entry cat='A46' int='45' />
    <entry cat='A47' int='46' />
    <entry cat='C13' int='47' />
    <entry cat='D01' int='48' />
  </map>

  <vgh value='[0:48]'>
    <node value='[0:23]'>
      <node value='[0:0]'/>
      <node value='[1:6]'/>
      <node value='[7:13]'/>
      <node value='[14:23]'/>
    </node>
    <node value='[24:48]'>
      <node value='[24:29]'/>
      <node value='[30:34]'/>
      <node value='[35:38]'/>
      <node value='[39:45]'/>
      <node value='[46:46]'/>
      <node value='[47:47]'/>
      <node value='[48:48]'/>
    </node>
  </vgh>
```

```
  </att>
</qid>
</config>
```

7.3. Example of Hierarchical Taxonomy (N. Corporation ISP Dataset Location Attribute)

L1-CODE	L1-J	L2-E	L3-E	L4-E
A01	Hokkaido	Hokkaido-Tohoku	East-Japan	Japan
A02	Aomori	Hokkaido-Tohoku	East-Japan	Japan
A03	Iwate	Hokkaido-Tohoku	East-Japan	Japan
A04	Miyagi	Hokkaido-Tohoku	East-Japan	Japan
A05	Akita	Hokkaido-Tohoku	East-Japan	Japan
A06	Yamagata	Hokkaido-Tohoku	East-Japan	Japan
A07	Fukushima	Hokkaido-Tohoku	East-Japan	Japan
A08	Ibaraki	Kanto	East-Japan	Japan
A09	Tochigi	Kanto	East-Japan	Japan
A10	Gunma	Kanto	East-Japan	Japan
A11	Saitama	Kanto	East-Japan	Japan
A12	Chiba	Kanto	East-Japan	Japan
A13	Tokyo	Kanto	East-Japan	Japan
A14	Kanagawa	Kanto	East-Japan	Japan
A15	Niigata	Tokai-Chubu	East-Japan	Japan
A16	Toyama	Tokai-Chubu	East-Japan	Japan
A17	Ishikawa	Tokai-Chubu	East-Japan	Japan
A18	Fukui	Tokai-Chubu	East-Japan	Japan
A19	Yamanashi	Tokai-Chubu	East-Japan	Japan
A20	Nagano	Tokai-Chubu	East-Japan	Japan
A21	Gifu	Tokai-Chubu	West-Japan	Japan
A22	Shizuoka	Tokai-Chubu	West-Japan	Japan
A23	Aichi	Tokai-Chubu	West-Japan	Japan
A24	Mie	Tokai-Chubu	West-Japan	Japan
A25	Shiga	Kansai	West-Japan	Japan
A26	Kyoto	Kansai	West-Japan	Japan
A27	Osaka	Kansai	West-Japan	Japan
A28	Hyogo	Kansai	West-Japan	Japan
A29	Nara	Kansai	West-Japan	Japan
A30	Wakayama	Kansai	West-Japan	Japan
A31	Tottori	Chugoku-Shikoku	West-Japan	Japan
A32	Shimane	Chugoku-Shikoku	West-Japan	Japan
A33	Okayama	Chugoku-Shikoku	West-Japan	Japan

A34	Hiroshima	Chugoku-Shikoku	West-Japan	Japan
A35	Yamaguchi	Chugoku-Shikoku	West-Japan	Japan
A36	Tokushima	Chugoku-Shikoku	West-Japan	Japan
A37	Kagawa	Chugoku-Shikoku	West-Japan	Japan
A38	Ehime	Chugoku-Shikoku	West-Japan	Japan
A39	Kochi	Chugoku-Shikoku	West-Japan	Japan
A40	Fukuoka	Kyusyu-Okinawa	West-Japan	Japan
A41	Saga	Kyusyu-Okinawa	West-Japan	Japan
A42	Nagasaki	Kyusyu-Okinawa	West-Japan	Japan
A43	Kumamoto	Kyusyu-Okinawa	West-Japan	Japan
A44	Oita	Kyusyu-Okinawa	West-Japan	Japan
A45	Miyazaki	Kyusyu-Okinawa	West-Japan	Japan
A46	Kagoshima	Kyusyu-Okinawa	West-Japan	Japan
A47	Okinawa	Kyusyu-Okinawa	West-Japan	Japan

8. Bibliography

[1] Sadiku, Matthew NO, Adebowale E. Shadare, and Sarhan M. Musa. "DATA MINING: A BRIEF INTRODUCTION." European Scientific Journal 11, no. 21 (2015).

[2] R. Agrawal and R. Srikant. “Privacy-preserving data mining,” ACM SIGMOD Record, vol. 29, 2000.

[3] L. Sweeney. “k-Anonymity: A Model for Protecting Privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, 2002, pp. 557–570.

[4] P.Golle. “Revisiting the Uniqueness of Simple demographics in the US Population,” Workshop on Privacy in the Electronic Society (WPES), October 30, 2006, Alexandria, Virginia, USA.

[5] Islam, Md Zahidul. "Privacy preservation in data mining through noise addition." PhD diss., University of Newcastle, 2007.

[6] P. Samarati and L. Sweeney, “Generalizing Data to Provide Anonymity when Disclosing Information,” Proc. ACM SIGMOD-SIGACT-SIGART Symposium (PODS), 1998.

[7] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. “Approximation algorithms for k-anonymity,” Journal of Privacy Technology, 2005.

[8] A. Gionis and T. Tassa. “K-Anonymization with Minimal Loss of Information ,” IEEE Transactions on Knowledge and Data Engineering,vol.21,NO 2, 2009.

[9] A. Meyerson and R. Williams. “On the complexity of optimal k-anonymity,” Proc. ACM SIGMOD-SIGACT-SIGART Symposium (PODS), 2004.

[10] G. Ghinita, P. Karras, P. Kalnis, and N.Mamoulis, “A framework for efficient data anonymization under privacy and accuracy constraints,” ACM Trans. Database Systems, vol. 34, 2009.

[11] A. Gionis, A. Mazza, and T. Tassa, “k-Anonymization revisited,” Proc. IEEE Int. Conf. on Data Eng. (ICDE), 2008.

[12] M. E. Nergiz and C. Clifton, “Thoughts on k-anonymization,” Journal of Data and Knowl. Eng., vol. 63, 2007.

[13] V. Iyengar, “Transforming data to satisfy privacy constraints,” Proc. Int. Conf. on Knowl. discovery and data mining (KDD), 2002.

[14] R. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymization,” Proc. IEEE Int. Conf. on Data Eng. (ICDE), 2005.

[15] K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Incognito: efficient full-domain k-anonymity,” Proc. SIGMOD, 2005, pp. 49–60.

[16] L. Sweeney,”Achieving k-anonymity privacy protection using generalization and suppression,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.

[17] P.Samarati,” Protecting Respondents’ Identities in Microdata Release,”IEEE Trans Knowledge and Data Eng., vol.13, no. 6, PP 1010-1027, NOV./Dec. 2001.

[18] K. LeFevre, David J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” Proc. IEEE Int. Conf. on Data Eng. (ICDE), 2006, pp. 25.

[19] X. Xiao and Y. Tao. “Anatomy: Simple and Effective Privacy Preservation,” In Proc. of VLDB, 2006, pp. 139–150.

[20] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu, “Utility-Based Anonymization Using Local Recoding,” Proc. Int. Conf. on Knowl. discovery and data mining (KDD), 2006, pp. 785–790.

[21] X.Xiao, Y.Tao, “Personalized privacy preservation,” SIGMOD '06 Proceedings of the 2006 ACM SIGMOD international conference on Management of data, Pages 229-240.

[22] Md. Nurul Huda, Shigeki Yamada and Noboru Sonehara, “On Enhancing Utility in K-Anonymization”, International Journal of Computer Theory and Engineering (IJCTE), Vol. 4, No. 4, August 2012.

[23] Willenborg, Leon, Waal, Ton de, “Elements of Statistical Disclosure Control”, Springer, 2001

[24] J. Han and M. Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, San Diego, CA 92101-4495,USA, 2001.

[25] C.Blake and C.Merz. UCI repository of machine learning databases, 1998.

[26] UTD Anonymization Toolbox, http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php

[27] Tian, Hongwei, and Weining Zhang. "Privacy-Preserving Data Publishing Based on Utility Specification." Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013.

[28] Jeff Sedayao, “Enhancing Cloud Security Using Data Anonymization”, IT@Intel White Paper, Cloud Computing and Information Security, June 2012

[29] Alufaisan, Yasmeen, and Alina Campan. "Preservation of centrality measures in anonymized social networks." Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013.

[30] Tassa, Tamir, Arnon Mazza, and Aristides Gionis. "k-Concealment: An Alternative Model of k-Type Anonymity." Transactions on Data Privacy 5.1 (2012): 189-222.

[31] Brickell, Justin, and Vitaly Shmatikov. "The cost of privacy: destruction of data-mining utility in anonymized data publishing." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.

[32] Friedman, Arik, Assaf Schuster, and Ran Wolff. "k-Anonymous decision tree induction." Knowledge Discovery in Databases: PKDD 2006. Springer Berlin Heidelberg, 2006. 151-162.

[33] Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." Security and Privacy, 2008. SP 2008. IEEE Symposium on. IEEE, 2008.

[34] Ayala-Rivera, Vanessa, Patrick McDonagh, Thomas Cerqueus, and Liam Murphy. "A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners." Transactions on Data Privacy 7, no. 3 (2014): 337-370.

[35] Kohlmayer, Florian, et al. "Flash: efficient, stable and optimal k-anonymity." Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE, 2012.

[36] Oguri, Hiroki, and Noboru Sonehara. "A K-Anonymity Method Based on SEM (Search Engine Marketing) Price of Personal Information." Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013.

[37] Gionis, Aristides, and Tamir Tassa. "k-Anonymization with minimal loss of information." Knowledge and Data Engineering, IEEE Transactions on 21.2 (2009): 206-219.

[38] Li, Jiuyong, et al. "Achieving k-anonymity by clustering in attribute hierarchical structures." Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, 2006. 405-416.

[39] He, Xianmang, et al. "Clustering-Based k-anonymity." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2012. 405-417.

[40] Li, Jiuyong, et al. "Anonymization by local recoding in data with attribute hierarchical taxonomies." Knowledge and Data Engineering, IEEE Transactions on 20.9 (2008): 1181-1194.

[41] Lindell, Yehuda, and Benny Pinkas. "Privacy preserving data mining." Journal of cryptology 15.3 (2002): 177-206.

[42] Vaidya, Jaideep, Christopher W. Clifton, and Yu Michael Zhu. Privacy preserving data mining. Vol. 19. Springer Science & Business Media, 2006.

[43] Aggarwal, C., and S. Phillip. "Privacy-Preserving Data Mining: Models and Algorithms, ch. A general survey of privacy-preserving data mining models and algorithms." (2008).

[44] Venkatasubramanian, Suresh. "Measures of anonymity." Privacy-Preserving Data Mining. Springer US, 2008. 81-103.

[45] Ciriani, Valentina, et al. "k-anonymous data mining: A survey." Privacy-preserving data mining. springer US, 2008. 105-136.

[46] Ienco, Dino, Ruggero G. Pensa, and Rosa Meo. "Context-based distance learning for categorical data clustering." Advances in Intelligent Data Analysis VIII. Springer Berlin Heidelberg, 2009. 83-94.

[47] UTD Data Security and Privacy Lab, http://www.cs.utdallas.edu/dspl/cgi-bin/index.php

[48] Mohammad Rasool Sarrafi Aghdam, and Noboru Sonehara, “Achieving High Data Utility K-Anonymization Using Similarity-Based Clustering Model”, IEICE Transactions on Information and Systems, Special Section on Security, Privacy and Anonymity of Internet of Things, VOL.E99-D, NO.8, PP. 2069-2078, August 2016.

[49] Mohammad Rasool Sarrafi Aghdam, and Noboru Sonehara, “ON ENHANCING DATA UTILITY IN K-ANONYMIZATION FOR DATA WITHOUT HIERARCHICAL TAXONOMIES” , International Journal of Cyber-Security and Digital Forensics (IJCSDF), Vol. 2, No. 2, PP. 12-22, July 2013.

[50] Mohammad Rasool Sarrafi Aghdam, and Noboru Sonehara, “EFFICIENT LOCAL RECODING ANONYMIZATION FOR DATASETS WITHOUT ATTRIBUTE HIERARCHICAL STRUCTURE”, The Second International Conference on Cyber Security, Cyber Peacefare and Digital Forensic (CyberSec2013), PP. 130-140, March 2013.

[51] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[52] Sadiku, Matthew NO, Adebowale E. Shadare, and Sarhan M. Musa. "DATA MINING: A BRIEF INTRODUCTION." European Scientific Journal 11, no. 21 (2015).Hand, David J., Heikki Mannila, and Padhraic Smyth. Principles of data mining. MIT press, 2001.

[53] Srinivas, K., B. Kavihta Rani, and A. Govrdhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering (IJCSE) 2, no. 02 (2010): 250-255.

[54] Act on Protection of Personal Information (APPI), http://law.e-gov.go.jp/htmldata/H15/H15HO057.html

[55] Loukides, Grigorios, and Jianhua Shao. "Data utility and privacy protection trade-off in k-anonymisation." In Proceedings of the 2008 international workshop on Privacy and anonymity in information society, pp. 36-45. ACM, 2008.

[56] Kim, Jay J., and William E. Winkler. "Masking microdata files." In Proceedings of the Survey Research Methods Section, American Statistical Association. 1995.

[57] Yancey, William E., William E. Winkler, and Robert H. Creecy. "Disclosure risk assessment in perturbative microdata protection." Statistics (2002): 01.

[58] Cormode, Graham, and Divesh Srivastava. "Anonymized Data: Generation, Models, Usage."

[59] HUANG, ZHEXUE. "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values." Data Mining and Knowledge Discovery 2 (1998): 283-304.

[60] Koperski, Krzysztof, Junas Adhikary, and Jiawei Han. "Spatial data mining: progress and challenges survey paper." In Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, pp. 1-10. 1996.

[61] Berkhin, Pavel. "A survey of clustering data mining techniques." In Grouping multidimensional data, pp. 25-71. Springer Berlin Heidelberg, 2006.

[62] Park, Hae-Sang, and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering." Expert Systems with Applications 36, no. 2 (2009): 3336-3341.

[63] Everitt, Brian S., Sabine Landau, Morven Leese, and Daniel Stahl. "Hierarchical clustering." Cluster Analysis, 5th Edition (2011): 71-110.

[64] El Emam, Khaled, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau et al. "A globally optimal k-anonymity

method for the de-identification of health data." Journal of the American Medical Informatics Association 16, no. 5 (2009): 670-682.

[65] Wang, Ke, Philip S. Yu, and Sourav Chakraborty. "Bottom-up generalization: A data mining solution to privacy protection." In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on, pp. 249-256. IEEE, 2004.

[66] Huda, Md Nurul, Shigeki Yamada, and Noboru Sonehara. "An Efficient k-Anonymization Algorithm with Low Information Loss." Recent Progress in Data Engineering and Internet Technology: 249.

[67] Tang, Qingming, Yinjie Wu, Shangbin Liao, and Xiaodong Wang. "Utility-based k-anonymization." In Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on, pp. 318-323. IEEE, 2010.

[68] Bedi, Rajneeshkaur, and Anjali Mahajan. "Identification of K-Tuples using K-Anonymity Algorithm to the Watermarking of Social Network Database." International Journal of Computer Applications 142, no. 14 (2016).

[69] Taha, Ayman, and Osman M. Hegazy. "A proposed outliers identification algorithm for categorical data sets." In Informatics and Systems (INFOS), 2010 The 7th International Conference on, pp. 1-5. IEEE, 2010.

[70] He, Zengyou, Shengchun Deng, and Xiaofei Xu. "An optimization model for outlier detection in categorical data." In International Conference on Intelligent Computing, pp. 400-409. Springer Berlin Heidelberg, 2005.

[71] Dewri, Rinku, Indrajit Ray, Indrakshi Ray, and Darrell Whitley. "On the Optimal Selection of k in the k-Anonymity Problem." In 2008 IEEE 24th International Conference on Data Engineering, pp. 1364-1366. IEEE, 2008.

[72] Ghinita, Gabriel, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. "Fast data anonymization with low information loss." In Proceedings of the 33rd international conference on Very large data bases, pp. 758-769. VLDB Endowment, 2007.