A FRAMEWORK FOR PRIVACY DIAGNOSIS AND PRESERVATION IN DATA PUBLISHING

MOHAMMAD REZA ZARE MIRAKABAD

UNIVERSITI SAINS MALAYSIA
2010

A FRAMEWORK FOR PRIVACY DIAGNOSIS AND PRESERVATION IN DATA PUBLISHING

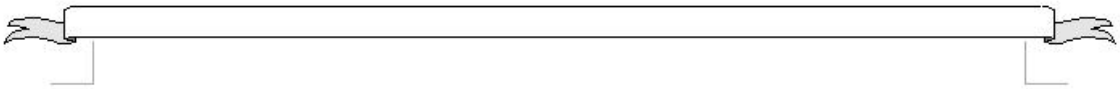
by

MOHAMMAD REZA ZARE MIRAKABAD

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

April 2010

DEDICATION



To:

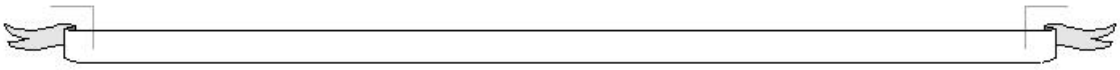
My dear wife, Zohreh, and
my lovely daughter, Parnian,
with much love and thanks

تقدیم به

همسر عزیزم زهره و

دختر گلم پرنیان،

همراه با عشق و قدردانی



ACKNOWLEDGEMENTS

I would like to thank all people who have helped and inspired me during my doctoral study.

First of all, I thank my supervisor Dr. Aman Jantan for his continuous advices during my PhD study. He showed me different ways to approach research problems and the need to be persistent to accomplish any goal. He taught me the way of research by mastering the field, mastering the research as well as mastering the writing.

I was delighted to become familiar with Prof. Stéphane Bressan through my internship and attachment to the project in the School of Computing at NUS. It is also my luck to have him as my supervisor. I am very thankful of him. His insights to the students and research is appreciated. He gave me good advices not only for writing and presenting papers, but also to take part in the academic societies. Dr. Ranaivo Malançon introduced him to me. Thank you Ranaivo Malançon.

My deepest gratitude goes to my family for their unflagging love and support throughout my life. I am indebted to all of them. I express my special thanks to my dear wife, Zohreh, and my lovely daughter, Parnian. I am certain that I could not have finished this study if they were not as remarkable as they are. When we arrived here, we started a new life, alone and in a completely new environment. They are totally perfect. I thank my late father; he worked industriously to support the family and spared no effort to provide the best possible conditions for me

to grow up and specially to study. He had never complained in spite of all the hardships he suffered in his life. Although he is no longer with us, he is forever remembered. I am sure he shares our joy and happiness in heaven. I don't know how I can thank my respectable mother to whom I owe many things. I have no suitable word that can fully describe her everlasting love to me and her patience. I remember her constant support when I encountered difficulties. The same goes to my father and mother in-law. I feel proud of them as they always worry about me and give me support and encouragement from far away.

Furthermore, many of my friends in Iran and also my new friends here have always been a constant source of encouragement during my graduate study. I admire their persistent and meticulous attitude.

Last but not least, thanks to God. You helped me in critical situations, when I was disappointed and I was close to give up. You were my last point of support and dependence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LISTINGS

LIST OF ABBREVIATIONS

adom	Active Domain
Bd$^+$	Positive Border
Bd$^-$	Negative Border
C$_{DM}$	Discernibility Metric
C$_{AVG}$	Average Discernibility Metric
DGH	Domain Generalizing Hierarchy
EMD	Earth Mover's Distance
IL	Information Loss
NCP	Normalized Certainty Penalty
PPDM	Privacy Preserving Data Mining
PPDP	Privacy Preserving Data Publishing
QId	Quasi-identifier
SQL	Structures Query Language
WHD	Weighted Hierarchical Distance

LIST OF SYMBOLS

$\|\ldots\|$	Cardinality of a set/multi-set
{...}	Set
{{...}}	Multi-set
$\subset$	Subset
$\subseteq$	Subset or equal

RANGKA KERJA UNTUK DIAGNOSIS DAN PEMELIHARAAN PRIVASI DALAM PENERBITAN DATA

ABSTRAK

Matlamat *pemeliharaan privasi* dalam *penerbitan data* ialah untuk menerbitkan data dengan melindungi maklumat sulit. Walaupun sekali pandang nampaknya membuang pengecam terus individu dapat melindungi ketanpanamaan individu tersebut, namun maklumat sulit boleh didedahkan dengan menyambung data itu kepada data luaran yang lain. *Pemeliharaan privasi* mengutarakan isu privasi ini dengan memperkenalkan prinsip *k-ketanpanamaan* dan *l-kepelbagaian*. Maka teknik pemeliharaan privasi iaitu algoritma k-ketanpanamaan dan *l*-kepelbagaian dapat mentransformasikan data ((misalnya melalui pengitlakan, penindasan atau penyerpihan) untuk melindungi identiti dan maklumat sensitif seseorang individu.

Kebelakangan ini, sebahagian besar usaha yang dilakukan untuk mengutarakan isu ini memberi tumpuan pada teknik pemeliharaan privasi. Namun demikian, tidak banyak usaha dilakukan untuk menghasilkan teknik, alat dan metodologi yang dapat membantu penerbit, pengurus dan juruanalisis data dalam penyelidikan dan penilaian risiko privasi. Justeru itu, disarankan idea penubuhan sebuah **pusat diagnosis privasi** yang menyediakan rangka kerja yang sewajarnya bagi pendiagnosisan risiko privasi dan lebih khusus lagi pendiagnosisan k-ketanpanamaan dan *l*-kepelbagaian. Masalah ini didapati merupakan suatu masalah penemuan pengetahuan yang dapat dipetakan kepada rangka kerja yang disarankan oleh Mannila dan Toivonen. Dengan memperkenalkan dan membuktikan sifat "monotonicity" yang wajar, algoritma "level-wise" yang wajar berdasarkan algoritma apriori dikemukakan dan dinilai.

Tambahan pula, model dan teknik yang disarankan untuk pemeliharaan privasi masih mempunyai beberapa kekurangan dan kelemahan. Secara khususnya, algoritma berasaskan penggugusan untuk k-ketanpanamaan boleh menyebabkan kehilangan maklumat yang tinggi. Dengan menunjukkan kekurangan kedua-dua gugusan kecil dan besar, k-ketanpanamaan **penggugusan dua fasa** disarankan. Gugusan akan menjadi cukup besar dan selepas itu gugusan besar dipisahkan kepada gugusan-gugusan yang sekecil mungkin dalam fasa seterusnya. Ini mengakibatkan kehilangan maklumat yang rendah. Tambahan pula, ditunjukkan bahawa perluasan algoritma k-ketanpanamaan bagi prinsip l-kepelbagaian tidak begitu mudah dilakukan. Ia boleh menyebabkan kehilangan maklumat yang tinggi atau tidak dapat dihentikan. Oleh sebab itu, disarankan l-kepelbagaian **penggugusan baldi** untuk menjamin penghentian dan juga kehilangan maklumat yang rendah.

Algoritma yang disarankan itu telah dilaksana dan menggunakan dua dataset sampel, iaitu Adults dan OCC, yang merupakan tanda aras de facto bagi algoritma pemeliharaan privasi. Dengan menganalisis keputusan yang diperoleh, keberkesanan dan kecekapan rangka kerja dan algoritma yang disarankan telah dapat dibuktikan secara eksperimen.

A FRAMEWORK FOR PRIVACY DIAGNOSIS AND PRESERVATION IN DATA PUBLISHING

ABSTRACT

Privacy preservation in *data publishing* aims at the publication of data with protecting private information. Although removing direct identifier of individuals seems to protect their anonymity at first glance, private information may be revealed by joining the data to other external data. *Privacy preservation* addresses this privacy issue by introducing *k-anonymity* and *l-diversity* principles. Accordingly, privacy preservation techniques, namely k-anonymization and *l*-diversification algorithms, transform data (for example by generalization, suppression or fragmentation) to protect identity and sensitive information of individuals respectively.

Most of the recent efforts addressing this issue have focused on privacy preservation techniques. However, not much effort has been made to address devising techniques, tools and methodologies to assist data publishers, managers and analysts in their investigation and evaluation of privacy risks. Hence, the idea of a **privacy diagnosis centre** is proposed that offers the necessary framework for diagnosing privacy risk and specifically k-anonymity and *l*-diversity. It is shown that this problem is a knowledge discovery problem that can be mapped to the framework proposed by Mannila and Toivonen. By introducing and proving the necessary monotonicity properties, necessary levelwise algorithms based on the apriori algorithm are presented and evaluated.

Moreover, proposed models and techniques for privacy preservation still have some deficiencies and drawbacks. Specifically, clustering-based algorithms for k-anonymization may result in high information loss. By showing the deficiencies of both small and big clusters, **two-phase clustering** k-anonymization is proposed. It allows clusters to become sufficiently big, and big clusters are split to smallest possible clusters in the next phase, both result in lower information loss. In addition, it is shown that the extension of k-anonymization algorithms for some l-diversity principles is not straightforward. It may result in high information loss or can not terminate. Accordingly, **bucket clustering** l-diversification is proposed to guarantee both termination and low information loss.

The proposed algorithms are implemented and ran on two sample datasets, namely Adults and OCC, which have become de facto benchmarks for privacy preservation algorithms. Effectiveness and efficiency of the proposed framework and algorithms are proved experimentally by analyzing the results.

CHAPTER 1

INTRODUCTION

1.1 Overview

In the age of information with exponential growth in the number and variety of data collections containing person-specific information, there is also tremendous demand for person-specific data. It is not only necessary for applications such as data mining, cost analysis and fraud detection, but also for other fields of research such as health care, risk analysis, insurance stability and so on. These data recipients are usually named *third parties* in the process of data publishing.

Besides, organizations and professionals need to publish operational data in order to ensure business visibility and effective presence on the World Wide Web. Individuals publish personal data in the hope of becoming socially visible and attractive in the new electronic communication forums. Consequently, large amounts of data with high level of details in the numerous sources are publicly available, which is usually named *microdata*. This may make *privacy* of individuals at risk. For example, even though the data may locally seem to respect privacy, cross-referencing with external data and statistical inferences can disclose more information than intended. Hence, *privacy preservation* in *data publishing* has become one of the most important research problems during the last decade. It addresses protecting privacy of individual entities to whom the

data belong. Organizations and companies want or need to share person-level data with keeping as much details as possible while making sure the information is sufficiently protected. This protection is necessary to prevent identity and sensitive information disclosure while detailed information of individuals is published (Oliveira and Zaiane, 2007; Aggarwal and Yu, 2008).

Typically, person-specific data (microdata) is stored in a table where each row (tuple) corresponds to one individual. This table has 4 kinds of attributes:

- Attributes like "Name", "Social security number" or "Driving license number" that uniquely identify individuals. They are referred to as *identity*.
- Attributes like "Income" for bank customers or "Disease" for hospital patients that are important for data holder, but have to be remained private for individuals. These attributes are named *sensitive attributes*.
- Set of attributes like {Age, Gender, Zip-code} that can be used by combination to identify some individuals. They are named *quasi-identifiers*.
- Other attributes that do not fall into the previous categories and the data holder can publish them without considering any protection. These attributes are named *normal attributes* or *non-sensitive attributes*.

Any privacy preservation process is started by protecting direct identity of individuals. They are generally removed or replaced by random values, the process which is named de-identification. However, this may not be enough because such de-identified data can sometimes be joined with other public databases (which

are usually named external data) on some combination of attributes to re-identify individuals who are supposed to remain anonymous. In literatures, for example (Sweeney, 2002b; Machanavajjhala et al., 2006; LeFevre et al., 2006a), this kind of attack is named *re-identification*, *joining attack* or *cross-referencing*. Accordingly, the attributes combination which make this join possible and result in identity leakage are named *quasi-identifier* (Byun et al., 2006a; Sweeney, 2002b; Xu et al., 2006b). As a real case, Sweeney [2002] empirically showed that 87% of the people in the U.S. can be uniquely identified by the combination of "Gender", "Date of birth", and "Zip-code"; thus, {Gender, Date of birth, Zip-code} forms a quasi-identifier for the U.S. population (Sweeney, 2002b). She could reveal disease of William Weld, governor of the state of Massachusetts, by joining public voter list and published medical database on this set of attributes, as shown in Figure 1.1. The uniqueness of such attributes combination leads to cross-referencing where data is re-identified by joining to publicly available datasets.

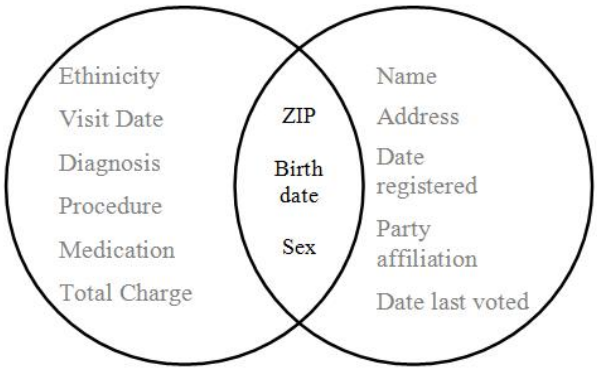


Figure 1.1: Linking (join attack) for re-identifying data (Sweeney, 2002b)

Example 1 (Private table and de-identification).

Consider Table 1.1 showing patient records of a fictitious hospital. Its de-identified version is also shown in Table 1.2. Although it seems de-identified data is protected at first glance, an adversary still can attack the data and in which the "Name", and subsequently the "Disease" of the patients may be disclosed. In fact, this is feasible when only one joinable record exists in an external table with one of the records in the published data. Such external table can be a voter list such as shown in Table 1.3 below. In this example, one can reveal "Alice" by using combination of attributes {Zip, Gender, Age} and joining two tables since value of these attributes set is unique for her.

Table 1.1: Patient records of a fictitious hospital

Name	**Age**	**Gender**	**Zip**	**Disease**
Alice	21	Female	17651	Cancer
Jack	22	Male	17652	Flu
Jan	23	Male	17661	HIV
Bob	24	Male	17662	HIV

Table 1.2: De-identified table of patient records

Age	**Gender**	**Zip**	**Disease**
21	Female	17651	Cancer
22	Male	17652	Flu
23	Male	17661	HIV
24	Male	17662	HIV

Table 1.3: Voters list as external data

Name	**Address**	**Gender**	**Age**	**Zip**
...	... ...	...	...	...
Alice	No15, Lakeside Street	Female	21	17651

1.2 Motivation

To prevent cross-referencing or join attacks, initial data is modified before release. The modification has to keep as much details in the data as possible, which is referred to as *data utility*; while still ensuring the information is sufficiently de-identified, which is referred to as *data privacy*. Intuitively, data privacy can be enhanced by hiding more data values, but it decreases data utility. On the other hand, revealing more data values increases data utility, but it normally decreases data privacy. Thus, as also proposed by Li and Li (2009), it is necessary to devise solutions that best address both utility and privacy requirements of data.

Two main principles of privacy preservation in data publishing are known as *k-anonymity* and *l-diversity*. k-anonymity, as defined and used in (Aggarwal et al., 2006, 2005; Bayardo and Agrawal, 2005; LeFevre et al., 2005; Samarati, 2001; Samarati and Sweeney, 1998), guarantees that any record in the released data is indistinguishable from at least $k-1$ other records with respect to the quasi-identifier. Hence, a join with a k-anonymous table would give k or more matches and create confusion. Then, an individual is hidden in a crowd of size k and has k-anonymity. The process of modification resulting in this principle is named *k-anonymization*. This requirement is typically enforced through generalization, where real values are replaced with "less specific but semantically consistent values" (Sweeney, 2002a). For example, Table 1.4 shows 2-anonymous version of the tuples of Table 1.2 after generalization. It is clear that even an adversary knows "Alice" as one of the two persons in the first group, however, he/she now cannot

infer which one is exactly “Alice” and therefore which disease she has contracted, i.e. “Cancer” or “Flu”.

Table 1.4: 2-anonymous version of Table 1.2

Age	Gender	Zip	Disease
[21-22]	*	1765*	Cancer
[21-22]	*	1765*	Flu
[23-24]	Male	1766*	HIV
[23-24]	Male	1766*	HIV

While k-anonymity prevents identity of individuals from being revealed in published data, it fails to protect sensitive information of individuals. For example, if an adversary knows “Bob” is one of the last two tuples of Table 1.4, even though he/she cannot understand which one belongs to “Bob”, he/she can infer that he has been infected by “HIV” with 100 percent confidence. This is due to the fact that both persons in his group have the same disease.

l-diversity (Li et al., 2007; Machanavajjhala et al., 2006; Xiao and Tao, 2006a; Iyengar, 2002) aims at privacy preservation by preventing inferences of unwanted information. It guarantees that one cannot associate an object with the sensitive information beyond a certain probability. This is achieved by ensuring that values of sensitive attributes are well represented as per the *l*-diversity principle announced in (Machanavajjhala et al., 2006). The process of modification resulting in this principle is named *l-diversification*. Table 1.5 shows 2-diverse version of the Table 1.2. In the table, each group of tuples has 2 different sensitive values (*disease* in this example), thus the actual disease of patients cannot be inferred with probability more than $\frac{1}{2}$.

Table 1.5: 2-diverse version of Table 1.2

Age	Gender	Zip	Disease
[21-23]	*	176**	Cancer
[21-23]	*	176**	HIV
[22-24]	Male	176**	Flu
[22-24]	Male	176**	HIV

The term *privacy* concerned in this study is about the anonymity of private information for individuals while the detailed information is published. It addresses protection of *identity* and *sensitive information* of individuals in the published data. The term *privacy preservation* is a process of data modification to achieve the demand level of the privacy before one publishes the data.

Most of the recent efforts addressing the issue of privacy focused on privacy preservation. However, fewer efforts have been made to devise techniques, tools and methodologies that assist data publishers, managers and analysts in their investigation and evaluation of privacy risks. This is the motivation for proposing and introducing idea of **privacy diagnosis**. In fact, we have to know the privacy risks before any modification could be done. Thus, measuring the privacy level that exist in the data and showing privacy threats to the data holder are necessary. *Privacy diagnosis*, as an upstream of privacy preservation, tries to answer the questions about the existing level of anonymity, and privacy threats in the data, based on different privacy aspects and principles.

Besides, although some of the new efforts in anonymization exploit clustering technique, the attention to the methods and capability and enhancement that they

may offer is not completely studied. They may result in *low data quality* or *high information loss*. Therefore, this study tries to provide an enhanced modification of clustering-based algorithm for **k-anonymization**. Moreover, some of the recently proposed approaches trying to extend k-anonymization to support l-diversity fail to address various principles such as frequency l-diversity. It can not be done straightforwardly as they may not successfully terminate. This is the motivation of the research to propose an algorithm for **frequency l-diversification**.

1.3 Goals, objectives and scope

This research aims at both *privacy diagnosis* and *privacy preservation* which are two important phases in *safe data publishing* process. The former is used by data holder before starting any modification to show privacy risks. The latter can be used to improve data privacy for achieving the level of privacy demanded, if it is not satisfactory. In the first phase, one needs to analyze the data to measure privacy risks and show to data holder. In fact, we propose a framework to investigate privacy risks before any modification methods can be exploited. We show the necessity of this important phase of the data publishing which is currently still not addressed by most researches. For the privacy preservation phase, we will introduce algorithms for both k-anonymity and l-diversity principles using clustering-based techniques.

To achieve the above goals, we investigate the state-of-the-art k-anonymity and l-diversity principles to address following objectives:

- To introduce a **privacy diagnosis centre** for measuring k-anonymity and *l*-diversity level in data. Such diagnosis centre will prepare:
  - Good understanding of the quasi-identifiers for the given dataset.
  - Understanding the effect of different choices of *k* for the given dataset.
  - Possible minimal suppression for achieving a desired level of k-anonymity.
  - Understanding the threat of attribute disclosure in the given dataset.
  - Measuring the level of guaranteed diversity by different notions of diversity.
- To propose a clustering-based algorithm for k-anonymization resulting in local recoding with minimum information loss.
- To extend clustering k-anonymization algorithm to be applicable for various instances of *l*-diversity principles.

To address the above objectives, we propose *a framework for privacy diagnosis and preservation in data publishing* as described in the next section.

For the sake of simplicity, only one attribute is considered as sensitive attribute in this study. However, as explained in (Gal et al., 2008), definitions and algorithms can be straightforwardly extended to multiple sensitive attributes. Moreover, for measuring the anonymity, monotonicity property, lemmas and proofs are given for k-anonymity and 3 representative principles of *l*-diversity. The same properties and proofs have to be considered for various *l*-diversity principles, if necessary.

1.4 Privacy diagnosis and preservation framework

The general framework of privacy diagnosis and preservation in data publishing is given in Figure 1.2. This framework consists of two major parts, namely **Privacy Diagnosis Centre** and **Privacy Preservation Module**. The diagnosis centre will be used for measuring of the k-anonymity and *l*-diversity principles level and diagnosing privacy threats, whereas preservation module is for modifying data to guarantee desired k-anonymity and *l*-diversity level. The initial data (private data) and privacy parameters, if any, are given by data holder as input. Then the framework is used either for measuring the level of target privacy principle and showing privacy risks, or for applying privacy preservation methods for protecting privacy of published data. Hence, the output may be privacy principles level, privacy threats, (e.g. by giving positive and negative borders), generalized data according to demand privacy level, or any combination of them.

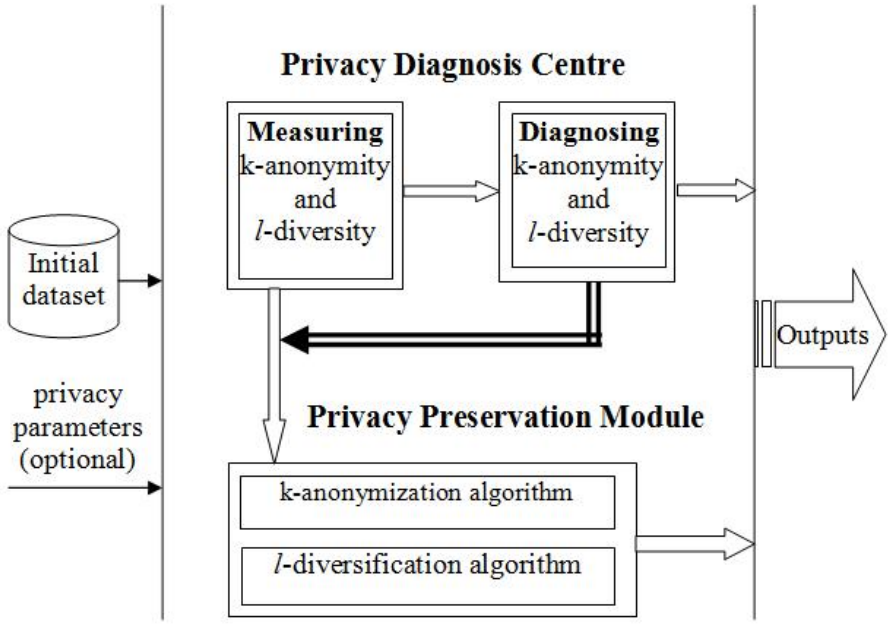


Figure 1.2: General framework for *privacy diagnosis* and *privacy preservation*

1.5 Thesis contributions

Below are the key contributions of this thesis:

I) Privacy diagnosing:

We propose the idea of a **privacy diagnosis centre** that offers the necessary framework for the exploratory analysis of the data and various publication scenarios. Such a diagnosis centre should answer to questions about existing level of k-anonymity and l-diversity of data. It also should explore the data to indicate privacy threats by giving subset of attributes that can be published safely and/or jeopardize privacy.

II) Enhancing k-anonymization by two-phase clustering:

We propose a clustering based k-anonymization method and show its efficiency and effectiveness. This greedy clustering algorithm considers defects of both small and big clusters. A **two-phase clustering** k-anonymization algorithm will be introduced accordingly.

III) Enhancing l-diversification by bucket clustering:

We show deficiency of recent proposed l-diversification methods which are trying to extend k-anonymization for supporting various principles of l-diversity. We benefit from two already proposed anonymization algorithms, namely *greedy k-member clustering* (Byun et al., 2006a) and *Anatomy* (Xiao and Tao, 2006a). We also show how an algorithm can be devised to achieve frequency l-diversity with less information loss while guaranteeing successful termination. Accordingly, a **bucket clustering** frequency l-diversification algorithm is introduced.

1.6 Structure of the thesis

The rest of this thesis is organized as follows.

In **Chapter 2**, some of the basic notions of anonymity and diversity and state-of-the-art studies in this field are reviewed. Related works in both identity disclosure and attribute disclosure are covered by concentrating on different aspects of the problem considered by each proposed technique.

Chapter 3 describes three main proposals of the thesis in different subsections. Firstly, devised **privacy diagnosis centre** is introduced by proposing the necessary framework for measuring k-anonymity. The diagnosis centre is advanced to aim at diversity diagnosis. Secondly, **two-phase clustering** k-anonymization algorithm is introduced. It includes *relaxing phase* allowing clusters to become big and *splitting phase* dividing big clusters, both results in less information loss. Finally, frequency l-diversity modification process is addressed by introducing **bucket clustering** l-diversification algorithm. It exploits some criteria resulting in less information loss meanwhile assuring termination.

Chapter 4 presents experimental results of the implemented algorithms followed by their analysis and discussion. Benefits and effectiveness of algorithms are explained by some examples and cases, especially by the vast range of questions that can be answered by diagnosis centre. Then, algorithms are run with a set of actual datasets and their effectiveness and efficiency are empirically shown. Each part is followed by analysis of the results and discussion.

Chapter 5 concludes this dissertation and outlines the directions for the future researches.

CHAPTER 2

LITERATURE REVIEW

In this chapter, state-of-the-art studies in the field of *privacy preservation* and *privacy diagnosis* in *data publishing* are surveyed. In Section 2.1, process of data publishing is explained to show the privacy violation and need of privacy preservation, when the data is published for third parties. Then, definition of the privacy preservation principles following their features and techniques are explained in Section 2.2. In Section 2.3, lack of the privacy measuring is shown that leads to proposing the privacy diagnosis centre. The chapter is summarized with justifying the need of *a framework for privacy diagnosis and preservation in data publishing* in Section 2.4. The outline of the chapter is shown by the taxonomy in Figure 2.1.

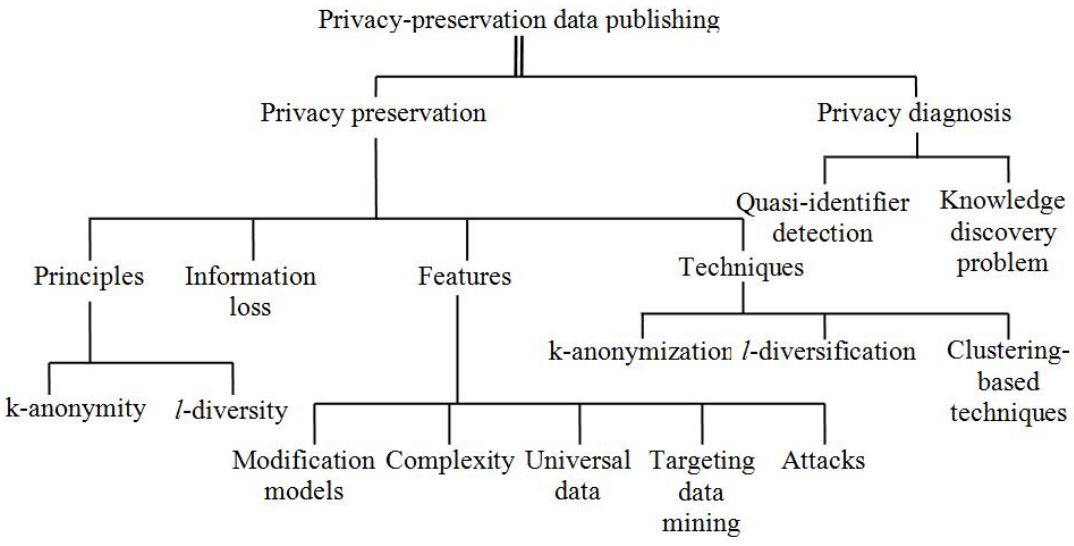


Figure 2.1: Taxonomy of related studies in privacy diagnosis and preservation

2.1 Data publishing

A typical scenario of data publishing is shown in Figure 2.2. There are three important actors in the process of data publishing, namely *individual entities*, *data owner* and *third party*. Individuals are any kind of entities, specifically people, that data belongs to them and contains their properties. They are usually named microdata in literatures. Data owner (or data holder) is one who collects the data and has this right to use them without worry about privacy concern. Third parties (or data recipients) are other data recipients such as research centers or organization that need the data for more exploration and analysis, or to extract new knowledge from it. For example, when a hospital collects the data and publishes it for investigation of an epidemic of a disease, "patients" are the individuals, "hospital manager" is the data owner and another "medical center" that receives the data is the third party.

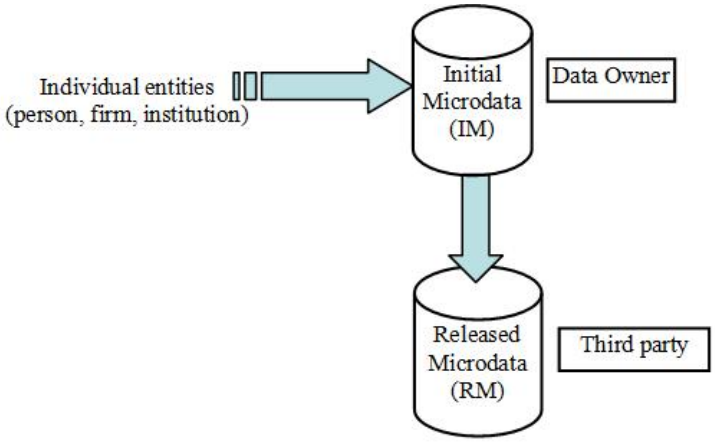


Figure 2.2: Data publication process

As a very simple privacy issue, direct identity of individuals such as "Name" and "Social security number" are generally removed or replaced by random values

before publishing to third parties. However, this de-identification process is not enough because of existence of *quasi-identifiers* that can result in disclosure of identity and sensitive information of individuals. More attentions need to be considered before publishing the data to prevent privacy breaches. *Privacy preservation* exactly aims at this problem by introducing methods and techniques to publish the data and guarantee anonymity. In fact, in the process of safe data publishing by privacy preservation, some modifications are necessary to convert the initial microdata (that data owner has) to released data (that data recipient is received), as shown in Figure 2.3

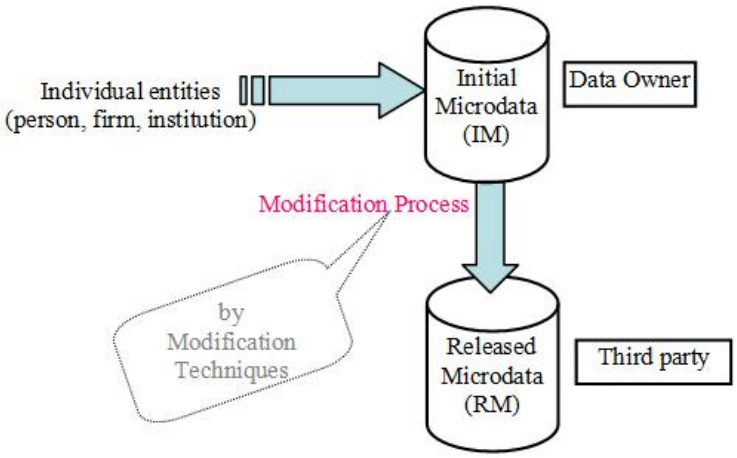


Figure 2.3: Privacy preservation in data publishing

2.2 Privacy preservation

In data publishing, privacy preservation exactly aims at the problem of sharing data with guarantee of anonymity. It considers how data owner can publish data without compromising the privacy of individuals and business entities reflected in the data. It addresses privacy concerns in data published to third parties with the

smallest possible modification in the initial data and simultaneously guaranteeing anonymity of individuals. In this section, main principles, important features and introduced techniques for privacy preservation are explained. Necessary definitions and algorithms are also given to show their deficiency and drawback leading to the proposing of the privacy preservation algorithms.

2.2.1 Principles

k-anonymity and l-diversity are two main principles of privacy preservation in data publishing. The process of data modification resulting in these principles is named k-anonymization and l-diversification respectively. These two principles and their variations will be reviewed in next subsections.

2.2.1.1 k-anonymity

In k-anonymity (Sweeney, 2002b), data privacy is guaranteed by ensuring that any record in the released data is indistinguishable from at least $k-1$ other records with respect to the quasi-identifier. Clearly, a join with a k-anonymous table would give rise to k or more matches and creates confusion. Thus, an individual is hidden in a crowd of size k giving k-anonymity. It also means that the identity disclosure risk is at most $\frac{1}{k}$ for *join* class of attacks. This aspect of privacy preserving is also referred to as *protecting identity disclosing* in some studies, such as (Truta et al., 2005, 2006). The main objective of k-anonymity is to anonymize a table so that nobody can make high-probability associations between records in the table and the corresponding entities.

Almost all literatures ((Sweeney, 2002b; LeFevre et al., 2005; Lodha and Thomas, 2006) to cite a few) mention that the quasi-identifier and non-sensitive attributes are determined based on background information such as previous data releases, knowledge of potential adversary, or the content of externally available data. Therefore, we assume identity attributes of the table has been removed and sensitive attributes are given by data holder based on his/her background and domain knowledge. Moreover, non-sensitive attributes do not play any role from the privacy preservation point of view and we can ignore them without loss of generality. Hence, given table has two subsets of attributes: *quasi-identifier* attributes and *sensitive attributes*. It can be shown by $T\{Q_1, Q_2, \ldots, Q_d, S_1, S_2, \ldots, S_m\}$ where $\{Q_1, Q_2, \ldots, Q_d\}$ is quasi-identifier set (QA) and $\{S_1, S_2, \ldots, S_m\}$ is sensitive attributes set (SA). Privacy preservation is about protecting re-identification of individuals' identifier based on their quasi-identifier and also preventing disclosure of sensitive attributes values.

As a simple definition, table T{QA,SA} is k-anonymous with respect to a quasi-identifier QA if and only if each distinct set of values in QA appears at least k times in T (Sweeney, 2002b; Byun et al., 2006a; Lodha and Thomas, 2006). It means every record in a k-anonymous table is indistinguishable from at least $k-1$ other records with respect to the quasi-identifier set. A group of records that are indistinguishable to each other is named an *equivalence class*.

Definition 2.1 (**Equivalence class with respect to a set of attributes**).

Given an instance r[1] of a relation R[2] and a set of attributes $Q \subseteq R$; $e \subseteq r$ is an equivalence class with respect to Q if and only if e is the multi-set of tuples in r that agree on the values of their attributes in Q. The empty equivalence class is ignored.

According to this definition, equivalence classes are the equivalence classes of the relation "having the same values for the attributes in Q" on tuples. The notion induces a partitioning of r. This notion is used in (Byun et al., 2006a; Li et al., 2006; Wong et al., 2006), to cite a few.

In this study, $r(Q,s)$ refers to the instance r of R in which $s \in R$ is the sensitive attribute[3], $Q \subseteq R$ is the set of non-sensitive attributes and $s \notin Q$. $r(R)$, or r, is used when the sensitive attribute doesn't exist or is not targeted, for the sake of simplicity.

Definition 2.2 (**k-anonymity**).

Given an integer k, an instance r of a relation R is k-anonymous with respect to $Q' \subseteq Q$ if and only if the cardinality of every equivalence class with respect to Q' is greater than or equal to k and r is not k+1-anonymous.

This definition of k-anonymity is compatible with but not identical to the definitions given in other papers such as (Sweeney, 2002b; Byun et al., 2006a; LeFevre et al., 2006a). This is a recursive definition that chooses k to be exactly

[1] r is a multi-set (i.e. it can contain duplicates).

[2] R is both the name of a relation and its schema (i.e. a set of attributes).

[3] This work is easily extended to multiple sensitive attributes (combinations of attributes).

the minimum cardinality of an equivalence class with respect to Q'. Without this recursion ("not $k+1$ anonymous"), an instance which is k-anonymous would also be k-1-anonymous. With the recursive definition, it is not the case.

Enforced by the k-anonymity requirement, it is guaranteed that even though an adversary knows that a k-anonymous table T contains the record of a particular individual and also knows values of the quasi-identifier attributes, he/she cannot determine which record in T corresponds to the individual with a probability greater than $\frac{1}{k}$.

The k-anonymity requirement is typically enforced through *generalization*, where real values are replaced with "less specific but semantically consistent values" (Sweeney, 2002b) from the domain of each attribute. This process is usually named *data modification*. There are various ways to modify the values of each domain. For instance, "Zip" codes '47907' and '47903'can be generalized to '4790*' (i.e., replacing least significant digit by * to cover a set of values), or even may be generalized to '*', that is a range covering every possible values. This is usually referred to as *value suppression*.

2.2.1.2 *l*-diversity

Although k-anonymity is helpful for protecting identity of individuals, a k-anonymous table can still be attacked to disclose sensitive information of individuals. For instance, if all patients of an equivalence class have the same disease, one can understand disease of a victim, though it is not possible to infer his/her actual

identifier. It motivated researchers to consider more sophisticated models to protect the association of individuals to sensitive information rather than k-anonymity. This anonymity principle is named l-diversity (Loukides and Shao, 2007; Xiao and Tao, 2006a; Kifer and Gehrke, 2006; Machanavajjhala et al., 2006; Wong et al., 2006). l-diversity is defined with respect to the sensitive attributes. It requires that each value of the sensitive attribute s in an equivalence class with respect to Q be "well-represented". Different instances of l-diversity differ in their realization of the property of "well-represented"-ness. For example, as shown in (Xiao and Tao, 2006a), when the number of sensitive values of an attribute in a class of tuples is l, sensitive value of an individual can be inferred with probability $\frac{1}{l}$, not $\frac{1}{k}$, even the data is k-anonymous with $k > l$.

While k-anonymity prevents identification, l-diversity aims at protecting sensitive information. Iyengar (2002) characterizes k-anonymity and l-diversity as *identity disclosure* and *attribute disclosure*, respectively (Iyengar, 2002). While the former tries to prevent disclosing of identity of individuals, the latter tries to protect association of individuals to their sensitive values. l-diversity guarantees that one cannot associate an object with sensitive information beyond a certain probability. This is achieved by ensuring that values of sensitive attributes are "well represented" as per the l-diversity principle declared in (Machanavajjhala et al., 2006).

Different instances of this principle, together with corresponding transformation processes, have been proposed. For instance, *distinct l-diversity* (Li et al., 2007),

entropy l-diversity and *recursive (c,l)-diversity* (Machanavajjhala et al., 2006), *(α,k)-anonymity* (Wong et al., 2006), and *t-closeness* (Li et al., 2007) are some of the proposed instances. Confusingly, the name *l*-diversity is sometimes used by authors to refer to some of the above instances rather than to the general principle.

The simplest instance of *l*-diversity counts the number of distinct values of the sensitive attribute in each equivalence class and requires that it be bigger or equal to l. Distinct *l*-diversity has been defined and used in some previous studies such as (Li et al., 2007; Ghinita et al., 2007; Truta and Vinay, 2006).

Definition 2.3 (**Distinct *l*-diversity**).
An instance r(Q, s) of a relation R is distinct l-diverse if and only if for each equivalence class e with respect to Q:

$$|\{v|v \in dom(s) \wedge \exists t(t \in e \wedge t.s = v)\}| \geq l \ ,$$

where $dom(s)$ is the domain of the attribute s[4].

Another important and more applicable interpretation of *l*-diversity requires that each value of the sensitive attribute in each equivalence class e appears at most $|e|/l$ times in e. We call and refer to this form of *l*-diversity as "frequency *l*-diversity" in order to differentiate it from other definitions, although this name is not originally used by the authors using the notion (Xiao and Tao, 2006a; Ghinita et al., 2007; Wong et al., 2006).

[4]Two surrounding mid symbols "$|\cdots|$" are used to denote cardinality.

Definition 2.4 (**Frequency l-diversity**).

An instance $r(Q,s)$ of a relation R is frequency l-diverse if and only if for each equivalence class e with respect to Q and each possible value $v \in adom(s)$:

$$p(e,v) \leq \frac{1}{l} ,$$

where $adom(s) = \{v | v \in dom(s) \wedge \exists t(t \in e \wedge t.s = v)\}$, the active domain of s, and $p(e,v) = |\{\{t | t \in e \wedge t.s = v\}\}|/|e|$ (note that e is a multi-set [5].

Machanavajjhala et al. (2006) propose entropy l-diversity and recursive (c,l)-diversity. Recursive (c,l)-diversity assures that "the most frequent value of sensitive attribute in each equivalence class is not too frequent, and the less frequent doesn't appear too rare" (Machanavajjhala et al., 2006). Let m be the number of sensitive values in an equivalence class and r_i is the frequency of the i_{th} most frequent values. A table is (c,l)-diverse if and only if, for each equivalence class, $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$, that is, number of occurrences of the most frequent sensitive value in each equivalence class is less than the sum of the frequencies of the $m - l + 1$ least frequent sensitive values, multiplied by a user defined constant c.

Entropy l-diversity is another variant of l-diversity principle. It measures the closeness of the distribution of values of the sensitive attribute in each equivalence class to the uniform distribution. It requires its entropy (as used in information theory) be bigger than $log(l)$ for a given l.

[5] The opening and closing double curly brackets "$\{\{\cdots\}\}$" are used to denote a multi-set.

Definition 2.5 (**Entropy l-diversity** (Machanavajjhala et al., 2006))**.**

An instance $r(Q,s)$ of a relation R is entropy l-diverse if and only if for each equivalence class e with respect to Q:

$$H(e) \geq \log(l) ,$$

where $H(e) = -\sum_{v \in adom(s)} p(e,v)\log(p(e,v))$ *is the entropy of the equivalence class,* $adom(s)$ *is the active domain of s and* $p(e,v) = |\{\{t|t \in e \wedge t.s = v\}\}|/|e|$ *is the fraction of tuples in e with sensitive value equal to v (as in Definition 2.4). Consider* $p(e,v)\log(p(e,v))$ *is 0 if* $p(e,v)$ *is 0.*

Wong et al. (2006) proposed another model for protecting the association between individuals and sensitive information. They named their model (α,k)-anonymity (Wong et al., 2006), that actually is one extension of l-diversity problem by extending the *Incognito* (LeFevre et al., 2005). They defined a new notion, α-de-association requirement, as a value that shows which degree of diversity exists in the sensitive attribute of the tuples in an equivalence class. It shows that frequency of a sensitive value, s, in every equivalence class is less than α. Actually what is considered in (Wong et al., 2006) as sensitive information is only some of the values of sensitive attribute. Then, the problem that they addressed is preventing association between quasi-identifier and sensitive values instead of considering all values of the sensitive attribute. (α,k)-anonymity, is a special kind of frequency l-diversity but for the selected values of the sensitive attributes known to be sensitive values.