UNIVERSITY OF SOUTHERN QUEENSLAND

ACCESS CONTROL MANAGEMENT AND PRIVACY-PRESERVING

A Dissertation submitted by

Md Enamul Kabir
B.Sc. Honours, M.Sc.

For the award of
Doctor of Philosophy

2010

Certification of dissertation

I hereby declare that the work presented in this dissertation is my own and is, to the best of my knowledge and belief, original except as acknowledged in the text. It has not previously been submitted either in whole or in part for a degree at this or any other university.

Signed: Date:

(Md Enamul Kabir)

Signed: Date:

(Hua Wang)

(Principal Supervisor)

Acknowledgements

First of all, I would like to thank almighty Allah, for His guidance and strength.

I would like to express my sincere gratitude and appreciation to my supervisor Associate Professor Hua Wang for his continuous inspiration, encouragement, patience, and individual feedback throughout the course of my Ph.D. study. I feel very grateful and blessed to have worked under his supervision. Special thanks to him for granting me the ARC scholarship to pursue my Ph.D. study. I would also like to express my sincere gratitude and appreciation to my co-supervisor Dr. Richard Watson for his advice and suggestions.

I would also like to thank the Center for Systems Biology (CSBi) for granting me half-tuition fee support. I would also like to thank the University of Dhaka, Bangladesh for the scholarship for study abroad.

I would also like to express my sincere gratitude and appreciation to the present head of the department of Maths and Computing (HOD) Dr. Stijn Dekeyser, former HOD Associate Professor Ron Addie and Dr. Richard Watson for their co-operation.

I extend my sincere and deep appreciation to my beloved wife Siuly and my son Shadman Srijon for their patience and continuous support. They have always been here with love and compassion to comfort me.

At last, but not the least, I wish to express my appreciation to my adorable parents, my brothers and my friends and relatives for their prayers, love and encouragement.

Abstract

In recent years, the phenomenal technological developments in information technology have led to an increase in the capability to store and record personal data about customers and individuals. This has led to concerns that the personal data may be misused for a variety of purposes. In order to alleviate these concerns, a number of techniques have been recently proposed in order to perform data mining tasks that are privacy-preserving. Thus the field of privacy has seen rapid advances in recent years and in the data mining environment have led to increased concerns about privacy. In this thesis, we develop efficient, effective and realistic methods in the privacy-preserving data mining field focusing on three core techniques, namely access control, data anonymization and statistical disclosure control.

In Part I, this thesis presents a model for privacy preserving access control which is based on a variety of purposes. Conditional purpose is applied along with allowed purpose and prohibited purpose in the model. It allows users to use some data for certain purposes with conditions. The structure of the conditional purpose-based access control model (CPBAC) is defined and investigated through a practical paradigm with access purpose and intended purpose. An algorithm is developed to achieve the compliance computation between access purposes and intended purposes. According to this model, more information from data providers can be extracted while at the same time assuring privacy that maximizes the usability of consumers' data. This model extends traditional access control models to a further coverage of privacy preservation in the data

mining environment. Its interior is a new structure for managing collected data in an effective and trustworthy way. This structure helps enterprises to circulate clear privacy promises and to collect and manage user preferences and consent. Finally, we inject this model with the conventional well known role-based access control (RBAC) model as RBAC is still the most popular approach towards access control to achieve database security and is available in many DBMS. The notion of applying these mechanisms to allow web sites to publish a privacy policy, and implement more nuanced management of usage information and other personal information, ultimately allows (legitimate) use of information.

In Part II, this thesis presents a systematic clustering based k-anonymization technique to minimize the information loss and at the same time assure data quality. The proposed technique adopts a system to group similar data together and then anonymize each group individually. The structure of systematic clustering problem is defined and investigated through paradigm and properties. An algorithm of the proposed problem is developed and it is shown that the time complexity is in $O(\frac{n^2}{k})$, where n is the total number of records containing individuals and their private information. Experimental results show that the proposed method attains a reasonable dominance with respect to both information loss and execution time. A way out is also shown to illustrate the usability of the algorithm for incremental datasets. Finally we extend the systematic-clustering approach to the l-diversity model that assumes that every group of indistinguishable records contains at least l distinct sensitive attribute values. The whole procedure consists of the two steps, namely a clustering step for k-anonymization and an l-diverse step.

In Part III, this thesis presents two heuristic algorithms for microdata protection in Statistical Disclosure Control (SDC). The first heuristic microaggregation algorithm works by partitioning the microdata into clusters of at least k records in a systematic way and then replacing the records in each cluster with the cen-

troid of the cluster which we refer to systematic microaggregation for SDC. The structure of the systematic microaggregation problem is defined and investigated and an algorithm of the proposed problem is developed. Experimental results show that the systematic microaggregation attains a reasonable dominance with respect to both information loss and execution time than the most popular heuristic algorithm called Maximum Distance to Average Vector (MDAV). Finally it has shown that the systematic microaggregation is highly scalable.

The second heuristic algorithm, called pairwise-systematic (P-S) microaggregation easily captures extreme values in the dataset and works by adopting simultaneously two distant groups at a time with the corresponding similar records together in a systematic way. Extensive experimental studies are conducted to show the efficiency and the effectiveness of the algorithm. The performance of the P-S algorithm is compared against the most recent microaggregation methods. Experimental results show that the P-S algorithm incurs significantly less information loss compared to the latest microaggregation methods for all of the test situations. Finally we propose a new microaggregation method where centroid is considered as median. The new method guarantees that the microaggregated data and the original data are similar by using a statistical test.

Publications Based on this Thesis

Accepted/Published Manuscripts

1. Kabir, M.E., Wang, H., and Bertino, E. (2010), "A Role-involved Purpose-based Access Control Model", Information Systems Frontiers (Revisions).

2. Kabir, M.E., Wang, H., and Bertino, E. (2010), "A Conditional Role-involved Purpose-based Access Control Model", Journal of Organizational Computing and Electronic Commerce (Revisions).

3. Kabir, M.E., Wang, H., and Bertino, E. (2010), "A Conditional Purpose-based Access Control Model with Dynamic Roles", Expert Systems with Applications (in Press).

4. Kabir, M.E., and Wang, H. (2010), "Microdata protection method through microaggragation: A median based approach", Information Security Journal: A Global perspective (in Press).

5. Kabir, M.E., Wang, H., and Yanchun, Z. (2010), "A Pairwise-Systematic Microaggregation for Statistical Disclosure Control", Proceedings of 10th International Conference on Data Mining (ICDM 2010), Sydney, Australia.

6. Kabir, M.E., Wang, H., and Bertino, E. (2010), "A Role-involved Conditional Purpose-based Access Control Model", Proceedings of IFIP EGES conference on E-Government and E-Services (EGES 2010) at the IFIP WCC world conference, Brisbane, Australia.

7. Kabir, M.E., and Wang, H. (2010), "Systematic Clustering-based Microaggregation for Statistical Disclosure Control ", Proceedings of International Conference on Data and Knowledge Engineering (ICDKE 2010), Melbourne, Australia.

8. Kabir, M.E., Wang, H., and Bertino, E. (2010), "Systematic Clustering Method for l-diversity Model", Proceedings of 21st Australasian Database Conference (ADC 2010), Brisbane, Australia.

9. Kabir, M.E., and Wang, H. (2009), "Conditional Purpose Based Access Control Model for Privacy Protection", Proceedings of 20th Australasian Database Conference (ADC 2009), Wellington, New Zealand.

Submitted Manuscripts

10. Kabir, M.E., Wang, H., and Bertino, E. (2009), "Efficient Systematic Clustering Method for k-anonymization", Acta Informatica. Submitted.

11. Kabir, M.E., and Wang, H. (2010), "Microaggregation for Statistical Disclosure Control: A Systematic Clustering-based approach", Applied Soft Computing. submitted.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The phenomenal technological developments in information technology have literally transformed our lives in recent years. The explosive growth of the Internet and e-commerce have enabled people to carry out daily activities online, for example, on line shopping, e-banking, and even consulting a doctor over the Internet. Such ubiquitous online activities imply that a vast amount of personal data is electronically produced and collected continuously. Over the last few decades, there has been a tremendous growth in the amount of private data collected about individuals. With the rapid growth in databases, networking, and computing technologies, such data can be integrated and analyzed digitally. On the one hand, this has led to the development of data mining tools that aim to infer useful trends from this data. This collected data can also be used for various purposes, ranging from scientific research to demographic trend analysis or marketing purposes. For instance, medical researchers may find out the factors associated with autism from a collection of autistic babies records, or government agencies can know the socioeconomic and demographic characteristics of its people and may make critical decisions based on various data collected by them. On the other hand, easy access to personal data poses a threat to individual privacy. Having observed many privacy related incidents [106, 107, 108, 109], individuals are afraid that their personal information might fall into the wrong hands and be abused against their will. As individuals are more concerned about their privacy,

they are becoming more reluctant to carry out their businesses and transactions online, and many organizations are losing a considerable amount of potential profits [66]. Research shows that on-line commerce was reduced by US$15 billion in 2001 due to individual privacy concerns. Thus the field of privacy[1] has seen rapid advances in recent years and in the data mining environment this has lead to increased concerns about privacy.

Data Privacy problems exist wherever uniquely identifiable data relating to a person or persons are collected and stored, in digital form or otherwise. The challenge in data privacy is to share data while protecting the personally identifiable information. Thus, personal data should be protected in such a way that only authorized users can access the data. The way of protecting personal data in such a way is called access control. Personal data can also be protected by anonymizing identifiable information before disclosing. This procedure of data protection is called anonymization. On the other hand, data can be modified in such a way that statistical results from the original data and the modified data are the same or at least similar. The process of producing modified data that produce similar statistical results of the original data is called Statistical disclosure Control (SDC).

In this thesis, we provide models and algorithms for protecting the privacy of individuals in data sets while still allowing users to mine useful trends and statistics. The thesis addresses problems from three areas, namely access control, data anonymization and statistical disclosure control (SDC). Specifically, access control enables DBMS to tightly control data access with respect to privacy requirements and preferences, data anonymization provides a way to guarantee privacy protection in data itself even if the control of access is not feasible, and the SDC is to control the risk that information about specific individuals can be extracted from amongst statistical summary results. We present formal models

[1]Privacy is defined as the right of an individual to decide when, how, and to what extent he/she would like to share his/her information

and develop mechanisms for realizing such models.

1.1 Overview and Motivation

1.1.1 Access Control

Access control is one of the fundamental security mechanisms for information systems. It determines the availability of resources to principles, operations that can be performed, and under what circumstances [110]. Figure 1.1 shows the major functional components of a typical access control system. The Policy Enforcement Point (PEP) interacts directly with users. When a user tries to access a resource, the PEP forms an appropriate access request that includes the attributes of the requester, the requested action, and the requested resource, and passes the request to the Policy Decision Points (PDP). The PDP looks up the access control policy that applies to the request, and returns a response to the PEP. The PEP then correspondingly permits or denies the user's action.

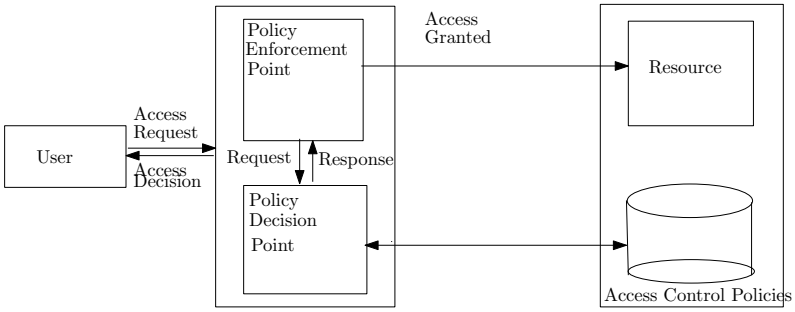


Figure 1.1: Major components of an access control system

As Privacy is one of the major issues to be handled in many environments, many privacy protecting access control models have been proposed [48, 67, 73, 104]. Privacy protection cannot be easily achieved by traditional access control models because of two reasons. The first reason is that while traditional access control models focus on which user is performing which action on which data object, privacy policies are concerned with which data object is used for which purpose(s). Fox example, a typical privacy policy such as "we will collect and use

customer identifiable information for billing purposes to enable us to anticipate and resolve problems with your service" does not specify who can access the customer information, but only states that the information can be accessed for the purpose of billing, customer service, and possibly some analysis. The second reason is that the comfort level of data usage varies from individual to individual. For example, some online consumers may feel that it is acceptable to disclose their purchase history or browsing habits in return for better service, such as site personalization [103] but other customers may believe that such techniques breach their privacy. Some organizations may have published privacy policies, which promise privacy protection practices on data collection, use and disclosure, but these practices may not be implemented. To maintain consistency between privacy policy and practices, privacy protection requirements in privacy policy should be formally specified.

The notion of purpose might play a major role in access control models in order to protect privacy. Thus an appropriate metadata model must be developed to support such privacy centric access control models. Byun et al. [54] proposed a purpose-based access control (PBAC), where purpose is used as the basis of access control. The PBAC model is based on intended purposes, which specify the intended usage of data and access purposes which specify the purposes for which a given data element is accessed. Extending this idea, in this thesis we propose an enhanced PBAC, called conditional purpose-based access control (CPBAC). The CPBAC model is based on a variety of purposes and conditional purpose is applied along with allowed purpose and prohibited purpose in the model. This allows users to use some data for certain purpose with conditions.

Role-based access control (RBAC) [95] has a significant impact on many access control systems. In recent years, RBAC has been widely used in databases system management and operating system products. RBAC is described in terms of individual users being associated with roles as well as roles being associated

with permissions (each permission is a pair of objects and operations). As such, a role is associated with users and permissions. For the current extensive use of the RBAC model in database systems, it is highly possible to analyze the access control model for privacy protection which supports purposes, conditions and obligations on the basis of the RBAC model. Based on RBAC, in this thesis, we have developed a Role-involved Conditional Purpose-based Access Control (RPAC) model, where access permission is a 3-tuple $\langle Object, Operation, AccessPurpose \rangle$ instead of 2-tuple $\langle Object, Operation \rangle$ and access purpose permission is assigned to roles. In the RPAC model, users are required to explicitly state their access purposes when they try to access data. Although this method is simple and easy to implement it requires complete trust in terms of the identity of users and thus, the overall privacy that the system is able to provide entirely relies on the users' trust worthiness. To overcome this problem, we have also developed a Conditional Role-involved Purpose-based Access Control (CPAC) model, where access purpose permission is assigned to Conditional Roles (CR). Users dynamically activate conditional roles in the CPAC model in accordance with the context attributes during the access purpose.

1.1.2 Data Anonymization

Publishing health, financial and personal information requires the data to be anonymized that the privacy of individuals in the database is protected. Anonymity is an important concept for privacy, and data anonymity is particularly crucial in public databases such as census data or health records collected by government agencies. Data anonymity can also be useful in the private sector, for example when an organization wishes to allow third parties to access its customer data. In such a case it cannot be guaranteed that the privacy policy of the data will always be respected by the third parties. Thus, organizations must assure customers'Privacy by removing all information that can link data items with individuals.

The traditional approach of de-identifying records is to remove identifying fields such as social security number or name. However, recent research has shown that a large fraction of the US population can be identified using non-key attributes (called quasi-identifiers) such as date of birth, gender and zip code [18]. A recent approach addressing this difficulty relies on the notion of the k-anonymity model [17, 18]. In this approach, the data privacy is guaranteed by ensuring that non-key attributes that leak information are suppressed or generalized so that, for every record in the modified table, there are at least $(k-1)$ other records that have exactly the same values for quasi-identifiers. The k-anonymity problem has recently drawn considerable interest from the research community and a number of algorithms have been proposed [27, 28, 29, 30, 31, 105]. In this thesis, we present a systematic clustering method for k-anonymization, where clusters form in a systematic way. This method has a time complexity of $\frac{O(n^2)}{k}$ in the clustering stage, where n is the total number of records that contain individuals' privacy elements. The proposed systematic clustering method outperforms the recent clustering based k-anonymization techniques. However the k-anonymity model may reveal sensitive information under two types of attacks, namely the homogeneity attack and the background knowledge attack [40]. To overcome this problem, we propose an enhanced systematic clustering method for the l-diversity model that assumes that every group of indistinguishable records contains at least l distinct sensitive attributes values.

1.1.3 Statistical Disclosure Control

Statistical Disclosure Control (SDC) in databases, also known as Inference control, is about protecting data so they can be published without revealing confidential information that can be linked to specific individuals to whom the data correspond [3]. This is an important application in several areas, such as official statistics, health statistics and e-commerce (sharing of consumer data). Since data protection ultimately means data modification, the challenge for SDC is to

modify data in such a way that sufficient protection is provided while keeping at a minimum information loss, i.e., the loss of accuracy sought by database users. Given an original microdata set $\boldsymbol{V}$, the purpose of microdata (individual data) SDC is to release a protected microdata set $\boldsymbol{V}'$ in such a way that:

- Disclosure risk (i.e., the risk that a user or an intruder can use $\boldsymbol{V}'$ to determine confidential attributes on a specific individual among those in $\boldsymbol{V}$) is low.
- User analyses (regressions, means, etc.) on $\boldsymbol{V}'$ and on $\boldsymbol{V}$ yield the same or at least similar results.

Microaggregation is a family of SDC techniques for continuous microdata. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if records correspond to groups of k or more individuals, where no individual dominates the group and k is a threshold value [8]. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. To obtain microaggregates in a microdata set with n records, these are combined to form g groups of size at least k. For each attribute, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting records can be published. The optimal k-partition is defined to be the one that maximizes within-group homogeneity; the higher the within group homogeneity, the lower the information loss. Since microaggregation replaces values in a group by the group centriod, the sum of squares criterion is common to measure homogeneity in clustering. The within-groups sum of squares SSE is defined as

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})' (x_{ij} - \bar{x_i}) \tag{1.1}$$

The lower the SSE, the higher the within-group homogeneity. Thus, in terms of sums of squares, the optimal k-partition is the one that minimizes SSE. For a microdata set consisting of p attributes, these can be microaggregated together or partitioned into several groups of attributes. The challenge in microaggregation is to form the groups such that the within-group homogeneity is at a maximum. Several taxonomies are possible to classify the microaggregation algorithms in the literature: i) fixed group size [6, 112, 113] versus variable group size [3, 8, 11, 112, 116, 117]; ii) exact optimal (only for the univariate case, [10, 13]) versus heuristic microaggregation; iii) continuous versus categorical microaggregation [19].

In this thesis, at first we propose a systematic clustering-based microaggregation method for SDC. The algorithm of systematic clustering is sometimes affected by extreme values. To overcome this problem, we propose another heuristic approach, called the pairwise systematic (P-S) microaggregation method to minimize the information loss. Both the algorithms are applicable for a multivariate fixed group size microaggregation on unprojected continuous data. Finally we propose a median based microaggregation method, where centroid is considered as median. The new method guarantees that the microaggregated data and the original data have the same distribution. The similarity of the data is conducted by using a statistical non-parametric test.

1.2 Objectives of the Thesis

Although the recent privacy-related regulations have made many organizations aware of the importance of privacy protection, such regulations are not the only incentive for organizations to protect individuals' privacy. For many businesses, especially e-commerce, consumers' concern for privacy is directly translated to a huge financial loss. A survey report from Forrester Research [66] states that individual privacy concerns reduced 15 billion dollars in e-commerce in 2001 alone. This thesis aims to help develop a comprehensive privacy-preserving DBMS. Some

efforts have already been reported that deal with a DBMS specifically tailored to support privacy policies. Although some follow-up effort has been made, the development of a privacy-preserving DBMS is still at a very preliminary stage. It is important to notice that a privacy-preserving DBMS may have to be combined with collected tools, such as a data anonymizer and metadata manager, in order to provide comprehensive platforms to support flexible and articulated privacy-preserving information management. The main objective of this thesis is to develop models and techniques for building a privacy-preserving DBMS in this regard. To investigate this, I will focus on three major tasks in a privacy preserving DBMS in my PhD research:

- Purpose-based access control;
- Data anonymisation;
- Microaggregation in SDC.

1.3 Organization of the Thesis

The thesis consists of eight chapters with three parts. Their precedence order is outlined and illustrated in Figure 1.2.

In the first part of this thesis, we address the issue of data privacy by developing access control techniques. In Chapter 2, we introduce the Conditional Purpose-based Access Control (CPBAC) model, which directly addresses the issue of individuals' control over their personal data. The CPBAC model can extract more information from data providers while at the same time assuring privacy. The key characteristic of the CPBAC model is that it allows users to use some data with certain conditions, and that multiple purposes can be associated with each data element. It exploits query modification techniques to support data access control based on the conditional purpose information. In Chapter 3, we inject the CPBAC with the conventional well known RBAC, as RBAC has

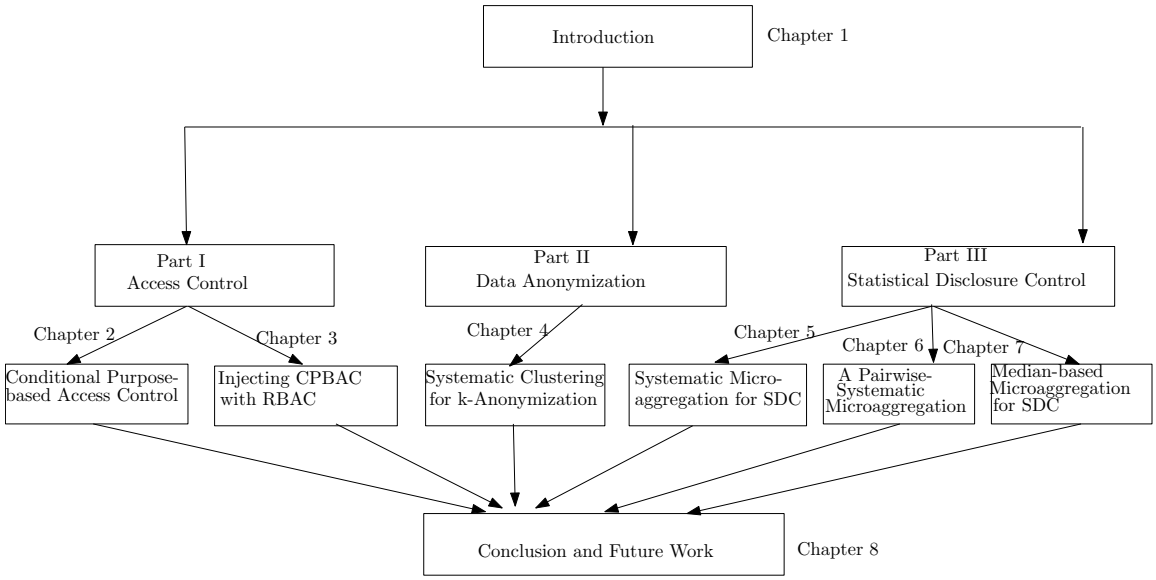


Figure 1.2: The structure of the thesis

been widely used in database system management and operating system products. Chapter 3 consists of two parts. In Section 3.2, we present a role-involved purpose-based access control (RPAC) model, where users are required to explicitly state their access purposes when they try to access data. In Section 3.3, we present a conditional role-involved purpose-based access control (CPAC) model, where users dynamically activate conditional roles in accordance with the context attributes. Based on the conditional role, access permissions are assigned that represent what can be accessed, and for what purpose, to roles under certain conditions.

In Part II, we address the issue of data privacy by developing efficient data anonymization techniques. We propose an efficient systematic clustering method for the k-anonymization in Chapter 4. The proposed technique adopts group similar data together and then anonymizes each group individually. We also extend this approach to the l-diversity model in this Chapter that assumes that every group of indistinguishable records contains at least l distinct sensitive attribute values.

In the last part of this thesis (Part III), we address the issue of data privacy

by developing microaggregation methods in SDC. In Chapter 5, we develop the systematic microaggregation method for SDC. In order to capture outliers in a dataset, we present a pairwise systematic (P-S) microaggregation method to minimize the information loss in Chapter 6. The proposed technique in this chapter adopts simultaneously two distant groups at a time with the corresponding similar records together in a systematic way and then anonymizes with the centroid of each group individually. Finally we present a median-based microaggregation method in Chapter 7, where the centroid is considered as the median. The new method guarantees the microaggregated data, and the original data are similar by using a statistical test. Another contribution of this chapter is that we propose a distance metric, called absolute deviation from median (ADM) to evaluate the amount of mutual information among records in microdata. Finally, conclusions and future work are indicated in Chapter 8.

Part I
Access Control

Chapter 2

Conditional Purpose-based Access Control

Data is collected for certain purposes. In order to protect information privacy, the notion of purpose must play a major role in an access control model. This chapter presents a model for privacy preserving access control which is based on a variety of purposes. Conditional purpose is applied along with allowed purpose and prohibited purpose in the model.

2.1 Introduction

With the increasingly extensive application of information technologies in people's daily life, Privacy preservation has become a challenging problem in the field of information security. Enterprises regularly collect customers' personal identification information along with other attributes during any kind of marketing activities. It is a natural expectation that the enterprise will use this information for various purposes, which leads to concerns that the personal data may be misused. Many enterprises collect, store and use huge amount of personal information. A study conducted by the Federal Trade Commission has shown that 97 percent of websites were collecting at least one type of identifying information such as name, e-mail address, or postal address of customers [56]. Privacy preservation in a data-mining environment has become a great concern both for enterprises and individuals. As individuals are more concerned about

their privacy, they are becoming more reluctant to carry out their businesses and transactions online, and many organizations are losing a considerable amount of potential profits [66]. Research has shown that on-line commerce was reduced by US$15 billion in 2001 due to individual privacy concerns. These reactions from individuals imitate an altering awareness about how data is managed. Therefore without a clear compromise between individuals and enterprises, data quality and data privacy cannot be achieved and many organizations are seriously thinking about privacy issues of consumers. By demonstrating good privacy practices, many businesses are now trying to build up solid trust with customers, thereby attracting more customers [51]. Considering the privacy of customers, enterprise has to develop a secure privacy policy to remove the fear of customers. Thus in an internal management system, a reliable, efficient, effective and secure privacy policy should be established depending on the customer's requirements.

A lot of work has been done in order to protect the privacy of individuals and this has shown that the notion of purpose should be used as the basis for access control for specifying a privacy policy [48, 49, 53, 54, 73, 79]. According to Yang et al. [99], a privacy policy ensures that data can only be used for its intended purpose (intended usage of data), and that an access purpose (intention for accessing data objects) is compliant with the data's intended purpose. During the last few years, rapid technological developments especially in the field of information technology, have directed most attention and energy to the privacy protection of Internet users. Unless customers' data are suitably protected, individuals' privacy can be breached revealing their personal information. On the other hand, these collected data sets are the most important tools for a wide range of studies. Again the data that is more protected usually loses data quality. Therefore, it is necessary to come with to a point where both data quality and data privacy are achieved. Although a significant number of works has been developed in this area [48, 49, 53, 54, 73, 76, 79], research has yet to be done to

Table 2.1: Hypothetical data base illustrating AIP and PIP

name	age	address	income	name$_{ip}$	age$_{ip}$	address$_{ip}$	income$_{ip}$
Alice	35	21, West St. TBA, QLD 4350	35000	⟨{G}, {Φ}⟩	⟨{Φ}, {G}⟩	⟨{G}, {A, S}⟩	⟨{G}, {M}⟩
Bob	29	45, Fay CT. TBA, QLD 4350	23000	⟨{G}, {Φ}⟩	⟨{G}, {M}⟩	⟨{G}, {A, S}⟩	⟨{G}, {A, M}⟩
Ron	56	20, Anita Dr. TBA, QLD 4350	56000	⟨{G}, {Φ}⟩	⟨{G}, {M}⟩	⟨{G} {A, S}⟩	⟨{G}, {A}⟩
Jak	48	25, Wuth St. 25, Wuth St.	48000 48000	⟨{G}, {Φ}⟩	⟨{G}, {M}⟩	⟨{G}, {A}⟩	⟨{G}, {A,M}⟩

G={General purpose}, A={Admin purpose}, S={Shipping purpose}, P={Purchase purpose}, M={Marketing purpose}, ip={Intended purpose}=⟨AIP, PIP⟩

remove the dilemma between data quality and data privacy.

One of the most popular approaches for protecting private information is the access control model. Access control is the process of limiting access to the resources of a system only to authorized users, programs, processes, or other systems [101]. Many privacy policy access control models have been proposed in order to protect the privacy of consumers. Byun et al. [53, 54] pointed out that privacy protection cannot be easily achieved by traditional access control models as it focuses on which user is performing which action on which data object. But a reliable privacy policy is concerned with which data object is used for which purpose. For example, a typical privacy policy such as "we will collect and use customer identifiable information for billing purposes and to enable us to anticipate and resolve problems with your service" does not specify who can access the customer information, but only states that the information can be accessed for the purpose of billing, customer service and possibly some analysis. Another complexity of privacy protection is that the comfort level of data usage varies from individual to individual. For example, some online consumers may feel that it is acceptable to disclose their purchase history or browsing habits in return for better service, such as site personalization [103]. Other customers, however, may believe that such techniques violate their privacy. Thus the notion of purpose must play a major role in access control models and an appropriate metadata model must be developed to support such privacy centric access control

models in order to protect data privacy. Byun et al. [54] developed an approach that is based on intended purposes, which specify the intended usage of data, and access purpose, and which in turn specify the purposes for which a given data element is accessed. Usually, during the data collection procedure customers are informed about the purposes of enterprises. Customers then decide whether their information could be used or not for a certain purpose. That means data providers are given an option for their data and for what purposes it may be used. If an individual mentions that his/her data could not be used for a certain purpose, then his/her information is not accessible for that purpose. Usually data providers are reluctant to use any part of their information for any purposes and so there is a possibility of losing information. However more information can be extracted from data providers by providing more options of using their information. It is possible to protect the privacy of individuals in this model, but there is a shortcoming of information loss. An intended purpose is divided (IP) into two parts: Allowed Intended Purposes (AIP) (explicitly allows to access the data for a particular purpose) and Prohibited Intended Purpose (PIP) (data access for particular purposes are never allowed). In order to recognize the model clearly, suppose that a company uses consumers' data for the purpose of General, Admin, Marketing and Shipping and consider the hypothetical database in Table 2.1.

In Table 2.1, the value of Alice's attribute income_{ip} is $\langle\{\text{G}\},\{\text{M}\}\rangle$, which means that Alice's income could be used for the General purpose but would be strictly prohibited to use for the Marketing purpose. If we take a query

SELECT name

FROM Table 2.1

FOR Marketing Purpose

it gives the name of Alice, Bob, Ron, Jak and if we have a query

SELECT name, age

FROM Table 2.1

FOR Marketing Purpose

it returns nothing because prohibited intended purposes override the allowed intended purposes. This model protects the privacy of consumers as it considers customers' requirements but it incurs more information loss. So a natural question arise

" Is it possible to extract information from PIP at least conditionally?"

The answer to this question is achieved in this chapter by adding a new term, conditional purpose, to the intended purpose. In order to extract more data and protect data privacy, conditional purpose plays a role in access control models. In this chapter, we address this goal by presenting a model of purpose management, which is a fundamental building block on which conditional purpose based access control can be developed. Our proposed model is based on access purpose and intended purpose. Both access purposes and intended purposes are specified with respect to a hierarchical structure that organizes a set of purposes for a given enterprise. A key feature of our proposed model is that it supports conditional purpose and prohibited purpose, thus allowing users to specify that data should be used conditionally or should not be used for a set of purposes.

Observing the challenges of privacy protection and the satisfaction of both enterprises and customers, we need a better model to extract more information from customers with privacy guarantees. To overcome this challenge, we propose a new access control model called conditional purpose-based access control (CPBAC) model. The access control model enables extracting information from PIP by giving conditions, which is called Conditional Intended Purpose (CIP). Our proposed model is helpful for enterprises to establish an ideal privacy policy and to manage data in a sensitive, effective and trustworthy way. It also helps policy makers and experts in the data-mining environment.

2.2 Related Work

This work is related to several topics in the area of privacy and security for data management, namely privacy policy specification, privacy-preserving data management systems and multilevel secure database systems. We now briefly discuss the most relevant approaches in these areas.

The most notable technique to protect privacy is the W3C's Platform for Privacy Preferences (P3P) that formally specifies privacy policy by service providers [74]. P3P provides a way for a web site to encode its data collection in a machine-readable format known as a P3P policy, which can be compared against a user's privacy preferences [99]. Byun et al. [54] pointed out that P3P does not provide any functionality to keep promises in the internal privacy practice of an enterprise. Thus it can be said that a striking privacy policy with an inadequate enforcement mechanism may place organizations at risk of reputation damage. The concept of a Hippocratic database introduced by Agrawal et al. [48] amalgamates privacy protection in a relational database system. A Hippocratic database includes privacy policies and authorizations that are associated with each attribute and each user's the usage purpose(s) [50]. Agrawal et al. [48] presented a privacy preserving database architecture called Strawman which based the access control on the notion of purposes, and opened up database-level research about privacy protection technologies. After that, purpose-based access control introduced by Byun et al. [53, 54] and Yang et al. [99], fine grained access control introduced by Agrawal et al. [49] and Rizvi et al. [78] are now widely used access control models for privacy protection. In IT systems the proposed Enterprise Privacy Authorization Language (EPAL) of IBM [67] is a language for writing enterprise privacy policies to run data handling practices. An EPAL policy defines hierarchies of data-categories, user-categories, and purpose [54]. A set of actions, obligations, and conditions are also defined by an EPAL policy.

A lot of works on multilevel secure relational databases [52, 55, 94, 96] pro-

vide many valuable insights for designing a fine-grained secure data model. In a multilevel relational database system, every piece of information is classified into a security level, and every user is assigned a security clearance [54]. LeFevre et al. [73] proposed an approach to enforce privacy policy in a database setting. This work focus on ensuring limited data disclosure, based on the premise that data providers have control over who is allowed to see their personal data and for what purpose. They introduced two models of cell-level limited disclosure enforcement and suggested an implementation based on query modification techniques. Byun et al. [54] present a comprehensive approach for a privacy preserving access control model. In their access control model, multiple purposes are to be associated with each data element and also support explicit prohibitions. This model is based on the notion of purpose as it plays a central role and is the basic concept on which access decisions are made. Massacci et al. [75] pointed out that most privacy-aware technologies use purpose as a central concept around which privacy protection is built.

All of these works proposed different approaches to protect the privacy of individuals through different models without considering to extract more information. Our aim is to preserve the privacy of individuals as well as extracting more information. With this aim, in this chapter we propose a model that has significantly improved the work of Byun et al. [54]. It has improved in three different ways. First, we introduce a conditional purpose in addition to explicit prohibitions that make data providers more flexible in giving information. Second, the enterprise can publish an ideal privacy policy to manage data in a sensitive, effective and trustworthy way, and third it reduces the information loss as it shows that we can extract more information from data providers.

2.3 Purpose, Access Purpose and Intended Purpose

Data is collected for a certain purpose. For instance, for a nation wide demographic survey in Australia, data may be collected to know the socioeconomic and demographic characteristics of all Australians. Each data access also serves a certain purpose. So it is a natural expectation that a privacy policy should be concerned about which data object is used for which purpose. Many authors have indicated that purpose is a central part in many privacy preserving access control model [48, 67, 73, 104].

2.3.1 Definition of Purpose

For preserving the privacy of customers, each and every data access must obey the privacy policies on which customers have conditionally or unconditionally agreed. A representative privacy policy for a data element includes purpose, retention, condition and obligation. This means that the particular data element can be conditionally or unconditionally accessed only for specific purposes with certain conditions. The retention indicates how long the data element can be reserved, and the obligation designates the actions that must be followed after an access to the data element is approved. So purpose is the most interesting thing to researchers as it directly shows how access to data elements has to be controlled. P3P defines purpose as "the reason(s) for data collection and use" and specifies a set of purposes [98]. In commercial surroundings purposes normally have hierarchical associations among them; i.e., generalization and specialization relationships. For instance, a group of purposes such as direct-marketing and third party marketing can be represented by a more general purpose, marketing. We borrow the purpose definition from [54].

Definition 2.3.1 (Purpose and Purpose Tree): A purpose describes the intentions for data collection and data access. A set of purposes, denoted as ω, is

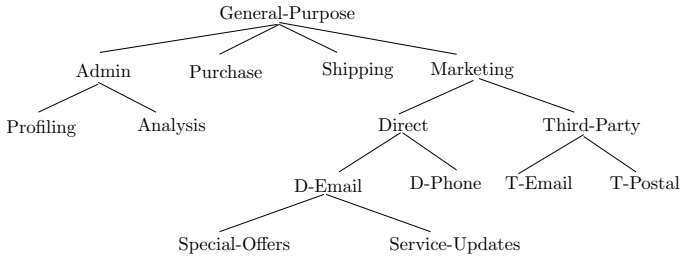


Figure 2.1: Purpose Tree

Table 2.2: Hypothetical data base illustrating AIP, CIP and PIP

name	age	address	income	name$_{ip}$	age$_{ip}$	address$_{ip}$	income$_{ip}$
Alice	35	21, West St., TBA, QLD 4350	35000	⟨{G}, {Φ}, {Φ}⟩	⟨{Φ}, {M}, {A}⟩	⟨{G}, {Φ}, {A, S}⟩	⟨{G} {A}, {M}⟩
Bob	29	45, Fay CT., TBA, QLD 4350	23000	⟨{G}, {Φ}, {Φ}⟩	⟨{G}, {M}, {Φ}⟩	⟨{G}, {M}, {A, S}⟩	⟨{G}, {M}, {A}⟩
Ron	56	20, Anita Dr., TBA, QLD 4350	56000	⟨{G}, {Φ}, {Φ}⟩	⟨{G}, {M}, {Φ}⟩	⟨{G}, {Φ}, {A, S}⟩	⟨{G}, {S}, {A}⟩
Jak	48	25, Wuth St., TBA, QLD 4350	48000	⟨{G}, {Φ}, {Φ}⟩	⟨{G}, {M}, {Φ}⟩	⟨{G}, {M}, {A}⟩	⟨{G}, {M}, {A}⟩

G={General purpose}, A={Admin purpose}, S={Shipping purpose}, P={Purchase purpose}, M={Marketing purpose}, ip={Intended purpose}=⟨AIP, CIP, PIP⟩

organized in a tree structure, referred to as Purpose Tree and denoted as Ω, where each node represents a purpose in ω and each edge represents a hierarchical relation between two purposes. Let r_i, r_j, be two purposes in Ω. We say that r_i is an ancestor of r_j (or r_j is a descendent of r_i) if there exists a downward path from r_i to r_j in Ω. Figure 2.1 is an example of a purpose tree.

Purposes, depending on their association with objects and subjects, may be called intended purposes or access purposes respectively.

Definition 2.3.2 (Access Purpose): An access purpose is an intension for accessing data objects, and it must be determined by the system when data access is requested. So access purpose specifies the purpose for which a given data element is accessed.

Definition 2.3.3 (Intended Purpose): An intended purpose is the specified usages for which data objects are collected. That is, purpose is associated with data and thus regulates data accesses as intended purpose.

According to our approach an intended purpose consists of the following three components.

Allowable Intended Purpose (AIP): This means that data providers explicitly allow accessing the data for a particular purpose. For example data providers may consider that their information can be used for marketing purposes without any further restrictions.

Conditional Intended Purpose (CIP): This means that data providers allow accessing the data for a particular purpose with some conditions. For example data providers may consider that their income information can be used for marketing purposes by hiding their personal identification information (e.g. id or name) or their income data can be revealed through generalization. or only the first letter of a name can be used for marketing purposes.

Prohibited Intended Purpose (PIP): This means that data providers strictly disallow accessing the data for a particular purpose. For example data providers may consider that their income information cannot be used for marketing purposes. In that case data a provider's income attribute is strictly prohibited to use for marketing purposes. An example of how AIP, CIP and PIP work is illustrated through a hypothetical database in Table 2.2.

So an intended purpose IP is a tuple $\langle AIP, CIP, PIP \rangle$, where AIP $\subseteq \omega$, CIP$\subseteq \omega$ and PIP$\subseteq \omega$ are three sets of purposes. The set of purposes implied by IP, denoted by IP* and the set of conditional purposes, denoted by IP$^\star_c$ are defined to be AIP$^\downarrow$-CIP$^\updownarrow$-PIP$^\updownarrow$ and CIP$^\downarrow$ -PIP$^\updownarrow$ respectively, where

R$^\downarrow$, is the set of all nodes that are descendants of nodes in R, including nodes in R themselves,

R$^\uparrow$, is the set of all nodes that are ancestors of nodes in R, including nodes in R themselves, and

R$^\updownarrow$, is the set of all nodes that are either ancestors or descendants of nodes in R, that is, R$^\updownarrow$=R$^\uparrow \cup$ R$^\downarrow$.

Definition 2.3.4 (Full Access Purpose Compliance): Let Ω be a purpose tree. Let IP= $\langle AIP, CIP, PIP \rangle$ and AP be an intended purpose and an access purpose defined over Ω, respectively. AP is said to be compliant with IP according to Ω, denoted as AP$_{\Leftarrow\Omega}$IP, if and only if AP$\in$ IP*.

Definition 2.3.5 (Conditional Access Purpose Compliance): Let Ω be a purpose tree. Let IP= $\langle AIP, CIP, PIP \rangle$ and AP be an intended purpose and an access purpose defined over Ω, respectively. AP is said to be conditionally compliant with IP according to Ω, denoted as AP$_{c\Leftarrow\Omega}$IP, if and only if AP$\in$ IP$^{\star}_{c}$.

The following example explains the definition of AIP, CIP and PIP.

Example 2.3.1 Suppose IP= $\langle$\{Admin, Direct\}, \{Third-party\}, \{D-mail\}$\rangle$, then IP* = \{Admin, Profiling, Analysis, D-Phone\} and IP$^{\star}_{c}$ = \{Third-party, T-Email, T-Postal\}, where subscript c indicates that customers information can be used for the purpose with some conditions.

2.3.2 Management of Intended Purpose

As discussed before, data providers are given three possible options to make use of their data, namely AIP, CIP and PIP. The CPBAC model builds the purpose hierarchy on both intended purpose and access purpose. When the user passes a request, the access control engine would verify whether the access purpose complies with the intended purposes of user's requested data, and permit or conditionally permit the access if it does, or otherwise deny the request. The key feature of this model is that it supports conditionally allowable and explicit prohibitions and organizes purposes in a hierarchy structure. Suppose that data providers are informed about the company's privacy policy which is compliant with the existing privacy laws and at the time of data collection process data providers already agreed with those policies. On the basis of purpose tree, privacy laws and policies, intended purpose with three levels are specified. Depending on

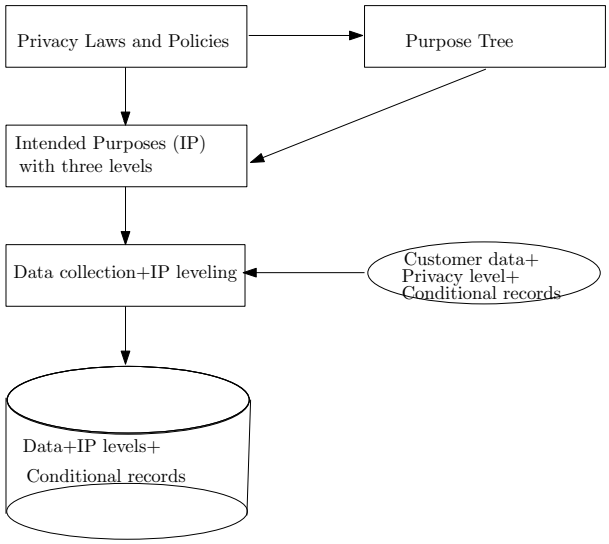


Figure 2.2: Intended Purpose Management

customer privacy level during the data collection process, intended purposes are associated with each data element along with the conditional records as some information will be disclosed conditionally. Finally, in the organization of the intended management process, the system receives data along with IP levels and all the conditional records. Thus, the privacy policy of each data item is predetermined and the intended purposes of data items are also predetermined. The management of intended purpose is shown in Figure 2.2.

Example 2.3.2 Suppose a company has established the following privacy policies:

- We use your information for purchasing purposes. This is just to provide services to you and to inform you of services that may suits you.
- We will disclose, conditionally disclose or will not disclose your information to third parties (e.g. external organizations). However, we will disclose if you allow us to do so.
- It is our policy not to make any use of the information of children under thirteen years old.

Table 2.3: Predetermined Intended Purposes

	Group 1	Group 2	Group 3
Name	$\langle\{G\},\{T\},\{\Phi\}\rangle$	$\langle\{G\},\{\Phi\},\{T\}\rangle$	$\langle\{G\},\{\Phi\},\{\Phi\}\rangle$
Address	$\langle\{G\},\{T\},\{\Phi\}\rangle$	$\langle\{G\},\{\Phi\},\{T\}\rangle$	$\langle\{G\},\{\Phi\},\{\Phi\}\rangle$
Phone	$\langle\{G\},\{T\},\{\Phi\}\rangle$	$\langle\{G\},\{\Phi\},\{T\}\rangle$	$\langle\{G\},\{\Phi\},\{\Phi\}\rangle$
Age	$\langle\{G\},\{T\},\{\Phi\}\rangle$	$\langle\{G\},\{\Phi\},\{T\}\rangle$	$\langle\{G\},\{\Phi\},\{\Phi\}\rangle$
Income	$\langle\{G\},\{T\},\{\Phi\}\rangle$	$\langle\{G\},\{\Phi\},\{T\}\rangle$	$\langle\{G\},\{\Phi\},\{\Phi\}\rangle$

G={General purpose}, T={Third-Party}, Φ={No restriction}.

- Sometimes web server administrators may collect some data (not your private information). We will not make any use of this information but the administrators may use this to provide better service to you.

In Table 2.3, Group 3 represents customers who have given consent for third-party marketing, Group 1 represents customers who have conditionally given consent for third-party marketing and Group 2 represents customers who have not given consent for third-party marketing.

2.4 Conditional Purpose-based Access Control (CPBAC)

In the CPBAC model data providers are asked three possible options for usage of each data item. Permissible usage means data providers allow use of their data, prohibited means data providers do not allow use of their data and conditional permissible usages means data providers conditionally allow use of their data item. Consider Table 2.4 that describes the intended purpose, types of data and possible data usages. For example, a data provider may select that his/her name and address is permissible for **Admin** purposes, address is not permissible for **Marketing** purposes but income information is conditionally permissible for **Marketing** purpose. That is, the data provider does not have any privacy concerns about the name and address when it is used for the purpose of administration, but great concerns about privacy of the address information (and so does not want to disclose the address) when it is used for the purpose of marketing,

Table 2.4: Intended purpose, data type and data usage type

Term	Description	Example
Intended Purpose	Intended usage of data specified by data provider	AIP, CIP, PIP
Data item	Types of data being collected (i.e. attributes)	Name, Age, Income
Data usage Type	Types of potential data usage (i.e. purpose)	Marketing, Admin

Table 2.5: Conditional records and intended purposes

	name	age	address	income
AIP	Alice	35	21, West St., TBA, QLD 4350	35000
CIP	A	30-40	West St., TBA, QLD 4350	30000-40000
PIP	$\star$	$\star$	$\star$	$\star$
AIP	Bob	29	45, Fay CT., TBA, QLD 4350	23000
CIP	B	20-30	Fay CT., TBA, QLD 4350	20000-30000
PIP	$\star$	$\star$	$\star$	$\star$
AIP	Ron	56	20, Anita Dr., TBA, QLD 4350	56000
CIP	R	50-60	Anita Dr., TBA, QLD 4350	50000-60000
PIP	$\star$	$\star$	$\star$	$\star$
AIP	Jak	48	25 Wuth St., TBA, QLD 4350	48000
CIP	A	50-60	Wuth St., TBA, QLD 4350	40000-50000
PIP	$\star$	$\star$	$\star$	$\star$

$\star$ means data providers are reluctant of any usage of their data items

but his/her income information can be used for marketing purpose with some conditions. Here the term "conditions" means that data provider is ready to release his/her certain information for certain purposes by removing his/her name or id or through generalization. This information is then stored in the database along with the collected data, and access to the data is tightly governed according to the data provider's requirements. By using the term condition, data providers feel more comfortable to release their data. Table 2.5 shows conditional records and intended purposes of the data providers in Table 2.2.

The design of intended purposes supports permissive, conditions and prohibitive privacy policies. This construction allows more squash and flexible policies in our model. Moreover, by using CIP and PIP, we can assure that data

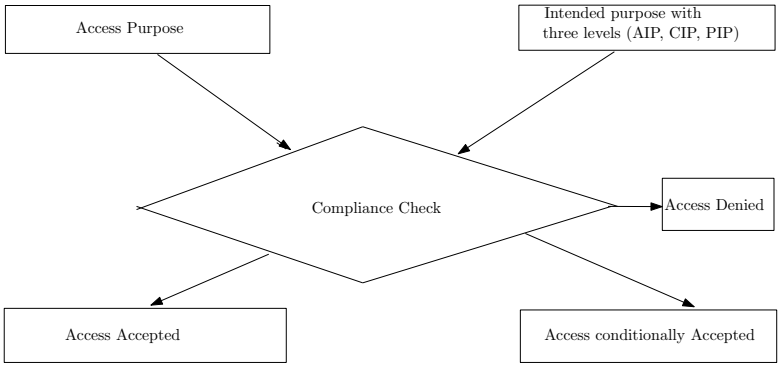


Figure 2.3: CPBAC Model

access for particular purposes is allowed with some conditions or never allowed. Note that an access decision is made based on the relationship between the access purpose and the intended purpose of the data. Access is allowed only if the access purpose is included in the implementation of the intended purpose; in that case the access purpose is compliant with the intended purpose. The access is accepted with conditions if the implementation of intended purpose includes the access purpose with conditions; in this case we say that access purpose is conditionally compliant with intended purpose. Access is denied if the implementation of the intended purpose does not include the access purpose, in this case access purpose is not compliant with the intended purpose. Figure 2.3 shows the structure of the CPBAC model. Suppose in the online marketing system an enterprise collects name, age, address and income of customers along with other information and the enterprise uses the customer's information for the purpose of admin, shipping, purchase and marketing. Consider the hypothetical database in Table 2.2.

In Table 2.2, the value of Alice's attribute income_{ip} is $\langle\{\text{G}\},\{\text{A}\},\{\text{M}\}\rangle$ which means that Alice's income could be used for general purposes but is strictly prohibited to be used for marketing purposes. It also means that Alice's income could be used for admin purposes by hiding her personal identification information or through generalization. Similarly, Bob, Ron and Jak's income information could be used conditionally for marketing purposes but their income information

is strictly prohibited for admin purpose.

2.5 Implementation

In our proposed model, users query the database using standard SQL statements. In this dissertation we assume that each query is connected with a specific purpose. The data accessible to each query varies depending on the data providers agreement and the purpose of the query. For example, any query against Table 2.2 with any purpose returns a result that is equivalent to the result of the query. As our proposed model directly reflect the information that is allowed, conditionally allowed or prohibited by each data provider, querying against these in the model does not violate privacy. This model is quite different from the conventional access control model as different sets of data may be returned for the same query depending on the purpose of the query and the data providers' agreements. Thus from the hypothetical database in Table 2.2, if we take the query

SELECT name, income

FROM Table 2.2

FOR Marketing Purpose,

then by using Table 2.5, we get the information in Table 2.6.

We can see from Table 2.6 that it gives name and income of Ron as he allows to disclose his name and income information for marketing purpose. It also shows two other incomes via generalization as they conditionally allowed to disclose their income. This clearly shows the utility of using our proposed model. It demonstrates that it can extract more information from data providers.

Theorem 2.5.1 Let p, q and r denote the probability that a data provider gives consent for a particular attribute for AIP, PIP and CIP respectively. Assuming that these probabilities remain the same from data provider to data provider,

Table 2.6: Filtering information

Ron	56000
Bob	20000-30000
Jak	40000-50000

then the conditional based access control model extracts more information than the model proposed by Byun et al. [54].

Proof Let n be the total number of data providers. If p and q are the probability that a given data provider gives consent for a particular attribute for AIP and PIP, then the average numbers of data providers who give consent for AIP is np. That means by using the model of [54], the average number of data providers who give consent for AIP of a particular attribute is np. If we use our model then the average number of data providers who give consent to disclose their data for a certain purpose with some conditions is nr. Therefore, by using the conditional based access control model, the total average number of data providers whose information is accessible is $(np+nr)$. Since n and p are both positive, $(np+nr)$ is always greater than np. This means that it is possible to extract more information from customers by using the conditional based access control model.

In our model, the collected data is used for different purposes on the basis of the data providers requirements. By using the CIP, both privacy and usability of data can be achieved as it filters out the values by performing a purpose compliance. Using a hypothetical database and the extracted outcome in Table 2.6, shows clearly that the data utility and data providers information is protected. Theorem 2.5.1 shows that our proposed model extracts more information while assuring privacy.

Table 2.7: Pt-table

p_id	p_name	parent	code	aip_code	cip_code	pip_code
1	A	-	0×200	0×3FF	0_c×3FF	0×3FF
2	B	1	0×100	0×130	0_c×130	0×330
3	C	1	0×080	0×080	0_c×080	0×280
4	D	1	0×040	0×04F	0_c×04F0	0×24F
5	E	2	0×020	0×020	0_c×020	0×320
6	F	2	0×010	0×010	0_c×010	0×310
7	G	4	0×008	0×00B	0_c×00B	0×24B
8	H	4	0×004	0×004	0_c×004	0×244
9	I	7	0×002	0×002	0_c×002	0×24A
10	J	7	0×001	0×001	0_c×001	0×249

subscript c is used to make a difference between aip_code and cip_code.

2.6 Access control

Among the various possible techniques to determine access purpose, in this chapter we utilize the method where the users are required to explicitly state their access purposes when they try to access data. That is, users provide an access purpose for each query they issue.

2.6.1 Compliance Check

Consider the purpose tree in Figure 2.1 and its encoding into a relation pt-table as shown in Table 2.7. The first column *p_id* represents the identification number of each purpose node, the second column *p_name* represents the name of each purpose node, and the third column *parent* is used to capture the hierarchical relationships among purpose nodes. The column code is the binary encoding of each purpose. For example, in Table 2.7 the purpose B is encoded as '0×100' in hexadecimal representation, while the purpose E is encoded as '0×020' in hexadecimal form. The last three columns *aip_code*, *cip_code* and *pip_code* are precalculated encodings of purpose implications. As we know, when a purpose r_i is used as an AIP, it means that every descendant of r_i, including r_i itself is allowed. For example, the purpose D in Figure 2.4 used as an AIP implies that access is allowed for the purpose of D as well as G, H, I and J. Thus, the *aip_code* of D contains the implied set of D, which is the sum of the encodings of D, G, H,

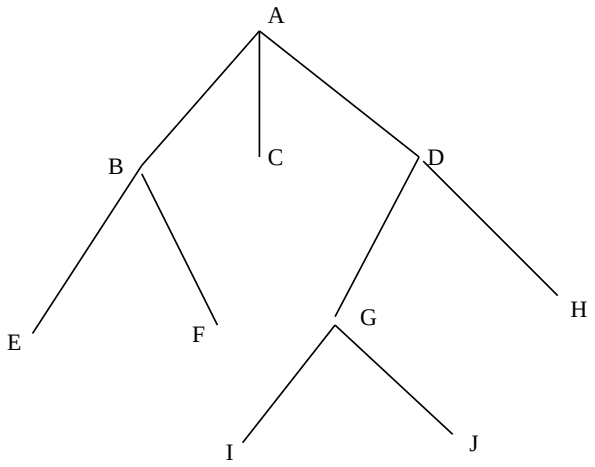


Figure 2.4: Purpose Tree Storage

I and J. Note that *aip_code* and *cip_code* of each purpose is the same, as in the long run both are allowed. The *pip_code* of a particular purpose r_j is computed similarly by summing the encodings of every descendant and ancestor of r_j with the encoding of r_j itself.

An access purpose is compliant with an intended purpose if and only if the access purpose is not prohibited by PIP and it is allowed by both AIP and CIP. Thus, the purpose compliance check can be done with two bitwise AND an operation as follows:

Given the encodings of an access purpose[1], AIP, CIP and PIP, say *ap_code*, *aip_code*, *cip_code* and *pip_code* respectively, the access purpose is fully compliant with the intended purpose if and only if

$$(\text{ap\_code} \ \& \ \text{pip\_code})=0 \wedge (\text{ap\_code} \ \& \ \text{aip\_code}) \neq 0$$

and the access purpose is conditionally compliant with the intended purpose if and only if

$$(\text{ap\_code} \ \& \ \text{pip\_code})=0 \wedge (\text{ap\_code} \ \& \ \text{cip\_code}) \neq 0$$

where, & is bitwise AND operator, $\wedge$ is logical AND operator and $\vee$ is logical OR operator. Conflicts among the AIP, CIP and the PIP for the same data element are resolved by applying the denial-takes-procedure policy where PIP overrides

[1]Access purposes are represented using the values in the *code* of the *Pt-table*

AIP and CIP, and CIP overrides AIP. The computation for a purpose compliance check is illustrated in Table 2.8.

2.6.2 Query modification

It is a natural expectation that privacy-preserving access control techniques ensure a query result containing only the data items that are allowed or conditionally allowed or completely prohibited for the access purpose of the query. This expectation is achieved in this chapter using query modification [97]. It is important to note that query modification provides powerful and flexible controls without requiring any alteration in the underlying mechanisms and that it is supported in a major commercial Data Base Management System [77]. Our query modification algorithm is outlined in Table 2.8.

The complexity of our query modification algorithm is in $O(n)$, where n is the number of attributes accessed by a given query. The method Modifying_Query is invoked only if the access purpose of the query is verified to be acceptable by the validate function. If the access purpose is unacceptable, then the query is rejected without further being processed. The query modification algorithm checks both the attributes referenced in the projection list and the attributes referenced in predicates. As the attributes in the projection list determine what data items will be included in the result relation of a query, it may seem enough to enforce privacy policy based only on the attributes in the projection list. However, the result of a query also depends on the predicates, and not enforcing privacy constraints on the predicates may introduce inference channels. The abounding algorithm filters out a tuple if any of its elements that are accessed are conditionally allowed or prohibited with respect to the given access purpose. For example, consider a query,

SELECT name, income, address

FROM Table 2.2

FOR Marketing Purpose.

Suppose there is a customer record of which name is allowed for marketing, and income is conditionally allowed for Marketing but the address is prohibited for this purpose, then our algorithm only excludes the address of this record from the query result, and income information is visible anonimyzing the customer's name or income information that is revealed via generalisation. Therefore, according to our proposed model, income information of this customer is still usable for marketing purposes instead of excluding other records.

The following example illustrates how our algorithm modifies queries. This example is a revised version of [54] where purpose encoding of marketing is assumed to be '0×200'. For the query

SELECT name, income

FROM table 2.2

FOR Marketing Purpose,

there are two modified queries, one for accessing allowable data items as follows:

SELECT name, income

FROM Table 2.2

WHERE Comp_Check$_1$('0×200', name_aip, name_pip)

AND Comp_Check$_1$('0×200', income_aip, income_pip).

and the other for conditionally allowable data items as follows:

SELECT name, income

FROM Table 2.2

WHERE Comp_Check$_2$('0$_c$×200', name_cip, name_pip)

AND Comp_Check$_2$('0$_c$×200', income_cip, income_pip).

2.7 Comparison

There is some related work on privacy preservation. The closest works related to this article are Hippocratic databases [48] and purpose based access control

Table 2.8: Query Modification Algorithm

```
Comp_Check1 (ap, aip, pip)
/⋆ This function is required for query modification ⋆/
Returns Boolean
1. if (ap & pip)≠= 0 then
2.    return False; 3. else if (ap & aip)0 then
4.    return False;
5. end if;
6. return True


Comp_Check2 (ap, cip, pip)
1. if (ap & pip)≠= 0 then
2.    return False; 3. else if (ap & cip)0 then
4.    return False;
5. end if;
6. return True




Modifying_Query (Query Q)
Returns a modified privacy-preserving query Q
1. Let R1,···, Rn be the relations referenced by Q
2. Let P be the predicates in WHERE clause of Q
3. Let a1,···,am be the attributes referenced in both
the projection list and P
4. Let AP be the access purpose encoding of Q
5.
6. for each Ri where i=1,.,n do
7.    for each aj which belongs to Ri do
8.       if (Comp_Check1 (AP, Ri.aip, Ri.pip)=False) then
9.          return ILLEGAL-QUERY;
10.       end if;
11.    end for;
12.      else if (Comp_Check2 (AP, Ri.cip, Ri.pip)=False) then
13.         return ILLEGAL-QUERY;
14.      end if;
15.   end for;
16. return Q without modified P;
```

model [54]. In this section we will compare our proposed model with these two models.

Agrawal et al. [48] proposed Hippocratic databases that incorporate privacy protection within a relational database system. The proposed technique uses privacy metadata, which consist of privacy policies and privacy authorizations stored in two tables. The authors proposed a strawman design for Hippocratic databases. This design identified the technical challenges and problems in designing such databases. But the authors did not consider the concept of purpose. By contrast, in our proposed model we investigated a more sophisticated concept of purpose. We used conditional purpose and the association of different purposes with a data element which are not considered in their work.

Byun et al. [54] provided a comprehensive framework for purpose and data management. They argued that in order to protect data privacy, the notion of purpose must play a major role in the access control model. The authors proposed approach is based on intended purposes, which specify the intended usage of data, and access purposes, which specify the purposes for which a given data element is accessed. They also argued that traditional access control models focus on which user is performing which action on which data objects but privacy policies are concerned with which data object is used for which purposes. The authors proposed a purpose based access control model (PBAC) that allows multiple purposes to be associated with each data element and also supports explicit prohibitions. Although their proposed model is designed on the basis of customers requirements and so does not violate privacy, the main drawback of this model is the information loss. In that model, customers are given only two options, whether their private data can be used or not for certain purposes, instead of giving more possible options. However, we strongly believe that by giving more options to customers data extractions can be easily achieved. Thus the proposed model in this chapter provides three more options that help enterprises to extract

more information from customers, while still assuring privacy. This criterion is achieved theoretically by Theorem 2.5.1 in Subsection 2.5. This clearly shows the utility and usability of our proposed model in an effective and trustworthy way.

2.8 Conclusion

Although privacy preserving desires a secure infrastructure and relies on access control technology, it is not a security problem but it is related to a data management problem. Purpose plays a significant role in the field of database management system privacy preserving techniques. In this chapter we introduced a conditional purpose-based access control (CPBAC) model for privacy protection in the database system that enables enterprises to operate as reliable keepers of their customers' data. The basic concepts of the proposed conditional based access control model are discussed and the possibility has been shown to extract more information from customers by providing a secure privacy policy. The study reveals that this model achieves better progress than the other access control models in the area of privacy preserving in a data mining environment. We also discussed the algorithm to achieve the compliance check between access purpose and intended purposes. The effect of the proposed access control model can be extremely useful for internal access control within an organization as well as for information sharing between organizations.

Our proposed approach provides a complete structure for a privacy preserving access control model. On the basis of this approach, we extend this model in Chapter 3 in the Role-based Access Control (RBAC) model.

Chapter 3

Injecting CPBAC with RBAC

Role-based access control (RBAC) has been widely used in database system management and operating system products because of its significant impact on access control systems. In Chapter 2 we proposed a conditional purposed-based access control (CPBAC) model for privacy protection in a relational database system. In this chapter we inject CPBAC with the conventional well known RBAC. This chapter consists of two parts. In the first part we present a role-involved purpose-based access control (RPAC) model and in the second part we present a conditional role-involved purpose-based access control (CPAC), where a conditional purpose is defined as the intention of data access or usages under certain conditions. The work presented in this chapter extends role-based access control models to a further coverage of privacy preservation in database management systems by adopting purposes and conditional intended purposes and to achieve a fine grained access control. The work in this chapter helps enterprises to circulate a clear privacy promise, and to collect and manage user preferences and consent.

3.1 Introduction

Access control is one of the most popular approaches for protecting private information. It is the process of limiting access to the resources of a system only to authorized users, programs, processes, or other systems [101]. Role based

access control (RBAC) proposed by Sandhu et al. [95] has been widely used in database system management and operating system products because of its significant impact on access control systems. RBAC is described in terms of individual users being associated with roles as well as roles being associated with permissions (each permission is a pair of objects and operations). As shown in Figure 3.1, a role is associated with users and permissions. A user in this model is a human being and a role is a job function or job title within the organization associated with its authority and responsibility. The RBAC model also includes a role hierarchy, a partial order defining a relationship between roles, to facilitate the administration tasks.

In Chapter 2, we developed a conditional purpose-based access control (CPBAC) model [68] that can extract more information from data providers while at the same time assuring privacy. The key characteristic of the CPBAC model is that it allows users to use some data with certain conditions, and multiple purposes can be associated with each data element. It exploited query modification techniques to support data access control based on conditional purpose information. However, RBAC is one of the most popular approaches towards access control to achieve database security and is available in many database management systems, so it needs to be addressed in CPBAC. To implement this, we need to expand the CPBAC model with the conventional well-known RBAC. Such an extension of CPBAC with roles which we refer to in the role-involved purpose-based access control (RPAC) model presented in this chapter. Both access purposes and intended purposes are specified with respect to a hierarchical structure that organizes a set of purposes for a given enterprise.

The importance of privacy preservation has been recognized for a long time, but the concept of privacy has not been supported in RBAC models [95, 102]. A security officer has to assign and check privacy issues if a role is associated with private information. Such a model significantly increases the management efforts

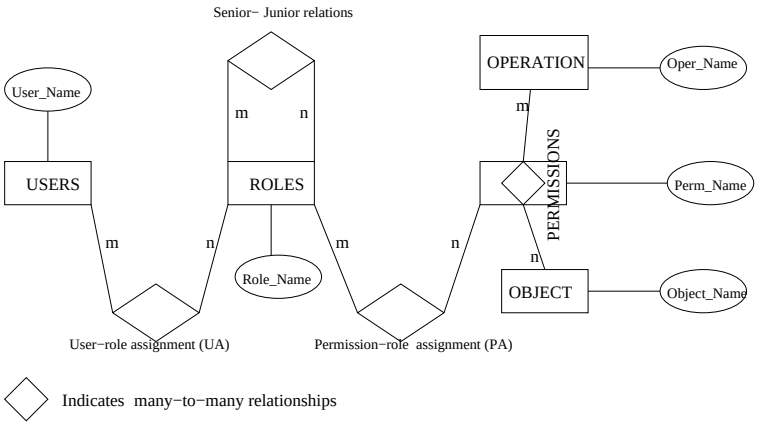


Figure 3.1: Role-based access control model

in a decentralized environments because of the variable private information with different individuals and the continuous involvement from security officers. This chapter provides a bridge of the gap between private preserving techniques and RBAC models.

As mentioned before, Our previous work in Chapter 2 exploited query modification techniques to support data access control based on the conditional purpose information[68] which is not associated with the role-based access control model. However, it is important to analyse access purposes with the RBAC model. For example, suppose the RBAC model is used in a university environment. Students and staff are two roles in the model. Chris, as a student, allows people to use her private information such as home address and home phone for marketing purposes in the university environment. It means the home address and phone cannot be used in other environments except the university and hence the university environment is a condition.

In this chapter we utilize RBAC which supports conditional purposes into our model. Thus the proposed RPAC model in this chapter has the following features:

- It satisfies data providers' requirements and allows users to use data with conditions. The data provider expresses their own privacy preferences through setting the intended purpose with three levels (AIP, CIP and PIP),

while the data owner is responsible for working out the policies for authorization of access purpose.

- Its algorithm utilizes RBAC to achieve the compliance computation between access purpose and intended purpose.
- It extracts more information from data providers by providing more possible options of using their information assuring privacy of private information which maximizes the usability of data.
- It determines the compliance computation between access purpose and intended purpose. Intended purposes are associated with the requested data objects during the access decision in the well-designed hierarchy of private metadata.

3.2 Role-involved CPBAC (RPAC)

The RBAC model is a landmark in the field of access control models and has become a NIST standard [95]. RBAC has been proposed as an alternative approach to discretionary access control (DAC) and mandatory access control (MAC)[1] both to simplify the task of access control management and to directly support function-based access control. It is preliminary designed to satisfy the need of simplifying the authorization management and directly presenting access control policies [100]. The key concept of the RBAC model is role which represents a certain job function or job title within the organization. The permission of performing certain operations on certain data is assigned to roles instead of single users. Users are thus simply authorized to play the appropriate roles, thereby acquiring the roles authorizations. When the user makes a request, the system activates specific roles predefined for him/her. Thus he/she gains the permission of operating directly or indirectly from roles, which considerably simplifies the

[1]MAC and DAC do not handle environments in which the originators of documents retain control over them even after their dissemination.

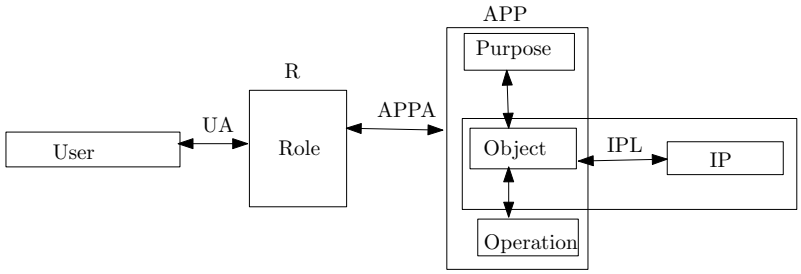


Figure 3.2: RPAC Model

authorization management. Because roles represent organizational functions, an RBAC model can directly support security policies of the organization. In the recent development of the privacy preserving data mining environment many researchers have recognised the importance of purpose, but in the RBAC model purpose is not yet fully investigated. Based on RBAC, the CPAC model extends mainly to the following aspects.

- The access permission is no longer a 2-tuple $\langle Object, Operation \rangle$, but a 3-tuple $\langle Object, Operation, AccessPurpose \rangle$ which is called the access purpose permission.

- The access purpose permission is assigned to roles and after the purpose compliance process, only the objects which are purpose compliant or conditionally compliant can be returned to the users.

In the RPAC model, the entity User is defined as a human being, a machine, a process, or an intelligent autonomous agent. The entity Role represents the working function or working title assigned within the organization according to different authorities and obligations. Roles are created for the various job functions in an organization and users are assigned roles based on their authority and qualifications. Users can be easily reassigned from one role to another. Roles can be granted new permissions as new applications and systems are incorporated and permission can be revoked from roles as needed. The entity Object stands for the data which the user requests and can be abstracted as a data set. The entity operation signifies a certain action that the user wants to perform on the

object. The entity Purpose represents all the possible access purposes in the system and IP signifies the intended purposes with three levels (AIP, CIP, PIP) attached to each data object. Permission is an approval of a particular operation to be performed on one or more objects. The RPAC model is illustrated in Figure 3.2. The formalized definition of the RPAC model is shown as follows:

Definition 3.2.1 (RPAC model):

- *User, Role, Operation, Object, Purpose* represent the set of users, roles, operations, objects and purposes.
- $IP=\{\langle aip, cip, pip\rangle | aip \subseteq \omega, cip \subseteq \omega, pip \subseteq \omega\}$ is the set of the object's intended purposes, where *aip* signifies the object's permitted intended purpose, *cip* is the conditionally permitted intended purpose and *pip* represents the object's forbidden intended purposes [68].
- $R=\{r | r \in Role\}$ is the set of roles.
- $APP=\{\langle o, opt, ap\rangle | o \in Object, opt \in Operation, ap \in Purpose\}$ is the set of access purpose permissions.
- $IPL=\{\langle o, ip\rangle | o \in Object, ip \in IP\}$ represents the set of data objects and their predefined intended purpose.
- $RH \subseteq Role \times Role$ is a partial order on roles, called the inheritance relationship among roles. We also define a partial order $\geq$ which is the transitive closure of *RH*. For example, $r_1 \leq r_2$ means r_1 inherits all permissions of r_2. Figure 3.3 is an example of role hierarchies of the Marketing department for a hypothetical company.
- $PT \subseteq Purpose \times Purpose$ is a partial order on purposes (generalization/ specialization) shown in the purpose tree. Figure 2.1 is an example of purpose tree.

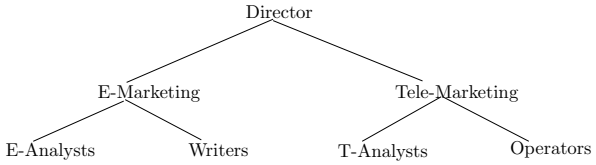


Figure 3.3: Example of Role Hierarchies in Marketing department

- *User Assignment* $UA \subseteq User \times Role$ is a many-to-many mapping relation between users and their assigned roles.

- *Access Purpose Permission Assignment* $APPA \subseteq Role \times APP$ is a many-to-many mapping relation between roles and access purpose permissions. It signifies the action that a certain role performs on a certain object on a certain access purpose.

- *Purpose Compliance* $PC \subseteq APP \bowtie IPL$ is a one-to-one relation between each access purpose permission and data object as well as its predefined intended purposes.

Now we are at the stage to provide function definitions to facilitate the discussion of the RPAC model.

- $assigned\_role : User \rightarrow 2^{Role}$, the mapping of a user u onto a set of roles. Formally,

  $assigned\_role(u) = \{r \in Role | \langle u, r \rangle \in UA\}$

- $assigned\_access\_purpose\_permission : Role \rightarrow 2^{APP}$, the mapping of a role r onto access purpose permissions. Formally,

  $assigned\_access\_purpose\_permission(r) = \{app \in APP | \langle app, r \rangle \in APPA\}$.

- $Purpose\_binding : Object \rightarrow IP$, the mapping of a data object o onto intended purposes ip with three levels, which means finding the bound intended purposes of the object.

- $Purpose\_compliance : AP \times IP \rightarrow \{True, Conditionally\,True, False\}$, is used to determine the compliance between the access purpose and the object's intended purposes [68]. Formally,

  $Purpose\_compliance(ap, ip)$=True if $ap \in IP^{\star}$,

  $Purpose\_compliance(ap, ip)$=Conditionally True if $ap \in IP_c^{\star}$.

In the RPAC model, the users are required to explicitly state their access purpose(s) when they try to access data. That is, the users present an access purpose for each query they issue. During the access decision process, the system combines the requested data with its intended purposes according to privacy metadata and sends the data whose intended purposes are fully compliant or conditionally compliant with the access purpose to the requester. As the model respects customers requirement regarding their data usages and also support RBAC, it prevents private information from disclosure.

3.2.1 Authorization and Authentication

Access purpose is the reason for accessing a data item and it must be determined by the system when a data access is requested. There are different possible methods for determining the access purpose [54]. First, users can be required to state their access purpose explicitly along with requests for data access. Most privacy policy access control models are based on this method. Second, the system registers a special access purpose for each application or stored procedure. This method however can not be used for complex applications or stored-procedures for the reason that requesters may access data objects for multiple access purposes. Lastly, access purposes can be dynamically determined based on attributes of users and the context of the system in addition to requested objects and actions. However, the key challenge for implementing this method is that it may be difficult to infer the access purposes both accurately and efficiently [54]. Among the various possible techniques to determine access purpose, in this chapter we

utilize the method where the users are required to explicitly state their access purposes when they try to access data.

In this model, access purposes are authorized to users through roles. Users are required to state their access purposes along with their queries and the system confirms the stated access purposes by ensuring that the users are indeed allowed to access data for the particular purposes they identified. Now we formally define access purpose authorization and its authentication.

Definition 3.2.2 (Access Purpose Authorization)

Let Ω be a purpose tree and ω be the set of purposes in Ω. Also let R be the set of roles defined in a system. An access purpose is authorized to a specific set of users by a pair $\langle ap, r \rangle$, where ap is a access purpose in ω and r is a role defined over R.

Usually in the typical situation, roles and access purpose are organized in a hierarchical structure. All users authorized for a role r_i are also authorized for any role r_j where $r_i \geq r_j$. Thus, activating a role r_i automatically activates all roles r_j, such that $r_i \geq r_j$. Similarly, authorizing an access purpose ap for a role r_i implies that the users belonging to r_i (or the users belonging to r_j, where $r_i \geq r_j$) are authorized to access data with ap as well as all the descendants of ap in the purpose tree. The access purpose authentication definition below confines the implications of access purpose authorizations.

Definition 3.2.3 (Access Purpose Authentication):

Let Ω be a purpose tree, ω be the set of purposes in Ω and R be the set of roles defined in a system. Suppose that an access purpose ap and a role r_i is activated by a user u. We say that ap is legitimate for u under r_i if there exists an access purpose authorization $\langle ap_l, r_i \rangle$, where ap_l in ω and r_i is a role defined over R such that $ap \in$ Descendants (ap_l) and the users belong to role r_i (or any descendants role of r_i.)

Table 3.1: Intended purposes table

Sl_No.	Table_ID	Table_Name	Cus_ID	Attr_Name	Intended_Purpose
1	1	Customer_info	22	Customer_Name	⟨{General}, {Admin}, {Shipping}⟩
2	1	Customer_info	25	Income	⟨{Marketing}, {Admin}, {Shipping}⟩
3	1	Customer_info	52	Address	⟨{Shipping}, {Admin}, {Marketing}⟩

Consider the purpose tree in Figure 2.1 and the role hierarchies of a Marketing department for a hypothetical company in Figure 3.3. Suppose that access purpose "Service-Updates" are assigned to the "E-Marketing" role, then the users who activate the role "E-Marketing" (or the two descendants roles) can access data for the purpose of "Service-Updates".

By accessing purpose authorization and authentication, users get access purpose permission from access the control engine. Now it is necessary to check whether users' access purposes are fully or conditionally compliant with data's intended purpose for the access decision. In the following section we discuss the compliance computation for access decision.

3.2.2 Access Decision

Usually data providers (customers) are reluctant for any use of their information. On the other hand, data users (enterprisers) want to make use of the collected data as much as possible. Therefore, a negotiating process is necessary between these two parties in order to protect privacy. Again the comfort level of privacy varies from individual to individual.

In our model customers are given three more possible options of using their data. These make them comfortable to release their data fully or conditionally and the private information will be protected. After data are collected, intended purposes with three different levels will be associated with data. As the intended purpose is assigned to every data element, an intended purposes table (IPT) is

Comp_Check$_1$ (ap, $\langle AIP, PIP\rangle$)
/$\star$ This function is required for access decision $\star$/
1. if $ap \in PIP^{\updownarrow}$ then
2. return False;
3. else if $ap \in AIP^{\downarrow}$ then
4. return True;
5. end if

Comp_Check$_2$ (ap, $\langle CIP, PIP\rangle$)
1. if $ap \in PIP^{\updownarrow}$ then
2. return False;
3. else if $ap \in CIP^{\downarrow}$ then
4. return True;
5. end if

Access Decision (ap, Object O)
/$\star$ IPT means intended purpose table $\star$/
1. For each tuple of IPT where Sl_No.= $i(i = 1$ to $n)$
2. c_id=$\prod_{Cus\_ID}$ ($\sigma_{Sl\_No.=i}$(IPT))
3. attr=$\prod_{Attr\_Name}$ ($\sigma_{Sl\_No.=i}$(IPT))
4. if O=$\prod_{Table\_Name}$($\sigma_{Sl\_No.=i}$ (IPT)),
attr $\in$ {A|A is one of O's attributes}
and c_id$\in$ $\prod_{O.Cus\_ID}$ (O)
5. ip=$\prod_{Intended\_Purpose}$($\sigma_{Sl\_No.=i}$(IPT))
6. if (Comp_Check$_1$ (ap, $\langle AIP, PIP\rangle$)= False)
7. O$\leftarrow$ $\prod_{attr_1,attr_2,\ldots,attr_n=null}$
($\sigma_{O.Cus\_ID=c\_id}(O)$)
8. else if Comp_Check$_2$ (ap, $\langle CIP, PIP\rangle$)= False
9. O$\leftarrow$ $\prod_{attr_1,attr_2,\ldots,attr_n=null}$
($\sigma_{O.Cus\_ID=c\_id}(O)$)
10. return O

Figure 3.4: Compliance computation and access decision algorithm

Table 3.2: Customer_info Table with AIP, CIP and PIP

name	age	address	income	name$_{ip}$	age$_{ip}$	address$_{ip}$	income$_{ip}$
Alice	35	21, West St., TBA, QLD 4350	35000	$\langle${G}, {Φ}, {Φ}$\rangle$	$\langle${G}, {M}, {A}$\rangle$	$\langle${G}, {Φ}, {A, S}$\rangle$	$\langle${G}, {A}, {M}$\rangle$
Bob	29	45, Fay CT., TBA, QLD 4350	23000	$\langle${G}, {Φ}, {Φ}$\rangle$	$\langle${G}, {M}, {Φ}$\rangle$	$\langle${G}, {M}, {A, S}$\rangle$	$\langle${G}, {M}, {A}$\rangle$
Ron	56	20, Anita Dr., TBA, QLD 4350	56000	$\langle${G}, {Φ}, {Φ}$\rangle$	$\langle${G}, {M}, {Φ}$\rangle$	$\langle${G}, {Φ}, {A, S}$\rangle$	$\langle${G}, {S}, {A}$\rangle$
Jak	48	25, Wuth St., TBA, QLD 4350	48000	$\langle${G}, {Φ}, {Φ}$\rangle$	$\langle${G}, {M}, {Φ}$\rangle$	$\langle${G}, {M}, {A}$\rangle$	$\langle${G}, {M}, {A}$\rangle$

G={General purpose}, A={Admin purpose}, S={Shipping purpose}, P={Purchase purpose}, M={Marketing purpose}, ip={Intended purpose}=$\langle$AIP, CIP, PIP$\rangle$, Φ={No restriction}.

Table 3.3: Conditional records and intended purposes for Table 3.2

	name	age	address	income
AIP	Alice	35	21, West St., TBA, QLD 4350	35000
CIP	A	30-40	West St., TBA, QLD 4350	30000-40000
PIP	⋆	⋆	⋆	⋆
AIP	Bob	29	45, Fay CT., TBA, QLD 4350	23000
CIP	B	20-30	Fay CT., TBA, QLD 4350	20000-30000
PIP	⋆	⋆	⋆	⋆
AIP	Ron	56	20, Anita Dr., TBA, QLD 4350	56000
CIP	R	50-60	Anita Dr., TBA, QLD 4350	50000-60000
PIP	⋆	⋆	⋆	⋆
AIP	Jak	48	25 Wuth St., TBA, QLD 4350	48000
CIP	A	40-50	Wuth St., TBA, QLD 4350	40000-50000
PIP	⋆	⋆	⋆	⋆

⋆ means information will not be disclosed.

formed. Consider a typical IPT table in Table 3.1 which consists of six columns, where Sl_No is the serial number, Table_ID is the identification of the original table, Cus_ID is the hidden attribute which is added when tables are created, Table_Name is the name of the table in the database and Attr_Name is the attribute name in the table. Thus the storage of intended purposes and data is separated. Data providers (customers) are able to control the release of their data by adding privacy levels into the IPT which will not affect data in the database.

After authorizing access purpose, users get access purpose permission from the access control engine. When data providers submit data, intended purposes with three different levels are defined. The access control engine needs a match process to finish the compliance computation fully or conditionally between access purposes and intended purposes. If the requester's access purpose is fully compliant with the intended purposes of requested data, the engine will release full data to the requester. On the other hand, if the access purpose is conditionally compliant, the engine will release conditional data to the requester, otherwise returned data will be null. Thus in this model the search engine needs to evaluate two compliance checks, the first one is for full compliance and the second one is for conditional compliance. The compliance computation and the access decision

Table 3.4: IPT table

Sl_No.	Table_ID	Table_Name	Cus_ID	Attr_Name	Intended_Purpose
1	3.2	Customer_info	1	Alice	⟨{G}, {Φ}, {Φ}⟩
2	3.2	Customer_info	1	age	⟨{G}, {M}, {A}⟩
3	3.2	Customer_info	1	address	⟨{G}, {Φ}, {A, S}⟩
4	3.2	Customer_info	1	iccome	⟨{G}, {A}, {M}⟩
5	3.2	Customer_info	2	Bob	⟨{G}, {Φ}, {Φ}⟩
6	3.2	Customer_info	2	age	⟨{G}, {M}, {Φ}⟩
7	3.2	Customer_info	2	address	⟨{G}, {M}, {A,S}⟩
8	3.2	Customer_info	2	income	⟨{G}, {M}, {A}⟩
9	3.2	Customer_info	3	Ron	⟨{G}, {Φ}, {Φ}⟩
10	3.2	Customer_info	3	age	⟨{G}, {M}, {Φ}⟩
11	3.2	Customer_info	3	address	⟨{G}, {Φ}, {A, S}⟩
12	3.2	Customer_info	3	income	⟨{G}, {S}, {A}⟩
13	3.2	Customer_info	4	Jak	⟨{G}, {Φ}, {Φ}⟩
14	3.2	Customer_info	4	age	⟨{G}, {M}, {Φ}⟩
15	3.2	Customer_info	4	address	⟨{G}, {M}, {A}⟩
16	3.2	Customer_info	4	income	⟨{G}, {M}, {A}⟩

G={General purpose}, A={Admin purpose}, S={Shipping purpose},
P={Purchase purpose}, M={Marketing purpose}, Φ={No restriction}.

algorithm of the model are illustrated in Figure 3.4. Method Comp_Check returns the result of the purpose compliance check (fully or conditionally) for the given intended purpose with three levels as described in Section 3.2. The Method Access Decision is based on the Comp_Check and the Intended_Purpose of a particular attribute in the IPT table.

Table 3.5: Table return to Russell

name	age	address	income
Alice	30-40	21, West St., TBA, QLD 4350	⋆
Bob	20-30	Fay CT., TBA, QLD 4350	20000-30000
Ron	50-60	20, Anita Dr., TBA, QLD 4350	56000
Jak	40-50	Wuth St., TBA, QLD 4350	40000-50000

⋆ means information will not be disclosed.

Consider a hypothetical customer_info table in Table 3.2. This table is created when data are collected and is based on the customer privacy preferences. Also assume that conditional records and intended purposes for Table 3.2 are available in Table 3.3. Suppose that Russell is an employee working in the Marketing deptaetment of a company and is trying to access customer_info for Marketing

purposes. Assume that the company is using the RBAC model for the privacy preserving access control model and when Russell activated his role, he got access purpose permission from the role for accessing customer_info for Marketing purpose. Based on the information in Table 3.2, the IPT Table for this example is given in Table 3.4. Thus, according to the customer_info table in Table 3.2 and the privacy level of customers associated with the intended purpose (IP), the set of purposes and the set of conditional purposes implied by IP are given by

$IP^{\star}(Sl\_No. = 1)$= AIP$^{\downarrow}$-CIP$^{\updownarrow}$-PIP$^{\updownarrow}$ ={Admin, Purchase, Shipping, Marketing, Profiling, Analysis, Direct, Third-Party, D-Email, D-Phone, T-Email, T-Postal, Special-Offers, Service-Updates}.

$IP_c^{\star}(Sl\_No. = 1)$= CIP$^{\downarrow}$ -PIP$^{\updownarrow}$={Φ}={Null}.

$IP^{\star}(Sl\_No. = 2)$=AIP$^{\downarrow}$-CIP$^{\updownarrow}$-PIP$^{\updownarrow}$={Purchase, Shipping}.

$IP_c^{\star}(Sl\_No. = 2)$=CIP$^{\downarrow}$ -PIP$^{\updownarrow}$={Marketing, Direct, Third-Party, D-Email, D-Phone, T-Email, T-Postal, Special-Offers, Service-Updates}.

$IP^{\star}(Sl\_No. = 3)$=AIP$^{\downarrow}$-CIP$^{\updownarrow}$-PIP$^{\updownarrow}$={Purchase, Marketing, Direct, Third-Party, D-Email, D-Phone, T-Email, T-Postal, Special-Offers, Service-Updates}.

$IP_c^{\star}(Sl\_No. = 3)$=CIP$^{\downarrow}$ -PIP$^{\updownarrow}$={Φ}={Null}, and so on.

So, when Russell hands a query

SELECT name, age, address, income

FROM Table 3.2

FOR Marketing Purpose,

then by using the IPT table in Table 3.4 and the algorithm in Figure 3.4 to perform the computation of the privacy protection access decision, the return table to Russell will be in Table 3.5. From Table 3.5, it can be said that customers' consent regarding the privacy level have been taken into account and the private information is protected.

Kabir and Wang [68] proposed an approach for safeguarding a consumer's privacy while allowing data mining and usage of data provided to an organization. The proposed CPBAC model satisfies the customer privacy requirement and allows users to use some data for certain purposes with conditions, thus extracting more information from data providers. The model is exploited by using query modification techniques to support access control based on purpose information. One of the main challenges is to implement the model with query modification techniques as individuals' privacy levels are ternary, not binary. On the other hand, the authors overlooked to implement the CPBAC model in RBAC. As RBAC is the most popular approach in access control models and is now used in many database management systems, it is essential to execute the CPBAC model in RBAC. By contrast, in this chapter we developed an RPAC model, where the CPBAC model is thoroughly investigated with roles.

3.3 A conditional Role-involved CPBAC (CPAC)

In Section 3.2, we developed a RPAC model where users are required to explicitly state their access purposes when they try to access data. Although this method is simple and easy to implement, it requires complete trust about the identity of users and thus the overall privacy that the system is able to provide entirely relies on the users' trust worthiness. To overcome this problem, this section presents a conditional role-involved purpose-based access control (CPAC) model, where users dynamically activate conditional roles in accordance with the context attributes. Based on the conditional role, access permissions are assigned that represent what can be accessed for what purpose to roles under certain conditions. On the other hand, conditional purpose is applied along with allowed purpose and prohibited purpose in the model. Access purpose is verified in a dynamic behavior, based on user attributes, context attributes and authorization policies. Intended purposes are dynamically associated with the requested data

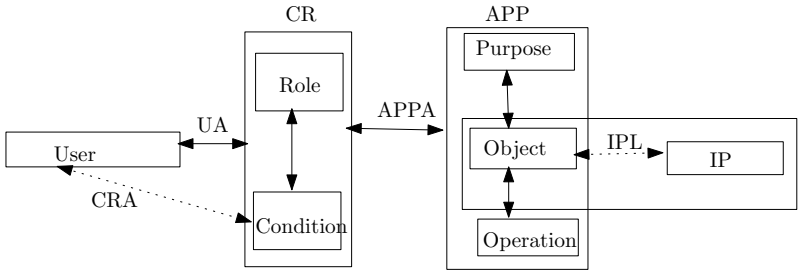


Figure 3.5: CPAC Model

object during the access decision. Access purpose authorization and authentication in the model are studied with the hierarchical purpose structure. The model separates authorization of access purpose from access decision, which improves the flexibility of private data control.

3.3.1 CPAC model

Based on RBAC, the CPAC model extends mainly in the following aspects.

- As in RPAC, the access permission is no longer a 2-tuple $\langle Object, Operation \rangle$, but a 3-tuple $\langle Object, Operation, AccessPurpose \rangle$ which is called the access purpose permission.

- Rather than static roles, the access purpose permission is assigned to Conditional Roles (CR). Users dynamically activate conditional roles in accordance with the context attributes during the access purpose and after the purpose compliance process, only the users which are purpose compliant or conditionally compliant can be returned to the users.

In the CPAC model, the privacy preserving access purpose is not explicitly associated with users but given to conditional roles by access purpose permissions. The activation of conditional roles is dynamically executed based on users' attributes and the state of the system. In this model, the NEW entity Condition is a predicated logic expression that role attributes and system attributes must satisfy. The CPAC model is illustrated in Figure 3.5, where the dotted lines represent dynamic associations. The formalized definition of the CPAC model is

shown as follows:

Definition 3.3.1 (CPAC model):

- *User, Role, Operation, Object, Purpose, Condition* represent the set of users, roles, operations, objects, purposes and conditions.
- $IP=\{\langle aip, cip, pip\rangle | aip \subseteq \omega, cip \subseteq \omega, pip \subseteq \omega\}$ is the set of object's in intended purposes, where *aip* signifies the object's permitted intended purpose, *cip* is the conditionally permitted intended purpose and *pip* represents the object's forbidden intended purposes [68].
- $CR=\{r | r \in Role, c \in C\}$ is the set of roles with condition expression.
- $APP=\{\langle o, opt, ap\rangle | o \in Object, opt \in Operation, ap \in Purpose\}$ is the set of access purpose permissions.
- $IPL=\{\langle o, ip\rangle | o \in Object, ip \in IP\}$ represents the set of data objects and their predefined intended purpose.
- $RH \subseteq Role \times Role$ is a partial order on roles, called the inheritance relationship among roles. We also define a partial order $\geq$ which is the transitive closure of *RH*. For example, $r_1 \leq r_2$ means r_1 inherits all permissions of r_2. Figure 3.3 is an example of role hierarchies of a Marketing department for a hypothetical company.
- $PT \subseteq Purpose \times Purpose$ is a partial order on purposes (generalization/ specialization) shown in the purpose tree. Figure 2.1 is an example of purpose tree.
- *User Assignment* $UA \subseteq User \times Role$ is a many-to-many mapping relation between users and their assigned roles.

- *Condition Role Assignment* $CRA \subseteq User \times CR$ is a many-to-many mapping relation between users and their conditional roles.

- *Access Purpose Permission Assignment* $APPA \subseteq CR \times APP$ is a many-to-many mapping relation between conditional roles and access purpose permissions. It signifies the action that a certain role performs on a certain object for a certain access purpose.

- *Purpose Compliance* $PC \subseteq APP \bowtie IPL$ is a one-to-one relation between each access purpose permission and data object as well as its predefined intended purposes.

Now we are at the stage to provide function definitions to facilitate the discussion of the CPAC model.

- $assigned\_role : User \rightarrow 2^{Role}$, the mapping of a user u onto a set of roles. Formally,

  $assigned\_role(u) = \{r \in Role | \langle u, r \rangle \in UA\}$

- $active\_condition\_roles : User \rightarrow 2^{CR}$, the mapping of a User s onto a set of condition roles.

- $assigned\_access\_purpose\_permission : CR \rightarrow 2^{APP}$, the mapping of a condition role cr onto access purpose permissions. Formally,

  $assigned\_access\_purpose\_permission(cr) = \{app \in APP | \langle app, cr \rangle \in APPA\}$.

- $Purpose\_binding : Object \rightarrow IP$, the mapping of a data object o onto intended purposes ip, which means finding the bounded intended purposes of the object.

- $Purpose\_compliance : AP \times IP \rightarrow \{True, ConditionallyTrue, False\}$, is used to determine the compliance between the access purpose and the object's intended purposes [68]. Formally,

$Purpose\_compliance(ap, ip)$=True if $ap \in IP^{\star}$,

$Purpose\_compliance(ap, ip)$=Conditionally True if $ap \in IP_c^{\star}$.

The determination of the access purpose is based on the enabling conditional roles dynamically and the association between access purpose and roles.

3.3.2 Authorization and Authentication

In a typical privacy policy it is determined who can access what under what conditions and for what purpose. As the existing role definitions are predefined for the access permission assignments, they may not adequately specify the set of users to whom we wish to grant an access purpose. The idea of conditional role is introduced which is based on the notion of role attribute and system attribute. On the basis of conditional role, access permissions are assigned that represent what can be accessed for what purpose to roles under certain conditions. Thus, users dynamically activate conditional roles according to their context attributes during the access process.

Definition 3.3.2 (Role Attributes): Role attributes are defined as the set of role properties linked to the grant of access purpose. Let *RoleAttribute* denote the set of role attributes. Every role $r \in Role$ is associated with a set of role attributes, denoted by $r.Attributes$=$\{r.attr_1, r.attr_2, ..., r.attr_n\}$. Each attribute $r.attr_i$ is associated with a finite domain of possible values, denoted as D_i. Let *RoleAttributevalue* denote the set of all possible role attribute values.

Definition 3.3.3 (System Attributes): System attributes are defined as the set of properties linked to the context of the access control system. The set of system attributes is denoted by $SystemAttribute$=$\{sysattr_1, sysattr_2, ..., sysattr_n\}$. The conditions of the access control system are specified by the values of the system attributes. Let *SystemAttributevalue* denote the set of all possible system attribute values.

Access purpose permission is directly assigned to conditional roles not in static roles. Thus the action of enabling conditional roles plays a significant part in the whole process of access purpose authorization. Again enabling a conditional role needs dynamic condition evaluation based on user attributes and system context, which is the difference between role activation in RBAC and conditional role activation in our model.

Definition 3.3.4 (Conditional Roles): A conditional role is a 2-tuple, denoted by $CR = \langle r, C \rangle$, where $r \in R$ represents the predefined static role (similar to the role attribute in RBAC) and $c \in C$ represents the conditions that the values of role attributes and system attributes must satisfy in the season. Note that C can be constructed from primitive constraints using $\wedge$ (AND), $\vee$ (OR), and $\neg$ (NOT). Let $X = RoleAttribute \cup SystemAttribute$, each $x \in X$ has a finite domain of possible values, denoted as $Domain(x)$. Each predicate in C defined over X is of the form $op_r = (x\ act\ value)$, where $x \in X, value \in Domain(x)$ and $act \in \{=, \neq, <, >, \leq, \geq\}$.

The condition C over X can be defined recursively as follows:

- Each predicate op_r can be a condition statement with the form $(x\ act\ value)$ which is called an automatic condition;
- If op_{r_i} and op_{r_j} are conditions then $op_{r_i} \wedge op_{r_j}$ is a condition, but $op_{r_i} \vee op_{r_j}$ consists of two separate conditions.

In the CPAC model, the relations between users and static roles are predefined by security administrators and access purpose permission is assigned to conditional roles, not to the static roles. Thus when a request arrives from a user, enabling conditional roles involve the following steps:

- Finding static roles for the user

Table 3.6: Conditional roles algorithm

Input: *user* is the one who requests an access
system is current system attributes
Output: conditional role set *enable_CR* activated by user

1. Let set *enable_CR* be a empty set of conditional roles
2. *user.Attribute* ← the attributes of *user*
3. *Role* ← assigned_roles(user)
4. for each *role* ∈ *Role*
5. initial *role.Attribute* with *user.Attribute*
6. for each $cr \in CR$ when $cr = \langle c, r \rangle$
7. if $role = r$
8. if $check\_condition(c, role, system) = True$
9. then $enable\_CR = enable\_CR \bigcup \{cr\}$
10. return *enable_CR*

- Finding all the conditional roles for the user. If in any season, the value of attribute r in the conditional role $\langle c, r \rangle$ equals static roles of the user, $\langle c, r \rangle$ is a conditional role.

- Recurring conditional roles that meet the condition evaluation during the season.

The complete algorithm for enabling conditional roles is shown in Table 3.6. Evaluating the conditions based on user attributes and system attributes is an important step in enabling condition roles. Here the function $check\_condition(c, role, system)$ returns *true*, if the condition logical expression is a tautology when each variable in condition express c is substituted with the values of corresponding attributes. We say that a user u with the static role *role* can activate a condition role cr_i in a system if the following conditions are satisfied:

- $role = cr_i[r]$;

- $check\_condition(c, role, system) = True$

As mentioned before, access purpose permissions are assigned to conditional roles, not to individual users. In the CPAC model, the relations between users and static roles are predefined by security administrators and it is assumed that

roles are already enabled when a request arrives. After roles are enabled, access purpose authorization can be simply implemented by authorizing access purpose permissions to certain roles based on the *Access Purpose Permission Assignment (APPA)*. *APPA* consists of predefined relations between *Conditional Roles* and *Access Purpose Permissions* in the policies determined by database administrators. Therefore, after a user enables a conditional role cr, he/she gets the access purpose permissions which are assigned to cr. Now we formally define access purpose authorization and its authentication.

Definition 3.3.5 (Access Purpose Authorization) Let Ω be a purpose tree and ω be the set of purposes in Ω. Also let R be the set of roles defined in a system. An access purpose is authorized to a specific set of users by a pair $\langle ap, cr \rangle$, where ap is a access purpose in ω and cr is a conditional role defined over X.

All users authorized for a conditional role cr_i are also authorized for any role cr_j where $cr_i \geq cr_j$. Thus, activating a conditional role cr_i automatically activates all conditional roles cr_j, such that $cr_i \geq cr_j$. Similarly, authorizing an access purpose ap for a conditional role cr_i implies that the users belonging to cr_i (or the users belonging to cr_j, where $cr_i \geq cr_j$) are authorized to access data with ap as well as all the descendants of ap in the purpose tree. The access purpose authentication definition below confines the implications of access purpose authorizations.

Definition 3.3.6 (Access Purpose Authentication): Let Ω be a purpose tree, ω be the set of purposes in Ω and R be the set of roles defined in a system. Suppose that an access purpose is ap and a role r_i is activated by a user u. We say that ap is legitimate for u under r_i if there exists an access purpose authorization $\langle ap_l, cr_i \rangle$, where ap_l in ω and $cr_i = \langle C_i, r_i \rangle$ is a conditional role defined over X such that $ap \in$ Descendants (ap_l) and the users belong to conditional role cr_i (or any descendants of the conditional role of cr_i.)

Consider the purpose tree in Figure 2.1 and the role hierarchies of a Marketing department for a hypothetical company in Figure 3.3. Suppose that access purpose "Service-Updates" are assigned to the $\langle E - Marketing, (Explevel > 10) \wedge (timeofday \geq 9) \wedge (timeofday \leq 17)\rangle$, assuming *Explevel* is defined as a role attribute and *timeofday* is defined as a system attribute. Then only the users who activate the role "E-Marketing" (or the two descendants role) with their $Explevel > 10$ can access data for the purpose of "Service-Updates" between 9 am and 5 pm.

Through conditional roles, access purpose authorization and authentication, users get access purpose permission from the access control engine.

Suppose that Russell is an employee working about 12 years in the Marketing department of a company. Assume that the company is using the RBAC model for a privacy preserving access control model and in the normal office hours (9 am to 5 pm), then only the users who have more than 10 years of working experience can access customer_ info for **Marketing purpose**. That means, in the office hours access purpose **Marketing** is authorized to users who are working in the Marketing department and have 10 years of experience. When Russell activates a role in the office hours, on the basis of his role attribute and the context of the system, he will get permission from the access control engine to access customer_ info for **Marketing purpose**. In other words, when Russell activates his role, the system reasonably infer that Russell is trying to use customer_ info for **Marketing purpose**. Thus after activating his role, Russell gets access purpose permission for accessing customer_ info for **Marketing purpose**.

3.4 Conclusion

In this chapter we injected CPBAC with the RBAC. In the first part of this chapter, we presented a RPAC model where users explicitly state their access purpose and in the second part we presented a CPAC where access purpose is

dynamically associated with roles. These models enable enterprises to operate as reliable keeper of their customers data. We analyzed algorithms to achieve the compliance check between access purpose and intended purposes. The effect of the proposed access controls can be useful for internal access control within an organization as well as information sharing between organizations, as many systems are already using RBAC mechanisms for the management of access permission. These techniques can also be used by enterprises to enforce the privacy promises they make and to enable their customers to maintain control over their data.

Part II
Data Anonymization

Chapter 4

Systematic Clustering for k-Anonymization

Privacy preservation of individuals has drawn considerable interest in data mining research. The k-anonymity model proposed by Samarati and Sweeney is a practical approach for data privacy preservation and has been studied extensively for the last few years. Anonymization methods via generalization or suppression are able to protect private information, but lose valued information. The challenge is how to minimize the information loss during the anonymization process. This chapter presents a clustering[1] based k-anonymization technique to minimize the information loss while at the same time assuring data quality. We refer to the challenge as a systematic clustering problem for k- anonymization which is analyzed in this chapter. The proposed technique adopts group similar data together and then anonymizes each group individually.

4.1 Introduction

In recent years, the phenomenal advances in technological developments in information technology have lead to an increase in the capability to store and record personal data about customers and individuals [32]. Data mining is a common methodology to retrieve and discover useful hidden knowledge and information

[1]Clustering partitions record into clusters such that records within a cluster are similar to each other, while records in different clusters are most distinct from one another.

from personal data. This has lead to concerns that personal data may be breached and misused. Therefore it is necessary to protect personal data through some privacy preserving techniques before conducting data mining. Thus privacy preserving is an important issue and has captured the attention of many researchers in the data mining research community.

One of the most important concepts for privacy is anonymity. Anonymity refers to a state where one's identity is completely hidden, and anonymity is oftentimes used as a synonym for privacy [34]. Anonymous data can protect individuals in two ways: firstly to protect identity privacy, for example by making it impossible to learn to whom a data record is related and secondly, through attribute privacy for example making it impossible to know about a particular property of individuals. In any database, specially where health records are collected by hospitals or government organizations, anonymity has a significant role to protect privacy as the information linked to individuals could be highly sensitive. In commercial databases where organizations would like to disclose an individual's data to third parties (e.g. external organizations), anonymity could be used to protect the privacy of individuals as in such cases an individual's privacy may not be respected. Thus, within organization individuals' data should be restricted in terms of access and anonymous, by removing all information that can directly link data items to individuals via generalization or suppression before disclosing, so that privacy is not beached. Such a process is referred to as data anonymization.

A contemporary approach dealing with data privacy relies on k-anonymity. The k-anonymity model proposed by Samarati and Sweeney [17, 18] is a simple and practical privacy-preserving approach to protect data from individual identification. The k-anonymity model works by ensuring that each record of a table is identical to at least $(k-1)$ other records with respect to a set of privacy-related features, called quasi-identifiers, that could be potentially used to identify

Table 4.1: Patients records in a hospital

ZipCode	Gender	Age	Education	Disease	Expense
4350	Male	24	9th	Flue	2000
4351	Male	25	10th	Cancer	3500
4352	Male	26	9th	HIV+	6500
4350	Male	35	9th	Diabetes	2000
4350	Female	40	10th	Diabetes	3200
4350	Female	38	11th	Diabetes	2800

Table 4.2: 3-Anonymization table

ZipCode	Gender	Age	Education	Disease	Expense
435⋆	Person	[21-30]	Educated	Flue	2000
435⋆	Person	[21-30]	Educated	Cancer	3500
435⋆	Person	[21-30]	Educated	HIV+	6500
435⋆	Person	[31-40]	Educated	Diabetes	2000
435⋆	Person	[31-40]	Educated	Diabetes	3200
435⋆	Person	[31-40]	Educated	Diabetes	2800

individuals by linking these attributes to external data sets [37]. Therefore, privacy related information can not be revealed from the k-anonymity protected table during a data mining process. For example, consider the patient diagnosis records in a hospital in Table 4.1, where the attributes ZipCode, Gender, Age and Education are regarded as quasi-identifiers. A diagnosis classifier can predict the patient's illness history based on attributes of ZipCode, Gender, Age and Education using these data. If the hospital simply publishes the table to other organizations for classifier development, those organizations might extract patients' disease histories by joining this table with other tables [36]. By contrast, Table 4.2 is a 3-anonymization version where data values of Table 4.1 in attributes ZipCode, Gender, Age and Education have been generalized as common values and the number of records in its two equivalence classes are both equal to three. It should be noted that the value of k in the k-anonymity model is specified by users according to the purpose of their applications. By enforcing the k-anonymity requirement, it is guaranteed that even though an adversary

knows that a k-anonymous table contains the record of a particular individual and also knows some of the quasi-identifier attribute values of the individual, he/she cannot determine which record in the table corresponds to the individual with a probability greater than $\frac{1}{k}$ [34]. This indicates that the larger the values of k, the less chance the adversary has of being able to determine personal identifiable information and the data is more protected. On the other hand, if the k-values are too large it incurs more information loss. Therefore, the k-value of the k-anonymization problem should not be too small or too large.

Usually, there are two methods to accomplish in k-anonymizing a dataset. The first one is suppression which involves not releasing an entire tuple or a value at all to the third party, which is just like deleting them. The other one is generalization which involves replacing the value or tuple with a less specific but semantically consistent value. For example, suppose the following five ages of individuals 51, 52, 53, 53, 55 exist. We can generalize attribute Age to age groups 50-55. On the other hand, we can also generalize them to in other set $5^{\star}$. However, we can suppress the age values by $\star$. Intuitively, generalization is better than suppression because of extracting at least some information. Undoubtedly, anonymization is accompanied by information loss. In order to be useful in practice, the dataset should stay as much informative as possible. Hence, it is necessary to consider deeply the tradeoff between privacy and information loss. To minimize the information loss due to k- anonymization, all records are partitioned into several groups such that each group contains at least k similar records with respect to the quasi-identifiers. Then the records in each group are generalized or suppressed such that the values of each quasi-identifier are the same. Such similar groups are known as clusters. In the context of data mining, clustering is a useful technique that partitions records into clusters such that records within a cluster are similar to each other, while records in different clusters are most distinct from one another [37]. Thus, the k-anonymity model can be addressed from the viewpoint of

clustering.

As discussed, a key difficulty of data anonymization comes from the fact that data quality and privacy are conflicting goals. Although it is possible to enhance data privacy by hiding more data values, it decreases data quality. By contrast, disclosing more data values increases data quality but decreases data privacy. Thus it is necessary to devise new k-anonymization approaches that best address both the quality and the privacy of the data. To overcome this challenge, this chapter proposes a new clustering method for k-anonymization. This method has a time complexity of $O(\frac{n^2}{k})$ in the clustering stage, where n is the total number of records that containing individuals concerning their privacy. However, the algorithm requires sorting the tuples in the dataset once, which alone takes $O(n*logn)$ time. According to this method, first exclude the number of records containing individuals who do not bother about the disclosure of personal identification information. Sort all records by their quasi-identifiers and partition all records into $[\frac{n}{k}]$ groups. Randomly select a record r from the first group to form the first cluster and the first records of the subsequent clusters will form in a systematic way. Then adjust the records in each group in a systematic way such that each group contains at least k records. Finally, distribute the records of individuals who do not bother about the disclosure to their closest clusters or these records constitute another cluster/clusters depending on the number of such records and the k-value. Note that the process of including such records causes no information loss. There are many clustering based k-anonymization techniques in the literature [34, 35, 36, 37, 45]. However, the proposed systematic clustering method differs from previously proposed clustering based k-anonymization methods in four different ways. First, our method endeavours to make all clusters simultaneously. By contrast, the methods proposed by Byun *et al.* [34] and Loukides and Shao [35] build one cluster at a time. Second, it takes less time than the previous two methods as only the first record randomly selects and the subsequently records

form in a systematic way. Third, since the first record of each cluster contains a non identical value, this method easily captures if there are any extreme values, and lastly the total information loss will be reduced as in the final step the process incurs no information loss. The performance of the proposed method is compared against the method proposed by Byun *et al.* [34]. The experimental results show that the proposed clustering method outperforms their method with respect to both information loss and computational efficiency.

4.2 Preliminaries Relating to k- Anonymization

The k-anonymity model has drawn considerable interest in the research community for the last few years and a number of algorithms have been proposed [26, 27, 28, 29, 30, 31, 42, 43]. However, these suffer from high information loss mainly due to reliance on pre-defined generalization hierarchies [27, 28, 29, 31] or total order [26, 30] imposed on each attribute domain. Some existing work on k-anonymization has attempted to capture usefulness by measuring the number of total suppressions [38], the size of the anonymized group [27, 30], the height of generalisation hierarchies [17, 34], or information loss through anonymization [39]. However, such metrics fail to detain security. In other works by Machanavajjhala *et al.* [40], and Truta and Vinay [40, 41] attempts have been made to enhance protection by enforcing anonymized groups. The intuition behind this is that if the values of a sensitive attribute of an anonymized group are quite diverse, then it is difficult for an attacker to breach privacy. However, these frequency-based criteria treat numerical attributes as categorical and thus protection is not captured adequately. For instance, l-diversity proposed by Machanavajjhala et al. [40] requires a sensitive attribute to have at least l distinct values in an anonymized group. Please refer to Ciriani et al. [26] for a survey of various k anonymization approaches.

4.2.1 Information Loss

Anonymization via generalization or suppression usually causes information loss. Now a natural question arises of how much information is lost due to anonymization. Thus the idea of information loss is used to measure the amount of information loss due to k-anonymization. There are various methods of conniving information loss [14, 27, 34, 37, 105]. The measurement of information loss in this article is based on the description given by Byun et al. [34]. Please also refer to Byun et al. [34] for more details.

Let η denote a set of records with r numeric quasi-identifiers $N_1, N_2, ..., N_r$ and s categorical quasi-identifiers $C_1, C_2, ..., C_s$. Let $\Im = \{\Omega_1, \Omega_2, ..., \Omega_p\}$ be a partitioning of η, such that $\cup_{i=1}^{p} \Omega_i = \eta$, Ω_i and Ω_j $(i \neq j)$ are pair wise mutually exclusive. To generalize the values of each categorical attribute $C_i (i = 1, 2, ..., s)$, let τ_{C_i} be the taxonomy tree defined for the domain of C_i.

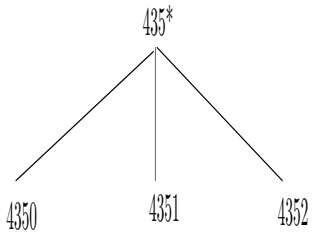


Figure 4.1: Taxonomy tree of ZipCode.

Consider a cluster Ω in η which consists of some numerical and categorical attributes. Let $N_{i_{max}}$, $N_{i_{min}}$ be the maximum and minimum values of the records in Ω and $\eta_{N_{i_{max}}}, \eta_{N_{i_{min}}}$ be the maximum and minimum values of the records in η with respect to numeric attribute $N_i (i = 1, 2, ..., r)$ and $\cup_{C_i}$ be the union set of values in Ω with respect to the categorical attribute $C_i (i = 1, 2, ..., s)$. Then the amount of information loss due to generalizing Ω, denoted by $IL(\Omega)$ is defined as

$$IL(\Omega) = \mid \Omega \mid .(\sum_{i=1}^{r} \frac{N_{i_{max}} - N_{i_{min}}}{\eta_{N_{i_{max}}} - \eta_{N_{i_{min}}}} + \sum_{j=1}^{s} \frac{H(\Lambda(\cup_{C_j}))}{H(\tau_{C_j})})$$

where $\mid \Omega \mid$ is the number of records in Ω, $\tau(\cup_{C_j})$ is the subtree rooted at the lowest common ancestor of every value in $\cup_{C_j}$ and $H(\tau)$ is the height of taxonomy

tree τ.

Suppose that the total number of records in η is partitioned into p clusters, namely $\Im = \{\Omega_1, \Omega_2, ..., \Omega_p\}$. The total information loss of η is the sum of the information loss of each $\Omega_i (i = 1, 2, ..., p)$. Therefore, the total information loss will be:

$$\begin{aligned} IL(\eta) &= \sum_{i=1}^{p} IL(\Omega_i) \\ &= \sum_{i=1}^{p} \mid \Omega_i \mid .(\sum_{k=1}^{r} \frac{N_{ik_{max}} - N_{ik_{min}}}{\eta_{N_{ik_{max}}} - \eta_{N_{ik_{min}}}} + \sum_{j=1}^{s} \frac{H(\Lambda(\cup_{C_{ij}}))}{H(\tau_{C_{ij}})}) \end{aligned} \quad (4.1)$$

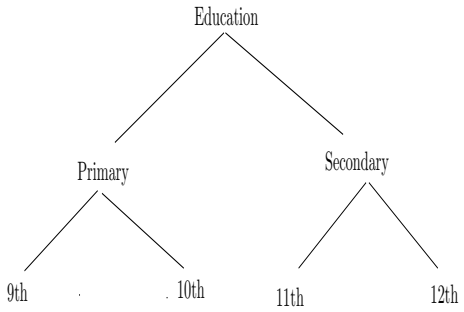


Figure 4.2: Taxonomy tree of Education.

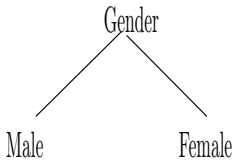


Figure 4.3: Taxonomy tree of Gender.

The main objective of clustering techniques is to construct the clusters in such a way that the total information loss of η will be minimum.

Example 4.2.1 Consider patient records in Table 4.1 and the 3-anonymization table in Table 4.2. The anonymized table consists of two clusters. The first cluster consists of the first three records and the second clusters consists of the last three records. Consider attributes ZipCode, Gender, Age, Education, where Age is a

quantitative variable and the others are categorical variables. Also consider the taxonomy tree of ZipCode, Education and Gender in Figure 4.1, Figure 4.2 and Figure 4.3 respectively. In the table the number of clusters is 2 and the size of each cluster is 3. In the first cluster the maximum and minimum values are respectively 26 and 24, and in the second cluster these values are respectively 40 and 35. Also the maximum and minimum values of all records are respectively 40 and 24. Then the total information Loss of the anonymized table in Table 4.2 will be

$$IL(\eta) = \mid 3 \mid (\frac{26-24}{40-24} + 1 + 1 + \frac{1}{2}) + \mid 3 \mid (\frac{40-35}{40-24} + 1 + 1 + \frac{2}{2}) \approx 14.81. \quad (4.2)$$

4.2.2 Clustering based techniques

Clustering based techniques are now used in k-anonymization to protect the privacy of sensitive attributes and there are various clustering techniques in the literature [30, 34, 35, 36, 37]. Byun et al. [34] introduced clustering techniques instead of equivalence class on k anonymization and proposed the greedy k-member clustering algorithm. This algorithm works by first randomly selecting a record r as the seed to start building a cluster, and subsequently selecting and adding more records to the cluster such that the added records incur the least information loss within the cluster. Once the number of records in this cluster reaches k, this algorithm selects a new record that is the furthest from r, and repeats the same process to build the next cluster. When there are fewer than k records not assigned to any cluster yet, this algorithm then individually assigns these records to their closest clusters. This algorithm has two drawbacks. First, it is slow. Second, it is sensitive to outliers. To build a new cluster, this algorithm chooses a new record that is the furthest from the first record selected for the previous cluster. If the data contains outliers, it is likely that outliers have a great chance of being selected. If a cluster contains outliers, the information loss of this cluster

increases. The time complexity of the algorithm is $O(n^2)$, where n is the number of records in the data set to be anonymized. Their experimental results showed that the k-member algorithm causes significantly less information loss than another k-anonymization technique called "Mondrian" proposed by LeFevre et al. [30].

Loukides and Shao [35] proposed another clustering technique for k- anonymization. Similar to k-member, this algorithm forms one cluster at a time. But, unlike the k-member algorithm, this algorithm chooses the seed of each cluster randomly. Also, when building a cluster, this algorithm keeps selecting and adding records to the cluster until the information loss exceeds a user defined threshold. If the number of records of a particular class is less than k, the entire cluster is deleted. With the help of the user-defined threshold, this algorithm is less sensitive to outliers. The time complexity of the algorithm is $O(\frac{n^2 log(n)}{c})$, where c is the average number of records in each cluster. However, this algorithm also has two drawbacks. First, it is difficult to decide a proper value for the user-defied threshold. Second, this algorithm might delete many records, which in turn cause a significant information loss. Chiu and Tsai [36] proposed another algorithm for k-anonymization that adapts the weighted feature c-means clustering. Unlike the previous two algorithms, this algorithm attempts to build all clusters simultaneously by first randomly selecting $\lfloor \frac{n}{k} \rfloor$ records as seeds. Then this algorithm allocates all records in the data set to their respective closest cluster and consequently updates feature weights to minimize information loss. This process is continued until the assignment of records to cluster stops changing. If some clusters contain fewer than k records, then those clusters should be merged with other large clusters to satisfy the k-anonymity requirement. One of the main drawback of this algorithm is that it can only be used for quantitative quasi-identifier. The time complexity of this algorithm is $O(\frac{t^2}{k})$, where t is the number of iterations needed for the assignment of records to clusters to converge.

To reduce the information loss and execution time, recently Lin and Wei [37] proposed an efficient one-pass k-mean clustering problem that runs in $O(\frac{n^2}{k})$. They showed that their algorithm performs better than the proposed algorithm of Byun et al. [34] with respect to both execution time and information loss. Like Chiu and Tsai's [36] algorithm, this algorithm forms all clusters at a time. According to their methods, first sort all records by their quasi-identifiers, then determine approximate number of clusters, by $p = \frac{n}{k}$, where k is the cluster size. Then randomly select p records as seeds to build p clusters. For each record r the algorithm finds the cluster that is closest to r, assigns r to that cluster and subsequently updates the center point. Finally, if some clusters contain more than k records remove excess records from those clusters that are dissimilar to most of the records and then add these records to other similar clusters (whose size less is than k). Although this method has less execution time there is still a chance of being affected by extreme values. Again if this algorithm first selects p records that come from the same equivalent class then the total information loss will be higher.

4.3 The New Systematic Clustering Method

As discussed before, clustering escorts to better data quality of the disclosed dataset as it partitions a set of records into groups such that records in the same group are more similar to each other than to records of other groups. If the records in a particular group are more similar, the group leads to a minimal generalization and thus incurs less information loss. In this respect, the problem of k-anonymization can also be considered as a clustering problem, where each equivalent class is a cluster and the size of each cluster is at least k. So the optimal solution of a clustering problem is to construct a set of clusters such that the total information loss will be at a minimum. In this section, we formally define and present our systematic clustering algorithm that minimizes information loss and

respects the k-anonymity requirement.

4.3.1 Systematic clustering problem

There are various clustering problems in the literature. Among them, the k-center clustering problem proposed by Gonzalez [45] aims to find k clusters from a given dataset such that the maximum inter-cluster distance (or radius) is minimized. Thus the optimum solution is to constitute p clusters $\{\Omega_1, \Omega_2, ..., \Omega_p\}$ in such a way that it minimizes the cost metric

$$MAX_{i=1,...,p} MAX_{j,k=1,...,|\Omega_i|} D(r_{i,j}, r_{i,k}), \tag{4.3}$$

where $r_{i,j}$ represents a data point in cluster Ω_i and $D(x, y)$ is a distance between two data points, x and y.

In the k-anonymity problem the only restriction is that the number of records in each equivalence class should be at least k and there is no such restriction about the number of clusters. So a clustering problem is to form in such a way that each cluster contains at least k similar records and the sum of information losses of all clusters is at a minimum. The proposed k member clustering problem of Byun et al.[34] satisfies this criterion but one of the most important problems of this algorithm is that it spends a lot of time selecting records from the input set. To reduce time of selecting records from the whole set, a systematic method of selecting records may be helpful. To apply a systematic method of selecting records, first of all it is necessary to sort all records in the whole data set with respect to quasi-identifiers. For example, consider the dataset in Table 4.1, where there are 6 records and suppose that the dataset is already sorted according to the quasi identifier attributes ZipCode, Gender, Age and Education. If the anonymized table follows 3-anonymity requirements, then the number of clusters should be $\frac{6}{3} = 2$. First select a record (say, the 2th record) from the first 3 records to form the first cluster. Then select $(2 + 3)th = 5$th record in a systematic way to form the second cluster. Now again select another record from the first 3 records (say,

3rd, not 2th, as it already selected) and calculate the information loss with both clusters using the equation (4.1). The information loss is 4.25 and 7.75, if this record is included in the first cluster and second cluster respectively. So, the 3rd record will be included in the first cluster as it causes the least information loss. Similarly, select $(3+3)th = 6$th record in a systematic way and include it in the the second cluster. Finally select the 1st and $(1+3)th = 4$th record and include these records respectively as first and second cluster as they will then cause least information loss. If the total number of records are not exactly divisible by the k-anonymity parameter, then the rest of the records will be included in the similar clusters where information loss is minimum and this process continues until the number of records in a particular cluster is k to satisfy the k-anonymity requirement. Thus we pretend k-anonymity problem to be a clustering problem, referred to as a systematic clustering problem.

Definition 4.3.1 (Systematic clustering problem) The systematic clustering problem is to find a set of clusters from a given set of n records such that each cluster contains at least k ($k \leq n$) records (where the records are selected in a systematic way and are included in a cluster that causes the least information loss) and that the sum of all intra-cluster distances is minimized. More specifically, if η be a set of n records and k the specified anonymization parameter, the optimal solution of the systematic clustering problem is a set of clusters $\Im = \{\Omega_1, \Omega_2, ..., \Omega_p\}$ such that:

1. $\Omega_i \cap \Omega_j = \Phi$, for all $i \neq j = 1, 2, ..., p$,
2. $\cup_{i=1,...,p}^{p} = \eta$,
3. for all $\Omega_i \in \Im, \mid \Omega_i \mid \geq k$, and
4. the total information loss obtained by using equation (4.1) is minimized.

In Definition 4.3.1, a set of clusters are constructed in such a way that the clusters are mutually exclusive, the sum of records of all clusters is equal to the total number of records and the size of each cluster is at least k which satisfies the criteria of k-anonymization. The problem tries to minimize the sum of all intra-cluster distances, where an intra-cluster distance of a cluster is defined as the maximum distance between any two records in the cluster. In the following subsection we formally design a systematic clustering algorithm.

Table 4.3: Systematic clustering algorithm

Input: a set η of n records containing individuals concerning their privacy, where $\eta_1, \eta_2, ..., \eta_n \in \eta$; the value k for k-anonymity
Output: a partitioning $\Im = \{\Omega_1, \Omega_2, ..., \Omega_p\}$ of τ
1. Sort all records in η by their quasi-identifiers;
2. Let p:= int$\lfloor \frac{n}{k} \rfloor$;
3. Get randomly k distinct records $r_1, r_2, ..., r_k$ from first 1 to k;
4. Let p_{ij} is the jth element in the ith cluster;
5. For $i = 1$ to p;
6. Let $p_{i1} := \eta_{[r_1+k(i-1)]}$;
7. Next i;
8. For $j := 2$ to k;
9. For $i := 1$ to p;
10. Let IL_i: = InfoLoss($\eta_{[r_j+k(i-1)]}$);
11. Let X:= Find cluster number with lowest IL_i;
12. where cluster size $\leq k$;
13. Add $\eta_{[r_j+p(i-1)]}$ to p_x;
14. Next i;
15. Next j;
16. Let $e := (n - pk)$;
17. Find extra element $E_1, E_2, ..., E_e \in E$;
18. For $k := 1$ to e;
19. For $m := 1$ to p;
20. Let IL_m := InfoLoss(E_k) in cluster m;
21. Next m;
22. Let X := Find cluster number with lowest IL;
23. Add E_k to p_x;
24. Next k;

4.3.2 Systematic clustering algorithm

Based on the information loss in equation (4.1) and the definition of a systematic clustering problem, we are now ready to discuss a systematic clustering algorithm. The general idea of the algorithm is as follows.

Note that for collecting medical data from patients it may be expected that some patients are not concerned about the privacy of their medical records and the other attributes. We would like to explore this opportunity because unnecessary anonimization may produce more information loss. Let q be the probability that a particular patient is not concerned about the privacy of medical records. Then out of n patients we can expect that on average nq patients are not concerned about their privacy. According to this method, first exclude the records of individuals who are not concerned about their privacy. Then sort all records by their quasi-identifiers and identify the equivalence class and the number of clusters by, $p = \frac{(n-nq)}{k}$, where k is the anonymity parameter for k-anonymization and round this as integer. Randomly select a record r_i from first k records as seeds to form the first cluster. If there are p clusters to be formed then select the $(r_i + k)$th, $(r_i + 2k)$th,..., $\{r_i + (p-1)k\}$th records in a systematic way to form the 2nd, 3rd, ..., pth cluster respectively. Select another record $r_j (j \neq i)$ from the first k records and add this record to the cluster which causes least information loss. Similarly, in a systematic way select (r_j+k)th, (r_j+2k)th,..., $\{r_j+(p-1)k\}$th records and add these records to their respective clusters that cause least information loss. If any cluster size is exactly k, stop adding records to that cluster and continue the same process until all records of first k records are finished. If $(n - nq)$ is not exactly divisible by k and there are still some records left, add these records to their closest clusters that incur least information loss. Finally distribute the nq records to their closest clusters or these nq records constitute separate cluster/clusters depending on their size. Note that these nq records incur no information loss. Since only the first record randomly selects and the subsequent records from in a systematic way, it has less execution time. Again usually the first record of each cluster contains a non identical value, so this algorithm easily captures if there are any extreme values. Moreover, this algorithm is adding some records that contain no information loss, so it is a natural expectation that the total

information loss will be reduced. The systematic clustering algorithm is shown in Table 4.3. In the algorithm it is assumed that all n individuals are concerned about their privacy.

Definition 4.3.2 (Systematic clustering decision problem) In a given data set of n records, there is a clustering scheme $\Im = \{\Omega_1, \Omega_2, ..., \Omega_p\}$ such that

1. $\mid \Omega_i \mid \geq k, 1 < k \leq n$: the size of each cluster is greater than or equal to a positive integer k, and

2. $\sum_{i=1}^{p} IL(\Omega_i) < c, c > 0$: the total information loss of the clustering scheme is less than a positive integer c.

where each cluster $\Omega_i (i = 1, 2, ..., p)$ contains the records that are more similar to each other, such that they require minimum generalization and thus cause least information loss. In the following subsection we are going to discuss some properties of the proposed systematic clustering algorithm.

4.3.3 Properties of the proposed algorithm

As discussed before, the proposed algorithm is designed in such a way that it finds a solution of k-anonymization in a greedy manner. This algorithm stops adding records in a particular cluster if the cluster size is exactly k. Again it always keeps in mind to add records that incur less information loss. Moreover, the records are selected in a systematic way that make the algorithm faster. With respect to this, this algorithm has the following desirable properties.

Theorem 4.3.1 Let n be the total number of input records and k be the specified k anonymity parameter. The time complexity of the systematic clustering algorithm in the clustering stage is in $O(\frac{n^2}{k})$.

Proof After sorting the records with respect to the quasi-identifiers, the systematic clustering algorithm determines the numbers of clusters by $p = \frac{n}{k}$. Then it selects the records as seeds in a systematic way to form all p clusters simultaneously. Thus for each tuple in the dataset, the algorithm needs to assign it to one of the p clusters, which has a complexity of $O(p)$. As a result, the assignment of all tuples to the clusters has a time complexity of

$$\begin{aligned} T &= O(Number\,of\,tuples * Number\,of\,clusters) \\ &= O(n * p) = O(n * \frac{n}{k}) = O(\frac{n^2}{k}). \end{aligned} \tag{4.4}$$

Therefore, the total execution time is in $O(\frac{n^2}{k})$.

Theorem 4.3.2 Let n be the total number of input records and q be the probability that a particular individual does not bother about the disclosure. Then the systematic clustering algorithm in fact work out the information loss of $(n - nq)$ individuals instead of all n individuals.

Proof If q be the probability that a particular individual does not bother about the disclosure. Then out of n individuals, nq individuals are not bothered about the disclosure. Assume that these nq records are in one separate cluster that causes no information loss. Also let $IL(\eta)$ and $IL(\eta_{all})$ be the total information loss due to k-anonymization for a systematic clustering algorithm and any other clustering algorithm respectively. According to the systematic clustering algorithm, the total information loss will be:

$$\begin{aligned} IL(\eta) &= IL(n) \\ &= IL(nq) + IL(n - nq) \\ &= 0 + IL(n - nq) = IL(n - nq). \end{aligned} \tag{4.5}$$

Thus, the systematic clustering algorithm actually calculates the information loss of $(n - nq)$ records instead of calculating the information loss of all n records.

Theorem 4.3.3 Let n be the total number of input records and k be the anonymity parameter in k-anonymization. Then according to the systematic clustering algorithm, the cluster size of any cluster is at least k but no more than $(2k-1)$.

Proof Let n be the total number of input records. According to systematic clustering, first select the initial seeds of all clusters in a systematic way and subsequently select adding more records to the clusters such that the added records incur the least information loss. Again this algorithm stops adding records to a particular cluster if the number of records is exactly k. So in the worst case, if there are $(k-1)$ records left and if all these records are included in a cluster that already contains k records, then the total number of records in that cluster will be $(k+k-1) = (2k-1)$. Therefore the maximum size of a cluster will be $(2k-1)$.

The properties discussed above show the utility of the proposed clustering algorithm with respect to both information loss and execution time. However it is necessary to check the efficiency of the algorithm by doing an experiment. In the following section the experimental results of the proposed algorithm are discussed.

4.4 Experimental Results

The objective of our experiment is to investigate the recital of our approach in terms of data quality and computational efficiency. To accurately evaluate our approach, the performance of the proposed systematic clustering algorithm is compared in this section with the k-member algorithm [34]. Byun et al. [34] showed that k-member algorithm causes significantly less information loss than "Mondrian", proposed by LeFevre et al. [30]. As it already evaluated that the k-member algorithm outperforms "Mondrian", in this chapter we compare our proposed algorithm with the k-member algorithm. Both experiments are implemented with Excel VBA programming language and run on a 3.20 GHz Pentium

(R) D CPU processor machine with 2GB of RAM. The operating system on the machine was Microsoft Windows XP Professional Version 2002.

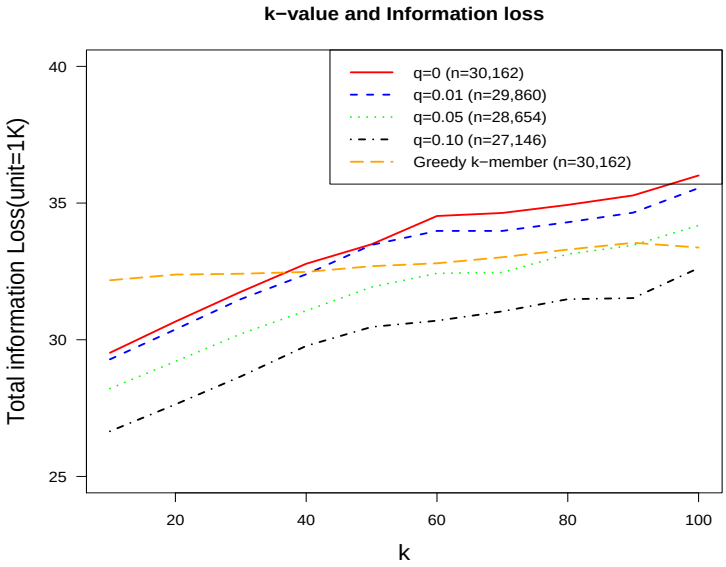


Figure 4.4: Information Loss

We utilized Adult dataset from the UC Irvine Machine Learning Repository [44] for both the experiments. It should be noted that the Adult dataset is considered as a standard benchmark for evaluating the performance of any k-anonymity algorithm. We deleted the records with missing values and retained only three of the original attributes, namely Age, Sex and Education as quasi-identifiers. Among these, Age is a numeric attribute but Sex and Education are the categorical attributes. The taxonomy trees for these two categorical attributes are based on Figure 4.2 and Figure 4.3 respectively.

The experiments are conducted as follows. First, the systematic clustering algorithm with three different scenarios ($q = 1\%$, $q = 5\%$, $q = 10\%$) and the k-member algorithm are run five times for every k value, and total information loss and execution time are collected for each run. Then, the average of every five runs using the same algorithm and k is computed and reported here.

Figure 4.4 shows the information loss of both the systematic clustering algorithm along with three levels ($q = 1\%$, $q = 5\%$, $q = 10\%$) and the k-member

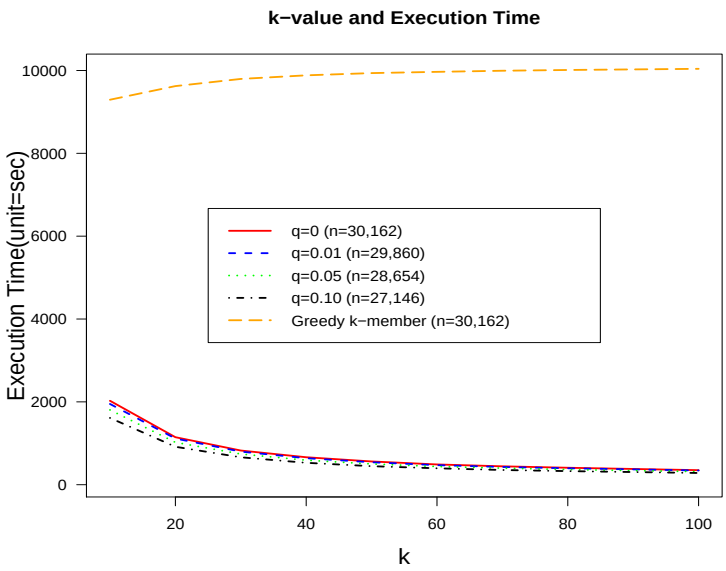


Figure 4.5: Execution Time

algorithm [34] with respect to the k-anonymity parameter. It shows that the total information loss of each of the algorithms increases as k increases. The logic behind this is that as k increases the clusters need to maximize generalization and this incurs more information loss. One of the most important criteria of choose a best clustering method is that it causes the least information loss. In this aspect, a systematic clustering method with $q = 10\%$ uniformly satisfies this criterion. That means that if at least 10% of individuals do not care about the disclosure then the systematic clustering method is the best choice as a clustering technique for k-anonymization in the data mining environment. However, as Figure 4.4 shows, for some moderate values of k ($k \leq 40$), the systematic clustering method always incurs less information loss even if all individuals are concerned about their privacy. In practice, the k-value of the k-anonymization problem should not be too small or too large as small values of k signify higher probability of disclosure and large values of k signify the more information loss. Thus in a realistic situation, $k \leq 40$ is reasonable for k-anonymity problems and in that case the proposed systematic clustering algorithm attains a reasonable dominance over the k-member algorithm.

On the other hand, Figure 4.5 displays the execution time of both algorithms. Figure 4.5 clearly shows that the execution time of the proposed systematic clustering algorithm with all different scenarios is much less than in the k-member algorithm. The greedy k- member algorithm takes too much time as it spends a lot of time selecting records from the input set. Again as expected, the execution time of the systematic clustering algorithm decreases with the increase of the probability that a particular individual does not care about the privacy as in this case the total number of records in the input set decreases. Thus it can be said that the proposed method is superior to the k-member algorithm in terms of both information loss and execution time.

As discussed before, a main challenge in data mining is to enable the legitimate usage and sharing of mined information while at the same time guaranteeing proper protection of the original sensitive data. Because of increasing concerns about the privacy of individuals, privacy preserving is an important issue and has captured the attention of many researchers in the data mining research community. Although k-anonymity is a proper solution of protecting sensitive attributes in a dataset, the main drawback of the method is the information loss. Thus, a natural question arises in this case: how to design a method in such a way that causes less information loss and execution time and at the same time satisfies the k-anonymity requirement. Based on this, an algorithm is developed in this chapter that uses the idea of clustering and incurs as little information loss as possible. As Figure 4.4 and Figure 4.5 show the proposed algorithm causes less information loss and execution time, and it demonstrates the flexibility and the usability of the proposed algorithm.

4.5 Anonymization for incremental Datasets

Anonymization based on k-anonymity models has been the focus of intensive research in the last few years. However the current techniques related to the

k-anonymity model are limited only where it is assumed that the entire dataset is available at the time of release (static data). This assumption leads to a shortcoming as data nowadays are continuously collected (thus continuously growing) and there is a strong demand for up-to-date data at all times [33]. In such a dynamic environment the proposed systematic clustering method can be easily applied without any modification of the previous anonymously released data.

Suppose that a hospital wants to publish its patient records for medical researchers and it is assumed that the hospital has already released the entire static data by using the systematic clustering method. The hospital can then infer the released data and this prior information can be used when a new record will be released. Assume that the hospital anonymized n records contain individuals concerning their privacy and build p clusters, where the information loss and the size of the ith cluster are respectively as f_i and N_i ($\sum_i^p N_i = n$). Thus, the total information loss is $f = \sum_i^p f_i$ and the proportion of information loss in the ith cluster is $P(f|i) = \frac{f_i}{n}$. Moreover, the probability that a new record will be included in the ith cluster is $P(i) = \frac{N_i}{n}$. Thus according to the Bays approach, the total probability of information loss is

$$P(f) = P(1)P(f|1) + P(2)P(f|2) + \cdots + P(p)P(f|p) = \sum_{i=1}^{p} \frac{N_i f_i}{nf} \tag{4.6}$$

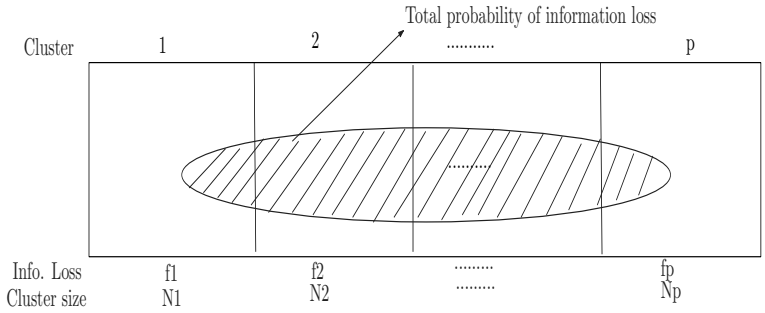


Figure 4.6: Bays Approach

The Bays approach is illustrated in Figure 4.6. Now suppose that the hospital wants to release a new record and assume that this record contains individuals' concerning their privacy (so needs anonymization and thus produce information loss), then the probability that this record will be included in the *ith* cluster is

$$P(i|f) = \frac{\frac{N_i f_i}{nf}}{\sum_{i=1}^{p} \frac{N_i f_i}{nf}} = \frac{N_i f_i}{\sum_{i=1}^{p} N_i f_i} \tag{4.7}$$

The higher probability indicates that the information loss will be higher if the new record is included in that particular cluster. So the new record should be included where the posterior probability is at a minimum. However, this is a preliminary idea of including a new record in a cluster. The easiest way to calculate the information loss of the new record with the existing clusters is to include the record with the cluster that causes the least information loss. As the preconstructed clusters based on the systematic clustering algorithm satisfy the k-anonymity requirement, the inclusion of the new record also respects the condition without any modification of preexisting clusters. Thus without any loss of generality the systematic clustering algorithm can be used for incremental datasets.

4.6 Systematic clustering for l-diversity

In Section 4.3, we have developed a systematic clustering method for k- anonymization. In this Section, we extend this approach to the l-diversity model that assumes that every group of indistinguishable records contains at least l distinct sensitive attributes values. The proposed technique adopts to group similar data together with l-diverse sensitive values and then anonymizes each group individually.

The proposed systematic clustering method outperforms the recent clustering based k-anonymization techniques. However the k-anonymity model may reveal sensitive information under two attacks, namely the homogeneity attack and the

Table 4.4: Patients records in a hospital

	ZipCode	Gender	Age	Education	Disease	Expense
1	4350	Male	24	9th	Flue	2000
2	4351	Male	25	10th	Cancer	3500
3	4352	Male	26	9th	HIV+	6500
4	4350	Male	35	9th	Diabetes	2000
5	4350	Female	40	10th	Diabetes	3200
6	4350	Female	38	11th	Diabetes	2800
7	4352	male	41	9th	Flue	2700
8	4352	Female	42	10th	Heart disease	4800
9	4352	male	43	10th	Cancer	5200

background knowledge attack [40]. For example, Jak and Ron are two antagonistic neighbors. Jak knows that Ron has been to hospital recently and tries to find out the disease that Ron suffers from. Jak finds the 3-anonymous table as in Table 4.5. He knows that Ron is 39 years old and lives in a suburb with postcode 4350. Ron must be record 4, 5 or 6. All three patients are suffering from diabetes. Jak knows for sure that Ron suffers from diabetes. Thus homogeneous values in the sensitive attribute of a k-anonymous group escape private information. Similarly k-anonymity does not protect individuals from a background knowledge attack. To overcome this problem, Machanavajjhala et al. [40] presented an l-diversity model to enhance the k-anonymity model. The l-diversity model assumes that a private dataset contains some sensitive attribute(s) which cannot be modified. Such a sensitive attribute is then considered disclosed when the association between a sensitive attribute value and a particular individual can be inferred with a significant probability. In order to prevent such inferences, the l-diversity model requires that every group of indistinguishable records contains at least l distinct sensitive attribute values; thereby the risk of attribute disclosure is kept under $\frac{1}{l}$. For example, records 4, 5 and 6 in Table 4.5 form a 3-diverse group. The records contain three values with equal frequencies of 33.33%, and no value is dominant. If we assume that $l = 2$, then although Table 4.5 is 3-anonymized but it is not a 2-diverse table as in the second equivalence class, the number of sensitive at-

Table 4.5: 3-Anonymization table

	ZipCode	Gender	Age	Education	Disease	Expense
1	435⋆	Person	[21-30]	Primary	Flue	2000
2	435⋆	Person	[21-30]	Primary	Cancer	3500
3	435⋆	Person	[21-30]	Primary	HIV+	6500
4	435⋆	Person	[31-40]	Secondary	Diabetes	2000
5	435⋆	Person	[31-40]	Secondary	Diabetes	3200
6	435⋆	Person	[31-40]	Secondary	Diabetes	2800
7	435⋆	Person	[41-50]	Primary	Flue	2700
8	435⋆	Person	[41-50]	Primary	Heart disease	4800
9	435⋆	Person	[41-50]	Primary	Cancer	5200

tribute value is only one (Diabetes). Thus it is necessary to devise new enhanced k-anonymization approaches (for example l-diversity) that best address both the quality and the privacy of the data. In Section 4.3, we developed a systematic clustering method for k-anonymization. However, as l-diversity is a more primitive and protected model than k-anonymization, it is necessary to extend the systematic clustering algorithm in the l-diversity model. This extension of the systematic clustering method to the l-diversity model is presented in this Section. It has been done in two steps. In the first step it develops some clusters that satisfy the k-anonymity requirements, called the clustering step for k-anonymization (same as described in Section 4.3). In the second step, it develops clusters that satisfy the l-diverse requirement on the sensitive attributes, called the l-diverse step. According to this step, first remove clusters in the first step that do not satisfy the l-diversity requirement. Then add the records contained in these clusters to other clusters that already satisfy the l-diversity requirement where they cause least information loss. Note that inclusion of new records to other clusters does not violate the l-diversity requirement. There are many clustering based k-anonymization techniques [34, 35, 36, 37, 45] that are available but to the best of our knowledge there is no such approach for the l-diversity model in the literature. Based on the leakages, this work is devoted to a systematic clustering method for the l-diversity model.

As discussed before, the problem of k-anonymization can be considered as a clustering problem, where each equivalent class is a cluster and the size of each cluster is at least k. However, the requirement for the l-diversity model is to satisfy at least l distinct sensitive attribute values in each equivalent class. Thus the optimal solution to a clustering problem is to construct a set of clusters that satisfy both k-anonymity and l-diversity requirements and the total information loss will be as minimal as possible. Now we formally define and present our systematic clustering algorithm that minimizes the information loss and respects the k-anonymity and l-diversity requirement.

4.6.1 Systematic clustering problem for l-diversity

In the k-anonymity problem the restriction is that the number of records in each equivalence class should be at least k and in the l-diversity model the restriction is that the number of sensitive attribute values in each equivalence class must be at least l distinct values, but there is no such restriction about the number of clusters in both cases. So a clustering problem is to form in such a way that each cluster contains at least k similar records, l distinct sensitive records and the sum of information losses of all clusters is as small as possible. To apply a systematic method to the l-diversity model of selecting records we need to follow two steps. The first one is the clustering step for k-anonymization and the second one is the l-diverse step. Suppose that we would like to apply a systematic clustering method to the l-diversity model for Table 4.4. Then in the clustering step for k-annualization first sort all records in the whole data set with respect to quasi-identifiers. There are 9 records in Table 4.4 and suppose that the dataset is already sorted according to the quasi identifier attributes ZipCode, Gender, Age and Education. If the anonymized table follows 3-anonymity requirements, then the number of clusters should be $\frac{9}{3} = 3$. Select a record (say, 2th record) from the first 3 records to form the first cluster. Then select $(2 + 3)th = 5$th and $(2 + 2 \times 3)th = 8$th records in a systematic way to form the second and third

cluster respectively. Now again select another record from the first 3 records (say, 3rd not 2th as it is already selected) and calculate the information loss with all of the three clusters using the equation (4.1). The information losses are respectively 5.10, 6.47 and 6.68, if this record is included in the first, second and third cluster. Thus, the 3rd record will be included in the first cluster as it causes least information loss. Similarly select $(3+3)th = 6$th and $(3+2\times 3)th = 9$th record in a systematic way and include them in the second and third cluster respectively. Finally select the 1st, $(1+3)th = 4$th and $(1+2\times 3)th = 7$th record and include these records in the first, second and third cluster respectively as they will then cause least information loss. If the total number of records is not exactly divisible by the k-anonymity parameter, then rest records will be included to similar clusters where information loss is at a minimum and this process continues until the number of records in a particular cluster is k to satisfy the k-anonymity requirement. Thus in the clustering step for k-anonymization, a set of clusters is built that satisfies the k-anonymity requirement. In the l-diverse step, the clusters will be formed in such a way that the number of distinct sensitive attribute values in each cluster is at least l. Note that if in the clustering step the table already satisfies the l-diversity requirement, the next step is not required. Suppose that $l = 3$, in this particular example, then the clusters that are obtained in the first step do not satisfy the l diversity requirement as the second cluster consists of only one distinct sensitive attribute value. Therefore, in the l-diverse step remove this cluster and the records contained in this cluster to other similar clusters that cause the least information loss. All of the three records in this cluster will be included in the third cluster as these records will then incur less information loss. Thus we get a table in Table 4.6 that satisfies both the 3-anonymity and 3-diversity requirements. The process of building the table by using a systematic method protects individuals' private information as well as sensitive attributes.

Table 4.6: 3-diversity table

	ZipCode	Gender	Age	Education	Disease	Expense
1	435⋆	Person	[21-30]	Primary	Flue	2000
2	435⋆	Person	[21-30]	Primary	Cancer	3500
3	435⋆	Person	[21-30]	Primary	HIV+	6500
4	435⋆	Person	[31-50]	Educated	Diabetes	2000
5	435⋆	Person	[31-50]	Educated	Diabetes	3200
6	435⋆	Person	[31-50]	Educated	Diabetes	2800
7	435⋆	Person	[31-50]	Educated	Flue	2700
8	435⋆	Person	[31-50]	Educated	Heart disease	4800
9	435⋆	Person	[31-50]	Educated	Cancer	5200

Definition 4.6.1 (Systematic clustering problem for l-diversity) The systematic clustering problem is to find a set of clusters from a given set of n records such that each cluster contains at least k ($k \leq n$) records (where the records are selected in a systematic way and included in a cluster that causes least information loss), the number of distinct sensitive attribute values is at least $l(l \geq 2)$ and that the sum of all intra-cluster distances is as minimal as possible. More specifically, if η be a set of n records and k & l are the specified anonymization and diversity parameter, the optimal solution of the systematic clustering problem is a set of clusters $\Im = \{\Omega_1, \Omega_2, ...\}$ such that:

1. $\Omega_i \cap \Omega_j = \Phi$, for all $i \neq j = 1, 2, ...$,
2. $\cup_{i=1,.....} = \eta$,
3. for all $\Omega_i \in \Im, \mid \Omega_i \mid \geq k$ & $l \geq 2$, and
4. the total information loss obtained by using equation (4.1) is minimized.

In Definition 4.6.1, a set of clusters is constructed in such a way that the clusters are mutually exclusive, the sum of records of all clusters is equal to the total number of records, the size of each cluster is at least k and the number of distinct sensitive attribute values is at least l to satisfy both the criteria of k-anonymization and l-diversity. The problem tries to minimize the sum of all

intra-cluster distances, where an intra-cluster distance of a cluster is defined as the maximum distance between any two records in the cluster. In the following subsection we formally design a systematic clustering algorithm for l-diversity.

4.6.2 Systematic clustering algorithm for l-diversity

Based on the information loss in Subsection (4.2.1) and the definition of the systematic clustering problem, we are now ready to discuss a systematic clustering algorithm for the l-diversity model. As discussed, the whole procedure consists of two steps, namely the clustering step for k-anonymization and the l-diverse step.

Clustering step for k-anonymization

The aim of this step is to develop a set of clusters from a given set of n records that satisfy the k-anonymity requirement. The algorithm of the clustering step for k-anonymization is the same as the algorithm for k-anonymity in Table 4.3.

l-diverse step

As discussed in the clustering step, we have some clusters that satisfy the k-anonymity requirement but may or may not satisfy the l-diversity requirement. Note that the l-diverse step is invoked only if in the first step, some of the clusters in the k-anonymization table do not satisfy the l-diversity requirement. If for a certain l-value, all clusters in the anonymized table satisfy the l-diversity requirement, the l-diverse step of the table is not required. According to this step, remove the clusters that do not satisfy l-diversity requirements and add the records contained in these clusters to other clusters that cause least information loss. As the existing clusters already satisfy the k-anonymity and the l-diversity requirement, inclusion of new records do not violate these requirement. The algorithm of the l-diverse step is illustrated in Table 4.7.

Definition 4.6.2 (Systematic clustering decision problem for l-diversity) In a

Table 4.7: l-diverse algorithm

Input: a partitioning $\Im_1 = \{\Omega_1, \Omega_2, ..., \Omega_{p_1}\}$ of τ that satisfy k-anonymity requirement, a partitioning $\Im_1^\star = \{\Omega_1^\star, \Omega_2^\star, ..., \Omega_{p_2}^\star\}$ of τ that satisfy both the k-anonymity and the l-diversity requirement, a set of sensitive attributes $S_i (i = 1, 2, 3, ...)$, and the value of l for l-diversity.

Output: a partitioning $\Im = \{\Omega_1, \Omega_2, ...\}$ of τ that satisfy the l-diversity requirement.
1. Let $\Upsilon = \{r_1, r_2, ...\}$={all records of $\Im_1$}
2. Let r_j is the jth record of Υ;
3. For $j = 1, 2, ...$;
4. For $i = 1, 2, ..., p_2$;
5. Let $\text{IL}_i^\star$:=InfoLoss($\Omega_i^\star$), $i = 1, 2, ...p_2^\star$;
6. Find the cluster $\Omega_i^\star$ with lowest $\text{IL}_i^\star$;
7. Add r_j to $\Omega_i^\star$;
8. Next j;

given data set of n records, there is a clustering scheme $\Im = \{\Omega_1, \Omega_2, ...\}$ such that

1. $\mid \Omega_i \mid \geq k, 1 < k \leq n$: the size of each cluster is greater than or equal to a positive integer k,

2. $l \geq 2$, the number of distinct sensitive attribute values in each cluster is at least 2, and

3. $\sum_{i=1} IL(\Omega_i) < c, c > 0$: the total information loss of the clustering scheme is less than a positive integer c.

where each cluster $\Omega_i (i = 1, 2, ...)$ contains the records that are more similar to each other with respect to k and l such that they require minimum generalization and thus causes least information loss.

4.7 Conclusion

In this chapter, we have proposed an efficient algorithm for k- anonymization to minimize information loss during the anonymization process and assure data quality. The proposed technique uses the idea of clustering and we refer to this

as the systematic clustering algorithm. The basic concepts of the proposed algorithm were discussed and investigated through example and properties. The time complexity of the developed algorithm is in $O(\frac{n^2}{k})$, where n is the total number of records containing individuals concerning their privacy. Finally a comparison was made on the proposed algorithm with the k- member algorithm proposed by Byun et al. [34] through experiment. For any k- anonymization algorithm, there are two significant criteria to judge the superiority of the algorithm, namely, information loss and execution time. The experimental results show that the proposed systematic clustering algorithm has a reasonable dominance over the k- member algorithm. This shows the utility and the efficiency of the proposed clustering algorithm. A way out was also shown to be used for continuously growing data without any violation of the k-anonymity requirement. Finally, we have proposed algorithms for the l- diversity model as an enhanced version of k-anonymity model. The proposed technique uses the idea of clustering and is implemented in two steps, namely the clustering step for k-anonymization and the l-diverse step.

Part III
Statistical Disclosure Control (SDC)

Chapter 5

Systematic Microaggregation for SDC

Microdata protection in statistical databases has recently become a major societal concern and has been intensively studied in recent years. Statistical Disclosure Control (SDC) is often applied to statistical databases before they are released for public use. Microaggregation for SDC is a family of methods to protect microdata from individual identification. SDC seeks to protect microdata in such a way that they can be published and mined without providing any private information that can be linked to specific individuals. Microaggregation works by partitioning the microdata into clusters of at least k records and then replacing the records in each cluster with the centroid of the cluster. This chapter presents a clustering based microaggregation method for SDC to minimize the information loss.

5.1 Introduction

In recent years, the phenomenal advance of technological developments in information technology enable government agencies and corporations to accumulate an enormous amount of personal data for analytical purposes. These agencies and organizations often need to release individual records (microdata) for research and other public benefit purposes. This propagation has to be in accordance with laws and regulations to avoid the propagation of confidential information. In other words, microdata should be published in such a way that it preserves

the privacy of the individuals. To protect personal data from individual identification, SDC is often applied before the data are released for analysis [2, 21]. The purpose of microdata SDC is to alter the original microdata in such a way that the statistical analysis from the original data and the modified data are similar and the disclosure risk of identification is low. As SDC requires suppressing or altering of the original data, the quality of data and the analysis results can be damaged. Hence, SDC methods must find a balance between data utility and personal confidentiality. Microaggregation is a family of SDC methods for protecting microdata sets that have been extensively studied recently [3, 4, 6, 9, 10, 115]. The basic idea of microaggregation is to partition a dataset into mutually exclusive groups of at least k records prior to publication, and then publish the centroid over each group instead of individual records. The resulting anonymized dataset satisfies k-anonymity [18], requiring each record in a dataset to be identical to at least $(k-1)$ other records in the same dataset. As releasing microdata about individuals poses a privacy threat due to the privacy-related attributes, called quasi-identifiers, both k-anonymity and microaggregation only consider the quasi-identifiers. Microaggregation is traditionally restricted to numeric attributes in order to calculate the centroid of records, but also has been extended to handle categorical and ordinal attributes [4, 6, 19]. In this chapter we propose a microaggregated method that is also applicable to numeric attributes.

The effectiveness of a microaggregation method is measured by calculating its information loss. A lower information loss implies that the anonoymized dataset is less distorted from the original dataset, and thus provides better data quality for analysis. k- anonymity [17, 18] provides sufficient protection of personal confidentiality of microdata, while to ensure the quality of the anonymized dataset, an effective microaggregation method should incur as little information loss as possible. In order to be useful in practice, the dataset should keep as much information as possible. Hence, it is necessary to seriously consider the tradeoff

between privacy and information loss. To minimize the information loss due to microaggregation, all records are partitioned into several groups such that each group contains at least k similar records and then the records in each group are replaced by their corresponding mean such that the values of each variable are the same. In the context of data mining, clustering is a useful technique that partitions records into groups such that records within a group are similar to each other, while records in different groups are most distinct from one another. Thus, microaggregation can be seen as a clustering problem with constraints on the size of the clusters. Many microaggregation methods derive from traditional clustering algorithms. For example, Domingo-Ferrer and Mateo-Sanz [3] proposed univariate and multivariate k-Ward algorithms that extend the agglomerative hierarchical clustering method of Ward et al. [20]. Domingo-Ferrer and Torra [114, 115] proposed a microaggregation method based on the fuzzy c-means algorithm [1], and Laszlo and Mukherjee [11] extended the standard minimum spanning tree partitioning algorithm for microaggregation [22]. All of these microaggregation methods build all clusters gradually but simultaneously. There are some other methods for microaggregation that have been proposed in the literature that build one cluster at a time. Notable examples include Maximum Distance [14], Diameter-based Fixed-Size microaggregation and centroid-based Fixed-size microaggregation [11], Maximum Distance to Average Vector (MDAV) [3, 6], MHM [7] and the Two Fixed Reference Points method [23]. Most recently, Lin et al. [24] proposed a density-based microaggregation method that forms records by the descending order of their densities, and then fine-tunes these clusters in reverse order.

All the works stated above proposed different microaggregation algorithms to form the clusters, where within clusters the records are homogeneous but between clusters the records are heterogeneous so that information loss is low. However, no single microaggregation method outperforms other methods in terms of infor-

mation loss. This work presents a new clustering method for microaggregation, where all clusters are made simultaneously in a systematic way. According to this method, sort all records by using a sorting function and partition all records into $[\frac{n}{k}]$ clusters, where n is the total number of records and k is the k-anonymity parameter. Randomly select a record r from the first k records to form the first cluster and the first records of the subsequent clusters form in a systematic way. Then adjust the records in each cluster in a systematic way such that each cluster contains at least k records. Performance of the proposed method is compared against the MDAV [3] as MDAV is the most widely used microaggregation method. The experimental results show that the proposed microaggregation method outperforms MDAV with respect to both information loss and computational efficiency.

5.2 Background

Microdata protection through microaggregation has been intensively studied in recent years. Many techniques and methods have been proposed to deal with this problem. In this section we describe some fundamental concepts of microaggregation. A microdata set $\boldsymbol{V}$ can be viewed as a file with n records, where each record contains p attributes on an individual respondent. The attributes in an original unprotected dataset can be classified in four categories which are not necessarily disjoint:

- **Identifiers**: These are attributes that unambiguously identify the respondent. Examples are passport number, social security number, and full name. Since our objective is to prevent confidential information from being linked to specific respondents, we will assume in what follows that in a pre-processing step, identifiers in $\boldsymbol{V}$ have been removed.

- **Quasi-identifiers**: A quasi-identifiers is a set of attributes in $\boldsymbol{V}$ that in combination can be linked with external information to re-identify (some

of) the respondents to whom (some of) the records in $\boldsymbol{V}$ refer. Unlike identifiers, quasi-identifiers cannot be removed from $\boldsymbol{V}$. The reason is that any attribute in $\boldsymbol{V}$ potentially belongs to a quasi-identifier depending on the external data sources available to the user of $\boldsymbol{V}$. As releasing microdata about individuals poses a privacy threat due to quasi-identifiers, microaggregation only considers the quasi-identifiers.

- **Confidential outcome attributes**: These are attributes which contain sensitive information about the respondent. Examples are salary, religion, political affiliation, and health condition.

- **Non-confidential outcome attributes**: Those attributes which contain non-sensitive information about the respondent. Examples are town and country of residence. Note that attributes of this kind cannot be neglected when protecting a dataset because they can be part of a quasi-identifier.

The purpose of microdata SDC can be stated more formally by saying that given an original microdataset $\boldsymbol{V}$, the goal is to release a protected microdataset $\boldsymbol{V}'$ in such a way that

1. Disclosure risk (i.e., the risk that a user or an intruder can use $\boldsymbol{V}'$ to determine confidential attributes on a specific individual among those in $\boldsymbol{V}$) is low.

2. User analysis (regressions, means, etc.) on $\boldsymbol{V}'$ and $\boldsymbol{V}$ yield the same or at least similar results. This is equivalent to requiring information loss caused by SDC to be low, i.e., that the utility of the SDC-protected data should stay high.

When we microaggregate data we should keep mind two goals: data utility and preserving privacy of individuals. For preserving the data utility we should introduce as little noise as possible into the data and preserving privacy data

should be sufficiently modified in such a way that it is difficult for an adversary to reidentify the corresponding individuals. Figure 5.1 shows an example of microaggregated data where the individuals in each cluster are replaced by the corresponding cluster mean. The figure shows that after aggregating the chosen elements, it is impossible to distinguish them, so that the probability of linking any respondent is inversely proportional to the number of aggregated elements.

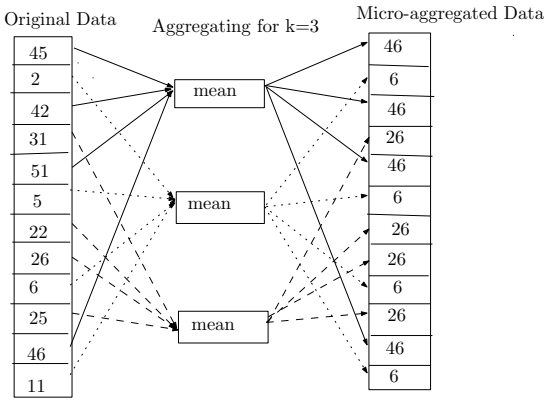


Figure 5.1: Example of Microaggregation using mean

Consider a microdata set T with p numeric attributes and n records, where each record is represented as a vector in a p-dimensional space. For a given positive integer $k \leq n$, a microaggregation method partitions T into g clusters where each cluster contains at least k records (to satisfy k-anonymity), and then replaces the records in each cluster with the centroid of the cluster. Let n_i denote the number of records in the ith cluster, and $x_{ij}, 1 \leq j \leq n_i$, denote the jth record in the ith cluster. Then, $n_i \geq k$ for $i = 1$ to g, and $\sum_{i=1}^{g} n_i = n$. The centroid of the ith cluster, denoted by $\bar{x}_i$, is calculated as the average vector of all the records in the ith cluster. In order to determine whether two records are similar, a similarity function such as the Euclidean distance, Minkowski distance or Chebyshev distance can be used. A common measure is the Sum of Squared Errors (SSE). The SSE is the sum of squared distances from the centroid of each cluster to every record in the cluster, and is defined as:

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i) \tag{5.1}$$

The lower the SSE, the higher the within cluster homogeneity and higher the SSE, the lower the within cluster homogeneity. If all the records in a cluster are the same, then the SSE is zero indicating no information is lost. On the other hand, if all the records in a cluster are more diverse, SSE is large indicating more information is lost. Thus SSE can be treated as a measurement of information loss due to microaggregation. In this chapter, we used SSE as a measure of information loss during the microaggregation process. Therefore, the microaggregation problem can be enumerated as a constraint optimization problem as follows:

Definition 5.2.1 (Microaggregation problem) Given a dataset T of n elements and a positive integer k, find a partitioning $\boldsymbol{G} = \{G_1, G_2, ..., G_g\}$ of T such that

1. $G_i \cap G_j = \Phi$, for all $i \neq j = 1, 2, ..., p$,
2. $\cup_{i=1}^{p} G_i = T$,
3. SSE is minimized,
4. for all $G_i \in T$, $\mid G_i \mid \geq k$ for any $G_i \in \boldsymbol{G}$.

The microaggregation problem stated above can be solved in polynomial time for a univariate dataset [10] but has been shown to be NP hard for multivariate dataset [13]. It is a natural expectation that SSE is low if the number of clusters is large. Thus the number of records in each cluster should be kept close to k. Domingo-Ferrer and Mateo-Sanz [3] showed that no cluster should contain more than $(2k - 1)$ records since such clusters can always be partitioned to further reduce information loss.

5.3 The Proposed Approach

This section presents the proposed systematic clustering-based algorithm for microaggregation that minimizes the information loss and satisfies the k-anonymity requirement. The proposed approach builds and refines all clusters simultaneously.

5.3.1 Sorting Function

According to the proposed approach, first sort all records with respect to the attributes, so it is necessary to define a sorting function to sort all the records in the dataset. Consider a microdata set T with p numeric attributes, namely $Y_1, Y_2, ..., Y_p$ and n records. Thus each record is represented as a vector in a p-dimensional space. To sort all the records with respect to the numeric attributes, we define the jth sorted record in the dataset T as follows:

$$SF_j = \sum_{i=1}^{p} (y_{ij} - \bar{y}_i), \quad j = 1, 2..., n. \tag{5.2}$$

where, y_{ij} is the jth record of the ith attribute and $\bar{y}_i$ is the centroid of the ith attribute. The sorting function (SF) stated above measures the distance between the records and their corresponding centroid. In this study, the SF is arranged in ascending order indicating records are arranged in order of magnitude. The lower the values of SF, the more the records are below their respective centroid and the higher the values of SF, the more the records are above their respective centroid. Thus the records in the dataset T, sorted in ascending order, based on the SF and the first and the last record, are most distant among all other records in the dataset T.

5.3.2 Systematic microaggregation algorithm

Based on the information loss measure in equation (5.1) and the definition of the microaggregation problem, we are now ready to discuss the systematic clustering-based microaggregation algorithm. The general idea of the algorithm is as follows.

Table 5.1: Systematic clustering-based microaggregation algorithm

Input: a dataset T of n records and a positive integer k
Output: a partitioning $\boldsymbol{G} = \{G_1, G_2, ..., G_g\}$ of T
where $g = |\boldsymbol{G}|$ and $G_i \geq k$ for $i = 1$ to g.
1. Sort all records in T in ascending order by using the SF
in equation (5.2);
2. Let g:= int$\lfloor \frac{n}{k} \rfloor$;
3. Get randomly k distinct records $r_1, r_2, ..., r_k$ from first 1 to k;
4. Let x_{ij} is the jth record in the ith cluster;
5. For $i = 1$ to g;
6. Let $x_{i1} := T_{[r_1+k(i-1)]}$;
7. Next i;
8. For $j := 2$ to k;
9. For $i := 1$ to g;
10. Let IL_i: = InfoLoss($T_{[r_j+k(i-1)]}$);
11. Let N:= Find cluster number with lowest IL_i;
12. where cluster size $\leq k$;
13. Add $T_{[r_j+k(i-1)]}$ to g_n;
14. Next i;
15. Next j;
16. Let $e := (n - gk)$;
17. Find extra element $E_1, E_2, ..., E_e \in E$;
18. For $k := 1$ to e;
19. For $m := 1$ to g;
20. Let IL_m := InfoLoss(E_k) in cluster m;
21. Next m;
22. Let N := Find cluster with lowest IL;
23. Add E_k to g_n;
24. Next k;

According to this method first sort all records in ascending order by using the sorting function in equation (5.2). Then identify the equivalence class and the number of clusters by, $g = \frac{n}{k}$, where n is the total number of records in the dataset T, and k is the anonymity parameter for k-anonymization. Round this as integer and randomly select a record r_i from the first k records as seed to form the first cluster. If there are g clusters to be formed then select the $(r_i + k)$th, $(r_i + 2k)$th,..., $\{r_i + (g - 1)k\}$th records in a systematic way to form the 2nd, 3rd, ..., gth cluster respectively. Select another record $r_j (j \neq i)$ from the first k records and add this record to the cluster which causes the least information loss. Similarly, in a systematic way, select $(r_j + k)$th, $(r_j + 2k)$th,..., $\{r_j + (g - 1)k\}$th records and add these records to their respective clusters that cause least infor-

mation loss. If any cluster size is exactly k, stop adding records to that cluster and continue the same process until all records of the first k records finish. If n is not exactly divisible by k and there are still some records left, add these records to their closest clusters that incur least information loss. A systematic microaggregation algorithm endeavors to build all clusters simultaneously, whereas most of the microaggregation algorithms in the literature build one/two cluster(s) at a time. The algorithm selects the first record randomly and the subsequent records form in a systematic way. As the records in the dataset T are arranged in ascending order and the first record of each cluster forms in every kth distance, the first record of each cluster contains a non identical value, so this algorithm easily captures if there are any extreme values in the dataset. The systematic microaggregation algorithm is shown in Table 5.1.

Definition 5.3.1 (Systematic clustering-based microaggregation decision problem) In a given dataset T of n records, there is a clustering scheme $\boldsymbol{G} = \{G_1, G_2, ..., G_g\}$ such that

1. $\mid G_i \mid \geq k, 1 < k \leq n$: the size of each cluster is greater than or equal to a positive integer k, and

2. $\sum_{i=1}^{g} IL(G_i) < c, c > 0$: the total information loss of the clustering scheme is less than a positive integer c.

where each cluster $G_i (i = 1, 2, ..., p)$ contains the records that are more similar to each other such that the cluster means are close to the values of the clusters and thus causes least information loss.

5.4 Experimental Results

The objective of our experiment is to investigate the recital of our approach in terms of data quality and the computational efficiency. This section experimen-

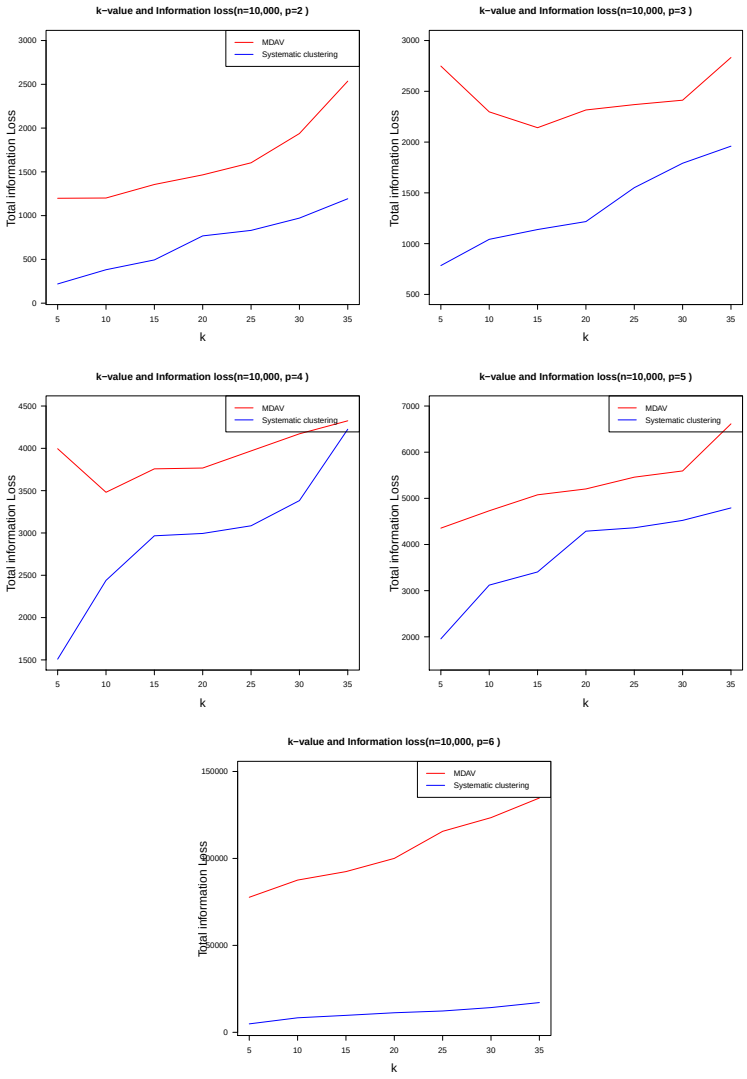


Figure 5.2: Information Loss comparison for no. of attributes between 2 and 6

tally evaluates the effectiveness and efficiency of the systematic clustering-based microaggregation algorithm. For this purpose, we utilize a real dataset CENSUS[1] containing personal information of 500 thousands American adults. The dataset has 9 discrete attributes.

To accurately evaluate our approach, the performance of the proposed algorithm is compared in this section with MDAV [3] as until now MDAV is the most widely used microaggregation method. For the experiment we have selected 10 thousands records randomly from the whole dataset and run the experiment for

[1]Downloadable at *http://www.ipums.org.*

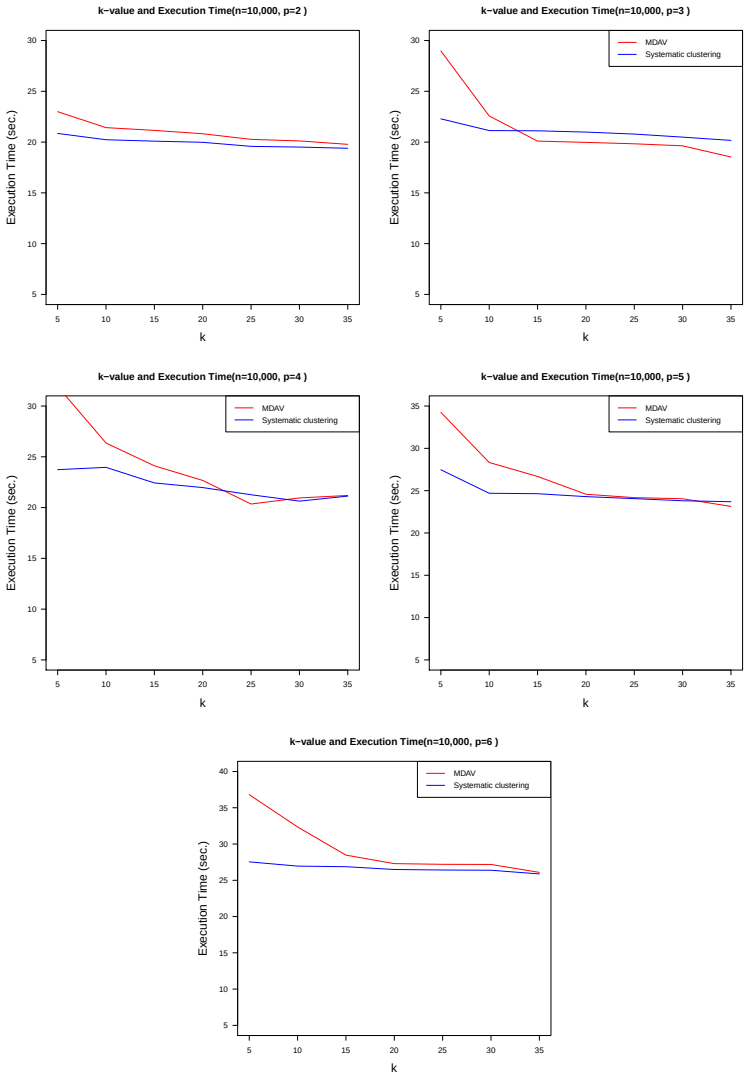


Figure 5.3: Running time comparison using census datset for no. of attributes between 2 and 6

$k = 5, 10, ..., 35$ and for different situations of number of attributes, $p = 2, 3, ..., 6$.

5.4.1 Data Quality and Efficiency

In this section, we report experimental results for the systematic clustering-based microaggregation algorithm for data quality and execution efficiency. In this chapter, SSE defined in equation (5.1) is used to measure the information loss due to microaggregation.

Figure 5.2 reports the information loss of both the MDAV and the systematic clustering-based microaggregation algorithms for increasing the values of k and

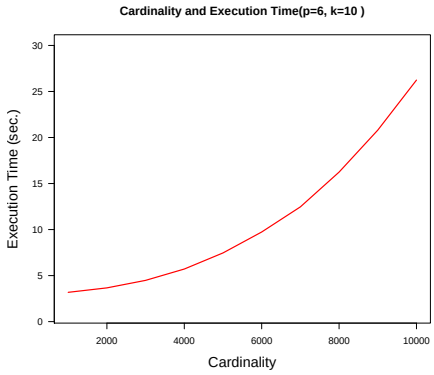


Figure 5.4: Cardinality and Runtime

p, where p is the number of attributes in the dataset. With the increase of k, the information loss is increasing for both the algorithms. As the figure illustrates, the systematic clustering-based microaggregation algorithm results in the lowest cost of the information loss for both all k and p values. The superiority of our algorithm over the MDAV algorithm results from the fact that our algorithm easily captures if there are any extreme values because of sorting function and the systematic way of selecting records in the clusters.

On the other hand, Figure 5.3 displays the execution (running) time of both the algorithms. In general the running time is decreasing with the increase of k in all scenarios. Figure 5.3 clearly shows that the running time of the proposed algorithm with all different scenarios is much less than the MDAV algorithm for almost all values of k. However, as shown in Figure 5.3, for some moderate values of k, the running time of the proposed algorithm is little bit more (in some situations) than the MDAV. We believe this is still acceptable in practice considering its better performance with respect to the information loss.

5.4.2 Scalability

Figure 5.4 shows the execution time behaviors of the systematic clustering-based microaggregation algorithm for various cardinalities with $p = 6$ and $k = 10$. For this experiment we used subsets of the Census dataset with different sizes. As shown, the running time increases almost linearly with the size of the dataset

for our proposed algorithm. Again the proposed algorithm introduces the least information loss for any p and k. This shows that our approach preserves the quality of the data and is highly scalable.

5.5 Conclusion

Microaggregation is an effective method of protecting privacy in microdata. This chapter has presented a new systematic clustering-based microaggregation method for numerical attributes. The new method consists of clustering individuals records in microdata in a number of disjointed clusters in a systematic way prior to publication and then publish the mean over each cluster instead of individual records. A comparison has been made between the proposed algorithm and the most widely used microaggregation method, called MDAV through an experiment. In the microaggregation problem, the performance of a method is judged by both information loss and running time. A method that incurs less information loss and has less execution time is a powerful method. The experimental results show that the proposed algorithm has a significant dominance over the MDAV method with respect to both information loss and execution time. Finally the results show through experiment that the proposed algorithm is highly scalable.

Chapter 6

A Pairwise-Systematic Microaggregation

In Chapter 5, we have developed a systematic clustering-based microaggregation method for SDC. The algorithm works well and has less execution time. However, the algorithm is sometimes affected by extreme values. If the dataset contains outliers, the systematic algorithm finally forces us to add those in the clusters whose size is less than k and that may cause more information loss. To overcome this problem, this chapter presents a pairwise systematic (P-S) microaggregation method to minimize the information loss. The proposed technique adopts a method that simultaneously forms two distant groups at a time with corresponding similar records together in a systematic way and then anonymizes with the centroid of each group individually.

6.1 Introduction

As discussed in the previous chapter, different microaggregation algorithms proposed in the literature form the clusters, where within clusters the records are homogeneous but between clusters the records are heterogeneous such that information loss is low. The level of privacy required is controlled by a security parameter k, the minimum number of records in a cluster. In essence, the parameter k specifies the maximum acceptable disclosure risk. Once a value for k has been selected by the data protector, the only job left is to maximize data util-

ity. Maximizing utility can be achieved by microaggregating optimally, i.e. with minimum within-groups variability loss. So the main challenge in microaggregation is to minimize the information loss during the clustering process. Although plenty of work has been done [3, 4, 6, 9, 10, 115], to maximize the data utility by forming the clusters, this is not yet sufficient in terms of information loss. So more research needs to be done to form the clusters such that the information loss is as low as possible. Observing this challenge, this chapter presents a new clustering-based method for microaggregation, where two distant clusters are made simultaneously in a systematic way. According to this method, sort all records in ascending order by using a sorting function so that the first record and the last record are most distant to each other. Form a cluster with the first record and its $(k-1)$ nearest records and another cluster with the last record and its $(k-1)$ nearest records. Sort the remaining records ($(n-2k)$, if dataset contains n records) by using the same sorting function and continue to build pair clusters at the same time by using the first and the last record as seeds until some specified records remain. Finally form one/two cluster(s) depending on the remaining records. Thus all clusters produced in this way contain k records except the last cluster that may contain at the most $(2k-1)$ records. Performance of the proposed method is compared against the most recent widely used microaggregation methods. The experimental results show that the proposed microaggregation method outperforms the recent methods in the literature in all test situations.

6.2 Previous Microaggregation Methods

Previous microaggregation methods have been roughly divided into two categories, namely fixed-size and data-oriented microaggregation [3, 7]. For fixed-size microaggregation, the partition is done by dividing a dataset into clusters that have size k, except perhaps one cluster which has a size between k and $(2k-1)$, depending on the total number of records n and the anonymity parameter k. For

the data-oriented microaggregation, the partition is done by allowing all clusters with sizes between k and $(2k-1)$. Intuitively, fixed-size methods reduce the search space, and thus are more computationally efficient than data-oriented methods [24]. However, data-oriented methods can adapt to different values of k and various data distributions and thus may achieve lower information loss than fixed-size methods.

Domingo-Ferrer and Mateo-Sanz [3] proposed a multivariate fixed-size microaggregation method, later called the Maximum Distance (MD) method [14]. The MD method repeatedly locates the two records that are most distant to each other, and forms two clusters with their respective $(k-1)$ nearest records until fewer than $2k$ records remain. If at least k records remain, it then forms a new cluster with all remaining records. Finally when there are fewer than k records not assigned to any cluster yet, this algorithm then individually assigns these records to their closest clusters. This method has a time complexity of $O(n^3)$ and works well for most datasets. Laszlo and Mukherjee [11] modified the last step of the MD method such that each remaining record is added to its own nearest cluster and proposed Diameter-based Fixed-size microaggregation. This method is however not a fixed size method because it allows more than one cluster to have more than k records.

The MDAV method is the most widely used microaggregation method [14]. MDAV is the same as MD except in the first step. MDAV finds the record r that is furthest from the current centroid of the dataset and the record s that is furthest from r instead of finding the two records that are most distant to each other, as is done in MD. Then form a cluster with r and its $(k-1)$ nearest records and form another cluster with s and its $(k-1)$ nearest records. For the remaining records, repeat this process until fewer than $2k$ records remain. If between k and $(2k-1)$ records remain, MDAV simply forms a new cluster with all of the remaining records. On the other hand, if the number of the remaining

records is below k, it adds all of the remaining records to their nearest clusters. So MDAV is a fixed size method. Lin et al. [24] proposed a modified MDAV, called MDAV-1. The MDAV-1 is similar to MDAV except when the number of the remaining records is between k and $(2k-1)$, a new cluster is formed with the record that is the furthest from the centroid of the remaining records, and its $(k-1)$ nearest records. Any remaining records are then added to their respective nearest clusters. Experimental results indicate that MDAV-1 incurs slightly less information loss than MDAV [24]. Another variant of the MDAV method, called MDAV-generic, is proposed by Domingo-Ferrer and Torra [6], where by the threshold $2k$ is altered to $3k$. If between $2k$ and $(3k-1)$ records remain, then find the record r that is furthest from the centroid of the remaining records and form a cluster with r and its $(k-1)$ nearest records and another cluster with the remaining records. Finally when fewer than $2k$ records remain, this algorithm then forms a new cluster with all the remaining records. Laszlo and Mukherjee [11] proposed another method, called Centroid-based Fixed-size microaggregation that is also based on a centroid but builds only one cluster during each iteration. This algorithm first find a record r that is furthest from the current centroid of the dataset and then finds a cluster with r and its $(k-1)$ nearest records. For the remaining records repeat the same process until fewer than k records remain. Finally add each remaining record to its nearest clusters. This method is not a fixed-size method as more than one cluster has more than k records.

Solanas et al. [16] proposed a variable-size variant of MDAV, called V-MDAV. V-MDAV first builds a new cluster of k records and then tries to extend this up to $(2k-1)$records based on some criteria. V-MDAV adopts a user-defined parameter to control the threshold of adding more records to a cluster. Chang et al. [23] proposed the Two Fixed Reference Points (TFRP) method to accelerate the clustering process of k-anonymization. During the first phase, TFRP selects

two extreme points calculated from the dataset. Let N_{min} and N_{max} be the minimum and maximum values over all attributes in the datasets respectively, then one reference point G_1 has N_{min} as its value for all attributes, and another reference point G_2 has N_{max} as its value for all attributes. A cluster of k records is then formed with the record r that is the furthest from G_1 and the $(k-1)$ nearest records to r. Similarly another cluster of k records is formed with the record s that is the furthest from G_2 and $(k-1)$ nearest records to s. These two steps are repeated until fewer than k records remain. Finally, these remaining records are assigned to their respective nearest clusters. This method is quite efficient as G_1 and G_2 are fixed throughout the iterations. When all clusters are generated, TFRP applies an enhancement step to determine whether a cluster should be retained or decomposed and added to other clusters.

Lin et al. [24] proposed a density-based algorithm (DBA) for microaggregation. The DBA has two different scenarios. The first state of DBA (DBA-1) repeatedly builds a new cluster using the k-neighborhood of the record with the highest k-density among all records that are not yet assigned to any cluster until fewer than k unassigned records remain. These remaining records are then assigned to their respective nearest clusters. The DBA-1 partitions the dataset into some clusters, where each cluster contains no fewer than k records. The second state of DBA (DBA-2) attempts to fine-tune all clusters by checking whether to decompose a cluster and merge its content with other clusters. Notably, all clusters are checked during the DBA-2 by the reverse of the order that they were added to clusters in the DBA-1. After several clusters are removed and their records are added to their nearest clusters in the DBA-2, some clusters may contain more than $(2k-1)$ records. At the end of the DBA-2, the MDAV-1 algorithm is applied to each cluster with size above $(2k-1)$ to reduce the information loss. This state is finally called MDAV-2. Experimental results show that the DBA attains a reasonable dominance over the latest microaggregation methods.

All of the microaggregation methods described above repeatedly choose one/two records according to various heuristics and form one/two cluster(s) with the chosen records and their respective $(k-1)$ other records. However there are other microaggregation methods that build all clusters simultaneously and work by initially forming multiple clusters of records in the form of trees, where each tree represent a cluster. Heuristics are then applied to either decompose a tree to reduce the cluster size to be fewer than $2k$ or merge trees to raise the cluster size to be greater than or equal to k. Instead of using trees, other methods may adaptively adjust the number of clusters to ensure that the size of each cluster is between k and $(2k-1)$.

The multivariate k-Ward algorithm [3] first finds the two records that are furthest from each other in the dataset and build two clusters from these two records and their respective $(k-1)$ nearest records. Each of the remaining record then forms its own cluster. These clusters are repeatedly merged until all clusters have at least k records. Finally the algorithm is recursively applied to each cluster containing $2k$ or more records. The k-Ward algorithm tends to generate large clusters, consequently increasing the information loss. For instance, this method could merge two clusters, each with $(k-1)$ records to form a large cluster of $(2k-2)$ records. The minimum spanning tree microaggregation method [11] first builds a minimal spanning tree (MST) of the dataset using the Prim method [25]. Then, as in the standard MST partitioning algorithm [22], the longest edge is recursively removed to form a forest of subtrees of the MST. However, unlike in the standard MST partitioning algorithm, the longest edge is removed only if both the resulting subtrees contain at least k nodes. Finally, another microaggregation method (such as MDAV) is applied to those groups containing more than $2k$ records. According to the experimental results reported by Laszlo and Mukherjee [11], this method has the same complexity as the multivariate k-Ward algorithm but causes less information loss. However, it still tends to generate large groups

and works well only if the dataset has well-separated clusters.

Domingo-Ferrer et al. [8] proposed a multivariate microaggregation method called μ-Approx. This method first builds a forest and then decomposes the trees in the forest such that all trees have sizes between k and $\max(2k-1, 3k-5)$. Finally, for any tree with a size greater than $(2k-1)$, find the node in the tree that is furthest from the centroid of the tree. Form a cluster with this node and its $(k-1)$ nearest records in the tree and form another cluster with the remaining records in the tree.

Hansen and Mukherjee [10] proposed a microaggregation method for univariate dataset called, HM. This method converts a dataset into a directed acyclic graph based on the ordering of the records and then transforms the microaggregation problem into the shortest path problem, which can be solved in polynomial time. This method cannot be applied directly to multivariate datasets since these only have a partial ordering among records. After that Domingo-Ferrer et al. [7] proposed a multivariate version of the HM method, called MHM. This method first uses various heuristics, such as nearest point next (NPN), maximum distance (MD) or MDAV to order the multivariate records. Steps similar to the HM method are then applied to generate clusters based on this ordering. Domingo-Ferrer et al. [115] proposed a microaggregation method based on a fuzzy c-means algorithm (FCM) [1]. This method repeatedly runs FCM to adjust the two parameters of FCM (one is the number of clusters c and another is the exponent for the partition matrix m) until each cluster contains at least k records. The value of c is initially large (and m is small) and is gradually reduced (increased) during the repeated FCM runs to reduce the size of each cluster. The same process is then recursively applied to those clusters with $2k$ or more records. Genetic algorithms (GAs) have also been applied to the microaggregation problem. Solanas et al. [15] encoded a partitioning of a dataset as a chromosome of n genes, where n is the number of records in the dataset and the value of the ith gene indicates

the cluster number of the ith record in the dataset. Since each cluster contains at least k records, each cluster number is an integer in the interval $[1, \lfloor \frac{n}{k} \rfloor]$. When generating the initial population of chromosomes and performing genetic operations on these chromosomes, special care must be taken to avoid generating a chromosome where any cluster numbers appear fewer than k or more than $2k$ times in their n genes. The experimental results showed that this method works well for small datasets ($n \leq 50$). Therefore they recommended first using a fixed-size microaggregation method such as MDAV to generate clusters with $k = 50$ and then applying GA for the real intended k value for each cluster. This two-step method was later studied by Martnez-Ballest et al. [12] and was also published in Solanas [14].

6.3 Information Loss

The notion of information loss is used to quantify the amount of information that is lost due to microaggregation. This section describes the measurement of information loss used to test the effectiveness of the P-S microaggregation method proposed in this chapter. Consider a microdata set T with p numeric attributes and n records, where each record is represented as a vector in a p-dimensional space. For a given positive integer $k \leq n$, a microaggregation method partitions T into g clusters, where each cluster contains at least k records (to satisfy k-anonymity), and then replaces the records in each cluster with the centroid of the cluster. Let n_i denote the number of records in the ith cluster, and $x_{ij}, 1 \leq j \leq n_i$, denote the jth record in the ith cluster. Then, $n_i \geq k$ for $i = 1$ to g, and $\sum_{i=1}^{g} n_i = n$. The centroid of the ith cluster, denoted by $\bar{x_i}$ is calculated as the average vector of all the records in the ith cluster.

In the same way, the centroid of T, denoted by $\bar{x}$, is the average vector of all the records in T. Information loss is used to quantify the amount of information of a dataset that is lost after applying a microaggregation method. In this chapter

we use the most common definition of information loss by Domingo-Ferrer and Mateo-Sanz [3] as follows:

$$IL = \frac{SSE}{SST} \tag{6.1}$$

where SSE is the within-cluster squared error, calculated by summing the Euclidean distance of each record x_{ij} to the average value $\bar{x}_i$ as follows:

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i) \tag{6.2}$$

and SST is the sum of squared error within the entire dataset T, calculated by summing the Euclidean distance of each record x_{ij} to the average value $\bar{x}$ as follows:

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})'(x_{ij} - \bar{x}) \tag{6.3}$$

For a given dataset T, SST is fixed regardless of how T is partitioned. On the other hand, SSE varies per dataset depending on the partition of the dataset. In essence, SSE measures the similarity of the records in a cluster. The lower the SSE, the higher the within cluster homogeneity and the higher the SSE, the lower the within cluster homogeneity. If all the records in a cluster are the same, then the SSE is zero indicating no information is lost. On the other hand, if all the records in a cluster are more diverse, SSE is large indicating more information is lost. In this chapter, we used SSE as a measure of similarity indicating a record will be included in a particular cluster if it causes least SSE among all other records in the dataset.

6.4 Pairwise-Systematic microaggregation algorithm

This section presents the proposed pairwise systematic algorithm for microaggregation that minimizes the information loss and satisfies the k-anonymity requirement. The proposed approach builds on and refines simultaneously two distant

Input: a dataset T of n records and a positive integer k.

Output: a partitioning $G = \{G_1, G_2, ..., G_g\}$ of T, where $g = |G|$ and $G_i \geq k$ for $i = 1$ to g.

1. Let $G = \Phi$, and $T' = T$;
2. Sort all records in T' in ascending order by using the SF in equation (5.2);
3. Find the first $f \in T'$ and the last record $l \in T'$;
4. Form a cluster G_1 containing f and its $(k-1)$ nearest records in T'; and another cluster G_2 containing l and its $(k-1)$ nearest records in T';
5. Set $G = G \cup \{G_1\} \cup \{G_2\}$ and $T' = T' - G_1 - G_2$;
6. Repeat steps 2-4 until $|T'| < 3k$;
7. If $2k \leq |T'| \leq (3k-1)$;
   (i) Go to step 2;
   (ii) Form a cluster containing the first record $f \in T'$; and its $(k-1)$ nearest records in T';
   (iii) Form another cluster with the remaining records in T';
8. else;
9. If $|T'| < 2k$;
   (i) Form a new cluster with all remaining records in T'.

Figure 6.1: P-S microaggregation algorithm

clusters at a time with the corresponding similar records together in a systematic way.

Based on the information loss measure in equation (6.1) and the definition of the microaggregation problem, we are now ready to discuss the Pairwise-Systematic (P-S) microaggregation algorithm.

According to this method, first sort all records of n in the dataset T in ascending order by using the SF in equation (5.2). Thus in the sorting dataset, the first record and the last record are the most distant to each other among all other pair records in the dataset T. The algorithm (see Fig. 6.1) repeatedly builds pair clusters using the first record and the last record in the sorting dataset and their corresponding $(k-1)$ nearest records until fewer than $3k$ records remain (see steps 2-6 of Fig. 6.1). The nearest records in a cluster are chosen in such a way that the inclusion of these records causes less SSE than the other records in the dataset. If between $2k$ and $(3k-1)$ records remain, then sort these records in ascending order by using the same sorting function in equation (5.2) and find

the first record f. Form a cluster with f and its $(k-1)$ nearest records, and another cluster with the remaining records (see step 7 of Fig. 6.1). Moreover, if fewer than $2k$ records remain, then form a new cluster with all remaining records (see step 9 of Fig. 6.1).

The P-S microaggregation algorithm stated above endeavor to repeatedly build two clusters simmultaneously in a systematic way. As the records in the dataset T are arranged in ascending order and the first record and the last record are most distant to each other, building clusters in this systematic way, the algorithm easily captures if there are any extreme values in the dataset.

Definition 6.4.1 (Pair-wise systematic clustering-based microaggregation decision problem) In a given dataset T of n records, there is a clustering scheme $\boldsymbol{G} = \{G_1, G_2, ..., G_g\}$ such that

1. $\mid G_i \mid \geq k, 1 < k \leq n$: the size of each cluster is greater than or equal to a positive integer k, and

2. $\sum_{i=1}^{g} IL(G_i) < c, c > 0$: the total information loss of the clustering scheme is less than a positive integer c.

where each cluster $G_i (i = 1, 2, ..., p)$ contains the records that are more similar to each other such that the cluster means are close to the values of the clusters and thus cause the least information loss.

6.5 Experimental Results

The objective of our experiment is to investigate the recital of our approach in terms of data quality. This section experimentally evaluates the effectiveness of the P-S microaggregation algorithm. The following three datasets [7], which have been used as benchmarks in previous studies to evaluate various microaggrega-

Table 6.1: Information loss comparison using Tarragona dataset

Method	$k = 3$	$k = 4$	$k = 5$	$k = 10$
MDAV-MHM	16.9326		22.4617	33.1923
MD-MHM	16.9829		22.5269	33.1834
CBFS-MHM	16.9714		22.8227	33.2188
NPN-MHM	17.3949		27.0213	40.1831
M-d	16.6300	19.66	24.5000	38.5800
μ-Approx	17.10	20.51	26.04	38.80
TFRP-1	17.228	19.396	22.110	33.186
TFRP-2	16.881	19.181	21.847	33.088
MDAV-1	16.93258762	19.54578612	22.46128236	33.19235838
MDAV-2	16.38261429	19.01314997	22.07965363	33.17932950
DBA-1	20.69948803	23.82761456	26.00129826	35.39295837
DBA-2	16.15265063	22.67107728	25.45039236	34.80675148
P-S	**5.494040549**	**8.329209112**	**10.8749404**	**17.01194228**

Table 6.2: Information loss comparison using Census dataset

Method	$k = 3$	$k = 4$	$k = 5$	$k = 10$
MDAV-MHM	5.6523		9.0870	14.2239
MD-MHM	5.69724		8.98594	14.3965
CBFS-MHM	5.6734		8.8942	13.8925
NPN-MHM	6.3498		11.3443	18.7335
M-d	6.1100	8.24	10.3000	17.1700
μ-Approx	6.25	8.47	10.78	17.01
TFRP-1	5.931	7.880	9.357	14.442
TFRP-2	5.803	7.638	8.980	13.959
MDAV-1	5.692186279	7.494699833	9.088435498	14.15593043
MDAV-2	5.656049371	7.409645342	9.012389597	13.94411775
DBA-1	6.144855154	9.127883805	10.84218735	15.78549732
DBA-2	5.581605762	7.591307664	9.046162117	13.52140518
P-S	**1.782851535**	**2.54581108**	**2.698883298**	**4.967556756**

tion methods, were adopted in our experiments.

1. The "Tarragona" dataset contains 834 records with 13 numerical attributes.

2. The "Census" dataset contains 1,080 records with 13 numerical attributes.

3. The "EIA" dataset contains 4,092 records with 11 numeric attributes (plus two additional categorical attributes not used here).

To accurately evaluate our approach, the performance of the proposed P-S microaggregation algorithm is compared in this section with various microaggregation methods. Tables 6.1-6.3 show the information losses of these microaggregation methods. The lowest information loss for each dataset and each k

Table 6.3: Information loss comparison using EIA dataset

Method	$k = 3$	$k = 4$	$k = 5$	$k = 10$
MDAV-MHM	0.4081		1.2563	3.7725
MD-MHM	0.4422		1.2627	3.6374
NPN-MHM	0.5525		0.9602	2.3188
μ-Approx	0.43	0.59	0.83	2.26
TFRP-1	0.530	0.661	1.651	3.242
TFRP-2	0.428	0.599	0.910	2.590
MDAV-1	0.482938725	0.671345141	1.666657361	3.83966422
MDAV-2	0.411101515	0.587381756	0.946263963	3.16085577
DBA-1	1.090194828	0.84346907	1.895536919	4.265801303
DBA-2	0.421048322	0.559755523	0.81849828	2.080980825
P-S	**0.213174523**	**0.32351185**	**0.435562877**	**1.044292097**

value is shown in bold face. The information losses of methods DBA-1, DBA-2, MDAV-1 and MDAV-2 are quoted from [24]; the information losses of methods MDAV-MHM, MD-MHM, CBFS-MHM, NPN-MHM and M-d (for $k = 3, 5, 10$) are quoted from [7]; the information losses of methods μ-Approx and M-d (for $k = 4$) are quoted from [8], and the information losses of methods TFRP-1 and TFRP-2 are quoted from [23]. TFRP is a two-stage method and its two stages are denoted as TRFP-1 and TRFP-2 respectively. The TFRP-2 is similar to the DBA-2 but disallows merging a record to a group of size over $(4k - 1)$.

Tables 6.1-6.3 show the information loss for several values of k and for the Tarragona, Census and for the EIA datasets respectively. The information loss is compared with the P-S microaggregation algorithm among the latest microaggregation methods listed above. Information loss is measured as $\frac{\text{SSE}}{\text{SST}} \times 100$, where SST is the total sum of the squares of the dataset. Note that the within-groups sum of squares SSE is never greater than SST so that the reported information loss measure takes values in the range [0,100]. Tables 6.1-6.3 illustrate that in all of the test situations, the P-S microaggregation algorithm causes significantly less information loss than any of the microaggregation methods listed in the table. Essentially, the P-S microaggregation algorithm causes less than 50% information loss compared to any of the previous methods listed above and for any of the datasets. This shows the utility and the effectiveness of the proposed algorithm.

6.6 Conclusion

Microaggregation is an effective method in SDC of protecting privacy in microdata and has been extensively used world-wide. The level of privacy required is controlled by a parameter k, often called anonymity parameter for k-anonymization that is basically the minimum number of records in a cluster. Once the value of k has been chosen, the data protector and the data users are interested in minimizing the information loss. This chapter has presented a new Pairwise-Systematic (P-S) microaggregation method for numerical attributes. The new method consists of pairwise clustering individual records in microdata in a number of disjointed clusters in a systematic way using a sorting function prior publication and then publishing the mean over each cluster instead of individual records. A comparison has been made of the proposed algorithm with the most widely used microaggregation methods through experimenting with the three benchmark datasets (Tarragona, Census and the EIA). The experimental results show that the proposed algorithm has a significant dominance over the recent microaggregation methods with respect to information loss. Thus the proposed method is very effective in preserving the privacy of respondents' contributions to microdata sets and can be used as a microaggregation method in SDC.

Chapter 7

Median-based Microaggragation for SDC

In this chapter, we introduce a new microaggregation method, where the centroid is considered as median. The new method guarantees the microaggregated data and the original data to be similar by using a statistical test. Another contribution of this chapter is that we propose a distance metric, called absolute deviation from median (ADM) to evaluate the amount of mutual information among the records in microdata.

7.1 Motivation

As stated before, the rationale behind microaggregation is to divide the dataset into some groups, where each group contains at least k records. For each variable, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximum similarity. Once the procedure has been completed, the resulting dataset can be published. Now a natural question arise, which relates to what centroid value should be used instead of individual records in each group, as the common center values that describe a set of values are mean, median and mode. The simplest answer is to use that value which apparently guarantees that the modified data and the original data are similar by using a statistical test. Previously the mean was used as a centroid value but it does not guarantee the similar modified data

and the original data. In this chapter we used median as the center value as it shows through a sign test that the modification has no effect and produces similar modified and original data. There are also some advantages of using median as the center value. Firstly, median is the appropriate measure for skewed distribution. If the records in each group follow skewed distribution, median should be used as the measure of central tendency. Mean is the appropriate measure for symmetric distribution; however, for symmetric distribution mean and median are equal and thus there is no difference between using mean or median as the center value. Secondly, mean is affected by extreme values, which means if a group contains any extreme values the total information loss will be increased. However, median is not at all affected by extreme values and lastly it is computationally more convenient to use median to measure the distortion of the original data. The distortion is measured as the difference between the original values and the modified values, but the sum of these differences is zero if mean is used as the center value. Thus sum of squares of differences is normally used to measure the distortion, if the mean is used as a center value that is computationally difficult. However, these sum of differences is not zero if median is used as a center value and the sum of absolute differences can be used to measure the distortion that is computationally less difficult. Using median as the center value produces a similar original but not the same data set, so there is still a chance of loss of information. Thus the effectiveness of a microaggregation method is measured by calculating its information loss. A lower information loss implies that the anonymized dataset is less distorted than the original dataset, and thus provides better data quality for analysis. As median is used as a measure of location to represent each group, in this chapter we propose the sum of absolute deviations from median (ADM) to measure the information loss that is always less than the SSE. That means that using ADM as a measure of information loss always produces less information loss than the SSE. Thus the proposed median based

microaggregation method has the following features:

- It divides the whole microdata set into a number of mutually exclusive and exhaustive groups prior to publication and then publishes the median over each group instead of individual records.
- It guarantees that the modification has no effect and the modified data and the original data are similar by using a statistical test.
- As microaggregated data causes information loss, it uses the sum of absolute deviations from median (ADM) as a measure of distortion that is always less than the so called distortion measure sum of squares of errors (SSE).

7.2 The Proposed Approach

Microdata protection through microaggregation has been intensively studied in recent years. Many techniques and methods have been proposed to deal with this problem. In this section we first describe some basic concept of microaggregation and a proposed approach of microaggregation.

When we microaggregate data we should keep in mind two goals, data utility and preserving privacy of individuals. For preserving the data utility we should introduce as little noise as possible into the data and preserving privacy data should be sufficiently modified in such a way that it is difficult for an adversary to re-identify the corresponding individuals. Figure 7.1 and Figure 7.2 show examples of microaggregated data where in Figure 7.1, the centroid is replaced by a mean and in Figure 7.2, the centroid is replaced by a median. Both the figures show that after aggregating the chosen elements, it is impossible to distinguish them, so that the probability of linking any respondent is inversely proportional to the number of aggregated elements.

Now it is necessary to check which figure shows similar original data and microaggregated data by using a statistical test. The sign test can be used to

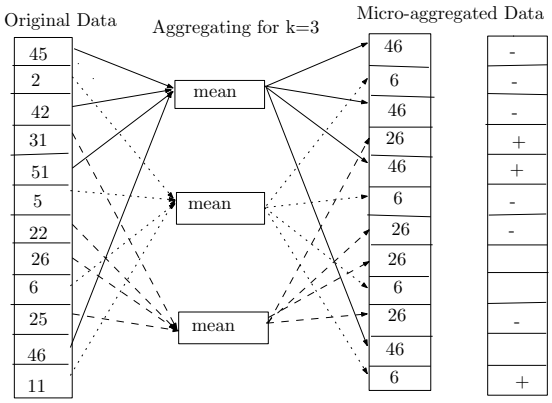


Figure 7.1: Example of Microaggregation using mean

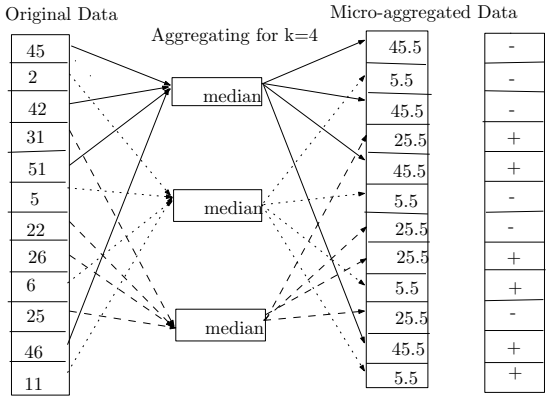


Figure 7.2: Example of Microaggregation using median

test the hypothesis that there is no difference between the distributions of original data and the microaggregated data. Both the figures consist of three groups and each group has four elements. The first group consists of the elements 45, 42, 51 and 46, the second group consists of the elements 2, 5, 6 and 11, and the third group consists of the elements 31, 22, 26 and 25, where in Figure 7.1 these values are replaced by their corresponding group mean and in Figure 7.2 these values are replaced by their corresponding group median. We would now like to test whether the original data and the micrioaggregated data are similar. Set up a null hypothesis H_0: the microaggregation method has no effect and the alternative hypothesis is H_a: the microaggregation method has an effect. Take the difference between microaggregated data from original data, give a "+" sign if the difference is positive and give a "−" sign if the difference is negative. We omit pairs for which there is no difference and count the number of positive

differences (X).

If we use median as centroid value then total pairs is, $n = 12$ (as no tie) and the number of positive sign is, $X = 6$. This is exactly what we would expect if there is no difference. Thus we can not reject H_0, as there is no evidence to support the hypothesis that the microaggregation method has an effect. This means the modification has no effect and both the microaggregated data and the modified data are similar. So, it can be concluded that using median as centroid value always gives a guarantee of producing similar original and modified data. This is true for any dataset as median is the middle most observations in a set of values.

By contrast, if we use mean as centroid value then total pairs is, $n = 12-3 = 9$ (as three tie) and the number of positive signs is, $X = 3$. This is not exactly what we would expect if there is no difference, which means we can not say anything unless we get p-value as the acceptance or rejection of H_0 depends on p-value. Thus, there is no grantee that the microaggregated data and the original data are similar. For some cases this may be true but it is not universally true for any particular dataset. Therefore, it can be concluded that using mean as centroid value does not give any guarantee of producing similar original and modified data for any dataset.

As discussed, the microaggregation method using median provides sufficient evidence that the modified data are similar to the original data, and in this chapter we propose to use median as the centroid point of each group. Thus before publishing, microdata should be partitioned into some groups so that within groups the records are closer to each other and between groups the records are more distant to each other, and then publish the median over each group instead of individual records.

7.3 Proposed distortion metric

Consider a microdata set T with p numeric attributes and n records, where each record is represented as a vector in a p-dimensional space. For a given positive integer $k \leq n$, a microaggregation method partitions T into g groups where each group contains at least k records (to satisfy k-anonymity), and then replaces the records in each group with the median of the group. Let n_i denote the number of records in the ith group, and $x_{ij}, 1 \leq j \leq n_i$, denote the jth record in the ith group. Then, $n_i \geq k$ for $i = 1$ to g, and $\sum_{i=1}^{g} n_i = n$. The centroid of the ith group, denoted by m_i, is calculated as the middle most (median) vector of all the records in the ith group. By using median, the microaggregated dataset produces similar data as the original dataset but not the same data and so there is still a chance of information loss. Information loss is used to quantify the amount of information of a dataset that is lost after applying a microaggregation method. To reduce the information loss it is necessary to form groups using a criterion of maximum similarity. This means the records in each group are closer to each other. To measure whether the records in each group are close to each other, in this chapter we use sum of absolute deviations from median (ADM) of each group and this is defined as

$$ADM = \sum_{i=1}^{g} \sum_{l=1}^{p} \sum_{j=1}^{n_i} \mid (x_{ilj} - m_{il}) \mid \tag{7.1}$$

where, x_{ilj} is the jth record of the lth attribute in the ith group and m_{il} is the median of the lth attribute in the ith group. As we replace each record by their corresponding group median, the distortion is measured by the difference between individual record and its median. The lower the distance, the closer the median is to its original value, and the higher the distance, the further the median is from its true value. We are only measuring the distance as it is of no interest to us whether the distance is positive or negative. Thus we take the absolute difference and the ADM is used to measure the information loss due to using the

median based microaggregation method. On the other hand, ADM could also be used to measure the homogeneity of the groups. The lower the ADM, the more homogeneous the records of the group are to each other.

Previously, the most common measure of information loss proposed by Domingo-Ferrer and Mateo-Sanz [3] was the Sum of Squares of Errors (SSE) and this is defined by

$$SSE = \sum_{i=1}^{g} \sum_{l=1}^{p} \sum_{j=1}^{n_i} (x_{ilj} - \bar{x}_{il})^2 \tag{7.2}$$

where p is the total number of numerical attributes in the dataset and $\bar{x}_{il}$ is the mean of the lth variable in the ith group. It should be noted that the sum of deviations from their mean of a set of observations is always zero, i.e. $\sum_{i=1}(x_i - \bar{x}) = 0$ and so the sum of squares of deviations from mean was used to measure the similarity of each group. As in this chapter we are taking the sum of deviations from median, i.e. $\sum_{i=1}(x_i - m)$, it always gives a value and so we do not need to square these deviations, but rather we take the absolute value of these deviations. Thus, given a homogeneity measure such as ADM and a security parameter k, which determines the minimum cardinality of the groups, the microaggregation problem can be enumerated as follows:

Definition 7.3.1 Given a dataset T of n elements and a positive integer k, find a partitioning $\boldsymbol{G} = \{G_1, G_2, ..., G_g\}$ of T such that ADM is a minimized subject to the constraint that $\mid G_i \mid \geq k$ for any $G_i \in \boldsymbol{G}$.

Once we get the homogeneous groups, the median value over each group is computed and replaces each of the original group values. Thus we get a microaggregated microdata set which could be published for general public use. This confirms that the microaggregated dataset is similar to the original data and preserves the privacy of individuals as well as increases the data utility.

7.4 Analysis of the Approach

As discussed, in this chapter we proposed a median based microaggregation method and proposed a distortion metric ADM to measure the homogeneity of the records in a group. In this section we will discuss some of the properties of the proposed approach and the metric.

Theorem 7.4.1 Suppose an attribute in a dataset consists of some groups and each group consists of records of at least k. Let the records of each group be replaced by the median of the corresponding group. Then the attribute consisting of the original records and the attribute consisting of the modified records (medians) have the same distribution.

group	1			2			...	g		
X	x_1	...	x_k	x_{k+1}	...	x_{2k}	...	$x_{(g-1)k+1}$	...	x_{gk}
M	m_1	...	m_1	m_2	...	m_2	...	m_g	...	m_g
sign	-	...	+	-	...	+	...	-	...	+

Figure 7.3: Values of a attribute

Proof Suppose an attribute in a dataset consists of n records that are exactly divisible by k. So the attribute consists of $g = \frac{n}{k}$ groups and each group consists of k records. Suppose the attribute consists of the values, $x_1, ..., x_k, x_{k+1}, ...x_{2k}$, $..., x_{(g-1)k+1}, ..., x_{gk}$, where the first group consists of first k-values, the second group consists of second k values, ..., and the last group consists of last k-values as shown in Figure 7.3. Also let $m_i (i = 1, ..., g)$ be the median of the ith group respectively, where m_i is the middle most observation of the ith group, when the values in ith group are arranged in order of magnitude. Thus the corresponding microaggregated values of the original values of the attribute are $m_1, ..., m_1, m_2, ..., m_2, ..., m_g, ..., m_g$, where first k-values consists of the first group, second k-values consists of the second group and so on, if median is replaced as the

centroid. Thus we get match pair data and let (X_i, M_i) be n pairs of observations. We wish to test,

$$H_0 : X \text{ and } M \text{ follow the same distribution,}$$

$$H_a : \text{The two distributions differ in location.}$$

Let $D_i = X_i - M_i$. Under H_0, both X and M come from the same distributions, so

$$P(D_i \text{ is positive}) = P(D_i \text{ is negative}) = \tfrac{1}{2}.$$

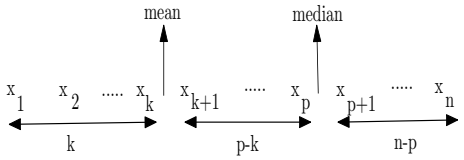
Let W be the total number of positive differences (D_i's). If X_i and M_i follow the same distribution then W follows Binomial distribution with parameters n and $\frac{1}{2}$. Suppose the values in each group are arranged in order of magnitude, thus for each group we get the first half as a positive sign and the other half as a negative sign. We omit pairs for which there is no difference, and this may be caused when k is an odd number. Thus finally the total number of positive sign is $\frac{n-g}{2}$, if n is odd and $\frac{n}{2}$, if n is even. That means, the number of positive signs and the number of negative signs would be the same whether k is even or odd. This is exactly what we would expect if there is no difference. Thus we can not reject H_0, which shows that the original values and the modified values of the attribute follow the same distribution and thus they are similar. Similarly this can be shown if n is not exactly divisible by k. This is true for each and every attribute in a microdata set. Thus if a microdata set is partitioned in to some groups and each record of a particular group is replaced by the corresponding median, then the microaggregated microdata set and the original dataset have the same distribution.

We will now show that the homogeneity measure ADM proposed in this chapter is always less than the so called homogeneity measure SSE. Before that we would like to discuss the following theorem.

Theorem 7.4.2 The sum of absolute deviations of a set of observations from their median is always less than the deviations from mean.

Proof Let $x_1, x_2, ..., x_n$ be a set of n observations. Let us assume that n is an even number and so $n = 2p$, where n is an integer. Thus median (m) lies between x_p to x_{p+1}. Also let $\bar{x}$ be the arithmetic mean which lies between x_k to x_{k+1}. Here we would like to show that

$$\sum_{i=1}^{n} \mid x_i - m \mid \leq \sum_{i=1}^{n} \mid x_i - \bar{x} \mid$$



Let us first take the absolute deviations from mean, say D_1

$$\begin{aligned} D_1 &= (\bar{x} - x_1) + (\bar{x} - x_2) + ... + (\bar{x} - x_k) \\ &= (x_{k+1} - \bar{x}) + (x_{k+2} - \bar{x}) + ... + (x_p - \bar{x}) \\ &= (x_{p+1} - \bar{x}) + (x_{p+2} - \bar{x}) + ... + (x_n - \bar{x}) \end{aligned} \tag{7.3}$$

and the absolute deviations from median, say D_2

$$\begin{aligned} D_2 &= (m - x_1) + (m - x_2) + ... + (m - x_k) \\ &= (m - x_{k+1}) + (m - x_{k+2}) + ... + (m - x_p) \\ &= (x_{p+1} - m) + (x_{p+2} - m) + ... + (x_n - m) \end{aligned} \tag{7.4}$$

Therefore,

$$\begin{aligned} D_1 - D_2 &= (\bar{x} - m)k - (\bar{x} + m)(p - k) - (\bar{x} - m)(n - p) \\ &+ 2(x_{k+1} + x_{k+2} + ... + x_p) \\ &= 2(x_{k+1} + x_{k+2} + ... + x_p - \bar{x}(p - k)) \\ &= 2[(x_{k+1} - \bar{x}) + (x_{k+2} - \bar{x}) + ... + (x_p - \bar{x}))] \end{aligned} \tag{7.5}$$

which is a positive quantity, so the sum of absolute deviations from median is always less than the deviations from mean. In other words,

$$\sum_{i=1}^{n} \mid x_i - m \mid \leq \sum_{i=1}^{n} \mid x_i - \bar{x} \mid.$$

Thus without any loss of generality, we can say that

$$\sum_{i=1}^{n} \mid x_i - m \mid \leq \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

This is true for every group in an attribute, for every attribute and for every dataset consisting of several numeric attributes. So,

$$\begin{aligned} \sum_{i=1}^{g} \sum_{l=1}^{p} \sum_{j=1}^{n_i} \mid (x_{ilj} - m_{il}) \mid &\leq \sum_{i=1}^{g} \sum_{l=1}^{p} \sum_{j=1}^{n_i} (x_{ilj} - \bar{x}_{li})^2 \\ \Longrightarrow ADM &\leq SSE \end{aligned} \tag{7.6}$$

Thus the proposed homogeneity measure in this chapter, ADM is always less than the SSE. In other words, ADM always incur less information loss than the SSE for any dataset.

7.5 Conclusion

This chapter has presented a new microaggregation method for numerical attributes. The new method consists of clustering individuals records in microdata in a number of disjointed groups prior to publication and then publishing the median over each group instead of individual records. We have shown by using a statistical test that the produced microaggregated data and the original data have the same distribution. As it produces a similar dataset, the statistical results also produce similar results as the original dataset. In addition, in this chapter we proposed a distortion metric to measure the homogeneity of the records in a group. The metric, called ADM can be used to measure the amount of information loss due to microaggregation. We have shown that ADM always produces less information loss than the previous information loss metric. This method of microaggregation can be extremely useful for researchers, experts and the associated people to analyse data accurately and efficiently as it protects the privacy of individuals as well as producing similar original data sets.

Chapter 8

Conclusions and future work

The privacy issue is not a new challenge in the field of information technology. Although much effort has been made in the past, it seems that many problems still remain open and they are getting more challenging. This is due to information technology becoming more intricate and directly involving many areas of our lives. In this thesis, we have discussed various issues of information privacy. We have focused on three broad areas, namely access control, data anonymization and statistical disclosure control.

With respect to data privacy, we first considered the problem of access control. Access control is used to control which parts of the data can be accessed by different users. Several models have been proposed for specifying and enforcing access control in databases [80]. The traditional access control models focus on which user is performing which action on which data object. But a reliable privacy policy is concerned with which data object is used for which purpose(s). Some organisations may have published privacy policies, which promise privacy protection practices on data collection, use and disclosure, but these practices may not be implemented. To maintain consistency between the privacy policy and the practices, privacy protection requirements in privacy policy should be formally specified. In specifying privacy policy, we use purpose as the basis of access control. In the first part of this thesis, we have presented a model for privacy preserving access control which is based on a variety of purposes. Con-

ditional purpose is applied along with allowed purpose and prohibited purpose in the model. This allows users using some data for certain purpose with conditions. We have called this the conditional purpose-based access control (CPBAC) model that enables enterprises to operate as reliable keepers of their customers data. Finally we injected the CPBAC model with the conventional well known role-based access control (RBAC). Based on RBAC, in this part we presented a role -involved purpose-based access control (RPAC) model, where users explicitly state their access purpose when they try to access data. We have also presented a conditional role-involved purpose-based access control (CPAC) model, where access purpose permission is assigned to conditional roles (CR). The CR is based on the notion of role attribute and system attribute. Users dynamically activate conditional roles in the CPAC model in accordance with the context attributes during the access purpose and after the purpose compliance process, so that only the users who are purpose compliant or conditionally compliant can be returned to the users.

Our propose model provides a comprehensive framework for the privacy preserving access control system, but much more work still remains to be done. Future work includes devising a high level language for a conditional purpose-oriented privacy policy which can be used to automatically manage the intended purposes of the data. Compatibility issues with P3P will also be investigated. We also plan to extend our model to cope with other elements of privacy such as obligations and complex conditions.

Role Engineering is a security-critical tasks for systems using role-based access control (RBAC). Different role-mining approaches have been proposed that attempt to automatically infer appropriate roles from existing user-permission assignments. Devising a complete and correct set of roles has been recognized as one of the most important an challenging tasks in implementing RBAC. In the CPAC model proposed in this thesis, we used the idea of CR which is based

on the role attribute and system attribute. Our future work is also to define the *conditional role mining problem* (CRMP) as the problem of discovering an optimal set of conditional roles from existing user permission.

Next, we considered the problem of data anonymization. Publishing data about individuals without revealing sensitive information is an important problem. The notion of privacy called k-anonymity has attracted a lot of research attention recently. In a k-anonymized database, values of quasi-identifying attributes are suppressed or generalized so that for each record there are at least $(k-1)$ records in the modified table that have exactly the same values for the quasi-identifiers. However, the performance of the best known approximation algorithms for k-anonymity depends on the information loss and the execution time. In the second part of this thesis, we introduced clustering as a technique to anonymize quasi identifiers before publishing them. We referred to this technique as a systematic clustering problem for k-anonymization. The proposed technique adopts group similar data together in a systematic way and then anonymises each group individually. Extensive experimental studies are conducted to show the efficiency and the effectiveness of the algorithm. The experimental results showed that the proposed systematic clustering algorithm has a reasonable dominance over the recent clustering algorithms for k-anonymization. At the end of the second part of this thesis, we extended the algorithm for the l-diversity model as an enhanced k-anonymity model. The proposed technique for the l-diversity model also uses the idea of clustering and was implemented in two steps, namely the clustering step for k-anonymization and the l-diverse step.

Recently many disparities of the k- anonymity model have been proposed in the literature to further protect the private information, e.g., t-closeness [46], and (α, k)-anonymity. Our further work will be to extend the systematic clustering algorithm to these models.

In the last part of this thesis, we addressed the problem of statistical disclosure

control (SDC)-revealing aggregate statistics about a population while preserving the privacy of individuals. In this thesis, we focused on microaggregation which is a family of SDC techniques for continuous microdata. Raw microdata (i.e., individual records) are grouped into small aggregates prior to publication. Each aggregate should contain at least k records to prevent disclosure of individual information. Fixed-size microaggregation consists of taking fixed-size microaggregates (size k). Data oriented microaggregation (with variable group size) was introduced recently. Regardless of the group size, microaggregation on a multidimensional dataset can be formed using univariate techniques on projected data or using multivariate techniques. We presented two heuristic algorithms of fixed size microaggregation to protect the privacy of individuals in microdata while allowing users to mine useful trends and patterns. The first heuristic consists of clustering individual records in microdata in a number of disjointed clusters in a systematic way prior publication and then publish the mean over each cluster instead of individual records. We refer to this heuristic as a systematic microaggregation for SDC. Experimental studies have been conducted and has shown that the proposed systematic clustering algorithm for microaggregation performs better in terms of both information loss and execution time over the most popularly used microaggregation method MDAV. It has also shown that the proposed algorithm is highly saleable. The second heuristic was adopted simultaneously to form two distant clusters at a time in a systematic way and then anonymized with the centroid of each cluster individually and we have referred to this as a Pairwise-Systematic (PS) microaggregation. Experimental studies on the proposed algorithm were conducted and compared with the most widely used microaggregation methods. To show the consistency of the algorithm, we used three benchmark datasets, namely Tarragona, Census and EIA. The experimental results show that the proposed algorithm has a significant dominance over the recent microaggregation methods with respect to information loss. Thus the proposed method is very

effective to preserve the privacy of respondents' contribution to microdata sets and can be used as a microaggregation method in SDC.

Finally, we introduced a new microaggregation method, where the centroid is considered as median. We showed by statistical test (sign test) that the microaggregated data and the original data have the same distribution, and thus the expected statistical results are similar to the original dataset. We also proposed a distortion metric to measure the homogeneity of the records in a group. The metric, called Absolute Deviation from Median (ADM), can be used to measure the amount of information loss due to microaggregation. As this method produces similar original datasets as well as protects the private information of individuals, it can be extremely useful for researchers, experts and associated people to analyse data accurately and efficiently.

The microaggregation problem can be regarded as a constraint single-objective problem, where the objective is to minimize the information loss, and the constraint is the k-anonymity requirement. Many variations of the k-anonymity model have recently been proposed to further protect data from identification, such as l-diversity [40], (α, k)-anonymity [47], m-confidentiality [118], (k, e)-anonymity [121] and (c, k)-safety [122]. An interesting direction for further investigation would be to formalize these models as a constraint multi-objective optimization problem, and develop new microaggregation methods based on it. The proposed microaggregation methods in this thesis would be a novelty in all of them.

Bibliography

[1] J.C. Bezdek. *Pattern recognition with fuzzy objective function algorithms.* Academic Publishers, 1981.

[2] J. Domingo-Ferrer, and V. Torra. Privacy in data mining. *Data Mining and Knowledge Discovery,* 11(2): 117–119, 2005.

[3] J. Domingo-Ferrer, and J. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering,* 14(1): 189-201, 2002.

[4] J. Domingo-Ferrer, and V. Torra. Extending microaggregation procedures using defuzzification methods for categorical variables. In *Proceedings of the 1st international IEEE symposium on intelligent systems,* 2002.

[5] J. Domingo-Ferrer, and V. Torra. Towards fuzzy *c*-means based microaggregation. In *Advances in Soft Computing: Soft Methods in Probability, Statistics and Data Analysis,* P. Grzegorzewski, O. Hryniewicz, and M. Gil, Eds. Heidelberg, Germany: Physica-Verlag, 2002, pp. 289-294.

[6] J. Domingo-Ferrer, and V. Torra. Ordinal, continuous and heterogeneous kanonymity through microaggregation. *Data Mining and Knowledge Discovery,* 11(2): 195-212, 2005.

[7] J. Domingo-Ferrer, A. Martnez-Ballest, J.M. Mateo-Sanz, and F. Sebe. Efficient multivariate data-oriented microaggregation. *The VLDB Journal,* 15(4): 355-369, 2006.

[8] J. Domingo-Ferrer, F. Sebe, A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Computer and Mathematics with Applications,* 55(4): 714-732, 2008.

[9] J.-M. Han, T.-T. Cen, H.-Q. Yu, and J. Yu. A multivariate immune clonal selection microaggregation algorithm. In *Proceedings of the IEEE international conference on granular computing,* pages 252-256, 2008.

[10] S. Hansen, and S. Mukherjee. A polynomial algorithm for optimal univariate microaggregation. *IEEE Transactions on Knowledge and Data Engineering,* 15(4): 1043-1044, 2003.

[11] M. Laszlo, and S. Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering,* 17(7): 902-911, 2005.

[12] A. Martinez-Balleste, A. Solanas, J. Domingo-Ferrer, and J.M. Mateo-Sanz. A genetic approach to multivariate microaggregation for database privacy. In *Proceedings of the IEEE 23rd international conference on data engineering workshop,* pages 180- 185, 2008.

[13] A. Oganian, and J. Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Commission for Europe,* 18: 345-354, 2001.

[14] A. Solanas. Privacy protection with genetic algorithms. In *Success in evolutionary computation: Studies in Computional Intelligence,* A. Yang, Y. Shan, and L.T. Bui, Eds. Heidelberg, Germany: Springer, 2008, vol. 92, pp. 215-237.

[15] A. Solanas, A. Martinez-Balleste, J. Mateo-Sanz, and J. Domingo-Ferrer. Multivariate microaggregation based genetic algorithms. In *Proceedings of the IEEE third international conference on intelligent systems,* pages 65-70, 2006.

[16] A. Solanas, A. Martinez-Balleste, and J. Domingo-Ferrer. V-MDAV: A multivariate microaggregation with variable group size. In *Proceedings of the 17th COMPSTAT Symposium of the IASC,* 2006.

[17] P. Samarati. Protecting respondent's privacy in microdata release. *IEEE Transactions on Knowledge and Data Engineering,* 13(6): 1010–1027, 2001.

[18] L. Sweeney. k-Anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10(5): 557-570, 2002.

[19] V. Torra. Microaggregation for categorical variables: A median based approach. In *Proceedings of the PSD, CASC Project International Workshop,* J. Domingo-Ferrer, and V. Torra, Eds. Heidelberg, Germany: Springer, 2004, vol. 3050, pp. 162-174.

[20] J.H.J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* 58(301): 236-244, 1963.

[21] L. Willenborg, and T.D. Waal. *Elements of statistical disclosure control.* Lecture notes in statistics, Vol. 155, Springer, 2001.

[22] C.T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers,* C-20(1): 68-86, 1971.

[23] C.-C. Chang, Y.-C. Li, and W.-H. Huang. TFRP: An efficient microaggregation algorithm for statistical disclosure control. *Journal of Systems and Software,* 80(11): 1866–1878, 2007.

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

[25] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to algorithms (2nd ed.).* The MIT Press, 2001.

[26] V. Ciriani, S.D.C. di Vimercati, S. Foresti, and P. Samarati. k-anonymous data mining: A aurvey. In *Privacy-preserving data mining: Models and algorithms,* C.C. Aggarwal, and P.S. Yu, Eds. Boston: Kluwer Academic Publishers, 2008, pp. 103–134.

[27] R.J. Bayardo, and R. Agrawal. Data privacy through optimal k-anonymization. In *Proceedings of the International Conference on Data Engineering,* pages 217–228, April 2005.

[28] B.C.M. Fung, K. Wang, and P.S. Yu. Top-down specialization for information and privacy preservation. In *Proceedings of the International Conference on Data Engineering,* pages 205–216, April 2005.

[29] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incogniti: Efficient full-domain k-anonymity. In *Proceedings of the ACM SIGMOD International Conference on Management of Data,* pages 49–60, June 2005.

[30] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the International Conference on Data Engineering,* pages 25, April 2006.

[31] L. Sweeney. Achieving k-anonymity privacy protection using generalization and supression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10(5): 571–588, 2002.

[32] J.W. Byun, and E. Bertino. Micro-views, or on how to protect privacy while enhancing data usability: concepts and challenges. *ACM SIGMOD Record*, 35(1): 9-13, 2006.

[33] J.W. Byun, Y. Sohn, E. Bertino, and N.Li. Secure anonymization for incremental datasets. In *Proceedings of the 3rd VLDB Workshop on Secure Data Management,* 2006.

[34] J.W. Byun, A. Kamra, E. Bertino, and N.Li. Efficient k-anonymization using clustering techniques. In *Proceedings of the International Conference on Database Systems for Advanced Applications,* pages 188–200, April 2007.

[35] G. Loukides, and J. Shao. Capturing data usefulness and privacy protection in k-anonymization. In *Proceedings of the Annual ACM Symposium on Applied Computing,* pages 370–374, March 2007.

[36] C.-C. Chiu, and C.-Y. Tsai. A k- anonymity clustering method for effective data privacy preservation. In *Proceedings of the International Conference on Advanced Data Mining and Applications,* pages 89–99, August 2007.

[37] J.L. Lin, and M.C. Wei. An efficient clustering method for k-anonymization. In *Proceedings of the International workshop on Privacy and anonymity in information society,* pages 46–50, March 2008.

[38] A. Meyerson, and R. Williams. On the complexity of optimal k-anonymity. In *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems,* pages 223–228, June 2004.

[39] J.Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.C. Fu. Utility-based anonymization using local recording. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pages 785–790, August 2006.

[40] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramanian. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the International Conference on Data Engineering,* pages 24–35, April 2006.

[41] T. Truta, and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In *Proceedings of the International Workshop on Privacy Data Management,* pages 94, September 2006.

[42] X. Sun, M. Li, H. Wang, and A. Plank. An efficient hash-based algorithm for minimal k-anonymity. In *Proceedings of the Australisian Computer Science Conference,* pages 101–107, January 2008.

[43] X. Sun, H. Wang, and J. Li. Priority Driven K-Anonymisation for Privacy Protection. In *Proceedings of the Australisian Data Mining Conference,* pages 73–78, November 2008.

[44] C.B.S. Hettich, and C. Merz. *UCI repository of machine learning databases,* 1998.

[45] T.Z. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science,* 38: 293–306, 1985.

[46] N. Li, and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the International Conference on Data Engineering,* pages 106–115, April 2007.

[47] R.C.-W. Wong, J. Li, A.W.-C. Fu, and K. Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pages 754–759, August 2006.

[48] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Hippocratic databases. In *Proeedings of the 28th International Conference on Very Large Databases,* pages 143-154, 2002.

[49] R. Agrawal, P. Bird, T. Grandison, J. Kiernan, S. Logan, and Y. Xu. Extending relational database systems to automatically enforce privacy policies. In

Proceedings of the 21st International Conference on Data Engineering, pages 1013-1022, 2005.

[50] S.S. Al-Fedaghi. Beyond Purpose-based privacy access control. In *Proceedings of the 18th Australian Database Conference,* pages 23-32, 2007.

[51] S. Barker, and P.N. Stuckey. Flexible access control policy specification with constraint logic programming. *ACM Transaction on Information and System Security,* 6(4): 501-546, 2003.

[52] E. Bertino, S. Jajodia, and P. Samarati. Data-base security: Research and practice. *Information systems,* 20(7): 537-556, 1995.

[53] J.W. Byun, E. Bertino, and N. Li. Purpose based access control of complex data for privacy protection. In *Proceedings of the 10th ACM Symposium on Access Control Model And Technologies,* pp. 102-110, 2005.

[54] J.W. Byun, E. Bertino, and N. Li. Purpose based access control for privacy protection in relational database systems. *VLDB Journal,* 17(4): 603-619, 2008

[55] D. Denning, T. Lunt, R. Schell, W. Shockley, and M. Heckman. The seaview security model. In *Proceedings of the IEEE Symposium on Research in Security and Privacy,* pages 218-233, 1988.

[56] United States Federal Trade Commission. Privacy online: fair information practices in the electronic marketplace, May 2000. Available at http://www.ftc.gov/reports/privacy2000/privacy2000.pdf

[57] S. Garfinkel. *Database Nation: The Death of Privacy in the 12st Century.* O'Reilly Media, Inc., 2000.

[58] A.I. Anton, Q. He, and D.L. Baumer. Inside JetBlues privacy policy violations. *IEEE Security and Privacy,* 2004.

[59] W. Chung, and J. Paynter. Privacy issues on the internet. In *Proceedings of the 35th Hawaii International Conference on System Sciences,* 2002.

[60] C. Potts. What is privacy? A presentation at the North Carolina State University E-Commerce Seminars, Oct 2001.

[61] A.F. Westin. *Privacy and Freedom.* Atheneum, New York, NY, 1967.

[62] Organization for Economic Co-operation and Development. OECD guidelines on the protection of privacy and transborder flows of personal data, 1980. Available at www1.oecd.org/publications/e-book/9302011E.PDF.

[63] United States Department of Health. Health insurance portability and accountability act of 1996. Available at http://www.hhs.gov/ocr/hipaa/.

[64] European Commission. Directive 95/46/ec on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Available at http://ec.europa.eu/justicehome/fsj/privacy/.

[65] S. Baker. Dont worry be happy. *Wired,* June 1994. Available at http://www.wired.com/wired/archive/2.06/nsa.clipper.html.

[66] Forrester Research. Privacy concerns cost e-commerce $15 billion, September 2001. Available at www.forrester.com.

[67] IBM. *The Enterprise Privacy Authorization Language (EPAL).* Available at http//www.zurich.ibm.com/security/enterprise-privacy/epal.

[68] M.E. Kabir, and H. Wang. Conditional Purpose Based Access Control Model for Privacy Protection. In *Proceedings of the 20th Australisian Database Conference,* pages 137-144, 2009.

[69] M.E. Kabir, H. Wang, and E. Bertino. Systematic Clustering Method for *l*-diversity Model. In *Proceedings of the 21th Australisian Database Conference,* pages 91-101, 2010.

[70] M.E. Kabir, and H. Wang. Microdata protection method through microaggragation: A median based approach. *Information Security Journal: A Global perspective,* 2010 (Accepted).

[71] M.E. Kabir, H. Wang, and E. Bertino. A Role-involved Conditional Purpose-based Access Control Model. *Proceedings of the IFIP EGES conference on E-Government and E-Services,* 2010.

[72] M.E. Kabir, and H. Wang. Systematic Clustering-based Microaggregation for Statistical Disclosure Control. In *Proceedings of the International Conference on Data and Knowledge Engineering,* 2010.

[73] K. LeFevre, R. Agrawal, V. Ercegovac, R. Ramakrishnan, Y. Xu, and D. DeWitt. Disclosure in Hippocratic databases. In *Proceedings of the 30th International Conference on Very Large Databases,* pages 108-119, 2004.

[74] M. Marchiori. The platform for privacy preferences 1.0 (P3P1.0) specification. *Technical report,* W3C, 2002.

[75] F. Massacci, J. Mylopoulos, and N. Zannone. Minimal Disclosure in Hierarchical Hippocratic Databases with Delegation. In *Proceedings of the 10th Europran Symposium on Research in Computer Security,* pages 438-454, 2005.

[76] OASIS: Core and hierarchical role based access control (rbac) profile of xacml v2.0. Available at http://www.oasis-open.org/.

[77] Oracle Corporation. The Virtual Private Database in Oracle9iR2. *An Oracle Technical White Paper, 2002.* Available at www.oracle.com.

[78] S. Rizvi, A.O. Mendelzon, S. Sudarshan, and P. Roy. Extending query rewriting techniques for fine-grained access control. In *Proceedings of the ACM SIGMOD Conference,* pages 551-562, 2004.

[79] C.S. Powers, P. Ashley, and M. Schunter. Privacy promises, access control, and privacy management. In *Proceedings of the 3rd International Symposium on Electronic Commerce,* pages 13-21, 2002.

[80] S. Castano, M. Fugini, G. Martella, and P. Samarati. *Principles of Distributed Database Systems.* AddisonWesley, 1995.

[81] D.F. Ferraiolo, D. R. Kuhn, and R. Chandramouli. *Role-Based Access Control.* Artech House, 2003.

[82] E. Bertino, E. Ferari, and A.C. Squicciarini. Trust negotation: Concepts, systems and languages. *IEEE Computing in Science and Engineering,* 6(4):27–34, 2004.

[83] X. Dong, A. Halevy, J. Madhavan, and E. Nemes. Reference reconciliation in complex information spaces. In *Proceedings of the ACM International Conference on Management of Data,* 2005.

[84] D.F. Ferraiolo, R.S. Sandhu, S. Gavrila, D.R. Kuhn, and R. Chandramouli. Proposed NIST standard for role-based access control. *ACM Transactions on Information and Systems Security,* 4(3):224–274, 2001.

[85] F. Chen and R. Sandhu. Constraints for role-based access control. In *Proceedings of the 1st ACM Workshop on Role-based access control,* 1996.

[86] C. Goh and A. Baldwin. Towards a more complete model of role. In *Proceedings of the 3rd ACM workshop on Role-based access control,* 1998.

[87] A. Kumar, N. Karnik, and G. Chafle. Context sensitivity in role-based access control. In *Proceedings of the ACM SIGOPS Operating Systems Review,* 2002.

[88] D. Lambert. Measures of disclosure risk and harm. *Journal of Official Statistics,* 9, 1993.

[89] C.K. Liew, U.J. Choi, and C.J. Liew. A data distortion by probability distribution. *ACM Transactions on Database Systems,* 10, 1985.

[90] S.P. Reiss. Practical data-swapping: The first steps. *ACM Transactions on Database Systems,* 9, 1980.

[91] J.F. Traub, Y. Yemini, and H. Wozniakowski. The statistical security of statistical database. *ACM Transactions on Database Systems,* 9, 1984.

[92] N. Adam, and J. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys,* 21, 1989.

[93] D. Dobkin, A.K. Jones, and R.J. Lipton. Secure databases: Protection against user influence. *ACM Transactions on Database systems,* 4, 1979.

[94] R. Sandhu, and S. Jajodia. Toward a multilevel secure relational data model. In *Proceedings of the ACM Transactional Conference on Management of Data,* pages 50-59, 1991.

[95] R. Sandhu, E.J. Coyne, H.L. Feinstein, and C.E. Youman. Role-based access control models. *IEEE Computer,* 29(2): 38-47, 1996.

[96] R. Sandhu, and F. Chen. The multilevel relational data model. *ACM Transaction on Information and System Security,* 1(1): 93-132, 1998.

[97] M. Stonebraker and E. Wong. Access control in a relational database management system by query modification. In *ACM CSC-ER Proceedings of the 1974 Annual Conference*, 1974.

[98] World Wide Web Consortium (W3C). Platform for Privacy Preferences (P3P). Available at http://www.w3.org/P3P.

[99] N. Yang, H. Barringer, and N. Zhang. A Purpose-Based Access Control Model. In *Proceedings of the 3rd International Symposium on Information Assurance and Security,* pages 143-148, 2007.

[100] H. Peng, J. Gu, and X. Ye. Dynamic Purpose-Based Access Control. In *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing with Applications,* pages 695-700, 2008.

[101] P.C.K. Hung. Towards a Privacy Access Control Model for e-Healthcare Services. In *Proceedings of the 3rd Annual Conference on Privacy, Security and Trust*, 2005.

[102] D.F. Ferraiolo, J.F. Barkley, and D.R. Kuhn. A Role-Based Access Control Model and Reference Implementation Within a Corporate Intranet. *ACM Transactions on Information and System Security,* 2(1), 34-64, 1999.

[103] A. Kobsa. Personalized hypermedia and interna tional privacy. *Communications of the ACM,* 2000.

[104] P. Ashley, C.S. Powers, M. Schunter. Privacy promises, access control, and privacy management. In *Proceedings of the 3rd International Symposium on Electronic Commerce,* 2002.

[105] V.S. Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining,* pages 279–288, 2002.

[106] A. Anton, Q. He, and D.L. Baumer. Inside JetBlue's privacy policy violations. In *Proceedings of the IEEE Security and Privacy*, 2004.

[107] W. Chung and J. Paynter. Privacy issues on the Internet. In *Proceedings of the 35th Hawaii International Conference on System Sciences*, 2002.

[108] L. Rnos. Toys "R" us sued for net privacy violations. *E-Commerce Times,* August 2000. Available at www.ecommercetimes.com/story/3957.html.

[109] United States Federal Trade Commission. Privacy initiatives: unfairness and deception. Available at www.ftc.gov/reports/privacy2000/privacy2000.pdf.

[110] T. Yu, D. Sivasubramanian, and T. Xie. Security Policy Testing via Automated Program Code Generation. In *Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies,* 2009.

[111] Q. Ni, A. Trombetta, E. Bertino, and J. Lobo. Privacy-aware role based access control. In *Proceedings of the 12th ACM symposium on Access control models and technologies,* pages 41-50, 2007.

[112] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proceedings of the 92th Symposium on Design and Analysis of Longitudinal Surveys,* pages 195–204, 1993.

[113] A. Hundepool, A. Van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand and S. Giessing. *μ-ARGUS version 4.0 Software and User's Manual.* Statistics Netherlands, Voorburg NL, May 2005. Available at http://neon.vb.cbs.nl/casc.

[114] J. Domingo-Ferrer, and V. Torra. Towards fuzzy c-means based microaggregation. In *Advances in Soft Computing: Soft Methods in Probability, Statistics and Data Analysis,* P. Grzegorzewski, O. Hryniewicz, and M. Gil, Eds. Heidelberg, Germany: Physica-Verlag, 2002, pp. 289-294.

[115] J. Domingo-Ferrer, and V. Torra. Fuzzy microaggregation for microdata protection. *Journal of Advanced Computational Intelligence and Intelligent Informatics,* 7(2): 153-159, 2003.

[116] J.M. Mateo-Sanz and J. Domingo-Ferrer. A method for data-oriented multivariate microaggregation. In *Statistical Data Protection,* J. Domingo-Ferrer, Eds. Luxemburg: Office for Official Publications of the European Communitites, 1999, pp. 89–99.

[117] G. Sande. Exact and approximate methods for data directed microaggregation in one or more dimensions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 10(5): 459–476, 2002.

[118] R.C. Wong, A.W.C. Fu, K. Wang, and J. Pei. Minimality Attack in Privacy Preserving Data Publishing. In *Proceedings of the 33rd International Conference on Very Large Data Bases,* pages 543–554, 2007.

[119] M. Templ. Statistical Disclosure Control for Microdata Using thr R-Package sdcMicro. *Transactions on Data Privacy,* 1(2): 67–85, 2008.

[120] K. Wang, P.S. Yu, and S. Chakraborty. Bottom-up Generalization: A Data Mining Solution to Privacy Protection. In *Proceedings of the fourth IEEE International Conference on Data Mining,* pages 249–256, 2004.

[121] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on aononymized tables. In *Proceedings of the 23rd International Conference on Data Engineering,* pages 116–125, 2007.

[122] D.J. Martin, D. Kifer, A. Machanavajjhala, and J. Gehrke. Worst-case background knowledge for privacy-preserving data publishing. In *Proceedings of the 23rd International Conference on Data Engineering,* pages 126–135, 2007.