A COMPARISON OF CLUSTERING ALGORITHMS FOR DATA ANONYMIZATION

ZAHRA BINTI MAHMOUD

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

A COMPARISON OF CLUSTERING ALGORITHMS FOR DATA ANONYMIZATION

ZAHRA BINTI MAHMOUD

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTERS IN COMPUTER SCIENCE

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Zahra Binti Mahmoud

Matric No: WGA150007

Name of Degree: Masters in Computer Science

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

A Comparison of Clustering Algorithms for Data Anonymization

Field of Study: Information Systems

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

I. ABSTRACT

Organizations today can easily store massive amounts of data as the cost of storage has significantly plummeted over the years. Data is used to help them raise their brand's value. However, as data becomes easier to store in mass amounts, the security risk also increases. In the last two years alone, multiple data leaks have been reported, the latest being from the Ministry of Education in Malaysia. Over the years, there has been extensive research on data security. Literature review showed that many researches have employed methods such as data encryption or privacy protection data publishing (PPDP). This thesis focuses more on the latter, as data encryption has proven to be more costly. Many of the literature also focused on using generalization and suppression to achieve the level of anonymity it required. However, a heavily suppressed or generalized data may paint a different picture instead. The objective of this thesis is to find a method of data anonymization that is efficient and produces the least percentage of information loss. By comparing multiple different types of PPDP, the researcher then determined that the clustering method is the best fit for this purpose. Next, multiple types of existing clustering algorithms are compared to determine which has the best performance. The researcher then created an enhanced method to do a final comparison– the researcher manipulated the distance function to show how cluster distance difference can affect the outcome of the anonymized dataset.

II. ABSTRAK

Adalah lebih mudah kini bagi organisasi-organisasi untuk menyimpan data dalam jumlah yang besar, kerana harga penyimpanan menjunam dalam beberapa tahun lepas. Data bagi mereka merupakan alat untuk membantu meningkatkan lagi nilai jenama mereka. Namun semakin mudah untuk menyimpan jumlah data yang besar, semakin tinggi risiko keselamatannya. Dalam dua tahun lepas sahaja terdapat beberapa laporan kebocoran data, paling terbaharu dari Kementerian Pendidikan di Malaysia. Dalam tahun-tahun sebelum ini, penyelidikan tentang keselamatan data adalah lanjut dan mendalam. Kebanyakan kajian lepas menggunakan kaedah seperti enkripsi data ataupun privacy protection data publishing (PPDP). Tesis ini tertumpu kepada PPDP, kerana didapati enkripsi data adalah pilihan yang lebih mahal. Kajian lepas juga banyak menggunakan generalisasi dan penyekatan data untuk mencapai kadar anonimisasi yang diperlukan. Tetapi data yang terlebih disekat ataupun digeneralisasi mungkin akan memberi gambaran yang salah. Pertama sekali tesis ini bertujuan untuk mencari kaedah anonimisasi yang cekap dan menjana peratusan kehilangan maklumat yang paling minima. Dengan membandingkan beberapa jenis PPDP yang berlainan, penyelidik mendapati bahawa kaedah kluster merupakan pilihan yang paling sesuai untuk kegunaan ini. Kemudian penyelidik membandingkan beberapa jenis algoritma kluster yang sedia ada untuk menentukan jenis manakah yang mempunyai prestasi yang terbaik. Akhirnya, satu lagi perbandingan dibuat -- penyelidik memanipulasi fungsi jarak untuk menunjukkan bagaimana perbezaan jarak boleh memberi kesan kepada hasil data yang dianomisasi.

III. ACKNOWLEDGEMENT

I would like to thank my supervisor, Dr Norjihan Abdul Ghani, for her guidance, patience and support in supervising me throughout this dissertation. Her advice has guided me through the beginning to the completion of this project.

Special thanks to my family and friends who was always there to support me and to give me ideas and advices. I appreciate all the comments, suggestions and encouragement throughout this project.

Thank you all who has been helping me throughout this dissertation.

IV. TABLE OF CONTENTS

LIST OF FIGURES

V. LIST OF TABLES

VI. LIST OF APENDICES

A. Keyword Combination and Searches Across Digital Database
B. Quality Assessment Form
C. Literature Review Results
D. Core studies/papers
E. Examples of Clustering Techniques

1 INTRODUCTION

Life after the advent of the Internet now means every person leaves a trail of digital exhaust. Every day, people churn out phone records, text messages, GPS data, browser history, email, tweets, Facebook status or Instagram posts – all of which will live forever even when deleted on their devices (Goodman, 2015). This amount of information not only helps companies in finding new customers, they also help to configure their current customers' preferences with laser-like accuracy. Data leakage used to only be possible when there are SQL injections, hacks, malwares, or trojans – a person is only hackable if they click on a link or a USB is inserted into a device. Fast forward to now, we have a group of companies most people have not heard of: data brokers. Data brokers are part of the data surveillance industry worth approximately US$156 billion a year (Goodman, 2015). For comparison, consider how Edward Snowden in June 2013 revealed to the world the size and scope of NSA's surveillance operations, shocking global citizens. However, note that the revenue of the data broker industry is twice the size of the US government's intelligence budget. To put it in perspective, the tools, techniques and infrastructures owned by the private sector can put government agencies to shame. These capabilities, in fact, allow them to extensively peer into an individual's personal life.

In these early decades of the information age, the flow of information is becoming more and more central to our daily lives. It has therefore become important that information transmission be protected against eavesdropping (as, for example, when one sends credit card information over the Internet) and against noise (which might occur in a cell phone transmission, or when a compact disk is accidently scratched) though most of us depend on schemes that protect information in

these ways, most of us also have a rather limited understanding of how this protection is done. Part of the aim of this dissertation is to introduce the basic concepts underlying this endeavour.

1.1 **Research Background**

Data brokers operate on information – they learn about us from our Internet service providers, mobile phone companies, banks, credit card issuers, credit bureaus, pharmacies, departments of motor vehicles, and even grocery stores. Furthermore, they are increasingly leveraging on our online activities – an individual's social network contains vital information in the form of a Like, Facebook poke, or even tweets. This information is tagged, geo-coded, and sorted for resale to advertisers and marketers. The business is so lucrative that even old-world retailers are now finding out about the potential of a secondary source of revenue, which is their customer data. They may find the data to be even more valuable than their primary product or service for sale. Many companies have begun to shift their data infrastructure from a cost center to a profit center, to tap into this new revenue stream. While credit bureaus like Experian or Equifax aren't new to most people, more and more new firms are now able to capture a massive amount of data on an individual, thanks to an increasingly connected online lifestyle.

The Acxiom Corporation of Little Rock from Arkansas, as an example, are constantly "collecting, collating and analyzing" more than 50 trillion unique data transactions each year, operating with more than twenty-three thousand computer servers. Acxiom had gathered the profiles of over 700 million consumers globally, and a staggering 96 percent of American households are represented in its data banks. Over fifteen hundred specific traits can be contained in just one profile, with information like race, gender, phone number, type of car driven, education level, number of children, the square footage of their home, portfolio size, recent purchases, age,

height, weight, marital status, politics, health issues, occupations and right or left-handedness, as well as pet-ownership down to its breed.

Data brokers like Acxiom aims to provide “behavioral targeting”, or what is alternatively called “predictive targeting” or “premium proprietary behavioral insights” on an individual. With their data banks, the individual can be understood with such precision that data brokers are able to sell the information they aggregate at the highest price to their buyers – usually marketers, advertisers and other companies using them for decision-making purposes.

The value of behavioral targeting is so high because it offers accuracy. Take for example, mass advertising a Pampers ad to a nineteen-year-old male college student. This would be a waste of a marketing budget. However, present the same information to a thirty-two-year-old pregnant housewife and it may just bring in hundreds of dollars of sale. As this example illustrates, data brokers are always segmenting people into groups or profiles, which only gets more and more specific. Doing this helps to maximize the value of digital intelligence they have collected.

In another example, on October 2017, a massive data leakage from several telco networks were published for sale on a Malaysian online forum, Lowyat.net. The leakage happened in 2012-2015 from a list of Malaysian telecommunication companies. Some well-known and huge companies were involved, including Maxis, Celcom, Digi and UMobile (BBC, 2017. Users who found their details on the online forum were incensed, but only months after the first incident, the list of organ donors in Malaysia were leaked in another data breach.

The most recent data breach had been in the Ministry of Education of Malaysia. Not only did this data leakage reveal student information such as IC (Identification Card) numbers, it also revealed parents’ and teachers’ information, as well as the relationship of an individual to another.

This was even worse than the previous telco breach, even though there were a smaller amount of people involved (Rozario, 2018).

These examples were some of the headlines in data breaches in Malaysia, happening in a span of only two years. They became the primary motivation for this research, as this study believes while there exist security systems for data protection, what is needed is a security system that protects data itself in the event of a breach.

1.2 Problem Background

In 2002, Sweeney proposed a method of anonymizing data. The method was ever-growing at the time, called k-anonymity. In this method of data protection, she introduced two ways of protecting data, using generalization and suppression. Generalization represents data in a categorical feature, while suppression stops any data from being released (Sweeney, 2002). However, over the years, data are also being sold and bought by researchers for their studies. If data is heavily generalized or suppressed, it no longer represents a clear picture for researchers to draw or form conclusions with or to prove their hypotheses.

When a set of data is collected, the k-anonymity protection model can classify the data into different types of identifiers, for example, sensitive attributes and quasi-identifiers. Quasi-identifiers are identifiers that if published, cannot be used to directly link to a person. This will be explained further in Chapter 2.

In 2006, Sweeney further developed the k-anonymity protection model and a technique called l-diversity was introduced. A year later, an enhanced version of l-diversity, called t-closeness, came into play.

Even with the advances in data protection, multiple breaches still resulted in data being leaked. Aside from that, the industry also started selling data to groups of professionals for profit. Thus in 2010, the Personal Data Protection Act (PDPA) released several guidelines for how these data can be used or re-used, as well as guidelines to regulate data collection, purchase and other legal concerns.

Despite all the progress in data protection, generalization and suppression remained the main method of execution.

> "*k*-anonymization techniques have been the focus of intense research in the last few years. An important requirement for such techniques is to ensure anonymization of data while at the same time minimizing the information loss resulting from data modifications." (Byun et al., 2007)

The idea to minimize data loss is from the fact that currently, all data that are analysed and collected can be published and sold for other use (repurposing). If a research requiring a certain level of data accuracy uses data that are heavily generalized or suppressed, the results of the research may also be imprecise, to a certain level.

1.2.1 Problem Statement

In most concepts of data anonymization, data are anonymized by categorizing a certain field or removing the more sensitive information. However, these resulted knowledge gained from analysing the datasets are no longer accurate. Therefore, there is a need to enhance the current anonymization techniques without losing most of its valuable information, while ensuring the efficiency of the techniques.

1.3 Research Questions

Based on the problem statement, the purpose of this research intends to answer three main questions. This work aims to introduce an enhanced algorithm for clustering big data for anonymization. These questions help define the goal and objective of this research. The research questions are as follows:

1. What are the current methods for anonymizing datasets?
2. Which method is most efficient in anonymizing datasets with minimal data loss?
3. Can the method be enhanced in terms of information loss and running time?

1.4 Research Objectives

According to the problem statement and research questions, the objectives of this research are as follows:

1. To investigate the current methods used in data anonymization.
2. To design and develop an enhanced method of data anonymization.
3. To evaluate the enhanced data anonymization in terms of information loss and efficiency.

1.5 Research Scope

This research will propose an enhanced method that can be used to run data anonymization in three different ways that will result in three different results – one that users can compare in terms of information loss or efficiency. The prototype built on that method can cater to three different methods of anonymization that can be used for comparison, on top of getting different degrees of information loss on the data that has been anonymized.

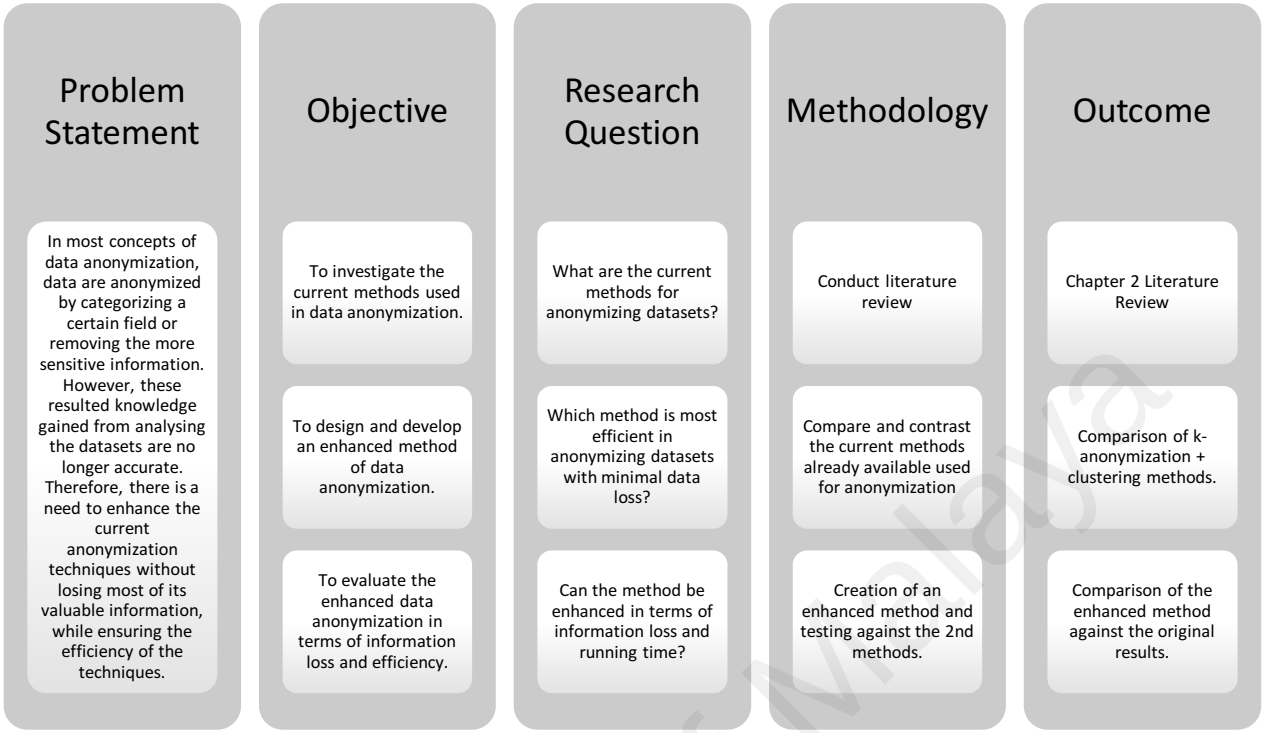

Figure 1.1 Research Questions to Objective Map

1.6 **Organization of Thesis**

This thesis consists of six chapters. Chapter 1 will describe the research background, problems, research questions and the scope of the research. Chapter 2 will discuss the literature on anonymization techniques and algorithms. Chapter 3 will explain how the research was carried out, as well as methods of data collection. Chapter 4 discusses the data analysis and the method of how the method was created. Chapter 5 discusses the results and discussion that was made from the prototype created, and finally, Chapter 6 discusses the conclusion.

2 LITERATURE REVIEW

Everyday our devices, sensors, and networks create new types of data en masse. Furthermore, the cost of storing these data has become negligible. On the flip side, there is growing public interest and demand in the reuse of these 'open data'. While the use of these open data can bring clear benefits for society at large, individuals, and organizations, it can only do so if each individual's rights are respected to protect their personal data and private life.

To keep the benefits and mitigate risks, anonymization can be a good strategy. To clarify, a dataset that has been truly anonymised to the point that individuals are no longer identifiable would no longer be compliant to European data protection laws. The task to create the underlying information to anonymize, however, is not a simple preposition, as previous case studies and research publications has demonstrated. Take for example, a dataset which is considered as anonymized can be combined with another dataset to identify one or more individuals.

This research discusses the topic of data anonymization in response to the need for privacy. Data privacy, or information privacy is an aspect of information technology that refers to an organization or an individual's ability to determine which data can be shared with third parties.

As discussed in Chapter One, the need for privacy has risen significantly as data grows exponentially bigger over the past years, and will continue to grow. The plummeting cost of storage for data also makes it more justifiable for companies and organizations to keep data rather than discard them after a predetermined amount of time. This practice leads to easier leaks without any proper security.

Encryption keys or network security works to protect data at the first level (against leaks), however there needs to be a contingency that protects the data itself if that fails. Data anonymization methods will be discussed further below.

There are two approaches to anonymization: the first is based on randomization while the second is based on generalization. This chapter will cover both this approach as well as what succeeds them.

Randomization is a family of techniques that changes the data's veracity, to remove the strength of the link between data and individual. Generalization, on the other hand, generalizes or "dilutes" the attributes of the data subjects – it does this by modifying the respective scale of the data or order of its magnitude. This chapter also covers other topics such as Anatomy and Clustering, the former is a successor of the generalization method, however approaches the subject quite differently and the latter is a current popular way of anonymizing datasets with minimal information loss.

2.1 **PDPA2010**

The Malaysian Personal Data Protection Act 2010 came into effect in 2013. The implication it brought to Malaysian businesses were added requirements and responsibilities relating to their employees', suppliers', and customers' personal data.

The "personal data" referred to in the Act generally means information from a subject that can be identifiable from it. This definition, therefore, covers important data like names, identity card numbers, phone numbers, and passport numbers. It could also include sensitive personal data like health or medical records of the subject, even their political leanings, religion, and criminal records.

Under the PDPA 2010, data users must comply with seven Personal Data Protection Principles. According to the requirements:

1. General: Only with the data subject's consent can Personal Data be processed.

2. Notice & Choice: Data subjects are to be informed of the type of data being collected, its purpose, sources, and the right to request access and correction, as well as the option and means by which the data subject is able to limit processing of their personal data – using written notice among other methods.
3. Disclosure: Without the data subject's consent, Personal Data may not be disclosed for any purpose other than which the data was disclosed at the time of the collection, or to any person other than that notified to the data user.
4. Security: The data subject is encouraged to take the practical steps in protecting their personal data from loss, misuse, modification, or unauthorized access or disclosure, alteration or destruction.
5. Retention: Data subject's personal data must not be kept longer than necessary once its purpose is fulfilled
6. Data Integrity: Data users are required to take reasonable steps in ensuring their personal data is kept as accurate, complete, not misleading, and up to date as possible
7. Access: Access to their personal data, as well as the ability to correct any inaccurate, incomplete, misleading, or outdated data, must be given to data subjects Data Anonymization

Data anonymization is essentially a type of 'sanitization' of information, aimed to protect privacy. The process involves removing personally identifiable information from data sets by encryption or removal, in order for individuals described in the data to remain anonymous – while reducing the risk of unintentional disclosure or exposing it to a situation that may enable evaluation and analytics post-anonymization (Rajendra, 2019).

An anonymized data is data from which a person can't be identified by the receiver of the information. Such data can be disseminated either in macro or microdata form.

Macrodata represents statistics of interests calculated over a sample population with aggregate values (De Capitani di Vimercati et al., 2015). They are measures that summarize one or more values of a respondent's properties or attributes (for example, individuals or organizations). Microdata, on the other hand, are specific data that relate to individual respondents.

A macro or micro data release may cause a leak of sensitive information not intended for disclosure. In this study, the researcher is more concerned with protecting microdata, as they are more vulnerable and pose higher risk of privacy breaches. It specifically needs to be protected from identity and attribute (respondent's sensitive information) disclosure.

Attributes in a microdata table can be categorized into four classes: identifiers, quasi-identifiers, sensitive attributes, and non-sensitive attributes. The attributes that univocally identify respondents, such as identification card or phone numbers are called identifiers. Attributes that may be linked to information from external sources are called quasi-identifiers, used to lessen the uncertainty over the identity of respondents – examples include date of birth, sex, or ZIP code number. Sensitive attributes are the remaining sensitive information in the table. Meanwhile non-sensitive attributes are any leftover information not categorized by the previous 3 classes (De Capitani di Vimercati et al., 2015).

To protect a microdata table, we first need to remove or encrypt explicit identifiers. However, a de-identified microdata table still may not guarantee complete anonymity because quasi-identifiers may allow respondents to be identified with its link to publicly available information (Ciriani, 2007).

For the purpose of this research, the table below will show an example of some attributes and where they stand as an identifier, as well as the general ways used to anonymize each table/cell.

Table 2.1 Identifier Types(taken from [https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html)](https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html)

Identifier Type	**Direct Identifier**	**Strong Direct Identifier**	**Indirect Identifier**	**Sensitive Attribute**	***Anonymization Method***
Personal Identification Number (IC Number)	X				*Remove*
Passport Number	X				*Remove*
Birth Certificate Number	X				*Remove*
Full Name	X				*Remove*
Email Address	X	X			*Remove*
Phone Number		X			*Remove*
Postal Code			X		*Remove/ Categorize*
District/Part of Town			X		*Categorize*
Municipality of Residence			X		*Categorize*
Region			X		*Categorize*
Major Region			X		*(Categorize)*
Municipality Type			X		*(Categorize)*
Audio File	X				*Remove*
Video File displaying person(s)	X				*Remove*
Photograph of person(s)	X				*Remove*
Year of Birth		X			*Categorize*
Age			X		*Categorize*
Gender			X		
Marital Status			X		
Household Composition			X		*Categorize*
Occupation		(X)	X		*Categorize*
Employment Industry			X		
Employment Status			X		
Education			X		*Categorize*
Field of Education			X		
Mother Tongue			X		*Categorize*
Nationality			X		*(Categorize)*
Workplace/Employer		(X)	X		*Categorize*
Vehicle Registration Number		X			*Remove*
Web Page Address		(X)	X		*Remove*
Student ID Number		X			*Remove*
Insurance Number		X			*Remove*
Bank Account Number		X			*Remove*
IP Address		X			*Remove*
Health-related information		(X)	X	X	*Categorize/ Remove*
Ethnic Group		(X)	X	X	*Categorize/ Remove*
Crime or Punishment		(X)	X	X	*Categorize/ Remove*
Political or Religious Alliance			X	X	*Categorize*

Table 2.1 shows a common guideline on how certain attributes are anonymized, mostly in terms of generalization and suppression. The following sections will be structured in the following way – the researcher first talks about identity disclosure where it will be an opening to k-anonymity, followed by attribute disclosure and the methods of protecting it, l-diversity and t-closeness.

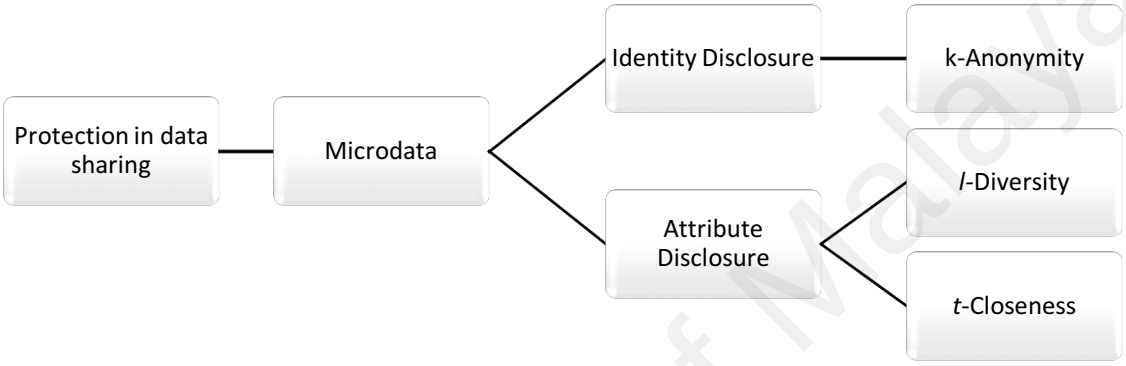

Figure 2.1 Solutions that will be illustrated in this section

The chapter continues the discussion with anatomy, differential privacy and clustering, as other methods of PDPP. The next section will describe how the identity disclosure and attribute disclosure looks like before talking about the concepts of k-anonymity, l-diversity, and t-closeness to make the transition flows more smoothly.

2.1.1 Identity Disclosure

Identity disclosure, also known as re-identification, refers to how an individual can be linked to a specific data entry (Benitez, 2010). This is an attack most serious since it carries legal consequences for data owners, as stipulated in many laws and regulations around the world. The definition itself suggests that an attacker will be able to learn about sensitive information in the data entry pertaining to the individual. To illustrate such scenario, observe Table 2.2, 2.3 and 2.4.

Table 2.2 Relational Table (Ticket Purchasing System)

Name	Phone Number	DoB	Sex	ZIP	Ticket Type	Payment
Hani	0198234578	1989/05/16	F	40000	Normal	CC
Zahra	0133221158	*1995/07/21*	*F*	*40150*	*Student*	*Cash*
Bob	0187623748	1990/12/01	M	60000	Normal	DC
Fara	0165472293	1956/05/22	F	34580	Senior Citizen	Cash
Azam	0198989324	1944/01/03	M	80000	Senior Citizen	Cash
Haziq	0122244896	1980/04/08	M	23450	Normal	CC
Emily	0133339876	1996/09/20	F	45670	Student	Cash

Table 2.3 De-identified version of a relational table 2.2

Name	Phone Number	DoB	Sex	ZIP	Ticket Type	Payment
		1989/05/16	F	40000	Normal	CC
		1995/07/21	*F*	*40150*	*Student*	*Cash*
		1990/12/01	M	60000	Normal	DC
		1956/05/22	F	34580	Senior Citizen	Cash
		1944/01/03	M	80000	Senior Citizen	Cash
		1980/04/08	M	23450	Normal	CC
		1996/09/20	F	45670	Student	Cash

Table 2.4 Malaysia voters' list.

Name	*Address*	*City*	*ZIP*	*DoB*	*Sex*
…	…	…	…	…	…
Zahra Mahmoud	*50, JSP U6/6, DSP*	*Shah Alam*	*40150*	*1995/07/21*	*F*
…	…	…	…	…	…

Based on Table 2.3 and 2.4, it is apparent that despite a de-identified table that removed names and phone numbers, it can still be linked to a public voters' list, including a single tuple

related to a female, with a home in the 40150 area, birthday 21st of July 1989. These combination of values (as long as unique in the external table too) uniquely identifies the corresponding tuple in the microdata table as an individual named Zahra Mahmoud, staying at JSP U6/6 and that she is a student.

To protect against a linking attack as just described, k-anonymity would require that any released tuple be indistinguishably related to a minimum number k of respondents (De Capitani di Vimercati et al., 2015). Since re-identification using linking attacks exploit quasi-identifying attributes, the requirement is translated as such: *Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents* (Samarati, 2001).

2.1.1.1 *k*-Anonymity

k-Anonymity works to adopt generalization and suppression techniques solely on the quasi-identifier attributes, which leaves sensitive and non-sensitive attributes untouched. Generalization will substitute original values with a general or categorical value. For example, to only release the year of birth instead of the full date. Meanwhile suppression removes information – it is especially handy in reducing generalization that may be required to guarantee *k*-anonymity when a small number of outliers (which are quasi-identifying values with less than *k* occurrences) would require more generalization. Both generalization and suppression may apply at different granularity levels, and approaches to this was proposed by the combination of the two in different ways. In fact, most available solutions are reliant on attribute generalization and tuple suppression.

Table 2.5 An example of a 3-anonymous table

Name	Phone Number	DoB	Sex	ZIP	Ticket Type
		1989/12/**	F	40***	Normal
		1989/12/**	F	40***	Normal
		1989/12/**	F	40***	Normal
		1956/05/**	M	80***	Student
		1956/05/**	M	80***	Senior Citizen
		1956/05/**	M	80***	Normal
		1996/09/**	F	45***	Student
		1996/09/**	F	45***	Normal
		1996/09/**	F	45***	Senior Citizen

Table 2.6 An example of a 2-diverse table

Name	Phone Number	DoB	Sex	ZIP	Ticket Type
		1989/**/**	F	401**	Normal
		1989/**/**	F	401**	Normal
		1989/**/**	F	401**	Normal
		1956/**/**	M	804**	Student
		1956/**/**	M	804**	Senior Citizen
		1956/**/**	M	804**	Normal
		1996/**/**	F	450**	Student
		1996/**/**	F	450**	Normal
		1996/**/**	F	450**	Senior Citizen

Table 2.5 depicts a 3-anonymous microdata table extracted from the Table 2.2. Note that the Payment attribute has been suppressed, because it isn't meant to be released. The quasi-identifiers in the table are attributes Date of Birth (DOB), Sex, and ZIP. Ticket Type is categorized as sensitive – typically, the museum is not authorized to disclose this information. By generalizing

several attributes, the 3-anonymous table was created – DOB only released the month and year of birth, also, just the first two digits of the ZIP was released. Any outlier tuple that did not satisfy the rule was suppressed.

By reducing the anonymized table's details, *k*-anonymity naturally leads to information loss. In order to achieve a balanced trade-off between data protection and the data's utility on the hands of recipients, computing a *k*-anonymous table is vital to minimize generalization and suppression.

2.1.2 Attribute Disclosure

While k-anonymity may be highly effective in protecting individuals' identities, it doesn't protect them from attribute disclosure (Domingo-Ferrer & Soria-Comas, 2015). To protect the association between respondents' identities and the values of their sensitive attributes, extending *k*-anonymity as an alternative has been proposed. Two well-known solutions are *l*-diversity and *t*-Closeness, both of which prevents attribute disclosure.

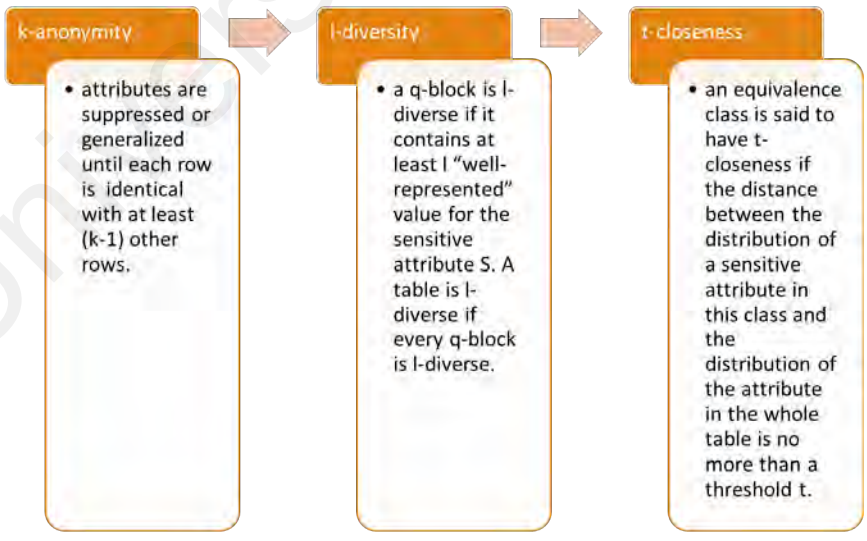

Figure 2.2 Anonymization techniques

2.1.2.1 l-Diversity

l-Diversity was proposed in order to conquer of *k*-anonymity's limitations. A method was introduced as an extension to *k*-anonymity, which ensures data privacy and avoids attribute disclosure without having to identify the attacker's background knowledge. It revolves on the notion that sensitive attributes in each group are "well-represented". The technique is actually k-anonymity that has been modified by incorporating the k-anonymity principle (Machanavajjhala et al, 2006).

If each of equivalence class in the table has at least "*l*" "well-represented" values for each sensitive attribute, then k-anonymous table can be said to be *l*-diverse (Machanavajjhala et al, 2006). "Well-represented" in this context can be explained using the following principles:

- Homogeneity (Inference) Attack: Since *k*-Anonymity doesn't impose restrictions on values that could be assumed by the sensitive attribute in an equivalence class, a given equivalence class may include a tuple with the same sensitive value. Let's say the recipient of the data knows an individual's quasi-identifier value represented in the table. The recipient can now identify the the corresponding individual's equivalence class, and therefore their sensitive attribute can be inferred.
- Background Knowledge Attack: This type of attack when a known fact on its own isn't a privacy disclosure, however when combined with other information can create a more precise inference on a target's sensitive information (Amiri et al, 2016).

In order to stave off both type of attacks, *l*-diversity will extend k-anonymity into each equivalence class by requiring "*l* well-represented" values for the sensitive attributes. "Well-represented" values here means requiring equivalence classes to have a minimum *l* different values for its sensitive attribute. Furthermore, as *l* increases, the background knowledge attacks will lose its

effectiveness. This is because more background knowledge is needed in order to associate sensitive attribute values to an individual (Machanavajjhala, 2006).

The l-diverse table keeps generalization and suppression for reducing information loss at a minimum – it can be computed using any algorithm computing an optimal k-anonymous table. This can be done by adding a control to enable checks on whether the diversity of the sensitive attribute values is being by each equivalence class in the table.

However, an l-diverse table remains subject to skewness and similarity attacks, and therefore can still cause disclosure of sensitive information. An *l*-diverse table is still in danger of disclosing sensitive information, considering its vulnerability to skewness and similarity attacks.

- Skewness attack: Might occur if the sensitive attribute's distribution of values in an equivalence class is different from a general (demographic/whole table) one. The difference in the distribution underlines the changes in the possibility of an individual in the equivalence class is related to a sensitive value (Li, Li & Venkatasubramanian, 2007).
- Similarity attack: Will occur when the values of the sensitive attribute in an equivalence class are semantically similar, instead of a syntactically different one, as an l-diversity requires (Machanavajjhala et al., 2006).

2.1.2.2 t-Closeness

A *t*-closeness technique decreases the interpreted data's granularity, essentially a betterment of *l*-diversity. Although the observer's knowledge may not be limited to the whole table which contains the datasets, the extent of their knowledge on specific data is limited. Thus, the correlation between the quasi-identifier and sensitive attributes is reduced. The distributions' distance can be measured with an Earth Mover's Distance (EMD). The EMD measures the distance between the values in a categorical attribute, according to a minimal level of generalization of these values in the domain hierarchy (Li, Li & Venkatasubramanian, 2007).

An equivalence class reaches *t*-closeness once the sensitive attribute distance in this class is not greater than the treshold, *t* with the attribute distance in the whole table. If all equivalence classes have t-closeness, then the table is acknowledged to have *t*-closeness (Li, Li & Venkatasubramanian, 2007).

After enforcing a *t*-closeness requirement, a skewness attack would be harmless. Considering knowledge about the value of a quasi-identifier now wouldn't affect the probabability to infer a target's sensitive value. Furthermore, it reduces a similarity attack's effectiveness: semantically similar values in an equivalence class can be present only due to the microdata table containing the same values.

2.1.2.3 Summary

Table 2.7 Summary of k-Anonymization Methods (Rajendran et al. 2017)

Methods	**Advantages**	**Disadvantages**
k-Anonymity	• It preserves against identity disclosure by inhibiting the links to a dataset with less than *'k'* values. This prevents the adversary from connecting a sensitive data with an external data • The cost incurred to establish this method is considerably less compared to another anonymity method such as cryptographic solution. • Algorithms of k-anonymity such as Datafly, Incognito and Mondrian are used extensively, especially in Privacy Preserving Data Publishing. It is also mentioned that clustering is incorporated in k-anonymity to enhance privacy preservation.	Prone to attacks such as: • Homogeneity Attack • Background Knowledge Attack
l-Diversity	• Provides greater distribution of sensitive attributes within the group, thus increasing data protection. • Protects against attribute disclosure, an enhancement of k-anonymity technique • The performance of l-diversity is slightly better than k-anonymity due to faster pruning by the l-diversity algorithm	• L-diversity can be redundant and laborious to achieve. • Prone to attacks such as skewness attack and similarity attack, as it is inadequate to avoid attribute exposure due to the semantic relationship between the sensitive attributes.
t-closeness	• It interrupts attribute disclosure that protects data privacy. • Protects against homogeneity and background knowledge attack. • It identifies the semantic closeness of attributes, a limitation of l-diversity	• Using Earth Mover's Distance (EMD) measure in t-closeness, it is hard to identify the closeness between t-value and knowledge gained. • Necessitates that sensitive attribute spread in the equivalence class to be close to that in the overall table.

Considering how powerful data is to an organization these days, using generalization or suppression to anonymize data would make for a certain amount of information loss, that may paint a different picture should the data be re-analyzed. And hence, the researcher has extended the search for different methods of anonymization. There were a few worth mentioning, but the researcher focuses on a single method of data perturbation- differential privacy as it was made known again in the research field as of late. Another method that adopted a different measure to generalization and suppression, Anatomy, is also discussed here. It was chosen as it was a simplistic method that could be incorporated easily into any algorithm. And lastly, the researcher discusses clustering. In the recent years, researchers has found a way to incorporate clustering methods with k-anonymization principles. This method does not require generalizing data or suppressing them and this will allow users with a better picture of what the data represent for a certain area.

After the introduction of k-anonymization, several schemes has been proposed, which do not rely on generalization hierarchies. For example, LeFevre et al. had transformed the k-anonymity problem into a partitioning problem. They had taken an approach consisting two steps:

Firstly, look for the the d-dimensional space's partitioning – d stands for the number of attributes in the quasi-identifier so each partition would have at least k-records. Next, records from the partitions are generalized until they have the same quasi-identifier value. While it may sound efficient, these steps are disadvantageous, since it demands a total order for each attribute domain. It renders almost impractical when dealing with cases that involve categorical data, because they do not have a meaningful order (LeFevre et al.,2006).

2.2 **Anatomy**

To overcome the defects of generalization, Xiao and Tao proposed a technique called Anatomy, in order to achieve a privacy-preserving publication that could capture the exact QI-distribution. To be more precise, anatomy releases a quasi-identifier table (QIT) and a sensitive table (ST), which separates QI-values from sensitive values. For example, Table 2.8 and 2.9 demonstrates the QIT and ST obtained from the microdata in Figure 2.

Table 2.8 The Quasi-Identifier Table

Row #	Age	Sex	ZIP	Group-ID
1	29	F	40000	1
2	*28*	F	40150	1
3	35	F	60000	1
4	59	F	34580	1
5	23	M	80000	1
6	22	M	23450	2
7	21	M	45670	2
8	43	M	40000	2
9	32	M	40150	2

Table 2.9 The Sensitive Table

Group-ID	*Ticket Type*	*Count*
1	Normal	4
2	Student	2
2	Senior Citizen	3

To simplify, anatomized tables can be understood as such: Firstly, based on a certain strategy, the tuples of the microdata are partitioned into several QI-groups. Then, the QIT and ST table are produced (Xiao, 2006).

The QIT table is created with the quasi-identifier partitioned to multiple groups and the sensitive value. Here, Ticket Type is placed on the sensitive table, together with the count of how many of those tickets are on the QIT group (Xiao, 2006).

Anatomy helps to protect privacy mainly because the QIT doesn't display sensitive values from any tuple, and therefore has to be guessed randomly from the ST (Xiao, 2006).

2.3 **Differential Privacy**

Di Differential Privacy was a definition developed in 2008 by Dwork, Nissim, McSherry and Smith. It adds noise to the data to make a dataset unrecognizable. And hence, it is a perturbed way of anonymizing data. Over the years, definition has had many contributions from others. In the effort of searching through all the techniques available, the researcher discusses Differential Privacy here as a method of data anonymization through perturbed means.

Picture two identical databases: the first has your information in it and second without. Differential Privacy guarantees the probability of a statistical query which will produce a given results that is almost the same, no matter if executed on either databases.

A way of looking at this is the ability to know if the data used will have a significant effect on the outcome of a query, and Differential Privacy provides this. If the data used had no effect, you may confidently contribute to the database knowing no harm can come to it. Below is an excerpt from the Apple iOS10 preview guide:

> "Starting with iOS10, Apple is using Differential Privacy technology to help discover the usage patterns of a large number of users without compromising individual privacy. To obscure an individual's identity, Differential Privacy adds mathematical noise to a small sample of the individual's usage pattern. As more people share the same pattern, general

patterns begin to emerge, which can inform and enhance the user experience. In iOS10, this technology will help improve QuickType and emoji suggestions, Spotlight deep link suggestions and Lookup Hints in Notes."

The primary mechanism in achieving this is introducing random noise to the aggregated data. Of course, the noise would be carefully designed. Apple had employed differential privacy for iOS10, where it had added noise into individual user inputs. This means that it is able to track, for example, your frequently used emojis, no matter how a user's emoji usage is masked.

2.4 **Clustering**

Cluster analysis, often used to explore inter-relationships among patterns, is an unsupervised learning method. It is a data analysis process that organizes inter-relationships into homogenous clusters. Cluster analysis differs from classification – which is known as supervised learning, because there are no apriori labelling of certain patterns available to be used in categorizing others, and infer the cluster structure of the whole data. The density of connections within a single cluster is referred to intra-connectivity. If the instances within a cluster are highly dependent on one another, it has a high intra-connectivity -- indicating a good clustering arrangement. Inter-connectivity, on the other hand, measures the connectivity between specific clusters. Ideally, individual clusters should be dependent of each other, therefore inter-connectivity should be at a low degree. Clustering techniques can be broadly classified into hierarchical, partition, density, grid and model-based clustering.

(Partition Based) In Partitioning based clustering, all objects are initially considered as a single cluster. The objects are divided into partitions with each partition representing a cluster. This algorithmn is especially effective in small to medium sized data points to find spherical-shaped clusters.

(Hierarchical Based) Hierarchical clustering can also be further divided using Agglomerative (top-down) and Divisive (bottom-up) techniques. In the Agglomerative approach, initially one object is selected and successively merges (agglomerates) with its closest similar pair based on similarity criteria until all the data forms a desired cluster. The Divisive approach starts with one cluster, which is then divided into additional clusters down the hierarchy, until the number of clusters formed are sufficient.

(Density Based) Based on density, clusters are formed. These density-based clusters will be separated from one another by regions of low-density objects – these are called noise, or outliers.

(Grid Based) The data space is partitioned into cells to form a grid like structure. Then working on each cell multi-resolution clustering is performed. Since Grid algorithms perform the clustering on the grid versus the database they have much faster processing power when compared to other algorithms.

(Model Based) Clusters are formed using models. An ideal fit between the model and data determined the cluster assignment. In the model-based clustering approach, the assumption is that data is produced by a combination of probability distributions in which each module characterizes a different cluster.

Table 2.10 describes the examples as well as the advantages and disadvantages to each approach.

Table 2.10 Advantages & Disadvantages of clustering methods (Ramesh, Nandini, 2017)

Method	Example	Advantage	Disadvantage
Partition	• k-means • k-mediods • k-modes	• Relatively scalable and simple • Suitable for well separated datasets with compact spherical clusters	• The concept of distance between points are ill-defined in high-dimensional spaces. • Pre-defined cluster count • Highly sensitive to the initialization phase, noise and outliers.
Hierarchical	• CACTUS • CURE • BIRCH • ROCK • Echidna • Wards • SNN • Chameleon •	• Embedded flexibility regarding the level of granularity. • Suitable for problems involving point linkages	• Correction not possible once the splitting/merging decision is made • Lack of interpretability regarding the cluster descriptors • Vague termination criteria. • Too expensive for high dimensional and massive datasets. • Highly ineffective in high dimensional spaces.
Density	• DBSCAN • OPTICS • DBCLASD • GDBSCAN • DENCLU • SUBCLU	• Discovery of arbitrary-shaped clusters with varying size. • Resistance to noise and outliers	• Highly sensitive to the setting of input parameters. • Poor cluster descriptors • Not suitable for high-dimensional datasets due to the dimensionality phenomenon.
Grid	• STING • CLIQUE • BANG • MAFIA • ENCLUS	• Efficient for large multidimensional spatial databases. • Insensitive to outliers and data input order.	• Need to tune grid size and density threshold • Can have high mining costs

Method	Example	Advantage	Disadvantage
	• PROCLUS		
Model	• EM • COBWEB • CLASSIT • SOM • SLINK	• Since models are a comparison, models can abstract away from details to capture a general insight.	• When generalized models are more complicated • Which model to compare with is a big exploration.

There are two prominent methods in the Partition-Based clustering approach: the k-means and the k-medoid technique. k-means, one of the oldest clustering algorithms, is a prototype in terms of a centroid, typically the mean of a group of points that is usually used on objects in a continuous n-dimensional space. k-medoid, on the other hand, is a prototype in terms of a medoid that can be applied to wide ranging types of data, considering it needs a proximity measure for a pair of objects. A centroid typically does not correspond to an actual data point. A medoid, however, has to be an actual data point. The most widely used clustering algorithm is the k-means.

Clustering methods have always been a prominent way of analyzing data. However, it was not paired with the use of k-anonymity until Byun et al.(2006) introduced the method of anonymization that incorporated using the clustering method together with k-anonymity in order to achieve a data anonymization method with less information loss.

Byun et al's (2006) approach runs on the key concept of viewing k-anonymization to be a clustering problem. Clustering would partition a set of objects into groups, making the objects similar to one another in that group compared to other groups' objects – based on a defined similarity criterion. Solving a k-anonymization problem optimally would be to have a set of equivalence classes where each are very similar to one another, since this will require minimal generalization. Typically, clustering problems would need to find a specific number of clusters in solutions. With k-

anonymity, though, there isn't a constraint on the number of clusters. Instead, it requires each cluster to contain at least k records.

Therefore, the k-anonymity problem was posed as a clustering problem, and referred to as a k-member clustering problem. Like most clustering problems, it was exponential to do an exhaustive search for an optimal solution of the k-member clustering. To be able to precisely characterize the problem's computational complexity, the k-member clustering problem was defined as follows:

Given n records, is there a clustering scheme $\varepsilon = \{e_1, \dots, e_\ell\}$ such that

1. $|e_i| \geq k, 1 < k \leq n$: the size of each cluster is greater than or equal to a positive integer k, and
2. $\sum_{i=1,\dots,l} IL(e_i) < c, c > 0$: the Total-IL of the clustering scheme is less that a positive constant c.

Theorem 1: The k-member clustering decision problems is NP-complete.

Theorem 2: Let *n* be the total number of input records and *k* be the specified anonymity parameter. Every cluster that the greedy k-member clustering algorithm find has at least *k* records, but no more that *2k-1* records.

Theorem 3: Let n be the total number of input records and k be the specified anonymity parameter. The time complexity of the greedy k-member clustering algorithm is in $O(n^2)$.

A comprehensive table highlighting some of the major crossovers between k-anonymity (means) algorithm that has been paired with the clustering method over the years is available in **Appendix E**. This table was taken from Arora, S. & Chana, I. (2014) for a thorough examination of the literature on this subject.

2.5 **Comparisons**

The researcher has compared four techniques to Data Anonymization. This subchapter will show the difference or comparison between all four methods. Observe Table 2.11 and Table 2.12.

Table 2.11 Comparison on the different methods of Anonymization

Advantages	Methods	Disadvantages
• Refer to Table 2.7 for a full list of advantages of each anonymity method.	Anonymity	• Refer to Table 2.7 for a full list of advantages of each anonymity method.
n/a	Anatomy	• can only be applied to limited applications.
• The original dataset does not need any modification • Noise is added to the results using mathematical calculations based on the data type and query type. • Noise added retains the usefulness of data.	Differential Privacy	• Designed for low-sensitivity query • Adaptive querying exposes more privacy • Data utility is reduced • Higher risk concerns will require more noise addition
• Refer to Table 2.10 for full advantages list of each method	Clustering	• Refer to Table 2.10 for full disadvantages list of each method

Table 2.11 shows the comparison overview of all 4 techniques mentioned in this chapter. However, anonymity and clustering can be divided into different methods in itself and the pros and cons of each method had been discussed in its own subchapter. From the table, we can see that to achieve the objectives that was set out, clustering is the best method as it does not employ any use of generalization or suppression nor does it perturb the data. From Table 2.12, the researcher also compared the complexity of each method as well as its ability to protect data against certain attacks, and it is obvious from there that clustering is the best method.

Table 2.12 Comparison on all methods on anonymization

Techniques	*Execution Time*	*Data Utility*	*Complexity*	*Computational Complexity*	*Protect Against*		
					Singling Out	*Linkability (Background Knowledge)*	*Inference (Homogeneity)*
k-anonymity	Low	Low	Very low	$O(k \log k)$	Yes	No	No
l-diversity	Low	High	Low	$O(n^2)k)$	Yes	No	May Not
t-closeness	High	High	Very high	$2^{O(n)O(m)}$	Yes	No	May Not
Anatomy	--				No	No	No
Differential Privacy	Medium	Medium	High	$O(n^{\frac{1}{2}})$	May Not	May Not	May Not
Clustering	Medium	Medium	Medium	$O\big((m+k)n\big)$ *this changes based on the model used	Yes	Yes	Yes

2.6 Variables

From the literature review, the researcher has identified certain variables that will be used to determine the outcome of the research. The variables are:

1. Dataset quality
2. Dataset volume
3. Distance function
4. Method Efficiency (Result)
5. Information Loss (Result)

2.7 Summary

In this chapter, the researcher has discussed four techniques of data protection; anonymization using generalization and suppression, anatomy, differential privacy and clustering.

The literature started with the work that Sweeney has introduced, the concept of k-anonymity and how to achieve that using generalization and suppression and ends with the introduction of achieving that same k-anonymity concept using clustering methods.

The following chapter will discuss a little more literature on the variables that will be used in the method.

3 METHODOLOGY

This chapter describes the methodology employed in this research. It explains the steps taken to address the objectives and questions in this study.

The first objective involves analysing a vast amount of literature from past studies on the subject of ‘privacy’ and ‘anonymization’. From there, the researcher laid out all the existing methods for data anonymization and was able to create a framework aimed to answer research questions two and three, and further satisfy the second and third objectives.

A prototype was then created to evaluate the method and later compared against a benchmark identified in the original study.

3.1 Introduction

A research design is a structure to plan and execute a particular research. It is a crucial part of the research as it includes all four important considerations: the strategy, conceptual framework, identification of what to study, and the tools and procedures used for collecting and analysing data.

Research design is divided into several types, for example, qualitative or quantitative research. For this study, the researcher used a quantitative research method. Quantitative research is a structured way of collecting and analysing data obtained from different sources. Quantitative research involves the use of computational, statistical and mathematical tools to derive results. It is conclusive in its purpose as it tries to quantify the problem and understand how prevalent it is by looking for projectable results to a larger population.

Aside from the quantitative research method, this research also uses historical design as it collects, verifies and synthesizes evidence from the past to establish facts that

defend or refute its hypothesis. In the context of social science, this method of research can be considered when the primary source of evidence comes from the collection of data from journals and articles –which the researcher has established as the main way of collecting data in this study, as opposed to interviews or questionnaires.

As part of data collection, this research also uses a systematic literature review to gather papers/articles/journals that has been written over a specified period of time pertaining to the subject matter.

3.2 Research Methodology

This research was carried out in multiple stages. As shown in Figure 3.1, the research was conducted following the research questions and objectives that was set in the beginning of the study.

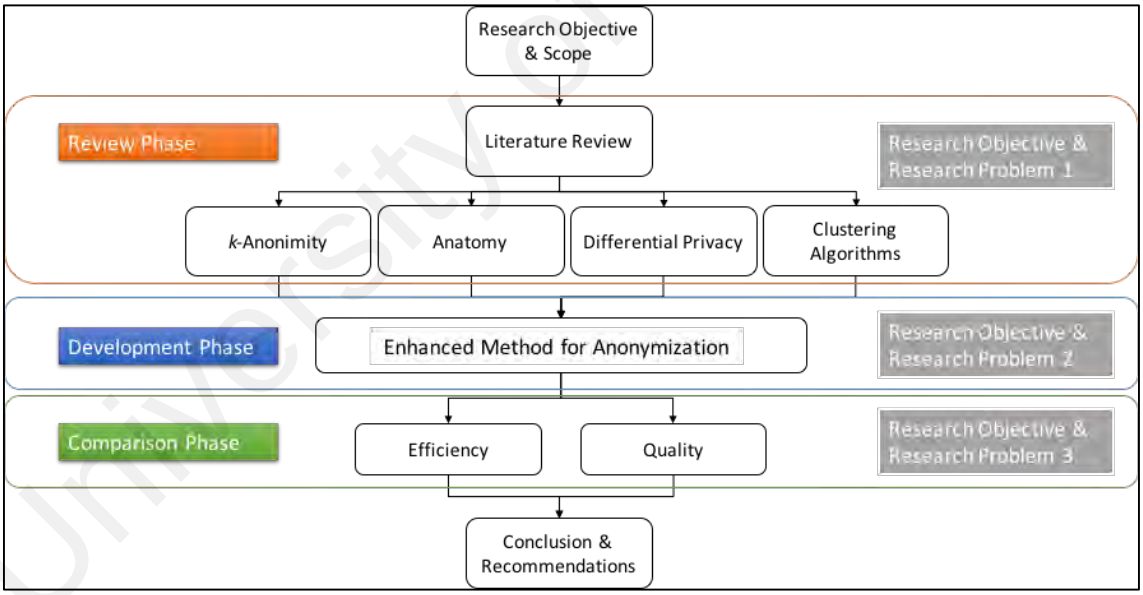

Figure 3.1 Research Methodology

3.2.1 Data Collection Method

The first part of the research was conducting a literature review. This study uses a combination of historical research and other methods for completion. It is therefore based on reviewing literature and not from interviews or questionnaires for data gathering. Therefore the researcher employed the use of a systematic literature review.

An effective literature review is one that "creates a firm foundation for advancing knowledge. It facilitates theory development, closes areas where a plethora of research exists and uncovers areas where research is needed" (Webster & Watson, 2002). A systematic literature review aims to present and evaluate literature related to the research topic by utilizing a thorough and auditable methodology. This research adopted a systematic literature review, a methodology proposed by Kitchenham (2004) and Kitchenham et al. (2008) as a way to recognize the correlation and history of big data, data privacy, and its solutions. This review consists of several activities, such as planning the review, data extraction, inclusion/exclusion/quality assurance phase, and finally, the write-up.

The planning activity focuses on developing the review protocol. It explains the workflow of the review conducted by the researcher. It also involves the identification of the research questions, the search strategy and evaluation of the resources, the inclusion and exclusion criteria, the quality assessment of the resources and the method of analysis. The second activity executes the defined protocol in the planning phase, while the explanation of the final report is elaborated in the final activity (Ijab et al., 2016). The data collection activity resulted in the papers to be analysed and used as a base for the framework that will be built in this research. These papers are also what is used to derive hypotheses meant to be confirmed or refuted in this research.

3.3 **Requirement Analysis**

Based on the systematic literature review, four techniques were narrowed down:

(a) anonymization using generalization and suppression

(b) anatomy

(c) differential privacy

(d) clustering algorithms

These four techniques were chosen because they were the most common techniques used over time. Generalization and suppression is the initial method of anonymization that was introduced, they were the pioneers. Anatomy followed suit shortly after employing a different method of anonymization in comparison, and hence, is included in the study. Differential Privacy was made known again when Apple used it in their iOS10. It was introduced a lot longer before that, but research went stagnant after the introduction, clustering algorithms is a method that was researchers has been able to combine with the concepts of k-anonymization and is a fairly new study.

To conduct this study, the researcher took the ADULT dataset from UCI Machine Learning Repository, as it was the benchmark for all anonymization practice published in journals and studies. To achieve the second objective, the dataset was run through four anonymization techniques to find one that satisfies the requirement of the study – one that has less information loss and another that is more efficient.

Table 3.1 Dataset Properties

Dataset Characteristics	Multivariate
Atrribute Characteristics	Categorical, Integer
Associated Tasks	Classification
Number of Instances	48842
Number of Attributes	14
Missing Values?	Yes
Area	Social

This study was conducted on a laptop with specifications of:

- Intel® Core(TM) i5-5200U CPU @ 2.20GHz 2.19 GHz
- 12GB RAM

and the program was written in JetBrains PyCharm3.2 on a Phyton 3.2 environment.

3.4 Variables Used in the Study

3.4.1 Independent Variables

Leedy (1997) defines an independent variable as one that potentially influences other (dependent) variables. Independent variables used in this study include dataset quality, dataset volume and distance function.

3.4.1.1 Dataset Quality

There are multiple characteristics that has to be looked into when searching for a dataset, like data accuracy and legitimacy, among others. The data for use has to be a clean set with all mandatory cells not left blank. Some fields such as gender or nationality are typically limited to a certain number of responses, and the data for use has to adhere to that set. In other words, there cannot be erroneous data as it will impact the final result.

Consistency is another important aspect of a good quality dataset. For example, a field asking for phone number should only include 10 digits and no more, with no letters or symbols to it. A phone number is a direct identifier and has to be removed when publishing anonymized dataset, thus would not be an issue in anonymization, but it is still a factor in determining a good dataset. The dataset also has to be reliable, it cannot contain a lot of trash data that was added just to increase count or a set that was created for random testing.

The dataset has to be able to be used to represent results over the benchmark in comparison to what was used in previous studies as well.

3.4.1.2 Dataset Volume

Since this study is not researching an algorithm/framework on a big data, hence, the volume of data required does not need to be huge. The dataset has to be, again, on par with what was used in the previous studies.

3.4.1.3 Distance Function

Distance Functions are functions that define a distance between each pair of elements of a set. This research consists of manipulating the distance functions in order to get the best performing results, both in terms of efficiency and minimizing information loss.

There are multiple types of distance functions that can be employed when running a clustering algorithm. An example is shown below:

Table 3.2 Distance Functions

Name	Definition	Formula
Euclidean Distance	Computes the root of square difference between co-ordinates of pair of objects	$Dis_{xy} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$
Manhattan Distance	Computes the absolute differences between coordinates of pair of objects	$Dis_{xy} = \left\|(x_{ik} - x_{jk})\right\|$
Chebychev Distance (maximum value distance)	Computed as the absolute magnitude of the differences between coordinate of a pair of objects.	$Dis_{xy} = max_k\left\|(x_{ik} - x_{jk})\right\|$
Minkowski Distance	Generalized metric distance *p=2, this becomes a Euclidean Distance, P=1, this becomes a city block distance	$Dis_{xy} = \left(\sum_{k=1}^{d}\left\|(x_{ik} - x_{jk})\right\|^{\frac{1}{p}}\right)^p$

The distance function affects the result of the anonymization in terms of efficiency as well as information loss, because it actually changes the way the next record is searched and added, as well as manipulates the way the next centroid is located. By manipulating the distance function, the data will be anonymized differently. Distance Functions for this particular study will be elaborated in Section 3.6.3.

3.4.2 Dependent Variables

Dependent variables can potentially be influenced by other (independent) variables, as defined by Leedy (1997). This study has two dependent variables: system efficiency and information loss.

3.4.2.1 Efficiency

One of the outcomes this research is investigating is the impact of the distance function manipulation to the total time it takes for the data to complete its anonymization phase.

3.4.2.2 Information Loss

Information Loss is a method to determine the accuracy of the data clustered into each group. In every clustering method, there exists outliers, which will be put in any of the closest neighbouring groups. These outliers will increase the percentage of information loss as it is now a record that does not satisfy the k-anonymity constraint. However, since the purpose of this study is to achieve minimal information loss, the researcher aims for a method that won't try to heavily generalize or suppress data – it is important to have a clearer picture of the data if it needs to be repurposed for any kind of analysis.

3.5 Research Hypotheses

3.5.1 Introduction

A statistical hypothesis is seen as an assertion concerning one or more population, where its plausibility is evaluated on the basis of the information obtained from sampling of the population (Bhattacharyya & Johnson, 1977). The concept behind hypothesis therefore lies in testing to determine if predicting about some feature of a population is strongly supported by the information obtained from the sample data.

3.5.2 Dataset Quality

The dataset for use needs to be of good quality, meaning it is complete, consistent, and accurate. A dirty dataset would not break the program, but would create one too many

outliers and thus severely compromising the final results. As it is, the framework creates a cluster containing records that satisfy a certain k-anonymity constraint. However, if there are many outliers, they will be added to any cluster, thus increasing the value of information loss within that cluster, defeating the purpose of that anonymization.

3.5.3 Dataset Volume

With dataset volume, the bigger the volume, the longer it takes to complete. The challenge for the system here is that the dataset, however big before it reaches the volume of big data, should be able to run without breaking/failing.

The assumption remains that the more records the system has to parse through, the longer it will take to complete its cycle. This is the efficiency based on the throughput itself.

3.5.4 Distance Function

Dissimilar to most distance functions used in neural networks or any other artificial intelligence study, there needs to be a distance function that can handle both numeric and categorical attributes. In k-anonymity problems, the data are person-specific records that consists of both types of data.

Table 3.3 Distance Function k-Anonymity

Category	**Definition**	**Function Formula**
Numerical Value	• *D* is a finite numeric domain • |*D*| is the domain size measured by the difference between the maximum and minimum values in *D*.	$\delta_N(v_1, v_2) = \frac{\lvert v_1 - v_2 \rvert}{\lvert D \rvert}$
Categorical Value	• *D* is a categorical domain • T_D be a taxonomy tree defined for *D* • *A(x,y)* is the subtree rooted at the lowest common ancestor of *x* and *y*	$t\delta_C(v_1, v_2) = \frac{H(A(v_i, v_j))}{H(T_D)}$

	• *H(R)* represents the height of tree *T*.	
Records (Maximum Distance)	*Distance between two records:* • $Q_T = \{N_1, \dots, N_m, C_1, \dots, C_n\}$ be the quasi-identifier of table *T* • $N_i (i = 1, \dots, m)$ is an attribute with a numeric domain • $C_j (j = 1, \dots, n)$ is an attribute with a categorical domain • $r_i[A]$ represents the value of attribute *A* in r_i • δ_N is the distance function for numerical values • δ_C is the distance function for categorical values	$\Delta(r_1, r_2) = \sum_{i=1,\dots,m} \delta_N(r_1[N_i], r_2[N_i]) + \sum_{j=1,\dots,m} \delta_N(r_1[C_j], r_2[C_j])$
(Median Distance)		$\Delta(r_1, r_2) = \frac{1}{2}\left(\sum_{i=1,\dots,m} \delta_N(r_1[N_i], r_2[N_i]) + \sum_{j=1,\dots,m} \delta_N(r_1[C_j], r_2[C_j]) \right)$
(Minimum Distance)		$\Delta(r_1, r_2) = \sum_{i=1,\dots,m} \delta_N(r_1[N_i], r_2[N_i]) - \sum_{j=1,\dots,m} \delta_N(r_1[C_j], r_2[C_j])$

For numeric attributes, the difference between two values describes the dissimilarity of the values naturally, which is also suitable for a k-anonymization problem. For categorical attributes, the difference cannot be enumerated in any specific order. Here, semantic relationships are the key. The relationship can be captured using a taxonomy tree. And so, the definition of the distance between both records has to capture both the numerical and categorical attributes.

Table 3.2 shows the distance functions used in this research. This function was adopted from Byun et al.(2007) in their clustering algorithm paper.

3.5.5 Efficiency

As mentioned in the previous section, one of the outcomes this research is investigating is the time it takes for a dataset to complete anonymization based on the manipulation of the distance from one centroid to another.

This research hypothesizes that the closer the next centroid is to another, the faster it will take the program to parse through and cluster the data.

3.5.6 Information Loss

Another factor the researcher looks at is the information loss within a cluster set. This research posits that the information loss, despite the distance of the next centroid to another, will not be much different – in other words, minimal.

Another hypothesis is that information loss may differ hugely between the nearest neighbour and furthest neighbour.

3.6 Summary

Table 3.4 Hypotheses List

Subject Matter	Hypotheses
Independent Variable	
Dataset Quality	The cleaner the dataset is, the less outliers will exists, and hence, the information loss percentage will be lower.
Dataset Volume	The volume of the dataset will influence how long the program will run, but, if counted on average, it should not effect the result.
Distance Function	The distance function has to cater to both numerical and categorical data; else, the program will fail.
Dependent Variable	
Efficiency (Runtime)	The closer the next centroid is to another, the faster it will take the program to parse through and cluster the data.
Information Loss	• The closer one centroid is to to the other, the bigger the information loss percentage. • The further one centroid is to the other, the smaller the information loss percentage. Information loss may differ hugely between the nearest neighbour and furthest neighbour. The concept behind this hypothesis is that by using the nearest neighbour, many more outliers may exist as the data parsing (searching) might overlook certain unique records when it simply looks randomly.

In this chapter, the researcher has discussed the methodology that was used that comes up to the framework. In short, after the literature review was carried out, the researcher narrowed down on clustering algorithms as it provides with the least information loss.

From there, a comparison was carried out against three common methods in the area to find one that suits the objective best. From that results, the researcher then created a method that allows the user to manipulate the distance function depending on the type of result the user may be looking for. Table 3.4 also lists all the hypotheses that will be tested and discussed in Chapter 5.

4 SYSTEM DESIGN AND DEVELOPMENT

4.1 Introduction

This chapter describes the design and development of the method of this research. Multiple figures have been created to show a step-by-step process on how the method works throughout this chapter.

4.2 System Design and Development

The outcome of this research has produced an enhanced method that manipulates distance to churn out anonymized data that can be used for anonymization, as shown in the figure below:

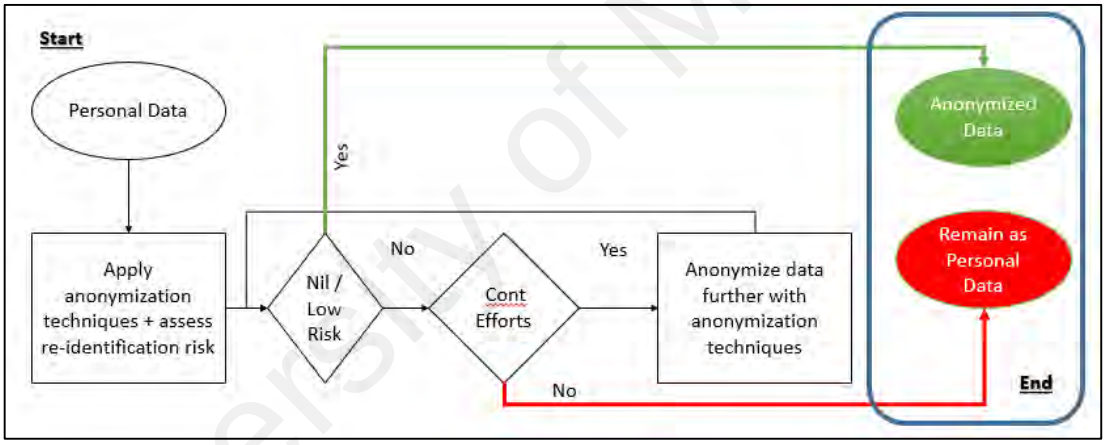

Figure 4.1 Anonymization Flow

Based on Figure 4.1, the anonymization process will only take place if the data received are still attached to the personal data. The researcher aims to find a method that will produce an anonymized table with the least information loss without sacrificing the speed of processing.

4.2.1 Components of the Method

This framework is built to process a dataset that contains personal data, thus its input is a dataset consisting of both numerical and non-numerical data. The first step to anonymizing data is to strip it off of any personal identifier such as NRIC, mobile number or passport number. Any identifier unique to the respondent has to be removed. The data will then be identified if there is a need for further anonymization, or if it satisfies a certain requirement to undergo another round of anonymization.

The requirement here being if the dataset is complete or has most people populated, is legal to use and was not obtained by illegal means, and that it fits the reason for repurposing as enacted in the PDPA2010 rules; that has been discussed in length in Chapter 2.

If the set fulfils all the requirements, it will go through another round of anonymization, which will be explained in the following sections. If, however, it does not meet one or any of the requirements, it will then be stored and remain as the original dataset.

4.3 Development Tools and Technologies

To explain this process further, consider the figure below. Figure 4.1 depicts the overall flow of how the anonymization process is carried out. The next figure shows a breakdown, or a more complete look of what happens in the framework.

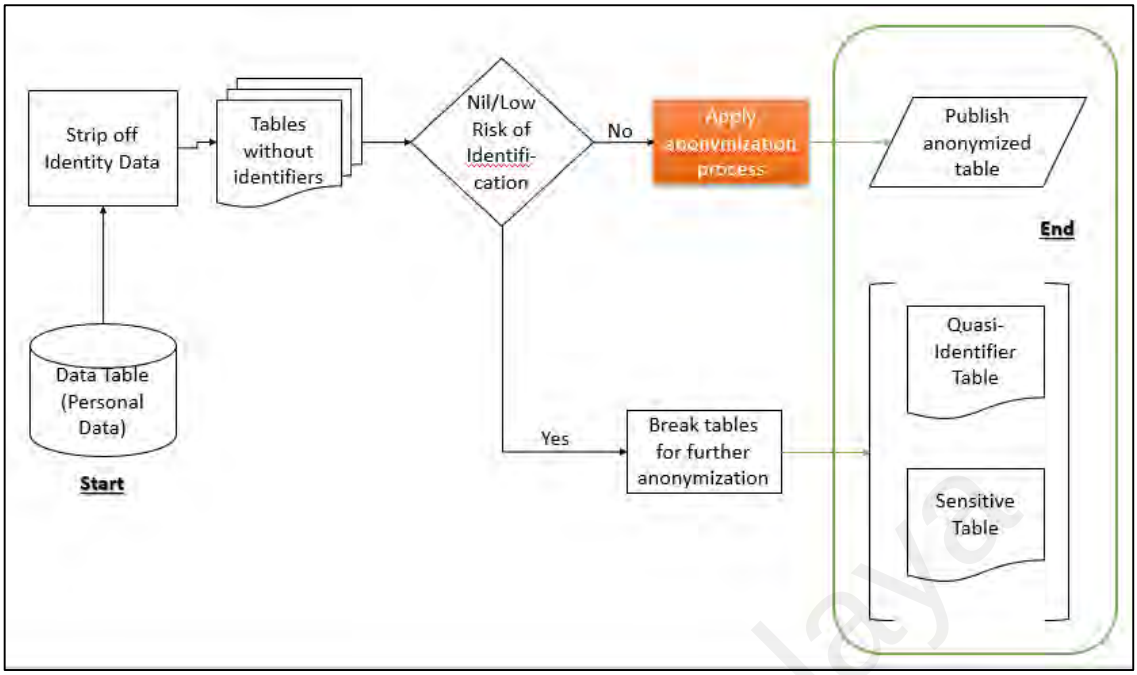

Figure 4.2 Anonymization Flow (Detailed)

The framework starts by stripping off any personal data in the table, for example, in a Malaysian data system it would be the NRIC or Phone Number, as these two attributes can be used to identify the person the data is related to. Next, with the table stripped of unique identifiers, the system would identify the risk of re-identification and assess if there is a need for anonymization. Data with low risk of re-identification can be further broken down into a quasi-identifier grouping and a sensitive table, to further protect the data when it is being stored. This uses the concept of anatomy in the storage of data.

Data with high risk of re-identification will go through an anonymization process that uses a k-anonymity concept as its base. This framework can be further broken down (on the orange box in figure 4.2; apply anonymization process) shown below:

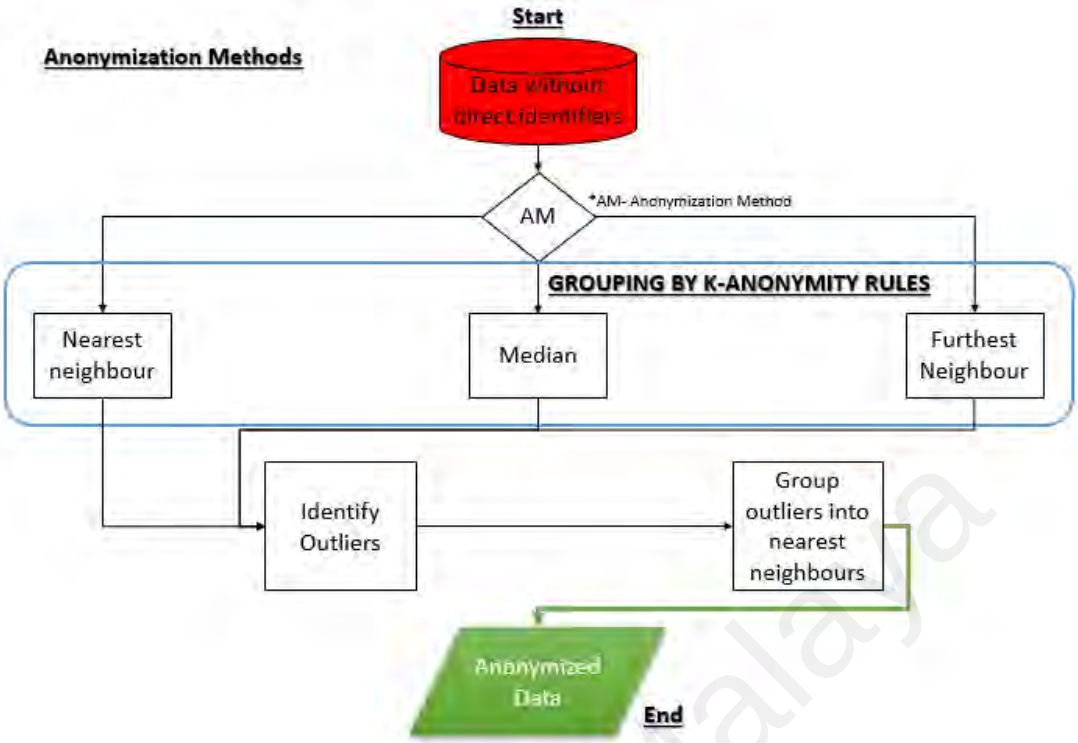

Figure 4.3 Enhanced Method

4.3.1 Clustering Method

The researcher created this system using three different methods of choosing the centroid for the next cluster. Using the same k-means concept to create clusters, the researcher created the system to satisfy the k-anonymity concept by creating the clusters to the size of the *k* constraint.

Each of the algorithm created starts the same: randomly selecting one record, *p* as the seed to start building a cluster, subsequently selecting and adding more records to the cluster such that the record added incurs the least information loss within the cluster. Once the number of records in the cluster reaches *k* (the anonymity constraint), the algorithm selects a new record for the new seed. Depending on the method chosen by the user, the new seed will choose records that are either the closest, the furthest or the median record, and the same process is repeated for the next cluster. Eventually, there will be less records left compared to the number of constraints imposed, and this is where the records (called outliers) will be assigned to the closest cluster available.

Considering the clusters are created to achieve the least information loss, the complexity in this method can be seen in the processing time. Generally, looking for records furthest from the first cluster will result in a longer wait compared to looking for the closest record to build the second cluster. However, looking for the closest records might also result in more outliers, especially if similar records are clumped together.

With the addition of every outlier in any cluster, the information loss percentage will increase, because these outliers are added to the nearest neighbour regardless of the similarity.

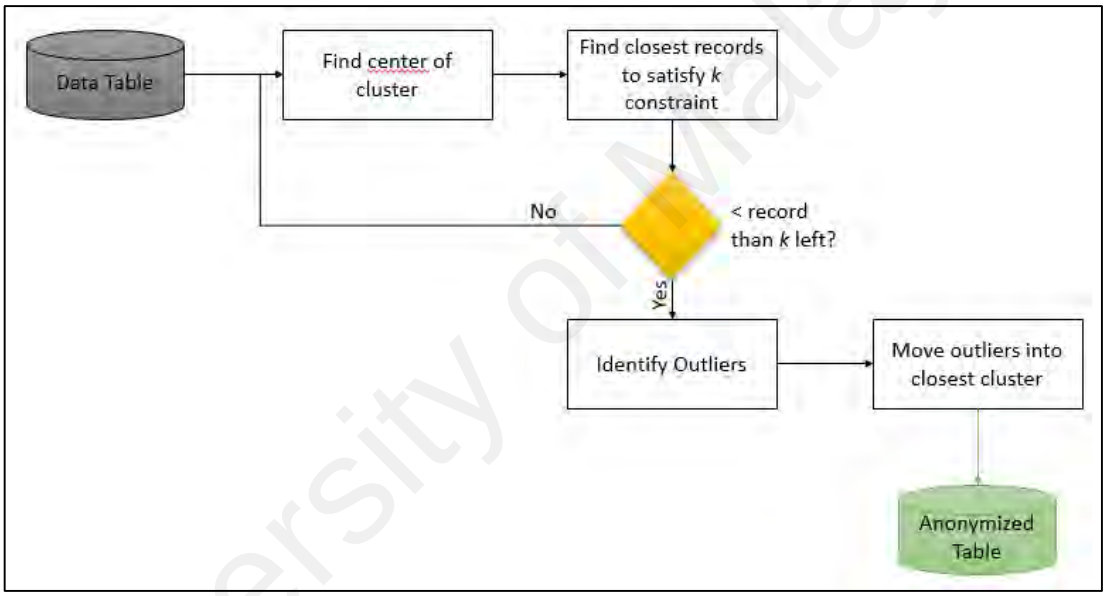

Figure 4.4 Method Visualized

From Figure 4.4, the researcher shows that the difference between the three algorithms or one that is being compared is how the complexity of looking for the distance between two points will compare to one another. It also shows that outliers being put into nearest neighbours will impact the information loss percentage, as a result.

Table 4.1 Algorithm Pseudocode

Nearest	Median	Furthest (Byun et al., 2006)
Input: a set of records R and a threshold value *k* Output: a set of clusters each of which contains at least *k* records		
$if(\|R\| \leq k)$ $return\ R;$ $end\ if;$ $Result = \emptyset;$ $p = a\ randomly\ picked\ record\ f$ $while\ (\|R\| > k)$ $p = closest\ record\ from\ p;$ $R = R - \{p\};$ $c = \{p\};$ $while\ (\|c\| < k)$ $p = find\_best\_records(R, c);$ $R = R - \{p\};$ $c = c \cup \{p\};$ $end\ while;$ $result = result \cup \{c\};$ $end\ while;$ $while\ (\|R\| \neq 0)$ $p = a\ randomly\ picked\ record\ f$ $R = R - \{p\};$ $c = find\_best\_cluster(result, p);$ $c = c \cup p;$ $end\ while;$ $return\ result;$ $end;$	$if(\|R\| \leq k)$ $return\ R;$ $end\ if;$ $Result = \emptyset;$ $p = a\ randomly\ picked\ record\ f$ $while\ (\|R\| > k)$ $p = middle\ distance\ record\ fro$ $R = R - \{p\};$ $c = \{p\};$ $while\ (\|c\| < k)$ $p = find\_best\_records(R, c);$ $R = R - \{p\};$ $c = c \cup \{p\};$ $end\ while;$ $result = result \cup \{c\};$ $end\ while;$ $while\ (\|R\| \neq 0)$ $p = a\ randomly\ picked\ record\ f$ $R = R - \{p\};$ $c = find\_best\_cluster(result, p);$ $c = c \cup p;$ $end\ while;$ $return\ result;$ $end;$	$if(\|R\| \leq k)$ $return\ R;$ $end\ if;$ $Result = \emptyset;$ $p = a\ randomly\ picked\ record\ f$ $while\ (\|R\| > k)$ $p = furthest\ record\ from\ p;$ $R = R - \{p\};$ $c = \{p\};$ $while\ (\|c\| < k)$ $p = find\_best\_records(R, c);$ $R = R - \{p\};$ $c = c \cup \{p\};$ $end\ while;$ $result = result \cup \{c\};$ $end\ while;$ $while\ (\|R\| \neq 0)$ $p = a\ randomly\ picked\ record\ f$ $R = R - \{p\};$ $c = find\_best\_cluster(result, p);$ $c = c \cup p;$ $end\ while;$ $return\ result;$ $end;$
Function *find_best_record(R,c)*		Function *find_best_cluster(C,p)*
Input: a set of records *R* and a cluster *c* Output: a record $p \in R$ such that $IL(c \cup \{p\}$ is minimal		Input:a set of clusters *C* and a record *p* Output: a cluster $c \in C$ such that $IL(c \cup \{p\}$ is minimal.
$n = \|R\|; min = \infty; best = null;$ $for(i = 1, \dots, n)$ $r = i - th\ record\ in\ \|R\|;$ $diff = IL(c \cup \{p\}) - IL(c);$ $if(diff < \min)$ $min = diff;$ $best = p;$ $end\ if;$ $end\ for;$ $return\ best;$ $end;$		$n = \|C\|; min = \infty; best = null;$ $for(i = 1, \dots, n)$ $c = i - th\ record\ in\ \|C\|;$ $diff = IL(c \cup \{p\}) - IL(c);$ $if(diff < \min)$ $min = diff;$ $best = c;$ $end\ if;$ $end\ for;$ $return\ best;$ $end;$

As highlighted in the pseudocode, the differences between the three methods are dependent on the distance between the clusters. The distance is calculated using the distance function, explained in Section 3.6.3.

4.3.2 Cost Functions

To ensure the computation is the same as the benchmark algorithm proposed by Byun et al, the researcher uses the same cost function to measure the information loss. The information loss metric measures the amount of distortion introduced by the generalization process to a cluster. To recap what has been discussed in Chapter 2, let:

$e = \{r_1, \dots, r_k\}$ be a cluster where the quasi-identifier consists of numeric attributes $N_1, \dots, N_m$ and categorical attributes $C_1, \dots, C_n$.,

T_{C_i} be the taxanomy tree defined for the domain of categorical attribute C_i,

MIN_{N_i} and MAX_{N_i} be the min and max values in *e* with respect to attribute N_i,

$\cup_{C_i}$ be the union set of values in *e* with respect to attribute C_i.

Then the amount of information loss occurred by generalizing *e*, denoted by *IL(e)*, is defined as:

$$IL(e) = |e|.\left(\sum_{i=1,\dots,m} \frac{(MAX_{N_i} - MIN_{N_i})}{|N_i|} + \sum_{j=1,\dots,n} \frac{H(A\left(\cup_{C_j}\right))}{H(T_{C_j})}\right.$$

Where,

$|e|$ is the number of records in *e*,

$|N|$ represents the size of numeric domain *N*,

$\wedge\left(\cup_{C_j}\right)$ is the subtree rooted at the lowest common ancestor of every value in $\cup_{C_j}$,

H(T) is the height of taxonomy tree *T*.

Using the definition, total information loss can be defined as $Total - IL(AT) = \sum_{e \in \varepsilon} IL(e)$, where ε is the set of all equivalence classes in the anonymized table AT. The cost function of this problem is the sum of all intra-cluster distances, where an intra-

cluster distance of a cluster is defined as the maximum distance between any two data points in the cluster (Byun et al., 2006).

4.4 Non-Functional Requirements

As highlighted in the research questions and research objective, the researcher aims to find an algorithm that allows for the least information loss without sacrificing the performance of the system. The result the researcher is looking for is one comparable or better than what was outlined in the k-member paper proposed by Byun et. al. The system, as mentioned, used that paper as a base, but changes the way anonymization works, to observe and compare results.

In terms of performance or efficiency, this system should overcome the complexity of the original k-member system, that efficiency is $O(n^2)$. By reducing the distance it has to search (despite still having to search one by one), efficiency increases when clusters don't need to search for the furthest record.

Furthermore, in terms of information loss or data integrity, the system already looks for records to satisfy the k-anonymity constraint. Information loss here is at its best, in comparison to normal generalization and suppression methods usually undertaken when it comes to anonymizing personal data.

4.5 Dataset

For the experiments, the researcher used the Adult dataset from the UC Irvine Machine Learning Repository, which is considered a de facto benchmark for evaluating the performance of k-anonymity algorithms.

This dataset was altered by removing records with missing values. The *quasi-identifier* values here are *{age, work, class, education, marital status, occupation, race, gender, native country}.* Age and education were treated as numerical values and the rest are treated as categorical values.

4.6 User Interface Design

User Interface Design is an important process in the design phase. The design must be easy to use and understand. The main function of this user interface is to show users the comparisons in using different distance functions when it comes to efficiency and information loss. The user can choose and upload any dataset to the system, and write the number of k-anonymity boundary they wish, and the prototype will take care of the rest.

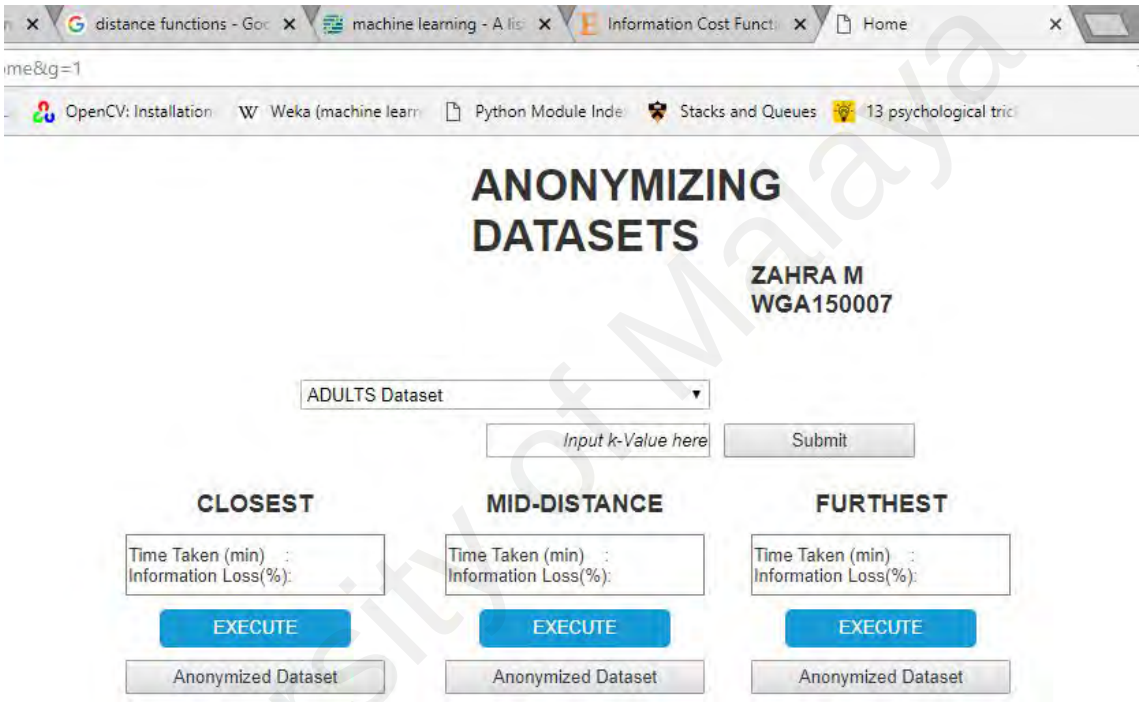

Figure 4.5 Front-End of the Comparison Home

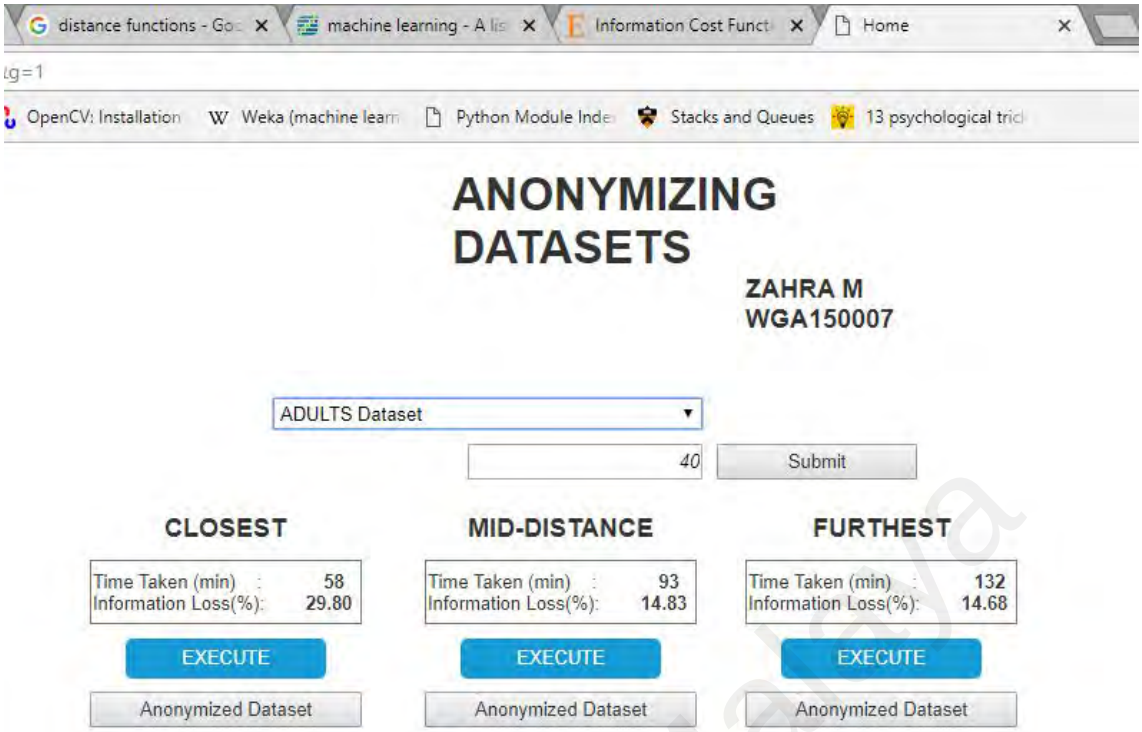

Figure 4.6 Results from Comparison of k=40

For example, Figure 4.6 shows what the interface would look like once the system is fully executed for k=40 on an ADULTS Dataset. Users can choose to execute any distance they wish or execute all three and compare the results if they wish. Users can also choose any file if they need immediate results, as opposed to best results for information loss. For example, once the shortest distance is completed, the anonymized dataset can be downloaded first and used for an initial study or initial assumption while waiting for the next dataset to be ready.

4.7 Summary

In this chapter, the researcher has discussed the design and development of the method. A few figure were drawn up to show the flow of the method. This chapter also shows the user interface that was created to be used for the system that users can use to execute the method. The UI gives the users the choice to execute different types of algorithm depending on the user result expectancy.

5 DATA ANALYSIS AND RESULTS

5.1 Introduction

This chapter will discuss the results of the experiments that has been carried out to test the researcher's theories in Chapter 3. The experiments were divided into two phases; one was the comparison between three different type of clustering methods to determine the one that matches the objectives best and the other is a comparison(manipulation) of three different distance function. The result are as follows.

5.2 Results

This research was originally carried out by executing a k-nearest neighbour against k-member and a one-time-pass k-means algorithm to find a base that could be used for developing the framework. The result of that experiment is as tabulated below:

Table 5.1 Comparison on Varying K (Runtime)

	KNN (k-Nearest Neighbour)	OKA (One Time Pass K-Means)	K Member
k-Value	Runtime (Minutes)		
10	59	176	198
20	51	139	167
30	49	112	154
40	47	99	132
50	44	87	97
60	39	69	75
70	36	64	69
80	32	59	60

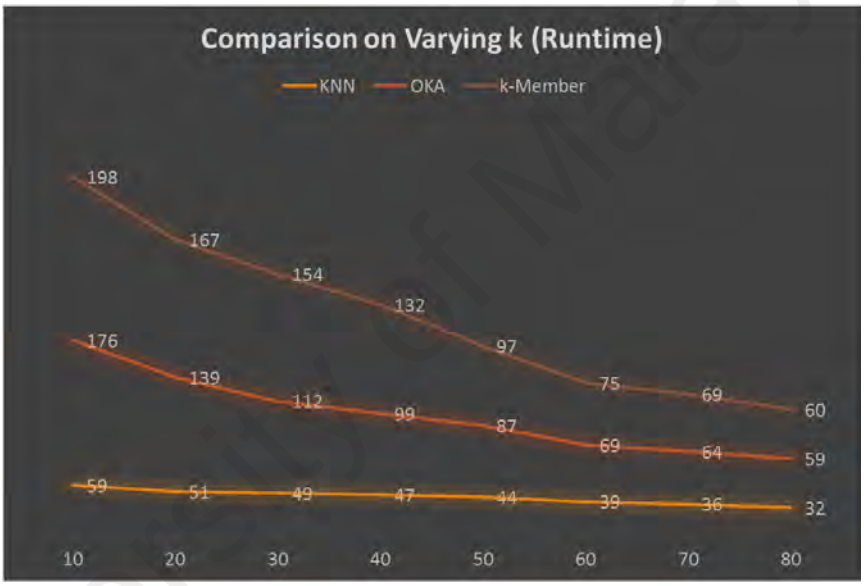

Figure 5.1 Comparison on Varying K (Runtime)

Based on the efficiency, the k-Nearest Neighbour ran the fastest, followed by a one-time pass k-means, and finally, k-member. The k-member algorithm justifiably takes the longest time as it builds its cluster slowly by adding each record as per calculation, and kNN is justifiably the fastest as it only considers the closest record to itself so there are no complex calculations. Next, we investigate the information loss factor.

Table 5.2 Comparison on Varying K (IL)

	KNN (k-Nearest Neighbour)	OKA (One Time Pass K-Means)	K Member
k-Value	Information Loss(%)		
10	11.93	16.84	14.32
20	11.95	16.88	14.44
30	11.99	16.90	14.65
40	12.01	17.02	14.68
50	12.04	17.33	15.03
60	12.05	17.54	15.67
70	12.18	17.67	15.81
80	12.20	17.69	15.90

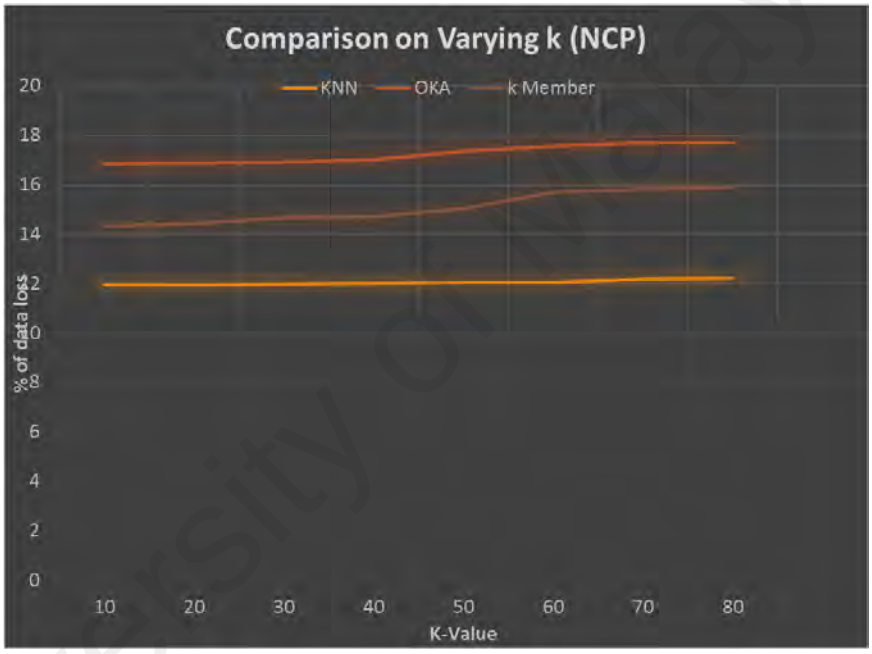

Figure 5.2 Comparison on Varying K (IL)

In terms of Information Loss, using Normalized Certainty Penalty to compute the penalty based on any generalization and suppression made to the cells, the kNN also comes first before the k-member algorithm. However, the kNN algorithm results differ in every iteration ran. Depending on how the data was constructed, the results in kNN is ever-changing as opposed to k-member.

Based on the two results, the k-member was a better fit to fulfil the research objective. And by using that result, a framework was built by switching certain elements that increased the efficiency without sacrificing Information Loss.

The program that was created was as discussed, based on the k-member algorithm that was created by Byun et al(2007). The researcher created a new framework that explored the way the next cluster is searched for. The researcher did not change the way the records were appended to the clusters, as it is already an optimum way to achieve least information loss. The result of the testing on the framework is tabulated below.

5.3 Comparison on System Efficiency

Table 5.3 Framework Comparison on Varying K (Runtime)

Runtime (Minutes)			
K-Value	**Nearest**	**Median**	**Furthest**
10	91	140	198
20	75	115	167
30	70	106	154
40	58	93	132
50	42	65	97
60	35	54	75
70	30	44	69
80	27	40	60

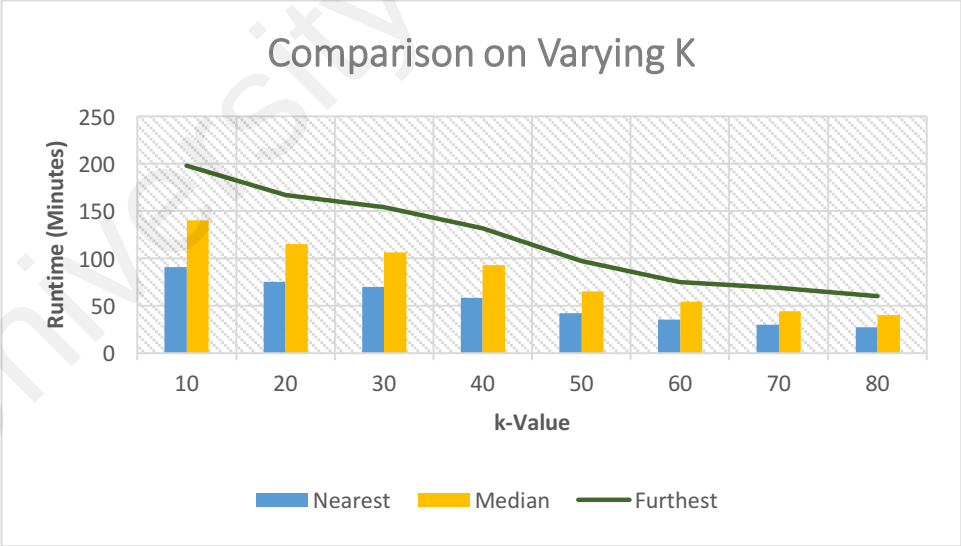

Figure 5.3 *Framework Comparison on Varying K (Runtime)*

Observe the graph: the green line is the result of the k-member as proposed by Byun et al. and the bar graphs underneath are the result of changing the method of searching for the next cluster. Looking for the next cluster saves quite a bit of time when it comes to

the running time, as compared to the work that was presented. As for the median distance, there is an increase in efficiency although not as good as the nearest distance.

5.4 Comparison on Information Loss

Table 5.4 Framework Comparison on Varying K (Information Loss)

Information Loss			
K-Value	**Nearest**	**Median**	**Furthest**
10	29.07	14.46	14.32
20	29.31	14.58	14.44
30	29.74	14.80	14.65
40	29.80	14.83	14.68
50	30.51	15.18	15.03
60	31.81	15.83	15.67
70	32.09	15.97	15.81
80	32.28	16.06	15.90

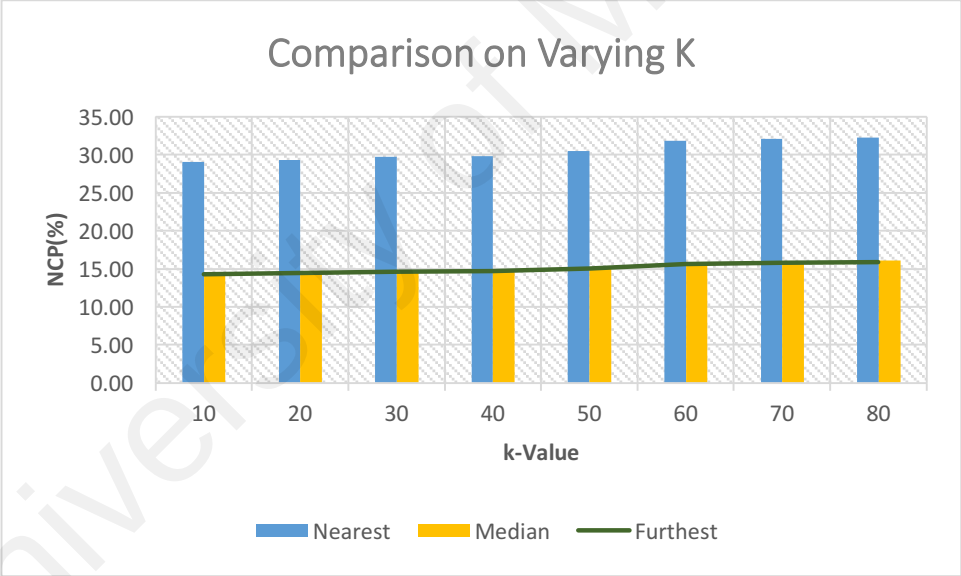

Figure 5.4 Framework Comparison on Varying K (Information Loss)

In terms of Information Loss, Figure 5.4 shows the result of the program based on the new framework. The green line shows the benchmark or the result that was based on the Byun et al k-member paper. The bar graph is the result of the changes made to that algorithm. Observe how using the nearest neighbour actually increases the normalized certainty penalty (NCP). The percentage is calculated based on assigning penalties to any

generalized data. The mid-distance algorithm doesn't fare as well as the original, but comes quite close to the original benchmark.

The research suggests that using the algorithm by updating the distance matrix to mid-distance is better in terms of efficiency without sacrificing Information Loss. The Information Loss doesn't fare as well to the benchmark, but it is a very close second, thus concluding that changing to a mid-distance fares better overall.

5.5 **Hypothesis Revisited.**

Back in Chapter 3, the researcher has hypothesized two results that is to be tested in this research.

Table 5.5 explains revisits the hypothesis that was describe in Chapter 3. The results of the experiment that is discussed earlier in this are those of the dependent variables – where we test to meet the objectives that has been set.

We can see from the results that the hypothesis of both dependent variable has been proven, where

- The closer the next centroid is to another, the faster it will take the program to parse through and cluster the data. (Efficiency – Data Throughput)

And

- The closer one centroid is to to the other, the bigger the information loss percentage (Information Loss Percentage – Data Quality)

The researcher has determined earlier on that the concept behind this hypothesis is that by using the nearest neighbour, many more outliers may exist as the data parsing (searching) might overlook certain unique records when it simply looks randomly.

Table 5.5 Hypothesis Revisited

Subject Matter	Hypotheses	Results – Revisit
	Independent Variable	
Dataset Quality	The cleaner the dataset is, the less outliers will exists, and hence, the information loss percentage will be lower.	When the program was ran with the data that was downloaded, without sanitation, it did not fail however caused a high information loss percentage.
Dataset Volume	The volume of the dataset will influence how long the program will run, but, if counted on average, it should not effect the result.	This was counted on average based on the number of runs that was done, along with the comparison of unsanitized against sanitized data, the volume of the data did not effect the throughput.
Distance Function	The distance function has to cater to both numerical and categorical data; else, the program will fail.	During the experiment, when the researcher inputs a distance function that can only calculate numerical data or categorical data, the system always encounters an error, and refuses to complete. Hence, if the distance function can't cater to both numerical and categorical data, the system does fail.
	Dependent Variable	
Efficiency (Runtime)	The closer the next centroid is to another, the faster it will take the program to parse through and cluster the data.	Based on Figure 4.3, it is evident that by manipulating the distance function to find the closest cluster, the runtime of the program is faster compared to a program that looks for the next centroid in the furthest record.
Information Loss	• The closer one centroid is to to the other, the bigger the information loss percentage. • The further one centroid is to the other, the smaller the information loss percentage. Information loss may differ hugely between the nearest neighbour and furthest neighbour.	Based on the results, the trend that can be seen is the difference between the nearest neighbour calculation to the furthest record centroid calculation. When the program looks for the nearest neighbour to make the centroid, the information loss percentage goes up due to the outliers that exists, as compared to if the system looks for the closest record.

5.6 **Summary**

The researcher has built a method on top of a k-member algorithm proposed by Byun et. al., who proposed clustering k-anonymity algorithm. The method uses their algorithm as a base and added different methods of finding next centre of a cluster to see how it would change the efficiency of the program. The change that was made on this algorithm is the distance where the next cluster is formed.

In comparison to the original algorithm, where the next cluster is formed furthest than the original, combining the k-mean framework on top of this k-member algorithm indeed increases the efficiency of the system.

6 DISCUSSION AND CONCLUSION

6.1 Introduction

This chapter concludes the study by summarizing the findings and gives suggestions for similar studies. This chapter also highlights the implication of the study.

6.2 Research Contributions

This study contributes to research literature as well as creating an anonymization framework that manipulates distance functions to create different clusters while maintaining a low rate of information loss, and thus, creating a dataset that is more valuable that can be used for data analysis in the future.

6.3 Implication of the study

This study explored how manipulating the distance function can affect the result of an anonymization, not only in how much faster or slower a data anonymization process can take, but also how differently the clusters of data are churned. It explores how the distance function affects the information loss of a certain cluster as well. From this study, it is found that by manipulating the distance of one centroid to another, there is a change in how fast the anonymization process happens, and the degree of information loss also differs. This is due to the outliers that are created (or left behind) after the clustering process is over. These outliers are then placed in the closest cluster, thus increasing the information loss percentage.

This study can be extended further by adapting to a much bigger dataset. It can also be adapted to cater to different types of distance functions, similar to how the researcher has used it.

6.4 **Research Achievements**

The research objective of this thesis is:

1. To investigate the current methods used in data anonymization.
2. To design and develop an enhanced method of data anonymization.
3. To evaluate the enhanced data anonymization in terms of information loss and efficiency.

The researcher has fulfilled all three objectives. To fulfill the first research objective, the researcher has carried out a systematic literature review in which the result has been discussed in Chapter 2. From that literature review, the researcher has found evidence to suggest that combining the k-anonymization concepts with clustering algorithms creates the best results in term of efficiency; where the researcher uses the running time as a benchmark, and the information loss percentage. Based on the review as well, the researcher has gathered the three most common ways of combining those methods for a new anonymization practice. From there, the researcher created a program comparison to see which method would most suit the researcher's purpose. To fulfill the second objective, the researcher then enhanced what the best method was in the previous result and finally, for the last objective, the testing was done against the previous method.

In studying previous methods, the researcher realized that despite k-Nearest Neighbor being the most efficient when it comes to processing speed, the Information Loss in comparison to k-member was higher, and also partly unstable as it is dependent on how the data is structured. And in k-member (the base of the current algorithm), based on the initial study, the researcher realized that the despite lacking in efficiency, the Information Loss is lesser as the clusters built are based on making sure it has the least information loss.

And thus, to increase the efficiency, the researcher created a framework that changes the way the next cluster centroid gets searched for.

The result was that when the algorithm gets changed to find the nearest neighbor (centroid) of the next cluster, efficiency is increased without sacrificing the information loss percentage by much.

6.5 **Limitations and Constraints**

The research was performed on a 2.20GHz Intel i5 processor machine with 12GB of RAM. The operating system on the machine was Microsoft Windows 10 and the implementation was built and run in JetBrains PyCharm 3.2 in a Python 3.2 environment. The result was dependent on the machine that was used for processing and the researcher suggests that the upgrade of the machine processing power would further improve the results.

Hence, results show a better performance compared to the original paper(by whom), as the specifications of the machine was significantly better than the original researcher's.

Another point to be considered is that much of currently available data are mostly unstructured. The unstructured data introduces an extra step, which is to identify and clean data that needed to be used.

Everything beyond the cleansing of data can be automated once it passes through the system, but definitions still need to be made, as is also the case for the categorical distance to be measured. The taxonomy tree also must be defined by the researcher and is not automated.

One other issue that can be mentioned here is the limited processing power to handle big amounts of data. This system can be further developed to be used on Big Data, which is where most research is heading towards, but processing Big Data requires a big processing power, which means a big machine.

6.6 **Suggestions for Future Research**

The system can be further developed to find a way to increase its efficiency by changing the way it looks for the records in the cluster. As it is currently, the system looks for the records that most satisfies the k-anonymity constraint by looking for the most dissimilar record and adding it one by one. This causes the complexity and the efficiency time to increase exponentially. Future researchers can find a way to find the records more efficiently.

As previously mentioned, this system can also be further developed to work in Big Data settings. The problem that may occur when running Big Data through this system can't yet be identified, as feeding copious amount of data may cause the system to fail, if not run for hours.

7 REFERENCES

Aggarwal, C. C., & Yu, P. S. (2008). A General Survey of Privacy-Preserving Data Mining Models and Algorithms. In C. C. Aggarwal & P. S. Yu (Eds.), *Privacy-Preserving Data Mining: Models and Algorithms* (pp. 11-52). Boston, MA: Springer US.

Aggarwal, G., Tom, #225, Feder, s., Kenthapadi, K., Khuller, S., . . . Zhu, A. (2006). *Achieving anonymity via clustering*. Paper presented at the Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Chicago, IL, USA.

Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus, 4*(1), 694. doi: 10.1186/s40064-015-1481-x

Amiri, F., Yazdani, N., Shakery, A., & Chinaei, A. H. (2016). Hierarchical anonymization algorithms against background knowledge attack in data releasing. *Knowledge-Based Systems, 101*, 71-89. doi: https://doi.org/10.1016/j.knosys.2016.03.004

Bayardo, R. J., & Rakesh, A. (2005, 5-8 April 2005). *Data privacy through optimal k-anonymization.* Paper presented at the 21st International Conference on Data Engineering (ICDE'05).

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association : JAMIA, 17*(2), 169-177. doi: 10.1136/jamia.2009.000026

Budgen, D., Burn, A. J., Brereton, O. P., Kitchenham, B. A., & Pretorius, R. (2010). Empirical evidence about the UML: a systematic literature review. *Software: Practice and Experience, 41*(4), 363-392. doi: 10.1002/spe.1009

Byun, J.-W., Kamra, A., Bertino, E., & Li, N. (2007, 2007//). *Efficient k-Anonymization Using Clustering Techniques.* Paper presented at the Advances in Databases: Concepts, Systems and Applications, Berlin, Heidelberg.

Ciriani, V., De Capitani di Vimercati, S., Foresti, S., & Samarati, P. (2007). κ-Anonymity. In T. Yu & S. Jajodia (Eds.), *Secure Data Management in Decentralized Systems* (pp. 323-353). Boston, MA: Springer US.

De Capitani di Vimercati, S., Foresti, S., Livraga, G., Paraboschi, S., & Samarati, P. (2015). *Privacy in Pervasive Systems: Social and Legal Aspects and Technical Solutions*. In F. Colace, M. De Santo, V. Moscato, A. Picariello, F. A. Schreiber & L. Tanca (Eds.), Data Management in Pervasive Systems (pp. 43-65). Cham: Springer International Publishing.

Domingo-Ferrer, J., & Soria-Comas, J. (2015). From t-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems, 74*, 151-158. doi: https://doi.org/10.1016/j.knosys.2014.11.011

Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. *Data Mining and Knowledge Discovery, 11*(2), 195-212. doi: 10.1007/s10618-005-0007-5

Dwork, C. (2008, 2008//). *Differential Privacy: A Survey of Results.* Paper presented at the Theory and Applications of Models of Computation, Berlin, Heidelberg.

Dwork, C. (2011). Differential Privacy. In H. C. A. van Tilborg & S. Jajodia (Eds.), *Encyclopedia of Cryptography and Security* (pp. 338-340). Boston, MA: Springer US.

Dwork, C. (2011, 22-25 Oct. 2011). *The Promise of Differential Privacy: A Tutorial on Algorithmic Techniques.* Paper presented at the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science.

El Emam, K., & Dankar, F. K. (2008). Protecting Privacy Using k-Anonymity. *Journal of the American Medical Informatics Association, 15*(5), 627-637. doi: 10.1197/jamia.M2716

Friedman, A., Wolff, R., & Schuster, A. (2008). Providing k-anonymity in data mining. *The VLDB Journal, 17*(4), 789-804. doi: 10.1007/s00778-006-0039-5

Ghinita, G., Karras, P., Kalnis, P., & Mamoulis, N. (2007). *Fast data anonymization with low information loss*. Paper presented at the Proceedings of the 33rd international conference on Very large data bases, Vienna, Austria.

Goodman, M. (2015). *Future Crimes: Inside the Digital Underground and the Battle for our Connected World.* New York: Anchor Books.

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., & Roth, A. (2014, 19-22 July 2014). *Differential Privacy: An Economic Method for Choosing Epsilon.* Paper presented at the 2014 IEEE 27th Computer Security Foundations Symposium.

Ijab, M. T., Bukhari, S., Norman, A. A., Hamid, S., & Ravana, S. D. (2016). *Role of social media in information-seeking behaviour of international students: A systematic literature review*. Aslib Journal of Information Management, 68(5), 643-666. doi: 10.1108/AJIM-03-2016-0031

Ke, W., Yu, P. S., & Chakraborty, S. (2004, 1-4 Nov. 2004). *Bottom-up generalization: a data mining solution to privacy protection.* Paper presented at the Fourth IEEE International Conference on Data Mining (ICDM'04).

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology, 51*(1), 7-15. doi: https://doi.org/10.1016/j.infsof.2008.09.009

Kitchenham, B., Pretorius, R., Budgen, D., Pearl Brereton, O., Turner, M., Niazi, M., & Linkman, S. (2010). Systematic literature reviews in software engineering – A tertiary study. *Information and Software Technology, 52*(8), 792-805. doi: https://doi.org/10.1016/j.infsof.2010.03.006

Leedy, P. D. (1997). Non-experimental research. *Practical research : planning and design*.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006, 3-7 April 2006). *Mondrian Multidimensional K-Anonymity.* Paper presented at the 22nd International Conference on Data Engineering (ICDE'06).

Li, N., Li, T., & Venkatasubramanian, S. (2007, 15-20 April 2007). *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity.* Paper presented at the 2007 IEEE 23rd International Conference on Data Engineering.

Lin, J.-L., Chang, P.-C., Liu, J. Y.-C., & Wen, T.-H. (2010). Comparison of microaggregation approaches on anonymized data quality. *Expert Systems with Applications, 37*(12), 8161-8165. doi: https://doi.org/10.1016/j.eswa.2010.05.071

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006, 3-7 April 2006). *L-diversity: privacy beyond k-anonymity.* Paper presented at the 22nd International Conference on Data Engineering (ICDE'06).

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008). *Privacy: Theory meets Practice on the Map*. Paper presented at the Proceedings of the 2008 IEEE 24th International Conference on Data Engineering.

Meyerson, A., & Williams, R. (2004). *On the complexity of optimal K-anonymity*. Paper presented at the Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Paris, France.

Presswala, F., Thakkar, A., & Bhatt, N. (2015). *Survey on Anonymization in Privacy Preserving Data Mining* (Vol. 2).

Rajendra, A. (2019). *Privacy and Security in Data-Driven Urban Mobility Utilizing Big Data Paradigms for Business Intelligence* (pp. 106-128). Hershey, PA, USA: IGI Global.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering, 13*(6), 1010-1027. doi: 10.1109/69.971193

Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., & Martínez, S. (2015). t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation. *IEEE Transactions on Knowledge and Data Engineering, 27*(11), 3098-3110. doi: 10.1109/TKDE.2015.2435777

Sweeney, L. (2002)(a). ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(05), 571-588. doi: 10.1142/S021848850200165X

Sweeney, L. (2002)(b). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,*

10(05), 557-570. doi: 10.1142/S0218488502001648

Terrovitis, M., Mamoulis, N., & Kalnis, P. (2011). Local and global recoding methods for anonymizing set-valued data. *The VLDB Journal, 20*(1), 83-106. doi: 10.1007/s00778-010-0192-8

Torra, V. (2004, 2004//). *Microaggregation for Categorical Variables: A Median Based Approach.* Paper presented at the Privacy in Statistical Databases, Berlin, Heidelberg.

Wang, J., Luo, Y., Zhao, Y., & Le, J. (2009, 25-26 April 2009). *A Survey on Privacy Preserving Data Mining.* Paper presented at the 2009 First International Workshop on Database Technology and Applications.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly, 26*(2), xiii-xxiii.

Xiao, X., & Tao, Y. (2006). *Anatomy: simple and effective privacy preservation.* Paper presented at the Proceedings of the 32nd international conference on Very large data bases, Seoul, Korea.

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information Security in Big Data: Privacy and Data Mining. *IEEE Access, 2*, 1149-1176. doi: 10.1109/ACCESS.2014.2362522

Zhong, S., Yang, Z., & Wright, R. N. (2005). *Privacy-enhancing k-anonymization of customer data*. Paper presented at the Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Baltimore, Maryland.

8 PUBLICATIONS

8.1 Papers

Title: Techniques for Big Data Protection
Journal: Online Information Review
Status: Revised and Submitted