UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

ON THE USE OF ECONOMIC PRICE THEORY TO DETERMINE THE OPTIMUM LEVELS OF PRIVACY AND INFORMATION UTILITY IN MICRODATA ANONYMISATION

by

Marek Piotr Zielinski

Submitted in fulfilment of the requirements for the degree
Philosophiae Doctor (Computer Science)

in the
Faculty of Engineering, Built Environment and Information Technology, University of Pretoria

June 2010

SUMMARY

On the use of Economic Price Theory to determine the optimum levels of privacy and information utility in microdata anonymisation

by

Marek Piotr Zielinski

Supervisor: Prof. M. S. Olivier

Department: Department of Computer Science

Degree: Philosophiae Doctor (Computer Science)

Abstract:

Statistical data, such as in the form of microdata, is used by different organisations as a basis for creating knowledge to assist in their planning and decision-making activities. However, before microdata can be made available for analysis, it needs to be anonymised in order to protect the privacy of the individuals whose data is released. The protection of privacy requires us to hide or obscure the released data. On the other hand, making data useful for its users implies that we should provide data that is accurate, complete and precise. Ideally, we should maximise both the level of privacy and the level of information utility of a released microdata set. However, as we increase the level of privacy, the level of information utility decreases. Without guidelines to guide the selection of the optimum levels of privacy and information utility, it is difficult to determine the optimum balance between the two goals.

The objective and constraints of this optimisation problem can be captured naturally with concepts from Economic Price Theory. In this thesis, we present an approach based on Economic Price Theory for guiding the process of microdata anonymisation such that optimum levels of privacy and information utility are achieved.

Key terms: information security; confidentiality; statistical databases; microdata; privacy; information utility; optimum balance; economic price theory; global recoding; microaggregation

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. Martin Olivier, for the guidance and advice he gave during the course of this study.

The work presented in this thesis was carried out while I was employed at SAP Research CEC Pretoria. I would like to thank colleagues from SAP Research for their support throughout this study. I would also like to thank Danie Kok, the Director of SAP Research CEC Pretoria, for giving me the opportunity to conduct this research work.

The support of SAP Research and the SAP Meraka Unit for Technology Development (UTD) towards this work is also hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and should not necessarily be attributed to SAP Research or the SAP Meraka Unit for Technology Development (UTD).

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

Many organisations, including government departments and businesses, collect and analyse data related to individuals. The data is then used to assist in the planning and decision-making activities of those organisations. For example, governments may collect data about individuals by means of a census. The data may then be analysed and released in the form of statistical data to assist in the assessment of population trends and to guide the development of government policies (Zielinski, 2006, 2007a).

However, when personal data is collected, analysed, or released in the form of statistical data, it is necessary to protect the privacy of the individuals whose data is used. This is necessary not only to ensure ethical conduct, but also to respect different privacy and data protection laws. This need is especially evident in environments where the data is of a highly sensitive nature, as it is in, for example, the medical environment (Gostin & Turek-Brezina, 1995), commerce (Rauhofer, 2008; Paul, 2001), or in the context of eParticipation (Zielinski, 2007a).

Statistical data can be disseminated in three different ways. These include dynamically queryable databases, tabular data, and microdata (Hundepool et al., 2007; Domingo-Ferrer, Sebe, & Solanas, 2008; Willenborg & De Waal, 2001). However, in this research work, we focus on microdata, since it used as the basis from which all other statistical data outputs are derived. A microdata set is the "raw data" itself; it is a set of records, where each record contains information on the entities represented in the database.

To protect the privacy of the respondents whose data is released, it is not sufficient to de-identify the microdata set by removing explicit identifiers (e.g. an ID number) from

the microdata set (Samarati, 2001; Skinner & Elliot, 2001). That is, a de-identified microdata set can still be manipulated and / or matched with external sources of data in an effort to re-identify individuals, or to disclose confidential data. Therefore, to protect privacy, a microdata set needs to be anonymised before it can be released.

1.2 Problem statement

To protect the privacy of the respondents in a microdata set, the microdata needs to be anonymised. As microdata is anonymised, data is removed (to some extent) from the identifying variables. As more data is removed from the identifying variables, it becomes increasingly difficult to infer sensitive data and to perform re-identification. Therefore, as microdata is anonymised, the level of privacy in the microdata increases. However, removing data from the identifying variables also reduces the accuracy and / or completeness of the released microdata. Therefore, as microdata is anonymised, its level of information utility also decreases. Consequently, as we increase the level of privacy in a microdata set, the level of information utility decreases, and vice versa.

Ideally, we would like to release microdata that has high levels of privacy and information utility. However, the protection of privacy implies that we should hide and obscure data. On the other hand, releasing usable and useful data implies that we should provide data that is accurate, complete and precise (Zielinski, 2007a, 2007b). Clearly, a conflict between the needs of privacy and information utility exists. This conflict needs to be resolved before a microdata set can be released.

Although a number of approaches have been proposed in the literature to address this conflict, we argue (in Chapter 5) that they are not completely appropriate for finding the optimum balance between privacy and information utility. Without guidelines to guide the selection of the optimum levels of privacy and information utility (taking into account the purpose for which the released data will be used), it is difficult to determine how to anonymise a microdata set such that it can be released with an optimum balance between the two conflicting goals. This difficulty may lead to cases where, in an effort to release a microdata set with a high level of information utility, the resulting level of privacy may be insufficient. Alternatively, a microdata set could be released with a level

of privacy that is far too high for a particular set of circumstances, leading to unnecessary loss in information utility. By guiding the selection of the optimum levels of privacy and information utility, we will be able to anonymise microdata without unnecessary loss in privacy or information utility, ensuring higher quality of the released microdata.

1.3 Research question, goal, and objectives of the study

This study will address the above problem by answering the following *research question*:

> *How can the process of microdata anonymisation be guided such that there will exist an optimum balance between privacy and information utility in the anonymised microdata?*

The above research question will be answered through the following two *research sub-questions*:

1. How should the optimum levels of privacy and information utility be determined?
2. How should a microdata set be anonymised such that the determined optimum levels of privacy and information utility are achieved?

Based on the above research question, *the goal of this study* is to propose a microdata anonymisation process that will anonymise microdata such that it will have an optimum balance between privacy and information utility.

Based on the above research sub-questions, *the objectives of this study* are:

1. To propose a process through which the optimum levels of privacy and information utility of a microdata set can be determined.
2. To propose a process that will anonymise a microdata set such it will possess the determined optimum levels of privacy and information utility.

1.4 Delimitation of scope

Statistical data, information security, privacy, and anonymisation techniques form the basis for the area in which our research problem exists. We will limit the scope of our study in these four aspects as follows.

Firstly, we will concentrate on only one type of statistical data, namely microdata. That is, we will not focus on other ways in which statistical data can be disseminated. Nevertheless, other types of statistical data will be briefly discussed in Chapter 2. We choose to focus on microdata because it forms the basis for deriving the other types of statistical data. In addition, we will focus on only the dissemination aspect of microdata production. Other steps of microdata production, such as collection and processing of the data, will not be addressed.

Secondly, since our focus is on the protection of privacy in microdata, our study is limited to only one aspect of information security, namely confidentiality. Other aspects of information security, such as integrity and availability (Pfleeger, 1997) will therefore not be considered.

Thirdly, we will restrict our focus to only one dimension of privacy, namely respondent privacy. This dimension of privacy aims to prevent the re-identification and disclosure of confidential data of the respondents whose records are released in a microdata set (Domingo-Ferrer & Saygin, 2009; Domingo-Ferrer, 2007). Other dimensions of privacy, such as owner privacy and user privacy, will not be part of the scope of this study, although they will be briefly discussed in Chapter 3.

Finally, we will only consider two microdata anonymisation techniques in this study, namely global recoding and microaggregation. These techniques are examples of non-perturbative and perturbative anonymisation techniques, respectively. We focus on these two techniques, since they are typically used to achieve *k*-anonymity (Samarati, 2001; Sweeney, 2002a, 2002b; Domingo-Ferrer & Torra, 2005). Combining local suppression with local or global recoding may also be used to achieve *k*-anonymity. However, we will not consider local suppression and local recoding, due to the

drawbacks in their use for *k*-anonymisation, which will be explained in Chapter 5. Other non-perturbative and perturbative techniques, as well as synthetic data generation techniques, are outside of the scope of this study.

1.5 Methodology

Our research question is answered through two research sub-questions and our goal is achieved through two objectives. In this Section, we discuss the methodology that will be applied to answer our research sub-questions and to achieve our objectives.

To answer our first research sub-question and to achieve our first objective, we aim to determine how best to allocate the available information in a microdata set between the released information (i.e. information utility) and the hidden information (i.e. privacy), so as to maximise the joint benefit of the data user and the data owner. This aim closely corresponds to the aim of the *utility maximisation problem of a consumer* in Economic Price Theory (or Microeconomics) (Besanko & Braeutigam, 2005; Dixit, 1990; J. Hirshleifer, Glazer, & D. Hirshleifer, 2005; Mansfield, 1985).

In Economic Price Theory, and particularly in the *utility maximisation problem of a consumer*, a consumer's "optimum" balance between the consumption of goods can be found when the consumer is constrained by prices of the goods as well as a budget available for purchasing the goods. Finding a solution to this problem is based on finding an optimum point for the consumer so as to maximise the (economic) utility, or satisfaction, that the consumer derives from consuming the goods. Therefore, to solve this problem, we need to determine how to allocate the consumer's income among different goods in order to maximise the consumer's utility (or satisfaction) gained from consuming the goods.

If we use these concepts from Economic Price Theory in our research work, we can consider our goods to be information utility and privacy, our budget to be the amount of data (in terms of information entropy) that is available in the non-anonymised microdata, and our economic utility to be the joint benefit of the data user and the data

owner. Hence, we will use Economic Price Theory as a basis for guiding the process of microdata anonymisation.

Therefore, to answer the first research sub-question and to achieve our first objective, we will apply an inter-disciplinary research approach. We will borrow concepts from Economic Price Theory and apply them in our discipline to propose an algorithm (in terms of high-level steps) that will determine the optimum levels of privacy and information utility of a microdata set. This algorithm will then be evaluated through a simulation. The evaluation will show how the input preferences, according to which the optimum levels are determined, impact the optimum levels of information utility and privacy.

To answer our second research sub-question and to achieve our second objective, we will propose two algorithms (in terms of high-level steps). These algorithms will determine how to anonymise a microdata set such that the optimum levels of information utility and privacy are achieved. The first algorithm will be for a non-perturbative anonymisation technique, namely global recoding. The second algorithm will be for microaggregation, which is a perturbative microdata anonymisation technique. Both algorithms will also be evaluated through a simulation, which will show how the optimum levels of information utility and privacy change the way in which a microdata set is anonymised.

1.6 Organisation of chapters

This thesis is organised as follows:

- Chapter 1 is a brief introduction to this research work.

- Chapter 2 contains a discussion of the main ways in which statistical data can be disseminated, with a specific focus on the dissemination of microdata.

- Chapter 3 discusses the concept of privacy. We focus specifically on the privacy of information, discuss why it needs to be protected, and show how privacy can be compromised in released microdata.

- Chapter 4 presents an overview of the techniques available for protecting privacy for the different ways in which statistical data can be disseminated. We place a greater emphasis on the techniques available for the protection of privacy in microdata.

- Chapter 5 investigates the conflict between privacy and information utility by discussing the appropriateness of existing approaches that address this conflict. We shall argue that existing approaches for addressing this conflict are not completely appropriate for finding the optimum balance between privacy and information utility. Consequently, we shall present recommendations for an appropriate solution.

- Chapter 6 uses the recommendations from Chapter 5 to propose a solution, in terms of a high-level algorithm, for determining the optimum levels of privacy and information utility in microdata anonymisation. The solution proposed in this Chapter will achieve the first objective stated in Section 1.3.

- Chapter 7 proposes a solution for determining how to anonymise microdata such that the optimum levels of privacy and information utility are achieved. The proposed solution will therefore achieve the second objective stated in Section 1.3. We consider two microdata anonymisation techniques for this purpose: global recoding and microaggregation.

- Chapter 8 contains concluding remarks as well as recommendations for future work.

CHAPTER 2

DISSEMINATION OF STATISTICAL DATA

2.1 Introduction

In this Chapter, we introduce the statistical data life cycle and discuss different ways in which statistical data can be disseminated. The main focus in this Chapter is on microdata, which is one form in which statistical data can be disseminated.

2.2 The statistical data life cycle

Statistical data is produced through a process known as the [statistical] data life cycle (Pongas & Vernadat, 2003). An example of a statistical data life cycle that is employed at statistical organisations, such as Eurostat, is shown in Figure 2.1.

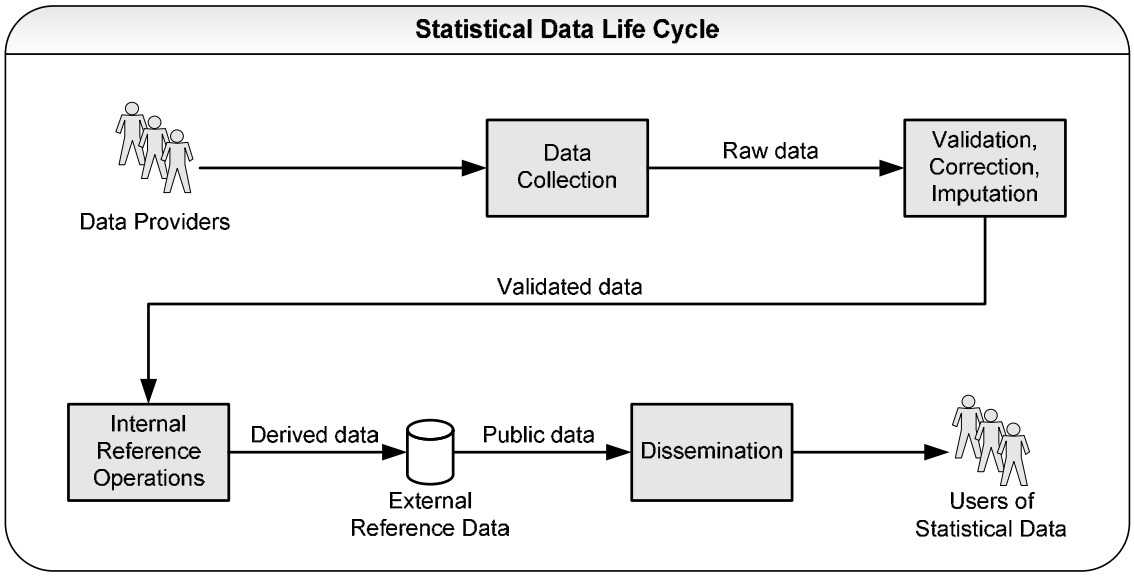


Figure 2.1 An example of a statistical data life cycle

The typical statistical data life cycle begins with data collection, during which data is collected from individuals and organisations by means of, for example, a survey or a census. This data is then validated and corrected. For those data items that are missing, invalid, inconsistent, or that are otherwise unusable, estimated values are provided through a process of data imputation (OECD, 2007). The estimated values are provided in such a way so as to ensure that a plausible and an internally-consistent record is created. The validated data is then used as the input to the process that creates the data sets used for internal processing by the statistical organisation. These data sets are used to derive the data that will form the External Reference Data. Different types of statistical data are then disseminated, or released to external users, by using the External Reference Data as the source. When data is disseminated to the public (i.e. the users of the statistical data), it is regarded as "safe" in the sense that it is considered to protect the privacy of the respondents.

In the context of the statistical data life cycle, we define the *data owner* as a person or an organisation that releases microdata about individuals. For example, a data owner may be a hospital that releases a microdata set that contains information on its patients. We also define the *data user* as a person or an organisation that requires the released microdata in order to perform specific types of data analysis.

2.3 Ways in which statistical data is disseminated

Statistical data can be disseminated in three main ways (Hundepool et al., 2007; Domingo-Ferrer et al., 2008; Willenborg & De Waal, 2001). These include:

- Dynamically queryable databases
- Tabular data, and
- Microdata

Dynamically queryable databases, which are also referred to as Statistical Databases (Adam & Wortmann, 1989), allow data users to access the External Reference Data of a statistical organisation by submitting statistical queries. The data users are able to retrieve only aggregate statistics (e.g. sums, averages, count) for a subset of the entities represented in the database.

Tabular data, which is also referred to as Macrodata (Ciriani et al., 2007), is a table that contains aggregate statistical information on one or more properties or attributes of the entities represented in the database. Therefore, data users access the External Reference Data of a statistical organisation indirectly through the aggregate statistical information of the tables published by the statistical organisation. The main difference between tabular data and dynamically queryable databases is that dynamically queryable databases allow data users to submit queries to create aggregate statistics, whereas users of tabular data are already provided with certain aggregate statistics without the need to submit queries to obtain the statistics. Table 2.1 shows an example of a Count tabular data that could be published about the number of men and women with specific diseases.

Microdata forms the basis from which all other statistical data outputs are derived (Hundepool et al., 2007). It is also the primary form in which data is stored, and hence microdata is the actual External Reference Data itself. A microdata set is a set of records, with each record containing data on the entities represented in the database. Table 2.2 shows an example of microdata set that contains a number of records of individuals with a specific disease. Note that the tabular data shown in Table 2.1 has been derived from this microdata set.

Microdata provides the greatest flexibility for statistical research, since microdata allows data users to create specific statistics that are useful for their particular research needs. That is, users of microdata are not restricted to the statistics that the statistical organisation publishes. For this reason, there has been an increasing demand for microdata from users, and many statistical organisations are making microdata available to meet this demand. In this research work, microdata will be the only type of statistical data that we will focus on. Hence, the rest of this Chapter concentrates on describing microdata in more detail.

Gender	Disease			Total
	Cancer	Hypertension	Heart disease	
Male	3	1	3	**7**
Female	0	2	1	**3**
Total	**3**	**3**	**4**	**10**

Table 2.1 An example of a Count Tabular data

Year of Birth	Marital Status	Gender	Zip Code	Disease
1967	Married	Male	40120	Cancer
1967	Divorced	Male	40322	Hypertension
1961	Widowed	Male	40322	Cancer
1962	Married	Male	40121	Heart disease
1965	Married	Male	40120	Heart disease
1977	Widowed	Male	40322	Heart disease
1984	Divorced	Male	40322	Cancer
1978	Widowed	Female	40120	Hypertension
1977	Divorced	Female	40321	Hypertension
1965	Married	Female	41454	Heart disease

Table 2.2 An example of a Microdata set

2.4 Microdata

A microdata set may be represented as a single data matrix, where the rows correspond to records of the entities of the database (e.g. an individual person or a respondent) and the columns correspond to the variables of each entity.

In the existing literature, different names for the different classes of variables of a microdata set are used by different authors. In this thesis, we shall adapt the naming conventions used by Willenborg and De Waal (2001). However, for completeness of this discussion, we also provide the alternative names used by other authors.

There are four, not necessarily disjoint, classes into which the variables of a microdata set can be classified. Before a microdata set is anonymised, the data owner should determine the class of each variable.

- *Direct identifiers.* These variables are those that uniquely identify a respondent in a microdata set. A person's Passport Number or ID Number are examples of a direct identifier. Direct identifiers are sometimes simply referred to as *Identifiers*, as has been done, for example, by Hundepool et al. (2007) and by Ciriani et al. (2007). Before a microdata set is anonymised, direct identifiers are removed from the microdata set.

- *Indirect identifiers.* These variables are not necessarily unique for each respondent in a given microdata set. However, the combination of the values of one or more indirect identifier of a single record may create a relatively rare, or even a unique combination in the given microdata set. Indirect identifiers are those variables on which an intruder will try to re-identify an individual respondent in a microdata set. Examples include the Date of Birth, or Zip Code of a person.

  Indirect identifiers are also sometimes referred to as *quasi-identifiers*, as has been done, for example, by Samarati (2001), or *key variables* (Hundepool et al., 2007). However, throughout this thesis, we shall refer to an indirect identifier as an *identifying variable*, as has been done by Willenborg and De Waal (2001).

- *Sensitive variables.* These variables are those that contain sensitive data of a respondent. For example, a sensitive variable can be a person's disease that he sought treatment for in a hospital. These variables are also referred to as *confidential outcome variables* (Domingo-Ferrer et al., 2008; Hundepool et al., 2007), since they contain confidential data about the respondents.

- *Non-sensitive, non-identifying variables.* These variables are those that do not fall into any of the above categories. These are also referred to as *non-confidential outcome variables* (Domingo-Ferrer et al., 2008; Hundepool et al.,

2007). An example of a non-sensitive, non-identifying variable may be a person's Gender. However, in combination with other variables, such as Marital Status, a person's Gender could also be an indirect identifier. Therefore, we mentioned earlier that the four variable classes are not necessarily disjoint.

Before a microdata set is anonymised, it is first de-identified, or "sanitised". That is, direct identifiers are removed from the microdata set. During microdata anonymisation, usually only the indirect identifying variables are modified, while the values of sensitive variables and non-sensitive non-identifying variables are preserved.

Now that the different variable classes have been introduced, we can provide a more formal definition of a microdata set. We define a de-identified non-anonymised microdata set as a matrix M (where each row contains attributes of a respondent and where p is the number of respondents whose data is released) as follows:

$$M = \begin{pmatrix} V_{11} & \cdots & V_{1n} & W_{11} & \cdots & W_{1m} & X_{11} & \cdots & X_{1q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ V_{p1} & \cdots & V_{pn} & W_{p1} & \cdots & W_{pm} & X_{p1} & \cdots & X_{pq} \end{pmatrix} \quad (1)$$

where:

- V_{ij} represents the j-th identifying variable of the i-th row,
- W_{ij} represents the j-th sensitive variable of the i-th row, and
- X_{ij} represents the j-th non-sensitive non-identifying variable of the i-th row.

2.5 Conclusion

In this Chapter, we presented different ways in which statistical data can be disseminated. We chose microdata as the statistical data type on which this research work is focused, since microdata forms the basis from which other types of statistical data are derived. We therefore discussed microdata in more detail and presented the different variable classes found in a microdata set. In the next Chapter, we will discuss the concept of information privacy and also show how privacy can be compromised in microdata.

CHAPTER 3

PRIVACY IN THE CONTEXT OF INFORMATION SECURITY

3.1 Introduction

The goal of information security is commonly regarded as the need to protect the confidentiality, integrity and availability of the assets (hardware, software, and data) of a computing system (Pfleeger, 1997). These three characteristics of information security are known as the CIA Triad. Confidentiality relates to ensuring that the assets of a computing system are accessible only to authorised parties. Integrity relates to ensuring that only authorised parties are able to modify the computing system assets and only in authorised ways. Finally, availability relates to ensuring that authorised parties are in fact able to access the assets of a computing system. Privacy, and specifically information privacy, relates directly to the confidentiality aspect of information security.

In this Chapter, we briefly discuss the concept of privacy by focusing on the privacy of information. We first provide a definition of information privacy and discuss its different dimensions. Thereafter, we discuss why information privacy needs to be protected. We conclude with a discussion of how privacy in microdata can be compromised when the confidentiality of microdata is not protected.

3.2 Privacy

3.2.1 Defining privacy

The concept of privacy not only relates to personal or physical privacy, but also to other aspects, such as information privacy. The way in which we perceive privacy depends on

what we regard as acceptable with reference to respecting one's privacy. However, this is highly influenced by one's culture, as has been noted by Räikkä (2008).

Therefore, privacy is a subjective and a difficult concept to define, since each person understands the concept of privacy in a different way. For example, Starr (1999, p. 200) states that "the essential interest in privacy is not control, but dignity – the protection of the individual from offensive and embarrassing disclosures". Then, at the other end of the spectrum, Schirmacher (1986) (as an example) argues that privacy should not be protected. On the contrary, he argues that "the destruction of privacy will not create an inhuman but a more humane society".

Nevertheless, our view is that many people would not allow for their data to be used for the creation, analysis, and release of statistical data. We also argue that, even if the data provided does not pertain to any offensive or embarrassing matters, many people may still desire that their privacy be protected.

A number of theories have been developed to define the concept of privacy. Tavani (2007) classifies privacy theories into four categories, as follows.

- *Non-intrusive theory*, which is centred on the theme of being free from intrusion. For example, the classic work on privacy by Warren and Brandeis (1890) falls into this category by defining privacy as a "right to be let alone".

- *Seclusion theory*, which focuses on the theme of being alone. If one chooses to define privacy using this theory, one can argue that the more isolated and the more alone a person is, the greater the privacy a person enjoys.

- *Control theory*, which is centred on the view that one has privacy if and only if one has control over information that pertains to oneself. For example, Fried (1990) argues that privacy is not the lack of information about ourselves, but it is rather the control one has over one's information.

- *Limitation theory*, which is centred on limiting or restricting access to information about oneself in certain contexts.

The non-intrusion and seclusion theories are most useful for defining privacy when it relates to physical access to people. The control and limitation theories of privacy are most useful in the case of *information* privacy.

Information privacy has three dimensions, depending on whose privacy we wish to protect (Domingo-Ferrer & Saygin, 2009; Domingo-Ferrer, 2007). The three dimensions are as follows.

- *Respondent privacy*, which focuses on the prevention of re-identification and disclosure of confidential data of the respondents whose records are released in a microdata set. Respondent privacy is usually sought when data is made available by the data owner (i.e. the one that collects the data) to data users. This dimension is the focus of this study.

- *Owner privacy*, which concentrates on preventing the disclosure of data in a database when two or more autonomous entities wish to compute queries across their databases, such that only the results of the query is revealed. It is usually the goal of privacy-preserving data mining (e.g. D. Agrawal and C. C. Aggarwal (2001); Bonchi, Malin and Saygin (2008); Clifton, Kantarcioglu and Vaidya (2005); Clifton, Kantarcioglu, Vaidya, Lin and M. Zhu (2002); Emekci, Sahin, D. Agrawal and Abbadi (2007); Kantarcioglu and Clifton (2004); Lindell and Pinkas (2002); Liu, Kantarcioglu and Thuraisingham (2008); Magkos, Maragoudakis, Chrissikopoulos and Gritzalis (2009); Vaidya and Clifton (2002)).

- *User privacy*, which aims to protect the privacy of queries to interactive databases, in order to prevent user profiling and re-identification (e.g. Chor, Kushilevitz, Goldreich and Sudan (1998); Ostrovsky and Skeith (2007); Domingo-Ferrer, Bras-Amoros, Wu and Manjon (2009)).

In this thesis, our goal in terms of protecting privacy is to ensure that, when microdata is released, confidential data about a particular respondent is not revealed. So, although users of the released microdata may be able to deduce confidential data about a group of respondents (e.g. a certain percentage of respondents have cancer), it should not be possible to deduce confidential data about any particular respondent. Hence, in this thesis, the limitation theory of privacy is the most useful one to define privacy. The ability of individuals or respondents to decide whether or not their data may be used or released is outside of the scope of this study. Therefore, we do not use the control theory of privacy for the purpose of defining privacy in this thesis. Furthermore, we focus on protecting respondent privacy. Therefore, the other two privacy dimensions are not part of the scope of this study.

Throughout the remainder of this thesis, when we refer to the concept of privacy, we will refer to *information* privacy, specifically in the context of dissemination of statistical data. In our research work, we define (information) privacy as the extent to which sensitive data (or information) about a respondent is protected when it is disseminated as statistical data. As it becomes more difficult to infer sensitive data about a respondent, the privacy of the respondent is increased. We will use this definition to quantify privacy in Section 6.5.

3.2.2 Why should information privacy be protected?

We argue that the need to protect information privacy stems from our desire to prevent certain information relating to us from being known to other people (or groups of people). But why does this desire exist? It may exist due to our need to protect our dignity or self-respect (as argued by Starr (1999) in the previous Section). It may also have its roots in the way in which we interact with other people as well as in the way in which we wish to be perceived by others. In this case, privacy allows us to create and maintain human relations, as has been argued by Gavison (1980), Gross (1971), and Rachels (1984). Räikkä (2008, p. 544) summarises these arguments quite well as follows:

> [The] control over editing one's self is very important, for it is through images of others that social relations are created and maintained. We are always

> concerned to control the ways in which we appear to others, and we tend to act to implement this concern in extremely subtle and sophisticated ways. The basic motive is to influence the reactions of others, and this is at the heart of all human social relations. We value privacy because our ability to control who has access to us, and who knows what about us, allows us to create and maintain the variety of relationships with other people that we wish to have.

We argue that certain data (e.g. our medical history) may violate our dignity or our self-respect, or it may allow others to perceive us in a way in which we would not wish to be perceived. Therefore, we argue that it is desirable to protect privacy by protecting certain data about ourselves.

The need to protect privacy was recognised as far back as the time of ancient Greece, during which the famous Hippocratic Oath was formulated. The Hippocratic Oath is an ancient ethical code for the practice of medicine, but one part of the oath relates to the protection of patient privacy. This part has been translated into English by Von Staden (1996) as follows:

> [W]hatever I may see or hear in treatment, or even without treatment, in the life of human beings – things that should not ever be blurted out outside – I will remain silent, holding such things to be unutterable.

However, in the present age, it is not only ethical obligations that require the protection of privacy. The need to protect privacy and to account for the disclosure of personal information has been passed as different laws in different countries. For example, in the US, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule was enacted, one aim of which was to ensure the protection of individuals' health information (OCR, 2003).

The desire to protect privacy manifests itself in many different environments where personal information, especially information of a sensitive nature, about individuals is collected, used and released. We conclude this Section by providing three example environments in which (information) privacy needs to be protected.

Information related to health care is likely to be the most confidential information that is maintained about an individual (Gostin & Turek-Brezina, 1995). It is likely that, for this reason, the medical environment is often seen as the most prominent example of an environment where protection of privacy is vital. Yet, medical records may, for example, contain data that can be an important source for scientific research work in epidemiology and health services research. This information source can be used to create knowledge related to different diseases, the safety of medical procedures and of pharmaceuticals, as well as the quality of medical care (F. G. Miller, 2008; Simon, Unützer, Young, & Pincus, 2000; Wald et al., 1994). Therefore, when data related to individuals' health is released for analysis, it is important to maintain the privacy of individuals whilst ensuring that the data is sufficiently useful for the required analysis.

As a second example, there is also a need to protect the privacy of consumers. For example, businesses may wish to exploit personal information by profiling consumers with the aim to improve their marketing strategy and to retain their customer base (Rauhofer, 2008). However, many consumers feel that they are tracked and exploited when their personal data is collected from different public and proprietary sources and is then mined to create consumer profiles (Paul, 2001).

As a final example, privacy should also be protected in the context of eParticipation. In one of our previous works (Zielinski, 2007a), we argued that statistical data can play a significant role in supporting the policy-making processes of a country. In particular, various statistical data related to citizens can be collected to serve as the basis for creating the knowledge necessary for guiding the development of policies. For example, collected statistical data may serve as a basis for assessing population trends, which may, in turn, serve to identify the challenges and opportunities that exist within the population. However, if citizens will not be assured that their privacy will be protected when their data is collected, processed and released for later use, then citizens may be reluctant to participate in the collection of their personal data. Such reluctance may reduce their participation in the policy-making process. It is, therefore, important that citizens are not discouraged from eParticipation processes due to privacy concerns. Therefore, to ensure that citizens take full advantage of eParticipation initiatives, techniques that protect the privacy of citizens should be developed and used when personal data is collected, analysed and released as statistical data.

3.3 How can privacy be compromised in microdata

Although direct identifiers can be removed from a microdata set through de-identification, the resulting microdata set is not necessarily anonymous (Samarati, 2001; Skinner & Elliot, 2001). That is, it is still possible for an intruder to compromise the privacy of the respondents in the microdata set. Privacy is compromised in a microdata set when *disclosure* occurs through the manipulation of the microdata set and / or by matching the microdata set with other sources of external data.

Two different forms of disclosure can occur (Lambert, 1993):

- *Identity disclosure*, which occurs when the identity of a respondent is revealed from the released microdata. This type of disclosure is also known as a re-identification. Even if the intruder is unable to obtain sensitive data from the re-identification, this type of disclosure may be sufficient to violate privacy.

  As an example of identity disclosure, consider a microdata set that is released about individuals with a criminal record, although the nature of their crime or other sensitive information is not released. If an intruder is able to identify an individual in the released microdata set, then the intruder may deduce that the particular individual has a criminal record. This deduction alone (i.e. the mere knowledge that an individual is present in the released microdata set) is sufficient to violate privacy.

- *Attribute disclosure*, which occurs when sensitive data about a respondent is obtained from the released microdata. This type of disclosure does not necessarily need to occur with identity disclosure.

  For an example of attribute disclosure without identity disclosure, consider the microdata set in Table 2.2, which may represent data about patients who were admitted to a certain hospital. From the microdata, we are able to deduce that all patients born in 1965 were admitted to the hospital with Heart Disease. Hence, we were able to infer sensitive data without the need to disclose the identity of

the respondents. That is, we did not need to determine the identities of the two respondents born in 1965 to determine the nature of their disease.

Attribute disclosure can also occur together with identity disclosure. Consider again the microdata set in Table 2.2. There is only one widowed female patient that lives in the geographical area with Zip Code 40120. If the geographical area has is relatively small population size, it may be possible for an intruder to determine the identity of this person, given the person's year of birth. Hence, identity disclosure occurs. Attribute disclosure can also occur because the intruder can deduce with certainty that the identified person suffers from Hypertension. This is possible since there is only one record of a female patient born in 1978 who is widowed and who lives in the geographical area with Zip Code 40120.

Disclosure risk is the risk that a given form of disclosure will occur when masked microdata is released (Chen & Keller-McNulty, 1998). Two main sources of disclosure risk for microdata exist (FCSM, 1994). The presence of highly visible records in a microdata set is one source of disclosure risk. Records that are relatively rare in a microdata set may contribute to easier re-identification. The other source of disclosure risk results from the possibility of matching the data in the microdata set with external sources of data. The greater the number of common variables between the microdata set and the external sources of data, the higher the possibility of linking between the two data sources.

Disclosure risk can be reduced through the application of statistical disclosure control techniques. These techniques aim to modify microdata (and other types of statistical data) such that an adequate level of privacy is provided. That is, statistical disclosure techniques aim to modify microdata such that the risk of disclosure is low, resulting in a microdata set that is considered "safe" for release. These techniques are discussed in the next Chapter.

3.4 Conclusion

In this Chapter, we discussed the concept of privacy with a specific focus on the privacy of information. When we release microdata, we need to ensure that information privacy is protected. This is required not only to ensure ethical conduct, but also to respect different privacy and data protection laws.

Information privacy is compromised when the released microdata is manipulated and / or matched with other sources of external data for the purpose of identity and / or attribute disclosure. The use of statistical disclosure control techniques can reduce the risk of disclosure. However, the risk of disclosure should be reduced in such a way so as to ensure that the resulting microdata is adequately useful for the purpose for which it is released. This challenge is further explored in Chapter 5, but first, in the next Chapter, we will discuss different techniques that may be used to reduce the risk of disclosure in microdata.

CHAPTER 4

PRIVACY PROTECTION IN MICRODATA

4.1 Introduction

In Chapter 2, we discussed the different ways in which statistical data can be disseminated, with a specific focus on microdata. In the previous Chapter, we discussed privacy in the context of information security and showed how privacy can be compromised when releasing microdata. In this Chapter, we present an overview of the techniques available for protecting privacy for the different forms in which statistical data can be disseminated, with a greater emphasis on the techniques available for the protection of privacy in microdata.

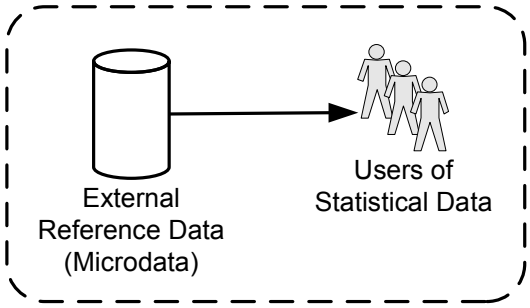
4.2 Approaches for statistical disclosure control

Approaches for statistical disclosure control depend on the way in which statistical data is disseminated. The approaches can be classified into four main classes as follows (Adam & Wortmann, 1989; Hundepool et al., 2007; Ciriani et al., 2007; Willenborg & De Waal, 2001):

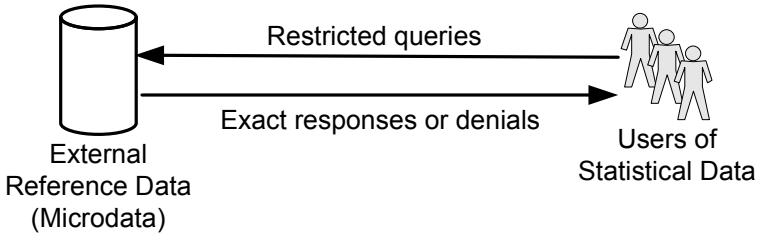
- Protection of data output through access control
- Protection of dynamically queryable databases
- Protection of tabular data, and
- Protection of microdata

These approaches are described next and are represented graphically in Figure 4.1.

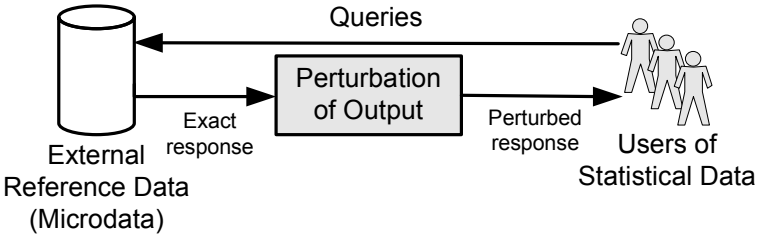
(a) Protection of data output through access control



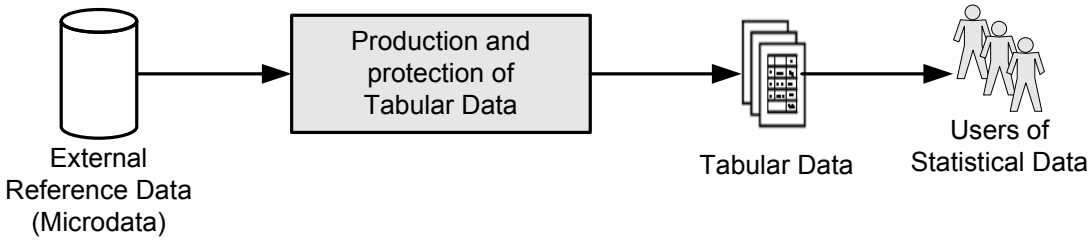
(b) Query restriction in dynamically queryable databases



(c) Output perturbation in dynamically queryable databases



(d) Protection of tabular data



(e) Protection of microdata

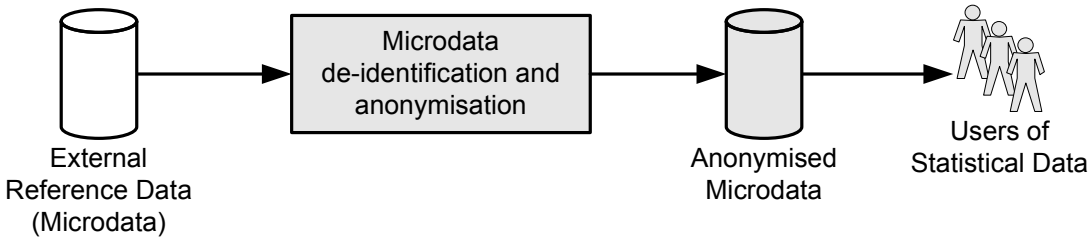


Figure 4.1 Approaches for statistical disclosure control

4.2.1 Protection of data output through access control

Some types of research require data with a high level of detail. The high level of detail that is required cannot be provided if statistical disclosure controls are used to protect the data. In such situations, a statistical organisation may consider to provide controlled access, as in Figure 4.1 (a), to de-identified, but non-anonymised data to a restricted group of users under certain conditions. The users are able to analyse data in a protected environment, under legal and administrative restrictions.

4.2.2 Protection of dynamically queryable databases

Two main approaches exist to protect statistical data that is disseminated through a dynamically queryable database: query restriction and output perturbation. Query restriction (Kenthapadi, Mishra, & Nissim, 2005; Nabar, Marthi, Kenthapadi, Mishra, & Motwani, 2006), shown in Figure 4.1 (b), involves restricting the statistical queries that can be made by the data user. Those queries that will allow a user to infer sensitive information about respondents are rejected. When a data user submits a new query, the answers of past queries made by the user are used to determine whether privacy will still be preserved. If privacy will still be preserved, the new query will be accepted; otherwise it will be rejected.

Output perturbation (Blum, Dwork, McSherry, & Nissim, 2005; Dinur & Nissim, 2003; Dwork, McSherry, Nissim, & Smith, 2006), shown in Figure 4.1 (c), preserves privacy by providing non-exact, or slightly modified, answers to the queries that a data user submits. One way in which output perturbation can be implemented is by modifying the exact result of a query at run time and providing the user with the perturbed answer. Another way to implement output perturbation is by perturbing the database itself and allowing the user to query only the perturbed database.

4.2.3 Protection of tabular data

Privacy can be preserved in tabular data by publishing the data such that there are no unsafe cells, or cells in the table that are associated with a relatively small number of

respondents (Figure 4.1 (d)). To protect privacy, tables that are to be published may be redesigned to remove the unsafe cells. Unsafe cells may also be suppressed or the data released in the table may be perturbed.

4.2.4 Protection of microdata

When microdata is released to the public, sampling is commonly employed (FCSM, 1994; Skinner, Marsh, Openshaw, & Wymer, 1994). Sampling protects microdata by releasing only a sample (or a subset) of the respondents in the original microdata set. Releasing only a sample of the respondents in a microdata set creates an uncertainty as to whether or not a certain respondent's data has been released. By creating this uncertainty, the risk of re-identification is reduced. However, sampling alone is insufficient to provide an adequate level of privacy. For this reason, other techniques are usually still used to protect microdata after sampling has occurred. Techniques that may be used to protect microdata (Figure 4.1 (e)) can be classified into two categories: masking techniques and synthetic data generation techniques (Ciriani et al., 2007; Domingo-Ferrer & Torra, 2005). Each category also has two sub-categories, as shown in Figure 4.2.

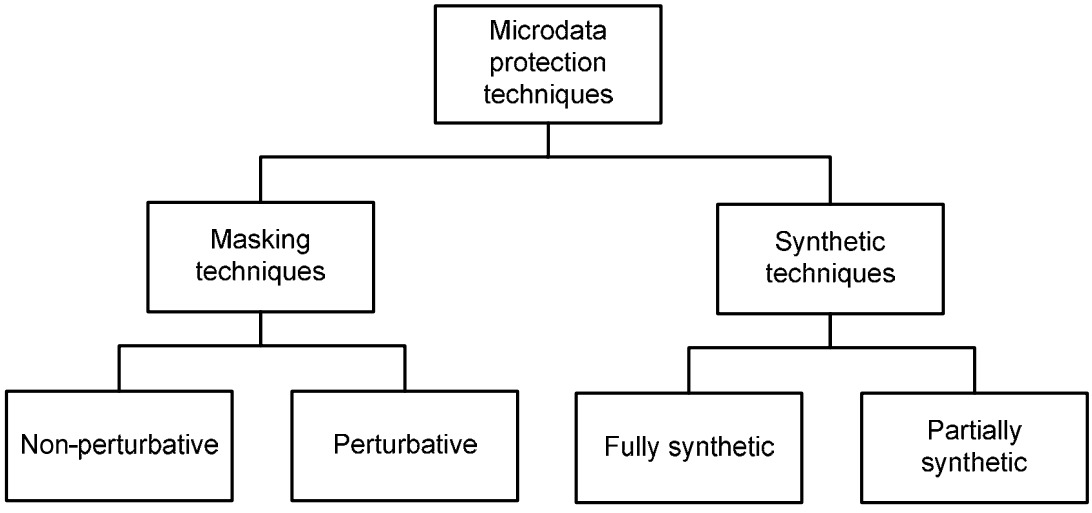


Figure 4.2 Categories of microdata protection techniques

Masking techniques either reduce the amount of data that is released (non-perturbative techniques), or modify the released data (perturbative techniques). Non-perturbative

techniques protect microdata by either partially suppressing data or by reducing the amount of detail of the released data. There is no distortion of the original data. Therefore, when a microdata set is protected by a non-perturbative technique, the released data is correct to the same degree as the original, unreleased microdata. However, the precision (or accuracy) and completeness of the released microdata is reduced, since the level of detail is reduced.

Perturbative techniques protect microdata by perturbing, or modifying, the data before it is released. The techniques do not limit the amount of data that is released, but distort the original data to prevent one from determining their exact values. Therefore, perturbative techniques alter the truth or the facts which the original data represents, although certain summary statistics may be preserved.

Synthetic data generation techniques aim to produce a microdata set that has artificial values. Nevertheless, the values are such that a conceivable microdata set is created by preserving the statistical properties of the original microdata set. That is, when the synthetic data has been generated, it should provide the same quality of statistical analysis as the original data. Since the released data has been generated, it is not related to the data that has been provided by specific respondents. Therefore, one could assume that no re-identification can take place. However, this assumption does not hold, since it is possible for even synthetic data to contain records that allow for re-identification (Reiter, 2005; Winkler, 2004). This is especially true for synthetic data that has been over-fitted to the original data.

Two sub-categories of synthetic data generation techniques exist: techniques that generate a completely new set of data (fully synthetic techniques) and techniques that merge the original and the synthetic data (partially synthetic techniques). Synthetic data generation techniques do not form part of our study. Therefore, we will not discuss these techniques further. Readers interested in a more detailed overview of synthetic data generation techniques are referred to the works of Ciriani et al. (2007) and Hundepool et al. (2007).

In our research work, we focus on only one non-perturbative anonymisation technique, namely global recoding. We also focus on only one perturbative anonymisation

technique, namely microaggregation. The reason for this focus is that these techniques are typically used to achieve *k*-anonymity (Samarati, 2001; Sweeney, 2002a, 2002b; Domingo-Ferrer & Torra, 2005). Combining local suppression with local or global recoding may also be used to achieve *k*-anonymity. However, we will not consider local suppression and local recoding, due to the drawbacks in their use for *k*-anonymisation, which will be explained in Chapter 5.

Recoding and microaggregation are discussed in the next two Sections. Readers who are interested in other non-perturbative and perturbative techniques are referred to the excellent comparative studies that were conducted by Ciriani et al. (2007), Domingo-Ferrer and Torra (2001), and Willenborg and De Waal (2001).

4.3 Recoding

Recoding is a non-perturbative technique that anonymises microdata by changing the original coding of a variable to coding that has a lower level of detail (Willenborg & De Waal, 2001).

A coding is defined as a combination of one or more categories of a variable, where a category is a disjoint partition of the domain of a variable. Given a variable, we partition its domain into two or more disjoint categories. A combination of one or more categories of the variable forms a coding. We also define a set of *k* codings $\{C_1,\ldots,C_k\}$ with which the variable can be released in the anonymised microdata. The codings are defined such that a coding $C_{(l+1)}$ has a lower level of detail than a coding C_l. A variable always has at least two codings, one which is used in the original microdata set (i.e. coding C_1), and one which consists of exactly one category (i.e. coding C_k).

For example, suppose we are given the variable *Year of Birth*, which can assume any valid year from 1961 to 1990. We may define codings with which the variable can be released as follows:

- C_1, which represents the non-recoded data of the variable. That is, the variable can assume any valid year from 1961 to 1990.

- C_2, in which groups of two categories from C_1 are combined into one to give us the categories "1961-1962", "1963-1964", ... , "1989-1990".
- C_3, in which groups of three categories from C_1 are combined into one to give us the categories "1961-1963", "1964-1966", ... , "1988-1990".
- C_4, in which groups of five categories from C_1 are combined into one to give us the categories: "1961-1965", "1966-1970", ... , "1986-1990".
- C_5: in which groups of ten categories from C_1 are combined into one to give us the categories "1961-1970", "1971-1980", "1981-1990".
- C_6: in which groups of fifteen categories from C_1 are combined into one to give us the categories "1961-1975", "1976-1990".
- C_7: where the variable can assume only one value "1961-1990".

The level of recoding that is performed depends on the type of coding chosen for a variable as well as on the number of records in which the variable is recoded. In *global recoding*, all variables are recoded to the same coding. For example, if the *Year of Birth* variable, in all records of a microdata set, is recoded to coding C_5 (as defined above), then the *Year of Birth* variable in all the released records can assume only the values "1961-1970", "1971-1980", or "1981-1990". On the other hand, if *local recoding* is used, then the variable can be recoded to different codings in different records of a microdata set. In this case, the *Year of Birth* variable could assume values from different categories in different records. This may complicate the analysis of locally recoded data as different categories co-exist in the released microdata set.

Generalization (Samarati, 2001; Sweeney, 2002a) may be regarded as a particular type of recoding. During generalization, the values of a variable are replaced by a more general form of the value, based on the *generalization hierarchy* that has been defined for that variable. The most specific values of a variable are at the leaves and the most general value is at the root of the hierarchy. A variable can be generalized by replacing the values represented at the leaf nodes by a more general form of the value, as indicated at their ancestor nodes at a higher level in the hierarchy.

An example of a generalization hierarchy for the variable *Marital Status* of a person is shown in Figure 4.3. In this example, the values "Married", "Widowed", and "Divorced" can be replaced by a more general value "Been_Married". In a similar way, the value "Single" can be replaced by the value "Never_Married". The values "Been_Married" and "Never_Married" themselves can be replaced by the value "Not_Released".

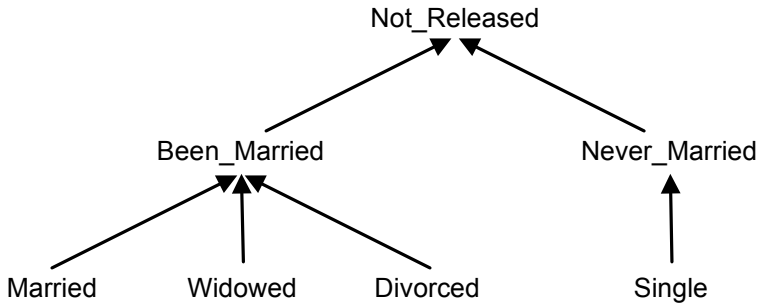


Figure 4.3 An example of a generalization hierarchy for the variable *Marital Status*

Two additional types of recoding include top- and bottom-coding. Top-coding requires that an upper limit (top-code) on an attribute be provided. The microdata is anonymised by substituting the upper limit for all values of a given variable, whose original value was greater than the upper limit. For example, a person's age can be released as being above 65, rather than releasing the exact value.

Bottom-coding is similar to top-coding. It requires that a lower limit (bottom-code) be provided on an attribute. When the microdata is anonymised, the lower limit is substituted for all values of a given variable, whose original value was less than the lower limit. For example, a person's income may be released as being below a certain amount, rather than releasing the exact value.

4.4 Microaggregation

Microaggregation is a perturbative technique that anonymises data in two steps. In the first step, clusters of similar records are created, such that each cluster has at least *k* records. Then, in the second step, every record in a particular cluster is replaced with the cluster's centroid value. This technique can be used for continuous numerical data

(Defays & Nanopoulos, 1993; Defays & Anwar, 1995; Domingo-Ferrer & Mateo-Sanz, 2002) as well as for categorical data (Torra, 2004). In the case of continuous data, a particular cluster's centroid value is the cluster's mean, and in the case of categorical data, a cluster's centroid value is the cluster's median.

Two classes of microaggregation exist, depending on the number of variables to which the technique is applied (Domingo-Ferrer, Oganian, & Torres, 2002). When microaggregation is applied to only one variable, it is referred to as *univariate* microaggregation. When more than one variable is microaggregated, then microaggregation is said to be *multivariate*.

Two approaches to multivariate microaggregation exist (Nin, Herranz, & Torra, 2008a). In the first approach, a multivariate microaggregation technique is applied to all the identifying variables of a microdata set. This technique ensures that all the identifying variables of the records in each cluster are the same, ensuring the property of *k*-anonymity. (The property of *k*-anonymity is discussed in Section 5.5.) In the second approach, the identifying variables of a microdata set are grouped into blocks of one or more variables. The grouping is such that each variable is part of only one block. Each block of variables is then microaggregated independently of the other blocks. If there is only one variable per block, then the multivariate microaggregation is reduced to performing univariate microaggregation in parallel on each variable. Although splitting the variables into blocks of several variables can increase the statistical utility of the released microdata set, the property of *k*-anonymity is not guaranteed anymore, which may reduce the level of privacy of the microdata.

The way in which a microdata set will be microaggregated depends on a number of factors, as identified by Nin et al. (2008a). These factors, which will be briefly discussed next, include the following:

- The type of microaggregation method or algorithm used,
- The way in which variables are split into blocks, and
- The least number of records required in each cluster (i.e. the parameter *k*).

The type of microaggregation method or algorithm used

Domingo-Ferrer and Mateo-Sanz (2002) have defined an optimal microaggregation as one that produces a *k*-partition that maximises the homogeneity within each group (or cluster). There exists a polynomial-time optimal algorithm that provides optimal microaggregation in the case of univariate microaggregation (Hansen & Mukherjee, 2003). However, in the case of multivariate microaggregation, this problem has been shown to be NP-hard (Oganian & Domingo-Ferrer, 2001). Therefore, heuristic microaggregation approaches have been proposed to improve the trade-off between computational complexity and information loss, such as, for example, the Maximum Distance to Average Vector (MDAV) algorithm (Domingo-Ferrer et al., 2006; Hundepool et al., 2005).

The way in which variables are split into blocks

The way in which a microdata set is microaggregated is also influenced by the way in which variables are split into blocks, which has been studied by Nin et al. (2008a, 2008b). It is influenced by the number of blocks into which the microdata set is split (and the number of variables in each block) as well as by the degree of correlation between the variables that have been grouped in each block. When microaggregation is performed on variables that are highly correlated, then the clusters will contain records that are similar to one another with respect to the variables that have been selected in the block. Therefore, as the degree of correlation between grouped variables increases, so does the information utility of the released microdata. However, as the number of blocks into which variables are split increases (thereby decreasing the number of variables per block), the anonymity that microaggregation provides decreases, independently of the way in which variables have been grouped.

The least number of records required in each cluster

The number of records *k* chosen per cluster has a direct effect on the privacy and information utility levels of a released microdata set. As a microdata set is microaggregated with a higher number of records per cluster, there are more records that have the same centroid value. Therefore, the discernability of the released records is

lower, leading to a lower disclosure risk. The loss in (unique) information therefore increases the level of privacy, but decreases the level of information utility. Nevertheless, it is still unclear in the literature what should be the optimum number of records *k* chosen per cluster to produce a microdata set with the optimum levels of information utility and privacy. We present a solution to this problem in Chapter 7.

4.5 Conclusion

In this Chapter, we briefly discussed the techniques available for protecting privacy when statistical data is disseminated. We focused primarily on the techniques that are used to protect microdata. In particular, we placed emphasis on the techniques of recoding and microaggregation, since *k*-anonymity is usually achieved through the use of these techniques.

Anonymisation of a microdata set removes (to some extent) data contained in the identifying variables. As more of this data is removed, re-identification and inference of sensitive data becomes more difficult. Therefore, as a microdata set is anonymised, the level of privacy in the microdata set increases. On the other hand, removing data from the identifying variables decreases the accuracy and / or completeness of the released microdata set. Therefore, as a microdata set is anonymised, the microdata set becomes less useful, which reduces the information utility level.

Clearly, a conflict between the needs of privacy and information utility arises. The protection of privacy implies that we should hide and obscure data. On the other hand, to release data that is useful requires that we provide data that is accurate, complete and precise (Zielinski, 2007b). This conflict between privacy and information utility is explored further in the next Chapter and a solution to this challenge is proposed in Chapters 6 and 7.

CHAPTER 5

THE CONFLICT BETWEEN PRIVACY AND INFORMATION UTILITY

5.1 Introduction

In the previous Chapter, we presented an overview of the techniques available for protecting privacy for the different forms in which statistical data can be disseminated, with a greater emphasis on protection of privacy in microdata.

To protect the privacy of the respondents in a microdata set, the microdata set needs to be anonymised. As microdata is anonymised, data is removed, to some extent, from the identifying variables. As more data is removed from the identifying variables, it becomes increasingly difficult to infer sensitive data and to perform re-identification. Therefore, the level of privacy in a microdata set increases as the set is anonymised. However, removing data from the identifying variables also reduces the accuracy and / or completeness of the released microdata. Therefore, as a microdata set is anonymised, its level of information utility also decreases.

Ideally, we would like to release microdata that has high levels of privacy and information utility. However, the protection of privacy implies that we should hide and obscure data. On the other hand, releasing usable and useful data implies that we should provide data that is accurate, complete and precise (Zielinski, 2007b). Clearly, a conflict between the needs of privacy and information utility exists. This conflict needs to be resolved before a microdata set can be released.

A number of approaches have been proposed in the literature to address the conflict between privacy and information utility. These include the *score*, R-U confidentiality maps, and *k*-anonymity, as identified by Domingo-Ferrer and Torra (2005). In this

Chapter, we present these approaches by discussing the extent to which they are appropriate in finding the optimum balance between privacy and information utility.

This Chapter is organised as follows. We first provide a definition of the "optimum" balance between privacy and information utility. We then discuss how the *score*, R-U confidentiality maps and *k*-anonymity are currently used to address the conflict between privacy and information utility. We shall argue that these approaches are not completely appropriate for finding the optimum balance between privacy and information utility. We conclude this Chapter by presenting recommendations for an appropriate solution that will determine the optimum balance between privacy and information utility when microdata is anonymised. These recommendations will be used to develop an appropriate solution, which will be presented in Chapters 6 and 7.

5.2 What is the "optimum" balance between privacy and information utility?

In our research work, we regard the optimum balance between privacy and information utility as one in which the levels of privacy and information utility are maximised while satisfying a set of constraints that capture the data owner's and the data user's preferences. These preferences refer to the preferences that exist between each identifying variable in the microdata set, as well as the preference between the resulting levels of privacy and information utility.

The preferences between each identifying variable in the microdata set are directly related to the usefulness of the data. The usefulness of the data should be considered from both the data user's and the data owner's points of view. In the case of the data user (whose main goal is to ensure utility of data), the preferences for identifying variables should reflect the extent to which each identifying variable will be useful for the data user's tasks. In the case of the data owner (whose main goal is to protect the privacy of the respondents in the microdata), the preferences are considered from a potential intruder's point of view, in terms of the perceived way in which an intruder may use the released data to infer sensitive information. In this case, the preferences for identifying

variables should reflect the extent to which each identifying variable would be useful for the intruder in inferring sensitive data.

Before a microdata set is anonymised, the preference between the resulting levels of privacy and information utility must be decided and agreed upon together by the data user and the data owner. That is, it is necessary to determine if protection of privacy is considered to be equally important as providing useful data, or if privacy should assume a greater or lower importance compared to information utility. For example, if the microdata is released to only a selected group of data users, under strict confidentiality agreements made with this group, then it is certainly possible that the data owner's preference for privacy may be lower when compared to cases where the microdata is made available to the public.

Therefore, the challenge of finding the optimum balance between privacy and information utility can be stated as an optimisation problem as follows: "Maximise privacy and information utility subject to the constraints imposed by the data user's and the data owner's preferences". In the next three Sections, we discuss the extent to which the *score*, R-U confidentiality maps, and *k*-anonymity are appropriate for finding the optimum balance between privacy and information utility.

5.3 The *score*

Domingo-Ferrer and Torra (2001) introduced the *score* as a way to evaluate the trade-off between information loss and disclosure risk. It was subsequently used in several other works, for example, by Medrano-Gracia et al. (2007), Nin et al. (2008a, 2008b), and Yancey, Winkler, and Creecy (2002). We use these works to define and discuss the *score*.

The *score* is defined as:

$$score = \frac{IL + DR}{2} \qquad (2)$$

where:

- IL refers to the Information Loss, computed as:

  $IL = 100\left(0.2\,IL_1 + 0.2\,IL_2 + 0.2\,IL_3 + 0.2\,IL_4 + 0.2\,IL_5\right)$, where:

  - IL_1 is the mean absolute error of the original microdata with respect to the protected data,
  - IL_2 is the mean variation of the attribute average vectors,
  - IL_3 is the mean variation of the attribute covariance matrices,
  - IL_4 is the mean variation of the attribute variance vectors, and
  - IL_5 is the mean variation of the attribute correlation matrices.

- DR refers to the Disclosure Risk, computed as:

  $DR = 0.25\,DLD + 0.25\,PLD + 0.5\,ID$, where:

  - DLD is the *distance-based linkage disclosure risk*, which is the average percentage of correctly linked records using distance-based record linkage. Distance-based record linkage (Pagliuca & Seri, 1999) is a record linkage method where the original record is linked to the closest protected record. For this purpose, the Euclidean distance (as an example) may be used.
  - PLD is the *probabilistic linkage disclosure risk*, which is the average percentage of correctly linked records using probabilistic record linkage. Probabilistic record linkage (Jaro, 1989) is a record linkage method where the original record is linked according to a criterion on coincidence vectors, which are defined from the available sets of original and protected records.
  - ID is the *interval disclosure risk*, which is the average percentage of protected values falling into the intervals around their corresponding original values.

The *score* is useful in that it allows us to regard the selection of a masking technique (for microdata protection) and the parameters of the technique as an optimisation problem (Domingo-Ferrer & Torra, 2005). For example, Sebe et al. (2002) applied a

masking technique to a microdata set, after which a post-masking optimisation procedure was applied to obtain an improved *score*.

The main drawback of the *score*, with reference to how appropriate it is for addressing the conflict between privacy and information utility, is that it is unable to take into account the way in which the released data will be used. That is, the *score* is unable to take into account the needs and preferences of the data user. When calculating the Information Loss (*IL*) measure, the *score* is unable to take into account the data user's preferences with regards to the identifying variables that will be useful for him. Therefore, the Information Loss measure does not truly reflect information utility for data users.

For example, one would assume that a microdata set with a high Information Loss measure will have low information utility. However, the level of information utility may in fact be relatively high for a certain data user, if most of the information loss occurred in the identifying variables that the specific data user does not require. Consequently, it is possible that an anonymised microdata set with a good *score* value (i.e. a microdata set with a good balance between privacy and information utility) will not provide the best (or optimum) levels of privacy and information utility for a particular user, since the particular user's needs were not taken into account.

The need to take into account this preference was one of the requirements we identified for the *optimum* balance between privacy and information utility. The *score* fails to take into account this requirement. Therefore, we do not use it in our research work.

It is also possible, as suggested by Nin et al. (2008b), to use different weights for *IL* and *DR*, to take into account the preferences between privacy and information utility. For example, we can use a higher weight for *DR* and a lower weight for *IL*, if we consider the preference for privacy to be higher than the preference for information utility. We can also use different weights for calculating *DR* itself. For example, if we perceive that linkage disclosure attacks are more likely than interval disclosure attacks, then we can use a lower weight for *ID* and higher weights for *DLD* and *PLD*. We also suggest that one avenue for future work would be to build upon the original definition of the *score*, such that it takes into account the preferences with regards to different identifying

variables. However, as stated by Nin et al. (2008b), these types of modifications would go against the philosophy of the *score* measure, since such modifications would assume some prior knowledge on the use of the data or on the behaviour of the intruders.

5.4 R-U confidentiality maps

R-U confidentiality maps (Duncan, et al., 2001; Duncan, Keller-McNulty & Stokes, 2001) provide a way in which to graphically represent the conflict between disclosure risk, R, and data utility, U. After the form of the disclosure risk, R, and the data utility, U, have been specified, the task is to determine how R and U are related to the parameter values of the specific masking technique chosen to anonymise a microdata set. An R-U confidentiality map is obtained by plotting, on a two-dimensional graph, a set of paired values, (R, U), which represent the disclosure risk and the data utility that correspond to various strategies for data release.

The graphical representation of the relationship between privacy and information utility allows one to easily determine how a particular masking technique and its parameters impact the balance between privacy and information utility. It is, of course, reasonable to expect that a microdata set should be released with a level of data utility U at which the disclosure risk R will be below the maximum tolerable risk. However, by using the R-U confidentiality map alone, it is still unclear where the optimum balance between R and U occurs. One does not know if the optimum balance occurs *exactly* at the point at which R is just below the maximum tolerable risk. However, it is also quite likely that the optimum balance may, in fact, occur at a lower risk level, much lower than the maximum tolerable risk. This is certainly possible when (R, U) pairs form an exponential graph. In such cases, reducing the utility level by a small factor may result in a relatively large reduction of the disclosure risk. Hence, the optimum balance between R and U may in fact occur lower than the maximum tolerable risk, but this is not known by just examining the R-U confidentiality map.

Nevertheless, we do not use R-U confidentiality maps in our research work. R-U confidentiality maps do not actually *determine* what the optimum balance between privacy and information utility is. That is, they can only *guide* the decision about how to

balance the needs of privacy and information utility, by graphically representing the relationship between privacy and information utility. However, the decision where to strike the balance between privacy and information utility is still left up to the user of the R-U confidentiality map.

5.5 *k*-anonymity

The concept of *k*-anonymity has been introduced by Samarati and Sweeney (Samarati, 2001; Sweeney, 2002a, 2002b). A microdata set satisfies the property of *k*-anonymity if every record in the microdata set is indistinguishable from at least *k* - 1 other records in the same microdata set, where *k* is greater than 1. The inability to distinguish between different records is based on the values of the identifying variables (or quasi-identifiers – an equivalent term commonly used in the literature on *k*-anonymity). That is, given a record with a particular set of values for the identifying variables, the same set of values will be present in the identifying variables of at least *k* - 1 other records in the same microdata set.

Since its introduction, there has been a considerable amount of research on the concept of *k*-anonymity. Research on *k*-anonymity has been mainly focused on proposing techniques for achieving *k*-anonymity, finding an optimal *k*-anonymisation, and proposing enhancements to *k*-anonymity to address its shortcomings. Research on *k*-anonymity has also focused on using *k*-anonymity (and its enhancements) for addressing the conflict between privacy and information utility. These research directions are discussed in the remainder of this Section.

5.5.1 Techniques used for achieving *k*-anonymity and for finding an optimal *k*-anonymisation

When *k*-anonymity was initially proposed, it was enforced by applying a combination of generalization (a type of Recoding – see Section 4.3) and suppression (Samarati, 2001; Sweeney, 2002a). Subsequently, generalization and suppression have also been used in many other algorithms for achieving *k*-anonymity (see below for references of example algorithms).

However, the use of generalization and suppression to achieve *k*-anonymity poses several challenges (Domingo-Ferrer & Torra, 2005). For example, given an identifying variable with *c* different categories, there are $2^c - c - 1$ possible generalizations. This poses the challenge of a high computational cost for finding an optimum recoding. Moreover, not all the recodings will be appropriate. For example, recoding a *Zip Code* to 401*0 does not correspond to a meaningful geographical location, as opposed to, for example, 4012*, which corresponds to a greater geographical area in which the city with the *Zip Code* 40120 is located. Therefore, the selection of the subset of appropriate generalizations from the $2^c - c - 1$ possible ones also poses a significant challenge. Moreover, when generalization is used on a continuous variable, the variable looses its numerical semantics and becomes categorical. For example, when a continuous variable that contains the age of a person is generalized, its values are replaced by a range of values (e.g. "32" may be replaced by the range "31 - 35"). This prevents data users from making inferences about the distribution of the original numerical values in the range (e.g. if the original values were mostly in the lower or upper half of a particular range).

Furthermore, applying only generalization may still leave a number of records that are relatively rare in the microdata set. If we were to attempt to reduce the scarceness of these records by further generalization, then it would lead to unnecessary loss of detail in other records, which were relatively common before. Therefore, local suppression is usually applied to these relatively rare records. However, it is not known how to optimally combine generalization with local suppression such that *k*-anonymity is achieved. Furthermore, if suppression is used, it is also unclear how the protected data can be analysed without special software that will be able to take the suppressions into account (e.g. through imputation).

Given the challenges associated with using generalization and suppression to achieve *k*-anonymity, microaggregation has been proposed as an alternative way of achieving *k*-anonymity (Domingo-Ferrer & Torra, 2005). Microaggregation provides several advantages over the use of generalization and suppression. For example, microaggregation does not create new categories (like recoding does), which does not complicate data analysis. In addition, microaggregation can be used to protect

continuous variables without changing them into categorical variables. Furthermore, microaggregation does not suppress data, which allows one to analyse a *k*-anonymised microdata set without the use of special software that takes suppressions into account.

Irrespective of the technique used to achieve *k*-anonymity, the research problem of optimal *k*-anonymisation aims to find an anonymisation that will produce the "best" *k*-transformed dataset, as determined by some cost metric (Bayardo & R Agrawal, 2005). For example, if the cost metric is the information loss that occurs as a result of the generalization and suppression applied, then an optimal *k*-anonymisation is an anonymisation that achieves *k*-anonymity with the least number of generalization and suppression combinations, so as to minimise information loss. Finding an optimal *k*-anonymisation has been proved to be NP-hard (Meyerson & Williams, 2004). Nevertheless, a number of polynomial time approximate algorithms have been developed, such as those proposed by Aggarwal et al. (2005), Fung, Wang and Yu (2005), Ghinita et al. (2007), Gionis and Tassa (2009), LeFevre, DeWitt and Ramakrishnan (2005, 2006a), Li et al. (2006), as well as Meyerson and Williams (2004).

Note that the research problem of finding an optimum *k*-anonymisation is different from the research problem of finding an optimum *k* value for *k*-anonymisation. In the former, we need to determine how to optimally achieve *k*-anonymity, when a given *k* value is already known. In the latter research problem, we aim to determine what is the optimum *k* value with which a microdata set should be *k*-anonymised.

5.5.2 Shortcomings and enhancements of *k*-anonymity

A record in a microdata set, which satisfies the property of *k*-anonymity, cannot be mapped back to the corresponding record in the original data set. Since there will be at least *k* records in the anonymised microdata set that can match any value of the identifying variables that an intruder uses, the best mapping that an intruder can perform is to map groups of *k* records in the anonymised microdata set to the original data set. Therefore, *k*-anonymity is able to prevent identity disclosure. However, *k*-anonymity is

unable to guarantee protection against attribute disclosure (Domingo-Ferrer & Torra, 2008).

To illustrate how attribute disclosure can occur in a microdata set that satisfies *k*-anonymity, consider the following example (Zielinski, 2007c). Let us assume that Table 5.1 represents a microdata set, which contains data on patients admitted to a hospital. This table has already been de-identified, as it does not contain any explicit identifiers. We anonymise the table such that it satisfies *k*-anonymity with $k = 3$. In order to do this, we must ensure that the values of the identifying variables (namely *Date of Birth*, *Race*, *Gender*, and *Zip Code*) of every record in the microdata set cannot be distinguished from the values of the identifying variables in at least 2 other records in the same microdata set. To achieve this, we have generalized the values for *Date of Birth*, by removing the month and day and keeping only the year of a person's date of birth. The resulting microdata set is shown in Table 5.2.

Date of Birth	Race	Gender	Zip Code	Disease
1967/01/01	Black	Male	40121	Cancer
1967/02/02	Black	Male	40121	Hypertension
1967/03/03	Black	Male	40121	Cancer
1968/04/04	White	Male	40242	Heart disease
1968/05/05	White	Male	40242	Heart disease
1968/06/06	White	Male	40242	Heart disease
1969/07/07	Black	Female	40373	Cancer
1969/08/08	Black	Female	40373	Hypertension
1969/09/09	Black	Female	40373	Hypertension
1970/10/10	White	Female	40404	Heart disease
1970/11/11	White	Female	40404	Cancer
1970/12/12	White	Female	40404	Hypertension

Table 5.1 A non-anonymised microdata set used to illustrate attribute disclosure

Date of Birth	Race	Gender	Zip Code	Disease
1967	Black	Male	40121	Cancer
1967	Black	Male	40121	Hypertension
1967	Black	Male	40121	Cancer
1968	White	Male	40242	Heart disease
1968	White	Male	40242	Heart disease
1968	White	Male	40242	Heart disease
1969	Black	Female	40373	Cancer
1969	Black	Female	40373	Hypertension
1969	Black	Female	40373	Hypertension
1970	White	Female	40404	Heart disease
1970	White	Female	40404	Cancer
1970	White	Female	40404	Hypertension

Table 5.2 A *k*-anonymised microdata set used to illustrate attribute disclosure

Since Table 5.2 satisfies *k*-anonymity, we would assume that it also protects the privacy of those individuals whose data is reflected in the table. However, this table does not protect the privacy of all individuals, because attribute disclosure can still occur. By examining Table 5.2 closely, we notice that every white male (admitted to the hospital), who was born in 1968 and who is living in the area with Zip Code 40242, has heart disease. Therefore, we are able to infer sensitive data (i.e. the type of disease) by using supposedly anonymised data. For example, suppose we know that our white male colleague has been admitted to hospital recently. Since he is our colleague, we also know that he was born in 1968 and lives in the area with Zip Code 40242. Based on this non-sensitive information, we can use Table 5.2 to infer sensitive information about him, namely that he suffers from heart disease. Although the example presented here is trivial, it does illustrate the limitation of *k*-anonymity in cases where the values of sensitive variables are not diverse.

To overcome this limitation, a number of enhancements of *k*-anonymity have proposed. For example, in one of our previous works (Zielinski, 2007c), we proposed a simple solution that combined *k*-anonymisation with association rule hiding. Given a non-anonymised microdata set, we proposed to first *k*-anonymise it. Thereafter association rule hiding was performed, which ensured that sensitive association rules cannot be

inferred from the anonymised table. During the process of association rule hiding, the values of sensitive variables (i.e. the *Disease* variable in Tables 5.1 and 5.2) of some records were suppressed. This approach ensured that sensitive information could not be inferred when there was a lack of diversity in the values of sensitive variables.

Other enhancements of *k*-anonymity have also been proposed. A critique of these enhancements is presented by Domingo-Ferrer and Torra (2008). An early enhancement of *k*-anonymity was proposed by Machanavajjhala, Gehrke, Kiefer and Venkitasubramanian (2006), and Machanavajjhala, Kiefer, Gehrke and Venkitasubramanian (2007), where the authors introduce the concept of *l*-diversity. A microdata set satisfies the property of *l*-diversity, if for every group of records that have the same value for the identifying variables, there are at least *l* "well-represented" values for the sensitive variable. In the context of *l*-diversity, "well-represented" can refer to:

- *Distinct l-diversity*, where there are at least *l* distinct values for the sensitive variable for every group of records that have the same values for the identifying variables.
- *Entropy l-diversity*, where for each group of a particular value of a sensitive variable, the entropy of the group is greater than or equal to *log*(*l*).
- *Recursive* (*c*, *l*)*-diversity*, where for each group of a particular value of a sensitive variable, those values that seldom occur (in the original microdata set), will occur more frequently (in the anonymised microdata set), and those values that appear often (in the original microdata set) will occur less frequently (in the anonymised microdata set).

However, *l*-diversity is also insufficient to prevent attribute disclosure, as noted by N. Li, T. Li, and Venkatasubramanian (2007). For example, if the values of a sensitive variable are semantically similar, in spite of satisfying *l*-diversity, attribute disclosure can still take place. For example, let us assume that a 3-diverse microdata set may have the following three types of values for a sensitive variable *Disease*: "colon cancer", "lung cancer", and "skin cancer". Then an intruder may still deduce that an individual linked to that group has cancer, although the intruder will not necessarily know what type of cancer the individual has.

Enhancements similar to *l*-diversity have also been proposed. For example, Truta and Vinay (2006) propose *p*-Sensitive *k*-anonymity, which is equivalent to Distinct *l*-diversity. A microdata set satisfies *p*-Sensitive *k*-anonymity if it satisfies *k*-anonymity and where each group of records with the same values for the identifying variables has at least *p* unique values for the sensitive variable. Both *p*-Sensitive *k*-anonymity and Distinct *l*-diversity have the limitation of presuming that the different values of a sensitive variable are assumed with similar frequencies. When this is not the case, achieving *p*-Sensitive *k*-anonymity and Distinct *l*-diversity may reduce the level of information utility significantly. Wong et al. (2006) also presented an approach that is equivalent to Recursive (*c*, *l*)-diversity, called (α, *k*)-anonymity. However, (α, *k*)-anonymity requires that the proportion of each sensitive value in each group is such that $0 \leq \alpha \leq 1$.

Another enhancement of *k*-anonymity is *t*-closeness, proposed by N. Li et al. (2007). A microdata set satisfies *t*-closeness if, for every group of records that have the same values for the identifying variables, the distance between the distribution of the value of the sensitive variable in the group, and the distribution of the value of the sensitive variable in the microdata set itself, is no more than a threshold *t*. Although *t*-closeness overcomes the limitations of *l*-diversity, ensuring the *t*-closeness property significantly reduces the information utility of a released microdata set. That is, when *t*-closeness is achieved, the correlations between the identifying and the sensitive variables are damaged (Domingo-Ferrer & Torra, 2008). This is because, by definition of *t*-closeness, the values of a sensitive variable will have the same distribution for any group of identical values for the identifying variables. The only way to decrease this damage on the correlation between the identifying and the sensitive variables is to relax the *t*-closeness property, by increasing the threshold *t*.

5.5.3 How appropriate is *k*-anonymity (and its enhancements) for addressing the conflict between privacy and information utility

The use of *k*-anonymity is often seen as a "clean way" of addressing the conflict between privacy and information utility (Domingo-Ferrer & Torra, 2005, 2008). It is seen as a "clean way" because, it is assumed that, if for a given *k* value, *k*-anonymity

will provide sufficient privacy, then it allows one to concentrate on only determining how to minimise information loss (or maximise information utility) such that the given level of *k*-anonymity will be achieved. However, we argue that if *k*-anonymity is used in this fashion, then it does not fully capture the objective of the optimisation problem. In this Section, we discuss the way in which *k*-anonymity is currently used to address the conflict between privacy and information utility, based on how it captures the objective and constraints of the optimisation problem (as stated in Section 5.2).

First of all, it is unclear (from the literature stemming from *k*-anonymity) how to determine the optimum value for *k* that will provide "sufficient privacy" for the particular set of circumstances in which anonymisation takes place. Before we can find the optimum value for *k*, we need to know what the optimum balance between privacy and information utility is for the given set of circumstances in which anonymisation takes place. Moreover, under the above assumption (i.e. that a certain *k* value is sufficient), when a certain *k* value is provided as input to anonymisation, it is provided without knowing if the given value will in fact lead to an optimum balance between privacy and information utility.

Under the assumption that a certain *k* value will provide sufficient privacy, the complexity of the optimisation problem is reduced to only maximising information utility when given a certain level of privacy that needs to be achieved (i.e. a *k* value for *k*-anonymity). However, we believe that such an assumption does not take into account the whole complexity of the optimisation problem (as stated in Section 5.2). That is, such an approach does not take into account that it is *both* privacy and information utility that have to be maximised in the optimisation problem.

When the above assumption is used to solve this optimisation problem, maximising privacy is *no longer an objective function of the optimisation problem*. Instead, under the above assumption, privacy is reduced to *only a constraint under which optimisation occurs*. When privacy becomes just a constraint under which optimisation occurs, then the optimisation does not necessarily lead to a truly optimum solution. Information utility is optimised only to satisfy a given level of privacy, rather than being optimised whilst being aware of the fact that the goal of maximising information utility is in direct

conflict with the goal of maximising privacy. In other words, information utility is optimised subject to a given level of privacy that is considered "sufficient".

Nevertheless, the given "sufficient" level of privacy may not necessarily be the optimum level, since the privacy level was decided upon through a means other than during the optimisation itself. This is not to say that, the optimum level of privacy will occur below the required "sufficient" or minimum level. It cannot occur below the minimum level, since otherwise the constraint of the minimum level of privacy would not be met. It is, however, possible that the optimum level of privacy will occur above the required minimum privacy level, but this will not be known unless privacy is optimised as well.

Note that we are not disputing the usefulness of *k*-anonymisation for anonymising microdata. We are, however, stating that when *k*-anonymisation is used to find the optimum balance between privacy and information utility, then the optimisation problem should be approached from both angles: the need to maximise both information utility and privacy. If this problem is approached from both these angles, then during the process of optimisation, the *k* value will actually be *calculated*. First, the optimum balance will be determined. Thereafter, in a second step, the optimum balance will be used to determine how the microdata should be anonymised. If *k*-anonymity is used as the anonymisation technique, then during the second step, the value for *k* will be calculated and then the microdata set will be *k*-anonymised with this value. In other words, the value for *k* will no longer be an input into the optimisation problem. The only input into the optimisation problem will be the constraints under which the optimisation should occur. These constraints are the preferences that were stated in Section 5.2.

It is, of course, also possible to have an input constraint requiring that the released microdata set should have a certain minimum level of privacy. However, even in such cases, the minimum level of privacy should not be expressed in terms of the *k* value, since it is possible to have two microdata sets satisfying the property of *k*-anonymity (with the same *k* value), whilst offering very different levels of actual privacy (i.e. in terms of how easy it is for an intruder to infer sensitive data). Consider for example, the microdata set in Table 5.2. Even though it satisfies the property of *k*-anonymity (with *k*

= 3), its actual level of privacy is much lower (in terms of an ordinal quantification of privacy) than a microdata set that also satisfies *k*-anonymity (with *k* = 3), but where the sensitive variables are different for every record of the microdata set.

The limitation of the way in which *k*-anonymity is used to address the conflict between privacy and information utility, as discussed above, relates to the objectives of the optimisation problem. Another limitation of *k*-anonymity, with regards to how it is currently used to address the conflict between privacy and information utility, is related to the definition of the constraints under which optimisation is performed.

In the original definition of *k*-anonymity, anonymisation is performed without taking into account the data user's preferences between the different identifying variables. Therefore, the anonymisation does not consider that information loss should be minimised in those identifying variables that a data user considers most useful. Some enhancements of *k*-anonymity have addressed this shortcoming, as discussed in the next Section. In a similar way, the original definition of *k*-anonymity also disregards the preferences between identifying variables that we perceive a potential intruder may have. That is, anonymisation does not necessarily ensure that the most information loss occurs in those identifying variables that are (perceived) to be most useful for a potential intruder. Furthermore, *k*-anonymity also does not take into account the preference between privacy and information utility. When we need to determine the optimum balance between privacy and information utility, these preferences should be taken into account as constraints under which the optimisation is performed. However, the original *k*-anonymity definition does not take these into account.

To summarise, although *k*-anonymity shows potential as a good way to address the conflict between privacy and information utility, we argue that the way in which it is currently used is not appropriate to address this conflict. That is, the way in which *k*-anonymity is currently used fails to find a truly optimum balance between privacy and information utility for two main reasons. The first reason relates to the way in which the objective of the optimisation problem is defined. That is, the objective of the optimisation problem focuses on only maximising information utility, such that a certain level of privacy (*k* value) is met. To find the optimum balance between privacy and information utility, the objective of the optimisation should focus on maximising

both privacy and information utility. The second reason relates to the way in which the constraints of the optimisation problem are defined. That is, the preferences between privacy and information utility, as well as the data user's preferences and the data owner's preferences between identifying variables are not taken into account when optimisation is performed.

5.5.4 Specific examples of how *k*-anonymity has been used to address the conflict between privacy and information utility

In this Section, we present a number of specific examples of how *k*-anonymity has been recently used to address the conflict between privacy and information utility.

Stark, Eder and Zatloukal (2006) propose a priority-driven anonymisation technique to achieve *k*-anonymity. The proposed technique allows specifying the degree of acceptable information loss for each variable seperately. Variables that are considered useful for the data user can be protected from extensive generalization. Those variables that have been assigned low priorities are generalized first. Variables that have been assigned higher priorities are only generalized when no other solution may be found to achieve *k*-anonymity. Although this approach is able to take into account the data user's preferences with respect to which variables will be useful to him, it is unable to take into account other constraints of the optimisation problem, namely the data owner's preferences between variables (from the perspective of a potential intruder) and also the preferences between privacy and information utility. Moreover, the optimisation problem is addressed by considering only the need to maximise information utility such that a certain level of *k*-anonymity is provided.

Other utility-based anonymisation approaches were also proposed. For example, LeFevre, DeWitt and Ramakrishnan (2006b) propose algorithms that will generate anonymous data such that the utility of the data is preserved with respect to the workload for which the data will be used. Xu et al. (2006) also study the problem of utility-based anonymisation and present a framework to specify the utility of variables. Zhang, Jajodia and Brodsky (2007) propose a model and an algorithm that will guarantee safety under the assumption that the intruder knows the disclosure algorithm

and the generalization sequence. Nevertheless, these works address the conflict between privacy and information utility from only one angle, namely the need to maximise information utility subject to a given *k* value (i.e. a level of privacy that is considered as "sufficient"). As we argued in the previous Section, considering the optimisation problem from this limited perspective does not lead to a truly optimum balance between privacy and information utility

In a more recent work, Gionis and Tassa (2009), study how to achieve *k*-anonymity with minimal loss of information (i.e. an optimum *k*-anonymisation). The authors provide an improvement on the best-known $O(k)$-approximation provided by Aggarwal et al. (2005) to an approximation of $O(\ln k)$. Nevertheless, the authors also do not consider the optimisation problem from the perspective of maximising both privacy and information utility. Instead, they aim to determine how to achieve *k*-anonymity such that information utility is maximised. That is, the algorithm proposed expects that the value for *k* will be provided as input. However, as we argued in the previous Section, if we are to obtain a truly optimum balance between privacy and information utility, by using *k*-anonymisation as the anonymisation technique, then the value for *k* should actually be calculated during the optimisation process.

Loukides and Shao (2008) consider how a *k*-anonymisation can be produced with an optimum trade-off between information utility and privacy. In that paper, the needs of both privacy and information utility are considered. The optimisation problem is addressed from both these angles when an optimal anonymisation is determined. However the proposed measure for information utility is based on the average amount of generalizations that each group of tuples incurs – the smaller this number, the higher the utility. This proposed measure does not consider the preferences that a specific data user may have between different identifying variables. Therefore, this measure will not be able to take into account the purpose for which the data user requires the data and hence does not provide a meaningful measure for information utility. Therefore, an anonymised microdata set will not necessarily have the optimum level of information utility for a specific data user and the purpose for which the data is released.

Although a number of approaches based on *k*-anonymity have been proposed to address the conflict between privacy and information utility, none are able to find a truly

optimum balance between privacy and information utility. The concept of *k*-anonymity itself is also currently being used inappropriately to address this conflict. Nor do we consider the *score* and R-U confidentiality maps to be appropriate. In the next Section, we present recommendations for an appropriate solution that will ensure that the optimum balance between privacy and information utility is achieved when microdata is anonymised.

5.6 Recommendations for an appropriate solution

When we consider the definition of the "optimum" balance between privacy and information utility provided in Section 5.2, it is clear that the current approaches for addressing the conflict between privacy and information utility are not appropriate. We now provide recommendations for a solution that will be appropriate for determining the optimum balance between privacy and information utility. The recommendations will be used as a basis for developing our solution, which is presented in Chapters 6 and 7.

To reiterate from Section 5.2, we consider the "optimum" balance between privacy and information utility as one in which the levels of privacy and information utility are maximised while satisfying a set of constraints that capture the data owner's and the data user's preferences. We stated our optimisation problem as follows "Maximise privacy and information utility subject to the constraints imposed by the data user's and the data owner's preferences".

We argue that if we are to find a truly optimal balance between privacy and information utility, then the goal of maximising *both* privacy and information utility should be regarded as *the objective function of the optimisation problem*. This stems from the fact that both privacy and information utility are desired, although they may be desired in different proportions. This is our recommendation with respect to the objective of the optimisation problem.

We also need to make recommendations that address the constraints under which optimisation should be carried out. These constraints should reflect the preferences

between privacy and information utility. The constraints should also reflect the data user's and the data owner's preferences between identifying variables. In the case of the data owner, the preferences between identifying variables should be considered from the perspective of a potential intruder (i.e. in terms of which identifying variables are considered to be most useful for an intruder in deriving sensitive data).

Therefore, a challenge exists to develop a solution that will appropriately capture the above objective and constraints and thereafter find the optimum balance between privacy and information utility. Moreover, once the optimum balance has been determined, the solution should also determine how to anonymise the microdata set such that the optimum levels are achieved. Therefore, the solution should have two components: an optimisation component, in which the optimum levels of privacy and information utility are determined, and an anonymisation component, during which the microdata set is anonymised.

In this research work, we propose such a solution. In Chapter 6, we address the optimisation aspect of the solution through the application of Economic Price Theory. The anonymisation aspect of the solution is addressed in Chapter 7, in which two anonymisation techniques are considered: global recoding and microaggregation.

5.7 Conclusion

When microdata is anonymised, it needs to satisfy two conflicting goals: privacy and information utility. In this Chapter, we discussed the conflict between privacy and information utility and also discussed the appropriateness of existing approaches for addressing the conflict. We argued that current approaches are not completely appropriate for finding the optimum balance between privacy and information utility and concluded with recommendations for an appropriate solution. These recommendations will be used in the next two Chapters.

CHAPTER 6

HOW TO DETERMINE THE OPTIMUM LEVELS OF PRIVACY AND INFORMATION UTILITY

6.1 Introduction

In the previous Chapter, we concluded that current approaches for balancing privacy and information utility in microdata anonymisation are not appropriate, since they do not necessarily lead to an *optimum* balance. We then made recommendations for an appropriate approach for balancing privacy and information utility. Specifically, our recommendation with respect to the objective of the optimisation problem was that both privacy and information utility should be maximised. With regards to the constraints under which optimisation should be carried out, we recommended that we should take into account the preferences between privacy and information utility as well as the data user's and the data owner's preferences between identifying variables.

In this Chapter, we use the recommendations from the previous Chapter to propose an approach for a microdata anonymisation process that will anonymise microdata such that optimum levels of privacy and information utility are achieved. The proposed approach is based on Economic Price Theory and hence we first describe how concepts from Economic Price Theory are applied in our approach. Thereafter, we present the proposed microdata anonymisation process. We evaluate the proposed microdata anonymisation process through a simulation that will show how the input preferences, upon which the optimum levels are determined, impact the optimum levels of privacy and information utility. In this Chapter, we focus on how to determine the optimum levels of privacy and information utility. In the next Chapter, we will describe how microdata should be anonymised such that the optimum levels are achieved.

6.2 Economic Price Theory

Economic Price Theory (also known as Microeconomics) uses three types of analytical tools. These include: constrained optimisation, equilibrium analysis, and comparative statics (Besanko & Braeutigam, 2005). Constrained optimisation is used to determine the optimum choice when given a set of limitations on how the choice can be made. As an example, the *utility maximisation problem of a consumer* uses constrained optimisation, in which we determine how a consumer should allocate his income among different goods in order to maximise his utility (or satisfaction) gained from consuming the goods. Equilibrium analysis is used to determine the conditions under which the market will reach equilibrium. For example, in supply and demand analysis, equilibrium analysis is used to determine how the supply and demand of a good determine the price and quantity of a good provided in the market. As another example, General Equilibrium Analysis aims to determine the equilibrium prices and quantities of goods of more than one market at a time. Finally, comparative statics are useful for examining how changes in external variables affect the state of internal variables in an economic system.

In our research work, we aim to determine how best to allocate the available information in a microdata set between the released information (i.e. information utility) and the hidden information (i.e. privacy), so as to maximise the joint benefit (economic utility) of the data user and the data owner. This closely corresponds to the *utility maximisation problem of a consumer*, because it is related to the theory of choosing: how much does an individual want to consume of each good? Therefore, in this research work, we use only one tool from Economic Price Theory, namely constrained optimisation.

In this Section, we provide a brief overview of the concepts from Economic Price Theory that were used as a basis for developing the solution to our research problem. This Section is based on the works of Besanko and Braeutigam (2005); Dixit (1990), J. Hirshleifer et al. (2005); and Mansfield (1985). Readers interested in more detail are referred to these works for more information on Economic Price Theory.

Let us suppose that a consumer has a particular *income I* that he wishes to spend on two *goods*, x_1 and x_2. Let us also suppose that the *prices* for x_1 and x_2 are p_1 and p_2, respectively. When the consumer spends all the income I on the purchase of the two goods x_1 and x_2, we can represent this as an equation known as the *consumer's budget*: $p_1x_1 + p_2x_2 = I$.

A consumer derives a certain amount of satisfaction from consuming different goods. This satisfaction is referred to as the consumer's (economic) *utility* (Mansfield, 1985). The utility derived from consuming the two goods x_1 and x_2 is determined by the consumer's own preference between the two goods, and can be represented by the utility function $U(x_1, x_2)$.

To find the *consumer's optimum point of consumption* of goods x_1 and x_2, we need to maximise the amount of (economic) utility derived from consuming these two goods subject to the constraints imposed by the consumer's budget (i.e. the constraints imposed by the prices for goods and by the consumer's income) and subject to the constraint that only non-negative amounts of goods can be consumed. This optimisation problem can be summarised as follows:

$$\max U(x_1, x_2) \qquad (3)$$

$$\text{s.t.} \quad p_1x_1 + p_2x_2 = I \text{ (Budget constraint)}$$

$$x_1, x_2 \geq 0$$

If the consumer is faced with the possibility of consuming more than two goods, we can generalise the optimisation problem as follows. Suppose that there are n goods $x_1, x_2, \ldots, x_n$. To find the consumer's optimum, we need to maximise the utility function $U(x_1, x_2, \ldots, x_n)$ subject to the constraint of the consumer's budget $p_1x_1 + p_2x_2 + \ldots + p_nx_n = I$, and also subject to non-negative consumption of goods. This can be written as:

$$\max U(x_1, x_2, \ldots, x_n) \qquad (4)$$

$$\text{s.t.} \quad p_1x_1 + p_2x_2 + \ldots + p_nx_n = I \text{ (Budget constraint)}$$

$$x_1, x_2, \ldots, x_n \geq 0$$

This optimisation problem can be solved by using the Lagrange Multipliers Method (Dixit, 1990; Bertsekas, 1982), which allows us to find the extrema of a function subject to one or more constraints.

To find the extrema points of the function *U*, we introduce a new function, the *Lagrangian*, in terms of the function *U*, the budget constraint, as well as the Lagrange multiplier λ. In this case, the Lagrangian (*L*) is defined as:

$$L(x_1,x_2,\ldots,x_n,\lambda) = U(x_1,x_2,\ldots,x_n) + \lambda[I - p_1x_1 - p_2x_2 - \ldots - p_nx_n] \quad (5)$$

At the optimum point, there are *first order conditions* that need to be satisfied:

$$\frac{\partial L}{\partial x_1} = 0 \; ; \; \frac{\partial L}{\partial x_2} = 0 \; ; \ldots ; \; \frac{\partial L}{\partial x_n} = 0$$

$$\frac{\partial L}{\partial \lambda} = 0$$

Based on these, we can obtain the optimum values for $x_1, x_2, \ldots, x_n$, by treating the conditions as (*n* + 1) equations in the (*n* + 1) unknowns $x_1, x_2, \ldots, x_n, \lambda$.

6.3 Quantification of information

A suitable way to quantify information is fundamental to our proposed solution. For this purpose, we use the concept of entropy (Shannon, 1948). Entropy provides us with a measure of the uncertainty of a random variable.

Given a random variable *X* in a data set, the entropy *H(X)* is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \, log(p_i) \quad (6)$$

where:

- *n* is the number of possible values that the variable *X* can assume in the particular dataset, and
- p_i is the probability that the variable *X* assumes the *i*-th possible value, in the particular dataset.

6.4 Quantification of Information Utility

The concept of information utility is subjective: what may be considered as useful information in one context could be considered not useful in another. Therefore, information utility is difficult to quantify because it must take into account a data user's preferences, which may vary according to what the user intends to use the data for.

In this research work, we quantify *Information Utility* (*Iu*) as the extent to which data is useful for a data user, for the purpose for which it is released. (Recall that, in this thesis, the data user is not the person who contributes the data, but is rather the person or organisation that will use the released data.) To measure Information Utility, we take into account the amount of data that is released for each identifying variable, as well as the usefulness of each released identifying variable for the user. The data user's preferences, with respect to the variables that will be useful for him, are used as a basis for measuring Information Utility. Therefore, Information Utility of the same data set will be calculated differently for each user of the data, based on each data user's needs and preferences.

In this research work, we propose that the Information Utility of an identifying variable is calculated as the product of the usefulness of that identifying variable to the data user and the amount of information entropy that is present when the identifying variable is released in the anonymised microdata.

To specify the usefulness of identifying variables, the data user is provided with 100 User Preference Points, which must be distributed over the identifying variables. The distribution of points is done in a similar way as has been proposed by Willenborg and De Waal (2001). Each identifying variable V_i is assigned a number of User Preference Points q_i such that:

- Those variables that are considered more useful by the data user are assigned more points over those variables that are considered to be less useful. Therefore, the more useful the variable is to the data user, the more points the data user should allocate to that variable.

- The sum of the User Preference Points allocated to each identifying variable must equal 100.

To calculate the amount of information that is released for a particular identifying variable, we use information entropy, as has been explained in the previous Section. However, an identifying variable can be anonymised to different degrees with a particular anonymisation technique. Therefore, this can lead to different amounts of information that can be released for the same identifying variable, depending on the degree to which the variable is anonymised. Therefore, when we calculate the amount of information in a particular identifying variable, we also need to specify the degree of anonymisation that has been applied to the variable.

To allow us to distinguish between the different degrees to which an identifying variable has been anonymised (with a particular anonymisation technique), we propose the notation $(V_i)^j$, where *i* indicates the number of the identifying variable, and *j* indicates the degree of anonymisation that has been applied to the variable.

What is meant by the "degree of anonymisation" depends on the particular anonymisation technique itself. For example, if the MDAV microaggregation algorithm (Domingo-Ferrer et al., 2006; Hundepool et al., 2005) is used to anonymise a particular identifying variable, then the degree of anonymisation will refer to the least number of records *k* that should exist in each cluster. In this case, the superscript *j* will be used to refer to the number of records *k* that exist in each cluster. Or, for example, if global recoding is used, then the superscript *j* will refer to the coding with which the variable is released.

More formally, if an identifying variable V_i has been anonymised (with a particular anonymisation technique) to the *j*-th degree, then the Information Utility of the released variable, $Iu\left((V_i)^j\right)$, is the product of the User Preference Points q_i allocated to the variable V_i and the amount of information entropy $H\left((V_i)^j\right)$ in the variable when it is anonymised to the *j*-th degree. That is,

$$Iu\left((V_i)^j\right) = q_i\, H\left((V_i)^j\right) = q_i\left(-\sum_{k=1}^{n} p_k\, log(p_k)\right) \quad (7)$$

where:

- n is the number of possible values that the variable V_i can assume (in the microdata) when the variable has been anonymised with a particular anonymisation technique to the j-th degree, and
- p_k is the probability that the variable V_i assumes (in the microdata) the k-th possible value when it has been anonymised with a particular anonymisation technique to the j-th degree.

When assigning User Preference Points to identifying variables, the data user should also take into account the meta-knowledge about a microdata set. For example, if the data user is given a subset of a microdata set that contains data only on children, then an identifying variable such as Martial Status is likely to be of little use. On the other hand, the Marital Status identifying variable can be quite useful if the subset contains data on adults (or adults and children). Therefore, the meta-knowledge about the microdata set should be considered to ensure that User Preference Points are assigned appropriately to identifying variables.

The interaction effect between identifying variables should also not be overlooked when assigning User Preference Points to identifying variables. Sometimes when, amongst all the identifying variables of a released microdata set, there are certain combinations of identifying variables, then the usefulness of those identifying variables may be greater when compared to cases where those combinations are not present.

For example, when given a subset of a microdata set about female patients who were admitted to a hospital, the usefulness of an identifying variable such as Age can be increased when it is released together with a variable such as the Number of Children of a patient. In this case, it may provide insight into the diseases suffered by mothers, in different age groups, as opposed to the diseases suffered by women in general. Moreover, the presence of the Number of Children identifying variable can now help

the data user make deductions about diseases suffered by underage mothers. This type of deduction would not be possible if the Number of Children identifying variable were not present. Hence, the value of the Age identifying variable may be greater to the data user when it is released together with the Number of Children variable. In such cases, the Age identifying variable should be assigned a greater number of User Preference Points.

6.5 Quantification of Privacy

The concept of privacy is also subjective: what may be considered as a sufficient level of privacy in one context could be considered as insufficient in another. Therefore, privacy is also difficult to quantify because it must take into account not only the sensitivity of the data, but also the ease with which an intruder will be able to infer sensitive data; both depend on the content of the data with which we are working.

We argue that one way in which privacy can be achieved is through the removal of data from the identifying variables. As more useful data (from the intruder's perspective) remains unreleased due to data anonymisation, it becomes increasingly difficult for the intruder to infer sensitive data. Therefore, we can think of privacy in terms of the amount of data that is removed from the identifying variables, taking into account the usefulness of each identifying variable in inferring sensitive data. As more data is removed from the identifying variables, the privacy level of the released data increases.

We will use this notion to quantify *Privacy* as the extent to which the unreleased, or hidden, data would be useful for an intruder for the purpose of inferring sensitive data. The greater the usefulness of the unreleased data for an intruder, the greater the privacy. Since we are measuring the usefulness of data, we can adapt our definition of Information Utility, to measure how useful the unreleased data would be for the intruder. Therefore, we define *Privacy* (*Priv*) as a measure of how useful the unreleased data would be for an intruder for the purpose of inferring sensitive data.

Although thinking about privacy in terms of the amount of data removed from useful identifying variables does not provide us with a cardinal quantification for privacy, it is

nevertheless useful for our research purpose as it provides us with an ordinal quantification. That is, we know that if a microdata set has less data for useful identifying variables (i.e. the variables that are useful for inferring sensitive data), it must have a higher privacy level than a microdata set with more data for useful identifying variables (ordinal quantification), although we are unable to determine the exact amount (cardinal quantification) of privacy that the two microdata sets possess.

In a similar way as for calculating Information Utility, the Privacy level of a variable is calculated as the product of the usefulness of that variable to the intruder and the amount of information entropy that is hidden (unreleased) when the variable is released in the anonymised microdata. The amount of information that is hidden is calculated as the difference in the amount of information present in the non-anonymised variable and the amount of information present in the released variable.

To specify the usefulness of identifying variables, we distribute 100 Intruder Preference Points over the identifying variables. The distribution of points is done in a similar way as has been proposed by Willenborg and De Waal (2001). Similarly to as in the case of quantifying Information Utility, when assigning Intruder Preference Points, the interaction of identifying variables should also be taken into account.

Each identifying variable V_i is assigned a number of Intruder Preference Points r_i such that:

- Those variables that we consider as being more useful to the intruder in inferring sensitive data are assigned more points over those variables that are considered to be less useful. Therefore, the more useful a variable is perceived to be to the intruder, the more points we should allocate to that variable.
- The sum of the Intruder Preference Points allocated to each identifying variable must equal 100.

More formally, if an identifying variable V_i is anonymised (with a particular anonymisation technique) to the *j*-th degree (as explained in the previous Section), then the Privacy of the released variable, $Priv\left((V_i)^j\right)$, is the product of the Intruder

Preference Points r_i allocated to the variable V_i and the difference $H(V_i) - H\left((V_i)^j\right)$. That is,

$$Priv(V_i)^j = r_i\left(H(V_i) - H\left((V_i)^j\right)\right)$$
$$= r_i\left(-\sum_{k=1}^{n} p_k \log(p_k) - \left(-\sum_{l=1}^{m} s_l \log(s_l)\right)\right) \quad (8)$$

where:

- n is the number of possible values that the non-anonymised variable V_i can assume (in the non-anonymised microdata),
- p_k is the probability that the non-anonymised variable V_i assumes the k-th possible value,
- m is the number of possible values that the variable V_i can assume when it has been anonymised (with a particular anonymisation technique) to the j-th degree, and
- s_l is the probability that the variable V_i assumes (in the microdata) the l-th possible value when it has been anonymised (with a particular anonymisation technique) to the j-th degree.

Without knowing the goals of the intruder, we are unable to predict the likely method of attack that an intruder may use. Hence, we are unable to assign Intruder Preference Points to variables in a way that would reflect the importance of the variables for the specific attack methods likely to be used by the intruder. This is made more difficult since there may be several types of intruders with different interests. Therefore, determining the usefulness of an identifying variable to an intruder (and hence the way in which privacy has been quantified) is limited to how well we can predict the behaviour of an intruder.

Nevertheless, we believe that we can make several assumptions about how useful an identifying variable will be to an intruder. Firstly, an intruder may attempt to match a microdata set with other microdata sets that are available to him, for example voter registration lists. Two microdata sets can be matched when there is one or more identifying variable that is common between the two microdata sets. When an

identifying variable appears in many microdata sets, its usefulness to the intruder is high, since the intruder can attempt to match many microdata sets on the common identifying variable. On the other hand, if an identifying variable is not likely to appear in other microdata sets, then its usefulness to the intruder may be lower, since the identifying variable can be used to match fewer microdata sets. Therefore, the higher the frequency of the identifying variable in different microdata sets, the greater the usefulness of the identifying variable. Hence, common identifying variables should be assigned more Intruder Preference Points over those identifying variables that are less common.

Secondly, we can make assumptions about the usefulness of an identifying variable based on the meta-knowledge about the identifying variables (i.e. based on knowledge about the identifying variables themselves and not on the content of the variables). For example, in a microdata set that contains information about the rural population of a region, the Zip Code may be considered to be more useful than in a microdata set that contains information about an urban population. This stems from the fact that rural areas are sparsely populated and hence a microdata set about a rural population may contain a greater number of relatively rare or even unique records per one unique Zip Code, when compared to a microdata set about an urban population. Hence, it may be easier for an intruder to use the Zip Code to re-identify individuals living in a rural area when compared to individuals living in an urban area. Therefore, meta-knowledge about identifying variables should also be taken into account when assigning Intruder Preference Points to identifying variables.

6.6 Using Economic Price Theory in Information Theory

To apply concepts from Economic Price Theory to find the optimum balance between privacy and information utility, we adapt the definitions from Economic Price Theory so that they are relevant for our research work as follows.

Consumer

For the purpose of this research work, we encapsulate the needs and preferences of both the data owner and the data user into the needs of one entity, known as the consumer in Economic Price Theory. The data owner primarily seeks to maximise privacy, while the data user primarily seeks to maximise the usefulness of the data. Therefore, the consumer seeks to maximise both the privacy and the information utility of the released data.

Goods

The optimisation problem in our research work can be thought of as maximising the amount of information utility and privacy that will exist in the anonymised microdata set (under a set of constraints). Hence, we refer to information utility and privacy as our (economic) goods, whose consumption we wish to optimise.

In simple terms, the optimisation problem can be seen as a process during which we need to determine, for each identifying variable, how much information needs to be released and how much information needs to be hidden. Therefore, for each released identifying variable, we need to measure two aspects: the resulting level of Information Utility and the resulting level of Privacy. Therefore, when given a microdata set with n identifying variables, we have $2n$ goods: n goods that will form the total level of Information Utility and n goods that will form the total level of Privacy.

Income

In Economic Price Theory, the consumer is given a particular income that must be distributed on the purchase of goods. In a similar way, in our research problem, we are given a particular amount of information entropy that exists in the identifying variables of a particular microdata set. The amount of information entropy that is available in the identifying variables of a non-anonymised microdata set determines the maximum information that can be released or hidden. That is, the amount of information entropy must be distributed between the entropy of the released information (i.e. information utility) and the entropy of the hidden information (i.e. privacy). Therefore, the income

in our research problem is the information entropy that exists in the identifying variables of the non-anonymised microdata. The income is equal to the sum of the entropies of the individual non-anonymised identifying variables. (This sum of entropies does not necessarily equal the joint entropy of the individual non-anonymised identifying variables. In Step 4 in Section 6.8, we explain why the income should be equal to the sum of the entropies of the individual non-anonymised identifying variables, as opposed to being equal to the joint entropy of the non-anonymised identifying variables.)

Prices

In economics, the price for a good is equal to the amount of money (income) required to obtain one unit of that good. By using this concept, the price for information utility of a particular identifying variable is the amount of information entropy (our income) that is required to obtain one unit of Information Utility (a good). In a similar way, the price for privacy of a particular identifying variable is the amount of information entropy (income) that is required to obtain one unit of Privacy (a good). The way in which the actual price amounts are determined is further explained in Step 3 in Section 6.8.

(Economic) Utility

As noted in Section 6.2, the amount of (economic) utility that is derived from consuming goods is determined by the consumer's preference between the different goods. In the case of the $2n$ goods defined above, this preference refers to the preferences that exist between each identifying variable in the microdata set, as well as the preference between the resulting levels of privacy and information utility. The (economic) utility, in our research problem, is the joint benefit that the data user and the data owner (i.e. our "consumer") derive from a released microdata set.

6.7 ANOPI

In this Section, we present our solution in terms of a microdata anonymisation process. We called the microdata anonymisation process ANOPI, or the ANonymisation with Optimum Privacy and Information utility. The purpose of ANOPI is to anonymise microdata by guiding the anonymisation process such that the optimum balance between privacy and information utility is obtained. The process is shown in Figure 6.1 and its two functions are described below.

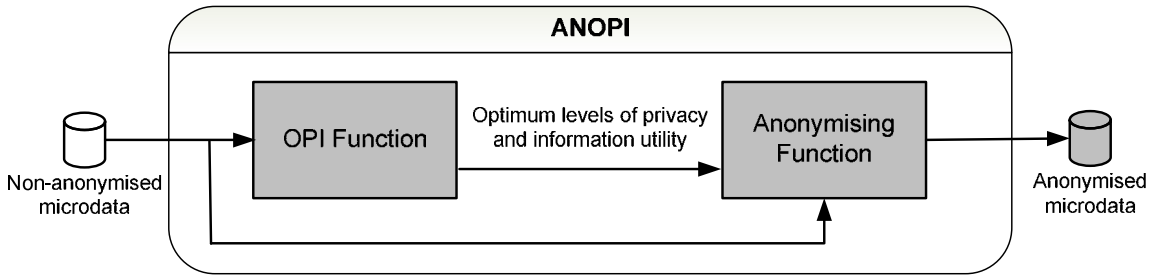


Figure 6.1 A representation of the ANOPI microdata anonymisation process

Given a non-anonymised microdata set, the ANOPI microdata anonymisation process first determines the optimum levels of privacy and information utility based on concepts from Economic Price Theory. This occurs in the OPI function (Optimum Privacy and Information utility function). Data anonymisation takes place in the Anonymising Function, which provides an anonymised microdata set as its output. The OPI function is described in the next Section, while the next Chapter describes the Anonymising function.

6.8 The OPI function

The OPI function (Optimum Privacy and Information utility function) is the function in which the optimum levels of privacy and information utility are determined. It only determines the optimum levels of privacy and information utility that a microdata set should possess, but it does not specify how the microdata should be anonymised to achieve the optimum levels. Therefore, the OPI function is independent of the anonymisation technique used to anonymise microdata.

The input to this function is a non-anonymised microdata set with n identifying variables $V_1, \ldots, V_n$. For each identifying variable V_i, we create two goods: x_i and y_i. For an identifying variable V_i, good x_i represents the variable's Information Utility level and good y_i represents the variable's Privacy level. Therefore, given a microdata set with *n* identifying variables, we have a set of 2*n* goods: $\{x_1, \ldots, x_n, y_1, \ldots, y_n\}$. The OPI function is performed in five steps as follows.

Step 1: Determine the consumer's preferences between different goods

This step is required to determine the preference between the need for privacy and the need for information utility. In this step, we also determine the preference between each identifying variable. These preferences are used to determine the amount of (economic) utility that can be derived.

To set the preference between information utility and privacy, we use the parameters α and β. The values of these parameters should be agreed on together by the data user and the data owner. The sum of the values set for α and β should be equal to 1. The difference between the values therefore reflects the difference in the importance of information utility and privacy of the released microdata. If information utility of the released microdata is seen (by both the data user and the data owner) as being more important than the privacy of the released microdata, then α should have a greater value than β. In that case, it implies that more (economic) utility is derived from information utility than from privacy, and the optimum balance between privacy and information utility will be determined based on this preference. In a similar way, if privacy of the released microdata is perceived as being more important than the usefulness of the released microdata, then α should have a smaller value than β. If the needs of privacy and information utility are equally important, then the value for α should equal to the value for β.

To set the preference between identifying variables, each variable V_i is allocated User Preference Points, q_i, and Intruder Preference Points, r_i. The way in which these points should be allocated has been explained in Sections 6.4 and 6.5, respectively.

Once we have set these preferences, we can determine the amount of (economic) utility that the consumer can derive. In order to do this, we use the Cobb-Douglas Utility Function (Besanko & Braeutigam, 2005). We use the Cobb-Douglas Utility Function because of its usefulness for measuring the consumer's (economic) utility that is derived from different goods.

Firstly, the function ensures that the marginal utility for every good is positive. That is, it ensures that the consumer always prefers more (as opposed to less) of each good. In our case, it ensures that, from each identifying variable, the data user and the data owner always prefer to derive the highest levels of information utility and privacy, respectively.

Secondly, the Cobb-Douglas Utility Function also ensures a diminishing marginal rate of substitution. In other words, it ensures that as the consumption of a particular good increases, the consumer is likely to give up further consumption of that good in order to consume more of the other available goods. For example, let us suppose that a consumer can drink only tea and coffee. After consuming many cups of tea, the satisfaction the consumer will gain from one more cup of tea is likely to be less than the satisfaction gained from consuming a cup of coffee. Therefore, after the consumer has consumed many cups of tea, he will likely give up an additional cup of tea in order to consume a cup of coffee. In the case of our research problem, the situation is similar. When the level of information utility of an identifying variable is already high, we are likely to consider increasing the level of privacy (instead of information utility) that can be derived from that variable.

When the Cobb-Douglas Utility Function is used in the context of our research work, the (economic) utility function *U* in this optimisation problem is derived as follows:

$$U(x_1,\ldots,x_n,y_1,\ldots,y_n) = \left(\prod_{i=1}^{n} x_i^{\frac{q_i}{100}}\right)^{\alpha}\left(\prod_{i=1}^{n} y_i^{\frac{r_i}{100}}\right)^{\beta} \quad (9)$$

where:

- α is the preference value for information utility,
- β is the preference value for privacy,
- q_i is the number of User Preference Points allocated to variable V_i, and
- r_i is the number of Intruder Preference Points allocated to variable V_i.

The amount of (economic) utility derived from a good that contributes to Information Utility is determined by the preference for information utility as well as the number of User Preference Points allocated to that variable. In a similar way, the amount of (economic) utility derived from a good that contributes to Privacy is determined by the preference for privacy as well as the number of Intruder Preference Points allocated to that variable. Note that we divide the User Preference Points and Intruder Preference Points by 100 to ensure that we represent the respective preferences as weights.

Step 2: Determine the consumer's income

To determine the consumer's income, *I*, we need to determine the amount of information with which we are presented in the identifying variables of the non-anonymised microdata. In order to do this, we calculate the sum of the information entropies that exist in each non-anonymised identifying variable of the microdata set. (We do not use the joint entropy $H(V_1,\ldots,V_n)$ of the identifying variables for calculating the income, as explained in Step 4.)

More formally, given a non-anonymised microdata set with n identifying variables $V_1,\ldots,V_n$, the consumer's income I is:

$$I = \sum_{i=1}^{n} H(V_i) \tag{10}$$

Step 3: Determine the price of each good

For each good x_i (i.e. those goods that represent the Information Utility level of a variable V_i), we determine its price s_i. Similarly, for each good y_i (i.e. those goods that represent the Privacy level for a variable V_i), we determine its price t_i.

To determine the price s_i for a good x_i, we need to determine the amount of information that is required such that the data user obtains one unit of Information Utility. Given an anonymisation technique, every possible degree of anonymisation of the variable V_i will yield the same price. Therefore, when we calculate the price, we can assume that the variable V_i has been anonymised to the j-th degree (taking into account the maximum value that j may have for a particular anonymisation technique with which the variable V_i is anonymised).

To calculate the price, we use (1) the amount of Information Utility that the user will obtain if V_i is anonymised to the j-th degree (with a certain anonymisation technique), and (2) the amount of information entropy $H\left((V_i)^j\right)$ that will exist in the anonymised V_i when it is anonymised to the j-th degree. The price s_i for a good x_i is calculated as follows (where q_i is the amount of User Preference Points allocated to V_i as described in Section 6.4):

$$s_i = \frac{H\left((V_i)^j\right)}{Iu\left((V_i)^j\right)} \tag{11}$$
$$= \frac{H\left((V_i)^j\right)}{q_i\, H\left((V_i)^j\right)}$$
$$= \frac{1}{q_i}$$

In a similar way, we can also determine the price t_i for a good y_i. In this case, we need to determine the amount of information that is required such that one unit of Privacy is obtained. To calculate the price t_i for a good y_i, we use (1) the amount of Privacy that will be obtained if V_i is anonymised (with a certain anonymisation technique) to the *j*-th degree, and (2) the amount of information entropy $H(V_i) - H\left((V_i)^j\right)$ that is lost from the variable when it is anonymised. The price t_i for a good y_i is calculated as follows (where r_i is the amount of Intruder Preference Points allocated to V_i as described in Section 6.5):

$$t_i = \frac{H(V_i) - H\left((V_i)^j\right)}{Priv\left((V_i)^j\right)} \tag{12}$$
$$= \frac{H(V_i) - H\left((V_i)^j\right)}{r_i\left(H(V_i) - H\left((V_i)^j\right)\right)}$$
$$= \frac{1}{r_i}$$

Step 4: Determine the consumer's budget

We determine the consumer's budget by using the income *I* obtained in Step 2, as well as the prices derived in Step 3. Using these, the consumer's budget becomes:

$$s_1x_1 + \ldots + s_nx_n + t_1y_1 + \ldots + t_ny_n = I \tag{13}$$

Since, in our research work, information can only be released or hidden, it implies that the *whole* value of the total information entropy must be allocated to either information utility (released information) or privacy (hidden information). Hence, it implies that the whole income must be spent and therefore the above formula is not an inequality.

Given the above formula for the consumer's budget, we can now explain why the income I should be equal to the sum of the entropies of the individual identifying variables, as opposed to being equal to the joint entropy of the identifying variables.

Let us assume that we would like to release the non-anonymised microdata. In this case, the budget is:

$$\frac{1}{q_1}x_1+\ldots+\frac{1}{qn}x_n+\frac{1}{r_1}y_1+\ldots+\frac{1}{r_n}y_n=I \tag{14}$$

Since each x_i and y_i are the Information Utility and Privacy levels of V_i, respectively, the budget equation can expanded as follows (by using Equations 7 and 8):

$$\begin{aligned}&\frac{1}{q_1}\cdot q_1 H(V_1)+\ldots+\frac{1}{q_n}\cdot q_n H(V_n)+\\&\frac{1}{r_1}\cdot r_1\left(H(V_1)-H(V_1)\right)+\ldots+\frac{1}{r_n}\cdot r_n\left(H(V_n)-H(V_n)\right)=I\end{aligned} \tag{15}$$

By simplifying, we obtain:

$$H(V_1)+\ldots+H(V_n)=I \tag{16}$$

Therefore, the income I is equal to the sum of the entropies of the individual identifying variables.

We also know that $H(V_1,\ldots,V_n)\leq H(V_1)+\ldots+H(V_n)$, with equality holding if and only if the identifying variables are independent. Therefore, if we set the income to equal to the joint entropy, then the income is equal to the sum of the entropies of the individual variables if and only if the variables are independent. That is, equality will hold if and only if the joint probability of the values of variables $V_1,\ldots,V_n$ is equal to

the product of the probabilities of the values of each individual variable $V_1, \ldots, V_n$. More formally, equality will hold if and only if $Pr(V_1, \ldots, V_n) = Pr(V_1) \cdot \ldots \cdot Pr(V_n)$, where $Pr(V_i)$ refers to the probability of the values of variable V_i.

However, if the identifying variables of a given microdata set are not independent, and we would set the income to equal to the joint entropy of the identifying variables, then we would not be able to release a non-anonymised microdata set, since it would not be possible to satisfy the budget equation. To explain this further, let us consider the following scenario.

We are given a microdata set (with non-independent identifying variables) and we would like to release it such that each identifying variable is not anonymised. Let us assume that we set the income to equal to the joint entropy of the non-anonymised identifying variables. That is, $I = H(V_1, \ldots, V_n)$.

When the microdata set is released with non-anonymised identifying variables, Equation 16 for the consumer's budget becomes $H(V_1) + \ldots + H(V_n) = H(V_1, \ldots, V_n)$. However, equality cannot hold in this case because the given identifying variables are not independent. Therefore, $H(V_1) + \ldots + H(V_n) > H(V_1, \ldots, V_n)$. Therefore, the available income (joint entropy) would be insufficient to ensure the feasibility of a solution where a non-anonymised microdata set is released.

Therefore, if we set the income to equal to the joint entropy of the non-anonymised identifying variables, a solution where the non-anonymised identifying variables are released would be unattainable. Hence, the assumption we made earlier that the income should equal to the joint entropy of the non-anonymised identifying variables is not correct, when given non-independent variables. Therefore, the income should equal to the sum of the entropies of the non-anonymised identifying variables.

Step 5: Optimise

In this step, the optimum values of each good are determined. This is done through optimisation, where we need to find the maximum value for $U(x_1,\ldots,x_n,y_1,\ldots,y_n)$ subject to the constraint $s_1x_1+\ldots+s_nx_n+t_1y_1+\ldots+t_ny_n=I$ and the constraint that only non-negative amounts of goods are consumed. We can rewrite this as:

$$\max\ U(x_1,\ldots,x_n,y_1,\ldots,y_n) \tag{17}$$

s.t. $s_1x_1+\ldots+s_nx_n+t_1y_1+\ldots+t_ny_n=I$ (Budget constraint)

$x_1,\ldots,x_n \geq 0$

$y_1,\ldots,y_n \geq 0$

To optimise, we use the Lagrange Multipliers Method. The Lagrangian (L) becomes:

$$L(x_1,\ldots,x_n,y_1,\ldots,y_n,\lambda) = U(x_1,\ldots,x_n,y_1,\ldots,y_n)+ \lambda[I-s_1x_1-\ldots-s_nx_n-t_1y_1-\ldots-t_ny_n] \tag{18}$$

At the optimum point, the first order conditions that need to be satisfied become:

$$\frac{\partial L}{\partial x_1}=0\ ;\ \ldots\ ;\ \frac{\partial L}{\partial x_n}=0$$

$$\frac{\partial L}{\partial y_1}=0\ ;\ \ldots\ ;\ \frac{\partial L}{\partial y_n}=0$$

$$\frac{\partial L}{\partial \lambda}=0$$

These are used to obtain the optimum values for $x_1,\ldots,x_n,y_1,\ldots,y_n$, by treating the conditions as $(2n + 1)$ equations in the $(2n + 1)$ unknowns $x_1,\ldots,x_n,y_1,\ldots,y_n,\lambda$.

6.9 Examples

In this Section, we provide two examples of using the OPI function of the ANOPI microdata anonymisation process. We first start with the simplest case, in which we need to find the optimum balance between privacy and information utility when given a microdata set with only one identifying variable. In the second example, we are given a

microdata set with two identifying variables. In both examples, we show how the optimum levels of Information Utility and Privacy change with different values for the input parameters. In the second example, we also represent the optimum values for Information Utility and Privacy graphically. These examples simulated the use of the OPI function with different input preferences. Through the simulation, we evaluated the OPI function by showing how the input preferences (according to which the optimum levels are determined) impact the optimum levels of Information Utility and Privacy.

6.9.1 Example 1 – A microdata set with one identifying variable

The simplest case to which our solution can be applied occurs when we are faced with the need to anonymise a microdata set with only one identifying variable. In this example, we are given a non-anonymised microdata set, as shown in Table A1 in the Appendix, with one identifying variable *Year of Birth*, which we shall denote as V_1. We create two goods: x_1 and y_1. Good x_1 represents the level of Information Utility in V_1 and good y_1 represents the level of Privacy in V_1.

Step 1: Determine the consumer's preferences between different goods

Let us assume that the data owner and the data user agree that the needs of information utility and privacy are equally important. Therefore, we assign equal values to α and β as follows: $\alpha = \frac{1}{2}$, and $\beta = \frac{1}{2}$. Furthermore, since there is only one identifying variable, the data user allocates all 100 User Preference Points to it. In a similar way, we allocate all 100 Intruder Preference Points to the single identifying variable. Therefore, the (economic) utility function is: $U(x_1, y_1) = \left(x_1^{\frac{100}{100}} \right)^{\frac{1}{2}} \left(y_1^{\frac{100}{100}} \right)^{\frac{1}{2}} = x_1^{\frac{1}{2}} y_1^{\frac{1}{2}}$

Step 2: Determine the consumer's income

In this example, we have only one identifying variable, and hence the income is:

$I = H(V_1) = 1.35$

Step 3: Determine the price of each good

The price for x_1 is $s_1 = \frac{1}{100} = 0.01$ and the price for y_1 is $t_1 = \frac{1}{100} = 0.01$.

Step 4: Determine the consumer's budget

The consumer's budget is defined by the equation: $0.01\,x_1 + 0.01\,y_1 = 1.35$

Step 5: Optimise

We now have all the required information to find the optimum values for Information Utility and Privacy. The optimisation problem can be stated as:

$$\max\ U(x_1, y_1) = x_1^{\frac{1}{2}}\ y_1^{\frac{1}{2}}$$

s.t. $0.01\,x_1 + 0.01\,y_1 = 1.35$ (Budget constraint)

$x_1 \geq 0$

$y_1 \geq 0$

When this is solved, we obtain the optimum values as follows: $x_1 = 67.51$, $y_1 = 67.51$. To demonstrate the effects of the change in the preference between information utility and privacy, we show in Table 6.1 how the optimum values for x_1 and y_1 change.

Information utility and privacy preference		Allocation of Points		Optimum values	
		User Preference Points	Intruder Preference Points		
α	β	q_1	r_1	x_1	y_1
0.25	0.75	100	100	33.757	101.27
0.5	0.5	100	100	67.513	67.51
0.75	0.25	100	100	101.27	33.76

Table 6.1 The effect of the change in information utility and privacy preference on the optimum values in Example 1

6.9.2 Example 2 – A microdata set with two identifying variables

We now present a more advanced example, in which we need to anonymise a microdata set with two identifying variables. We are given a microdata set, as shown in Table A2 in the Appendix, with the identifying variables *Year of Birth* and *Marital Status*. We shall denote these identifying variables as V_1 and V_2, respectively. We create four goods: x_1, x_2, y_1, and y_2. Goods x_1 and x_2 represent the Information Utility levels for the identifying variables V_1 and V_2, respectively. Goods y_1 and y_2 represent the levels of Privacy for the identifying variables V_1 and V_2, respectively.

Step 1: Determine the consumer's preferences between different goods

As in Example 1, let us assume that the data owner and the data user agree that the needs of information utility and privacy are equally important. Furthermore, let us assume that the data user prefers V_1 over V_2 and allocates 60 and 40 User Preference Points to the variables, respectively. Let us also assume that the data user and the data owner jointly decide to allocate 60 Intruder Preference Points to V_1 and 40 Intruder

Preference Points to V_2 to show the relative importance that an intruder would place on these two variables. Therefore, the (economic) utility function is:

$$U(x_1,x_2,y_1,y_2) = \begin{pmatrix} \frac{60}{100} & \frac{40}{100} \\ x_1 & x_2 \end{pmatrix}^{\frac{1}{2}} \begin{pmatrix} \frac{60}{100} & \frac{40}{100} \\ y_1 & y_2 \end{pmatrix}^{\frac{1}{2}}$$

Step 2: Determine the consumer's income

In this example, we have two identifying variables, and hence the income is:

$$I = H(V_1) + H(V_2) = 1.35 + 0.52 = 1.87$$

Step 3: Determine the price of each good

The prices for x_1, x_2, y_1, and y_2 are, respectively, $s_1 = \frac{1}{60} = 0.0167$; $s_2 = \frac{1}{40} = 0.025$; $t_1 = \frac{1}{60} = 0.0167$; $t_2 = \frac{1}{40} = 0.025$

Step 4: Determine the consumer's budget

The consumer's budget is defined by the equation:

$0.0167\,x_1 + 0.025\,x_2 + 0.0167\,y_1 + 0.025\,y_2 = 1.87$

Step 5: Optimise

The optimisation problem can be stated as:

$$\max\ U(x_1,x_2,y_1,y_2) = \begin{pmatrix} \frac{60}{100} & \frac{40}{100} \\ x_1 & x_2 \end{pmatrix}^{\frac{1}{2}} \begin{pmatrix} \frac{60}{100} & \frac{40}{100} \\ y_1 & y_2 \end{pmatrix}^{\frac{1}{2}}$$

s.t. $0.0167\,x_1 + 0.025\,x_2 + 0.0167\,y_1 + 0.025\,y_2 = 1.87$ (Budget constraint)

$x_1, x_2 \geq 0$

$$y_1, y_2 \geq 0$$

When this is solved, we obtain the optimum values as follows: $x_1 = 33.69$, $y_1 = 33.69$, $x_2 = 14.97$, $y_2 = 14.97$.

To demonstrate the effects of the change in the preference between information utility and privacy, and the effects of different allocation of User Preference Points and Intruder Preference Points to V_1 and V_2, we show in Table 6.2 (on the next page) how the optimum values change.

Information utility and privacy preference		Allocation of Points				Optimum values			
		User Preference Points		Intruder Preference Points					
α	β	q_1	q_2	r_1	r_2	x_1	y_1	x_2	y_2
0.25	0.75	25	75	25	75	2.92	8.77	26.32	78.96
0.25	0.75	25	75	50	50	2.92	35.09	26.32	35.09
0.25	0.75	25	75	75	25	2.92	78.96	26.32	8.77
0.25	0.75	50	50	25	75	11.70	8.77	11.70	78.96
0.25	0.75	50	50	50	50	11.70	35.09	11.70	35.09
0.25	0.75	50	50	75	25	11.70	78.96	11.70	8.77
0.25	0.75	75	25	25	75	26.32	8.77	2.92	78.96
0.25	0.75	75	25	50	50	26.32	35.09	2.92	35.09
0.25	0.75	75	25	75	25	26.32	78.96	2.92	8.77
0.5	0.5	25	75	25	75	5.85	5.85	52.64	52.64
0.5	0.5	25	75	50	50	5.85	23.39	52.64	23.39
0.5	0.5	25	75	75	25	5.85	52.64	52.64	5.85
0.5	0.5	50	50	25	75	23.39	5.85	23.39	52.64
0.5	0.5	50	50	50	50	23.39	23.39	23.39	23.39
0.5	0.5	50	50	75	25	23.39	52.64	23.39	5.85
0.5	0.5	75	25	25	75	52.64	5.85	5.85	52.64
0.5	0.5	75	25	50	50	52.64	23.39	5.85	23.39
0.5	0.5	75	25	75	25	52.64	52.64	5.85	5.85
0.75	0.25	25	75	25	75	8.77	2.92	78.96	26.32
0.75	0.25	25	75	50	50	8.77	11.70	78.96	11.70
0.75	0.25	25	75	75	25	8.77	26.32	78.96	2.92
0.75	0.25	50	50	25	75	35.09	2.92	35.09	26.32
0.75	0.25	50	50	50	50	35.09	11.70	35.09	11.70
0.75	0.25	50	50	75	25	35.09	26.32	35.09	2.92
0.75	0.25	75	25	25	75	78.96	2.92	8.77	26.32
0.75	0.25	75	25	50	50	78.96	11.70	8.77	11.70
0.75	0.25	75	25	75	25	78.96	26.32	8.77	2.92

Table 6.2 The effect of changes in different preferences on the optimum values in Example 2

6.9.3 A graphical representation of the changes to Information Utility and Privacy for a microdata set with two identifying variables

In this sub-section, we use the same microdata set as in Example 2, but we use graphs to show how the values for Information Utility and Privacy change as the input values for α and β change, with all possible assignment of the User and the Intruder Preference points to variables V_1 and V_2.

In each graph in Figures 6.2 to 6.6, we show the possible values for x_1, y_1, x_2, and y_2. In each graph, we show the allocation of User Preference Points only to variable V_1 on the X-axis. Since the sum of the User Preference Points allocated to each identifying variable must equal to 100 (see Section 6.4), the number of User Preference Points allocated to variable V_2 can be implied from the number of points allocated to variable V_1. Similarly, on the Y-axis we show the allocation of the Intruder Preference Points only to variable V_1, since the Intruder Preference Points allocated to variable V_2 can be implied. On the Z-axis, we show the levels of Information Utility (*Iu*) in the case of values for x_1 and x_2, and the levels of Privacy (*Priv*) in the case of values for y_1 and y_2.

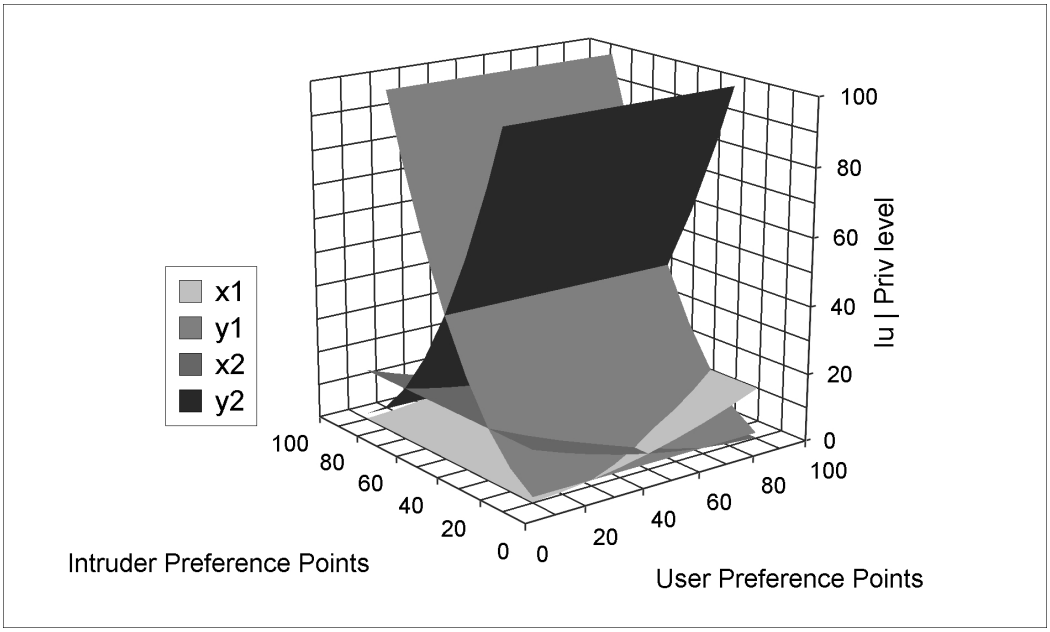


Figure 6.2 Possible Information Utility and Privacy levels when α = 0.1 and β = 0.9

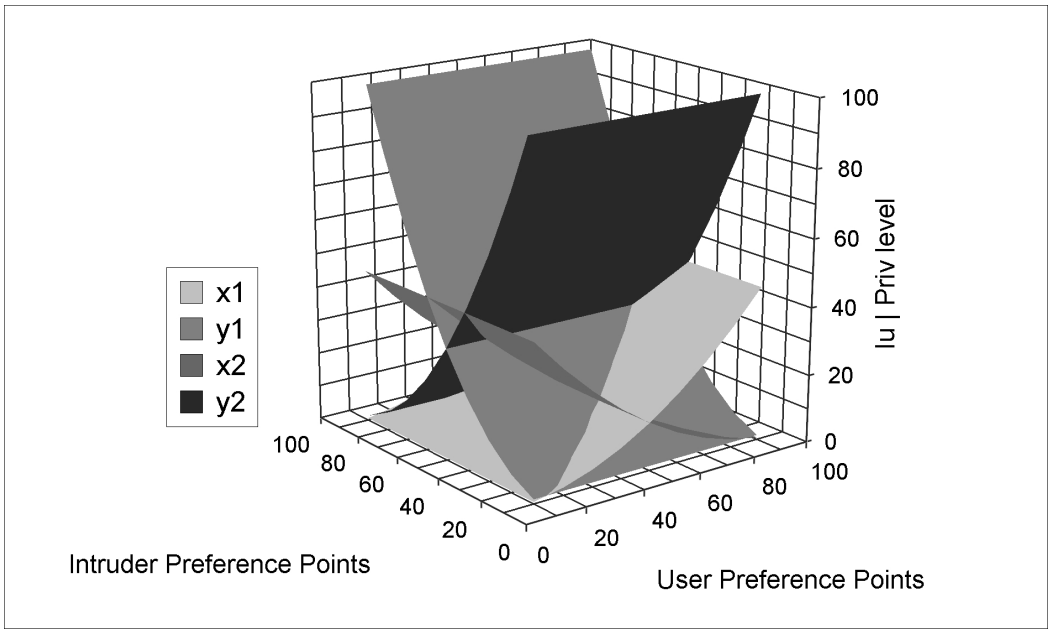


Figure 6.3 Possible Information Utility and Privacy levels when α = 0.3 and β = 0.7

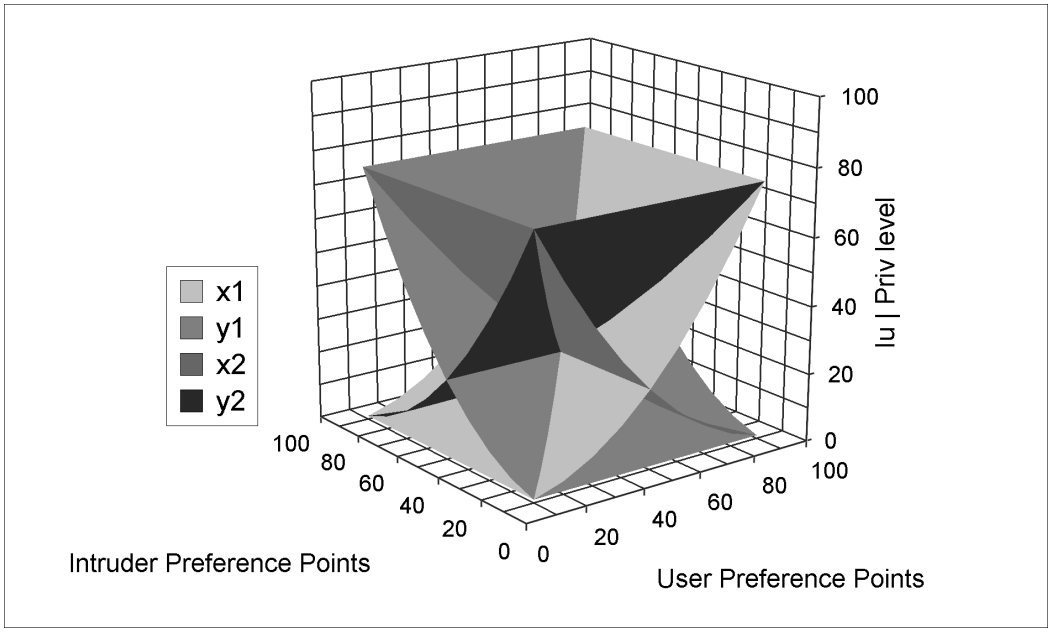


Figure 6.4 Possible Information Utility and Privacy levels when α = 0.5 and β = 0.5

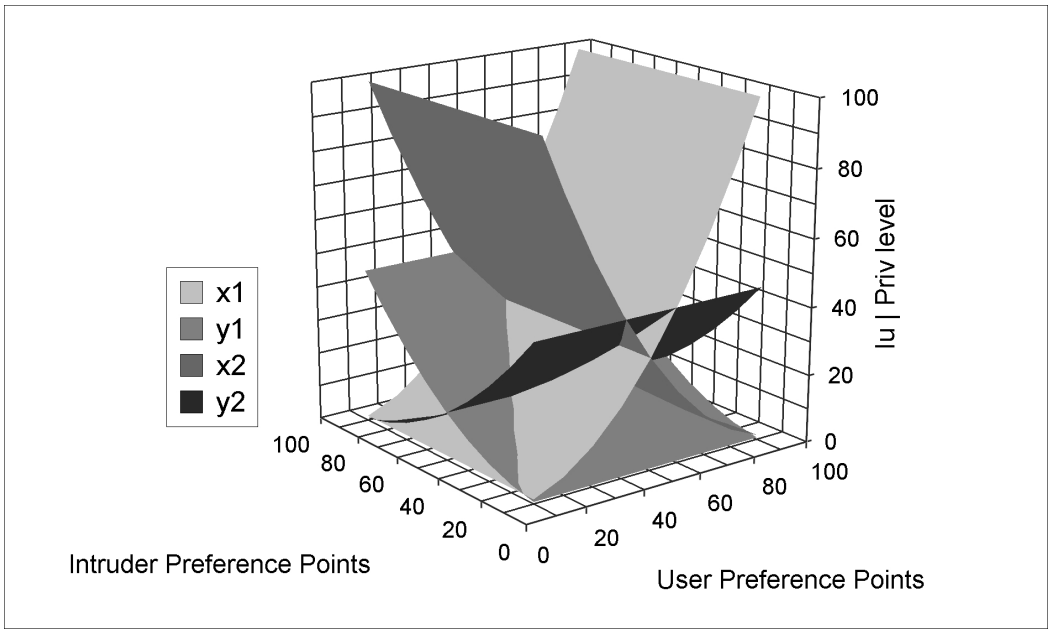


Figure 6.5 Possible Information Utility and Privacy levels when α = 0.7 and β = 0.3

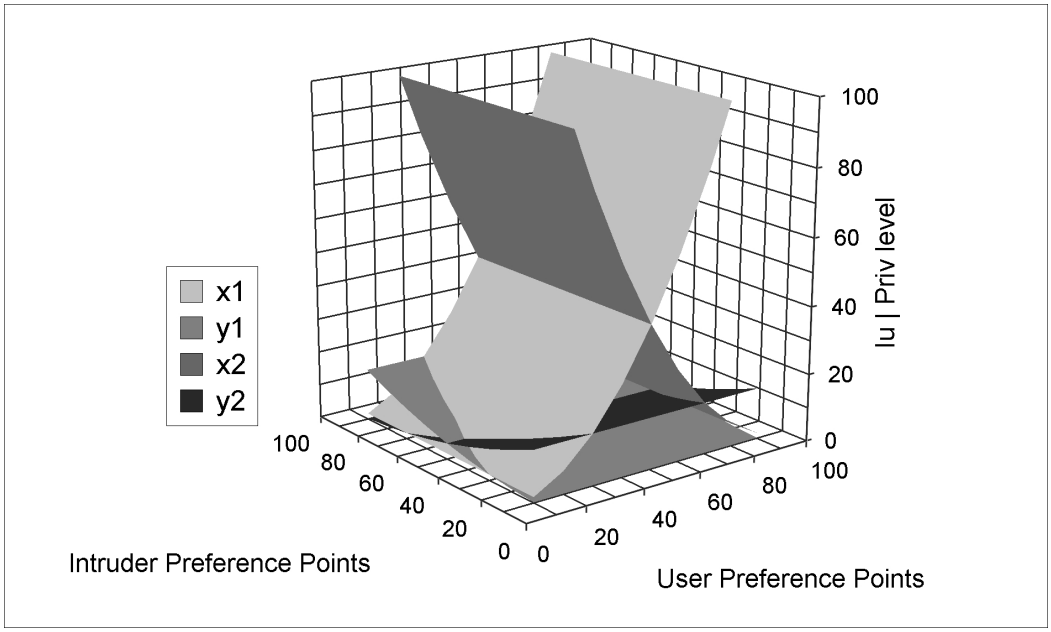


Figure 6.6 Possible Information Utility and Privacy levels when α = 0.9 and β = 0.1

6.10 Analysis of graphical results and simplification of the OPI function

In this Section, we discuss the results contained in the graphs from the previous Section. Based on this analysis, we are able to simplify the proposed solution, while still maintaining the original concepts on which the solution was based.

In line with our initial expectations, as the preference for information utility increases (i.e. as the value for α increases), the values for x_1 and x_2 are greater in comparison to the values for y_1 and y_2. The converse is true as the preference for privacy increases (i.e. as the value for β increases). Moreover, the values for x_1 and x_2 increase as more User Preference Points are allocated to the respective variables. In a similar way, the values for y_1 and y_2 increase as more Intruder Preference Points are allocated to the respective variables.

Upon closer inspection of the results that were used to create the graphs, we noticed that the values allocated to x_i are always a portion of the sum of the total entropy of the non-anonymised identifying variables. This portion is based on a ratio of the values allocated to α and q_i. Similarly, the values assigned to y_i are also a portion of the total entropy of the non-anonymised identifying variables, but this portion is based on a ratio of the values assigned to β and r_i. We summarise this finding in a more generalised form as follows.

Given a microdata set with n identifying variables $V_1, \ldots, V_n$, the optimum level of Information Utility x_i for a variable V_i can be calculated as:

$$x_i = \alpha \frac{q_i}{100} q_i \sum_{j=1}^{n} H(V_j) \qquad (19)$$

and the optimum level of Privacy y_i for a variable V_i can be calculated as:

$$y_i = \beta \frac{r_i}{100} r_i \sum_{j=1}^{n} H(V_j) \tag{20}$$

where:

- α is the preference value for information utility, as explained in Step 1 in Section 6.8,
- β is the preference value for privacy, as explained in Step 1 in Section 6.8,
- q_i is the number of User Preference Points allocated to variable V_i, as explained in Section 6.4,
- r_i is the number of Intruder Preference Points allocated to variable V_i, as explained in Section 6.5,
- n is the number of identifying variables in the non-anonymised microdata,
- $H(V_j)$ is the entropy of a non-anonymised identifying variable V_j.

Through the use of the above two simplified formulas for calculating the optimum levels of Information Utility and Privacy, we are able to produce equivalent results to those produced using the steps presented in Section 6.8. Moreover, by using the above two simplified formulas, we were also able to simplify the process through which our optimisation problem is solved and also reduce the complexity of the actual optimisation itself. Therefore, we were able to reduce the complexity involved in solving the optimisation problem, while still being able to make use of the richness of Economic Price Theory and the usefulness it provides in guiding the anonymisation process.

Through this simplification, we are now only required to set the preferences between privacy and information utility, as well as the preferences between each identifying variable, and thereafter use Equations 19 and 20 to determine the optimum values of information utility and privacy of each identifying variable.

6.11 Conclusion

In this Chapter, we proposed ANOPI, which is a microdata anonymisation process that anonymises microdata such that an optimum balance between privacy and information utility is obtained. We focused on the first function of ANOPI, namely the OPI function, which determines the optimum levels of information utility and privacy based on preferences between privacy and information utility, as well as preferences between the different identifying variables of a microdata set. After these optimum levels have been determined, we need to anonymise the microdata set such that those levels are achieved. This occurs in Anonymising function of ANOPI and is described in the next Chapter.

CHAPTER 7

HOW TO ANONYMISE MICRODATA TO ACHIEVE THE OPTIMUM LEVELS OF PRIVACY AND INFORMATION UTILITY

7.1 Introduction

In the previous Chapter, we introduced ANOPI, which aims to anonymise microdata by guiding the anonymisation process such that the microdata possesses optimum levels of privacy and information utility. ANOPI has two functions: the OPI function and the Anonymising function. The OPI function determines the optimum levels of privacy and information utility; it was discussed in detail in the previous Chapter. The Anonymising function anonymises microdata such that the identified optimum levels of privacy and information utility will exist in the microdata.

In this Chapter, we discuss how to anonymise microdata to achieve the optimum levels of privacy and information utility. The discussion focuses on the Anonymising function of ANOPI. Our discussion of the Anonymising function will be limited to only two anonymisation techniques, namely global recoding and microaggregation. As explained in Chapter 4, these two techniques are typically used to achieve *k*-anonymity. The Anonymising function will first be discussed when global recoding is used and thereafter when microaggregation is used. In both discussions, the Anonymising function will be simulated by using examples, which will show changes in the way in which a microdata set is anonymised as the optimum levels of information utility and privacy change.

7.2 Anonymising function applied with Global Recoding

Once the OPI function has determined the optimum levels of privacy and information utility that a released microdata set should possess, the microdata set needs to be anonymised such that those levels are achieved. This process of data anonymisation takes place in the Anonymising function of the ANOPI process.

The Anonymising function takes, as its input, the non-anonymised microdata set as well as the optimum values for privacy and information utility that were calculated in the OPI function. When global recoding is used as the anonymisation technique, the Anonymising function determines the codings with which each identifying variable should be released. Thereafter, it applies global recoding to recode each identifying variable according to the determined codings. Once the variables have been recoded, the anonymised microdata set is provided as the output of the ANOPI process.

Recall from Chapter 6 that the notation $(V_i)^j$ indicates that the variable V_i has been anonymised to the *j*-th degree by a certain anonymisation technique. Therefore, when the Anonymising function is applied with global recoding, *j* shall indicate the coding with which V_i should be released.

To determine the coding with which an identifying variable should be released, we must identify that coding of the variable at which the Information Utility and Privacy levels match the optimum levels.

It is not likely that the optimum Information Utility and Privacy levels obtained in the OPI function for an identifying variable will match the Information Utility and Privacy levels of any of the possible codings of that variable. This is because, when the optimum values are calculated, we use the *continuous* consumer's budget function, and thereby assume that any values for the Information Utility and Privacy levels are possible. The assumption is not valid, since the possible values for the Information Utility and Privacy levels do not form a continuous function.

Therefore, if the optimum levels of Information Utility and Privacy of an identifying variable do not match the Information Utility and Privacy levels of any of the possible codings of that variable, we need to choose that coding at which the levels are closest to the optimum solution. In order to do this, we need to choose a coding that has the least effect on the economic utility level *U*. That is, we choose a coding C_j such that $\left(\left| x_i - Iu\left((V_i)^j\right) \right| \right)^{\frac{q_i}{100}\alpha} \left(\left| y_i - Priv\left((V_i)^j\right) \right| \right)^{\frac{r_i}{100}\beta}$ is minimal (where α and β are the preference values for information utility and privacy, respectively, and where q_i and r_i are the User Preference Points and the Intruder Preference Points allocated to the variable V_i, respectively).

Since it may still occur that, after global recoding, there exist records which are relatively rare in the anonymised microdata set, it may be necessary to apply suppression to remove these rare records from the anonymised microdata set. However, if we apply data suppression to the anonymised microdata set, we will reduce the resulting level of information utility and increase the level of privacy. Therefore, we will further deviate from the optimum values of information utility and privacy that have been found in the OPI function.

The need to suppress relatively rare records from the anonymised microdata set is currently out of the scope of this study. However, for future work, we propose to expand our proposed solution by assuming that both global recoding and data suppression can be used in the microdata anonymisation process.

7.2.1 Examples of Applying the Anonymising function with Global Recoding

Continuation of Example 1 – A microdata set with one identifying variable

We now continue Example 1, presented in Section 6.9.1, to show how the Anonymising function is applied.

In Example 1, we were given a microdata set with one identifying variable, *Year of Birth*, which was denoted as V_1. Let us set the possible codings for V_1 as follows:

- C_1: non-recoded data; V_1 can assume any valid year from 1961 to 1990
- C_2: V_1 can assume the values "1961-1962", "1963-1964", ... , "1989-1990"
- C_3: V_1 can assume the values "1961-1963", "1964-1966", ... , "1988-1990"
- C_4: V_1 can assume the values "1961-1965", "1966-1970", ... , "1986-1990"
- C_5: V_1 can assume the values "1961-1970", "1971-1980", "1981-1990"
- C_6: V_1 can assume the values "1961-1975", "1976-1990"
- C_7: V_1 can assume the values "1961-1990"

The optimum values calculated were x_1 = 67.51 and y_1 = 67.51. In Figure 7.1, we show a graph of the set of possible values for x_1 and y_1 for each coding of V_1 as well as the optimum values. It is clear that we cannot obtain these exact optimum values, since the optimum levels of Information Utility and Privacy for V_1 do not match the Information Utility and Privacy levels of any of the possible codings of that variable. Therefore, we need to choose that coding at which the possible levels are closest to the optimum solution. This coding is C_4, which is the coding with which variable V_1 is released. The resulting anonymised microdata set is shown in Table A3 in the Appendix. In Table 7.1, on the next page, we show how the codings change as the optimum values for x_1 and y_1 change.

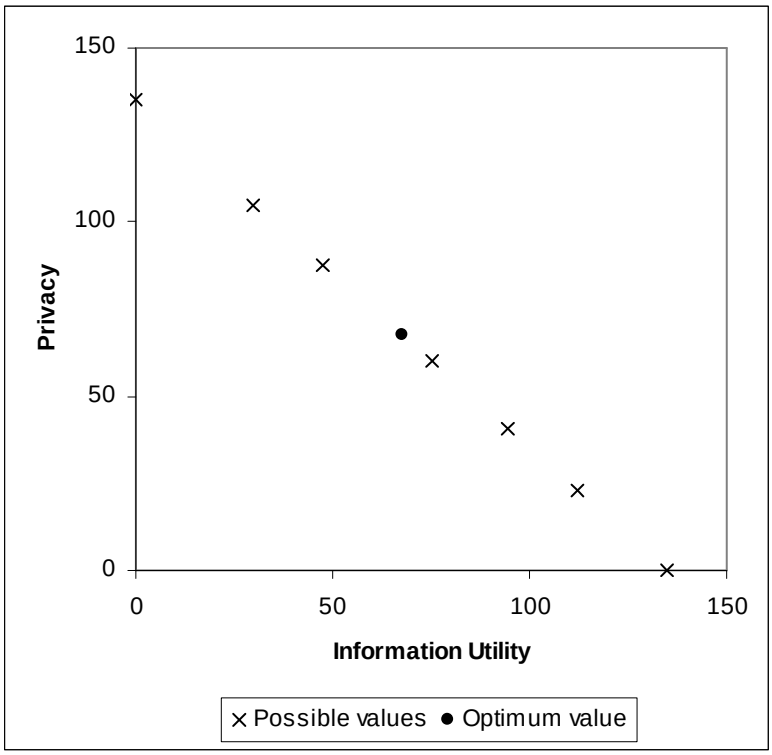


Figure 7.1 A graph of the set of possible values for x_1 and y_1 and the optimum value

Information utility and privacy preference		Allocation of Points		Released Coding
		User Preference Points	Intruder Preference Points	
α	β	q_1	r_1	V_1
0.25	0.75	100	100	C_6
0.5	0.5	100	100	C_4
0.75	0.25	100	100	C_3

Table 7.1 Codings at different optimum values of Information utility and Privacy in Example 1

Continuation of Example 2 – A microdata set with two identifying variable

In Example 2, we were given a microdata set with two identifying variables, *Year of Birth* and *Marital Status*, which were denoted as V_1 and V_2, respectively. Let the set of possible codings for V_1 be as in Example 1, and let the set of possible codings for V_2 be as follows:

- D_1: non-recoded data; V_2 can assume the values "Single", "Married", "Widowed", or "Divorced"
- D_2: V_2 can assume the values "Never_Married", for those values that were equal to "Single" in V_2, or "Been_Married", for those values that were equal to "Married", "Widowed" or "Divorced" in V_2
- D_3: V_2 only assume the values "Not_released"

The optimum values obtained in Example 2 in Chapter 6 were: $x_1 = 33.69$, $y_1 = 33.69$, $x_2 = 14.97$, $y_2 = 14.97$. Based on these values, we release V_1 with coding C_4 and V_2 with coding D_2. The resulting anonymised microdata is shown in Table A4 in the Appendix. In Table 7.2, on the next page, we also show how the optimum values affect the codings with which identifying variables in the microdata should be released.

Information utility and privacy preference		Allocation of Points				Released Coding	
		User Preference Points		Intruder Preference Points			
α	β	q_1	q_2	r_1	r_2	V_1	V_2
0.25	0.75	25	75	25	75	C_3	D_3
0.25	0.75	25	75	50	50	C_4	D_3
0.25	0.75	25	75	75	25	C_6	D_2
0.25	0.75	50	50	25	75	C_3	D_2
0.25	0.75	50	50	50	50	C_4	D_2
0.25	0.75	50	50	75	25	C_6	D_2
0.25	0.75	75	25	25	75	C_3	D_3
0.25	0.75	75	25	50	50	C_4	D_3
0.25	0.75	75	25	75	25	C_6	D_2
0.5	0.5	25	75	25	75	C_2	D_1
0.5	0.5	25	75	50	50	C_6	D_3
0.5	0.5	25	75	75	25	C_6	D_2
0.5	0.5	50	50	25	75	C_2	D_1
0.5	0.5	50	50	50	50	C_5	D_1
0.5	0.5	50	50	75	25	C_5	D_1
0.5	0.5	75	25	25	75	C_2	D_2
0.5	0.5	75	25	50	50	C_4	D_2
0.5	0.5	75	25	75	25	C_4	D_2
0.75	0.25	25	75	25	75	C_6	D_2
0.75	0.25	25	75	50	50	C_6	D_1
0.75	0.25	25	75	75	25	C_6	D_1
0.75	0.25	50	50	25	75	C_4	D_1
0.75	0.25	50	50	50	50	C_4	D_1
0.75	0.25	50	50	75	25	C_4	D_1
0.75	0.25	75	25	25	75	C_2	D_2
0.75	0.25	75	25	50	50	C_2	D_2
0.75	0.25	75	25	75	25	C_2	D_2

Table 7.2 Codings at different optimum values of Information utility and Privacy in Example 2

7.3 Anonymising function applied with Microaggregation

In this Section, we discuss the Anonymising function when it is applied with microaggregation. We shall assume that the MDAV (Maximum Distance to Average Vector) algorithm (Domingo-Ferrer et al., 2006; Hundepool et al., 2005) will be used for univariate and multivariate microaggregation. The MDAV algorithm clusters records in a microdata set such that each cluster will have at least *k* records. The greater the value for *k*, the greater the least number of records per cluster, and hence the greater the privacy and the lower the information utility of the anonymised microdata.

Therefore, to determine how to anonymise microdata such that it possesses the optimum levels of Information Utility and Privacy, we need to determine the optimum value for *k* at which the optimum levels of Information Utility and Privacy will occur. In this Section, we propose a procedure that can be followed to determine the optimum number of records *k* per cluster in both univariate and multivariate microaggregation.

In the case when microaggregation is used in the Anonymising function, the notation $(V_i)^j$ (introduced in Chapter 6) shall indicate that V_i has been microaggregated such that there are at least j records in each cluster of the released variable. Given the optimum levels of Information Utility, x_i, and Privacy, y_i, of each identifying variable V_i, we determine the optimum *k* value in three steps as follows.

Step 1: For every identifying variable, determine the optimum value for *k* that will satisfy the variable's optimum level of Information Utility

For the time being, we consider each identifying variable seperately, since the optimum level x_i of Information Utility is applicable only to a particular variable. Therefore, in this step, we assume that each variable V_i is microaggregated seperately. That is, we assume that we shall apply univariate microaggregation in parallel for each variable. For each variable V_i, we determine the variable's optimum value *k* independently from the other variables. If this assumption is not valid, that is, if some variables should be

grouped together and microaggregated as a group (i.e. with one *k* value), then we shall determine this optimum *k* value for the group in Step 3.

Let *m* be the number of records that exist in the microdata set. Since we do not remove records when the microdata set is anonymised through microaggregation, the number of records present in the non-anonymised microdata set is equal to the number of records present in the anonymised microdata set.

We assume that we are required to microaggregate V_i such that there are at least *k* records in each cluster of V_i. Therefore, the maximum number of clusters, and hence the maximum information entropy present in $(V_i)^k$, will occur if every cluster has exactly *k* records. Therefore, the maximum number of clusters that can exist in the microaggregated $(V_i)^k$ is $\frac{m}{k}$.

Of course, the MDAV algorithm can microaggregate a variable such that there is at least one cluster with more than *k* records. In that case, we will have fewer than $\frac{m}{k}$ clusters in V_i. However, before V_i has been microaggregated, we are unable to determine the *exact* number of clusters that will exist in V_i. We can only make assumptions about the maximum number of clusters and the maximum amount of information entropy that can exist in V_i.

Recall, from Equation 6, that the information entropy of $(V_i)^k$ is calculated as follows (Shannon, 1948):

$$H\left((V_i)^k\right) = -\sum_{j=1}^{t} p_j \, log\left(p_j\right) \qquad (6)$$

where:

- *t* is the maximum number of values that $(V_i)^k$ can assume, and
- p_j is the probability that $(V_i)^k$ assumes the *j*-th possible value.

If there can be a maximum of $\frac{m}{k}$ clusters in $(V_i)^k$, then there can be a maximum of $\frac{m}{k}$ different values that $(V_i)^k$ is able to assume, given that every one of the possible $\frac{m}{k}$ values is assumed at least once by $(V_i)^k$. Therefore, we can substitute *t* in Equation 6 with $\frac{m}{k}$.

Maximum information entropy in $(V_i)^k$ will occur when each of the possible $\frac{m}{k}$ values is assumed by $(V_i)^k$ with equal probability, which is equal to $\frac{1}{\frac{m}{k}}$. Therefore, assuming that we are aiming to achieve maximum information entropy, each of the possible $\frac{m}{k}$ values will have a probability of $\frac{1}{\frac{m}{k}}$. Therefore, we can substitute p_j in Equation 6 with $\frac{1}{\frac{m}{k}}$.

Therefore, Equation 6 becomes:

$$H\left((V_i)^k\right) = -\sum_{j=1}^{\frac{m}{k}} \frac{1}{\left(\frac{m}{k}\right)} \cdot log\left(\frac{1}{\left(\frac{m}{k}\right)}\right) \qquad (21)$$

where:

- *m* is the number of records that exist in the microdata, and
- *k* is the least number of records that exist in each cluster of the microaggregated $(V_i)^k$.

Therefore, the maximum information entropy that can exist in the microaggregated variable $(V_i)^k$ such that each cluster has at least *k* records, is:

$$H\left((V_i)^k\right) = -\sum_{j=1}^{\frac{m}{k}} \frac{1}{\left(\frac{m}{k}\right)} \cdot log\left(\frac{1}{\left(\frac{m}{k}\right)}\right) \tag{22}$$

$$= -\sum_{j=1}^{\frac{m}{k}} \frac{k}{m} \cdot log\left(\frac{k}{m}\right)$$

$$= -\frac{m}{k} \cdot \frac{k}{m} \cdot log\left(\frac{k}{m}\right)$$

$$= -log\left(\frac{k}{m}\right)$$

$$= log(m) - log(k)$$

Using Equation 19 (derived in Chapter 6), we know that $x_i = \alpha \frac{q_i}{100} q_i \sum_{j=1}^{n} H(V_j)$, where $H(V_j)$ refers to the information entropy contained in the non-anonymised variable V_j.

We also know that $x_i = q_i H\left((V_i)^k\right)$, based on the way in which Information Utility is defined in Equation 7 (in Chapter 6).

Therefore,

$$\alpha \frac{q_i}{100} q_i \sum_{j=1}^{n} H(V_j) = q_i H\left((V_i)^k\right) \tag{23}$$

Hence:

$$\alpha \frac{q_i}{100} \sum_{j=1}^{n} H(V_j) = H\left((V_i)^k\right) \tag{24}$$

Since $H\left((V_i)^k\right) = log(m) - log(k)$, it follows that:

$$\alpha \frac{q_i}{100} \sum_{j=1}^{n} H(V_j) = log(m) - log(k) \tag{25}$$

Therefore,

$$log(k) = log(m) - \alpha \frac{q_i}{100} \sum_{j=1}^{n} H(V_j) \tag{26}$$

Therefore, the optimum value k with which the variable V_i should be microaggregated (i.e. the least number of records that should exist in each cluster of the anonymised V_i) in order to achieve the optimum level of Information Utility is:

$$k = 10^{\left(log(m) - \alpha \frac{q_i}{100} \sum_{j=1}^{n} H(V_j) \right)} \tag{27}$$

where:

- m is the number of records that exist in the microdata set,
- n is the number of identifying variables in the (non-anonymised and anonymised) microdata set,
- α is the preference for information utility,
- q_i is the number of User Preference Points allocated to the variable V_i, and
- $H(V_j)$ is the amount of information entropy contained in the non-anonymised variable V_j.

We should also ensure that the k value obtained in Equation 27 is not greater than the number of records that exist in the microdata set. Therefore, if $k > m$, we set $k = m$. Moreover, it is possible that the calculated value for k is not an integer. Therefore, we round off the k value obtained in Equation 27 to the nearest non-negative integer value.

Step 2: For every identifying variable, determine the optimum value for *k* that will satisfy the variable's optimum level of Privacy

The amount of information entropy that should be released in V_i to satisfy the optimum level of Information Utility x_i may in fact be different from the amount of information entropy that should be released in V_i to satisfy the optimum level of Privacy y_i. Therefore, the optimum value for *k* calculated in Step 1 may not necessarily be the optimum value that will satisfy the required level of Privacy. Therefore, in this step, we determine the optimum *k* value that will satisfy the optimum level of Privacy in variable V_i.

We proceed similarly as in Step 1, but use the optimum level of Privacy y_i that should result from each variable V_i after it has been anonymised. As in the previous step, we shall, for the time being, consider each variable seperately because the optimum level of Privacy y_i is applicable only to a specific variable. Therefore, we shall assume (as in the previous Step) that each variable V_i is microaggregated seperately, as if univariate microaggregation would be applied to each variable in parallel. For each variable V_i, we determine the variable's optimum *k* value independently from the other variables. Should it be required that some variables should be grouped together and microaggregated as a group with one *k* value, the optimum value for *k* will be determined for the whole group in Step 3.

We have established in Step 1 (in Equation 22) that $H\left((V_i)^k\right) = log(m) - log(k)$, where *m* is the number of records that exist in the microdata.

Therefore, by substituting $H\left((V_i)^k\right)$ with $log(m) - log(k)$ in Equation 8, we obtain:

$$y_i = r_i\left(H(V_i) - \left(log(m) - log(k)\right)\right) \qquad (28)$$

As mentioned above, the optimum value for *k* calculated in Step 1 may not necessarily be the optimum value that will satisfy the required level of Privacy. Therefore, we are

merely using the fact that $H\left((V_i)^k\right) = log(m) - log(k)$ and the k used in Equation 28 does not refer to the k obtained in Step 1. That is, in this step, we do not base our calculations on the actual level of information entropy $H\left((V_j)^k\right)$ or on the k value that were obtained in Step 1.

From Equation 20 (derived in Chapter 6), we know that $y_i = \beta \frac{r_i}{100} r_i \sum_{j=1}^{n} H(V_j)$.

Therefore, Equation 28 can be rewritten as:

$$\beta \frac{r_i}{100} r_i \sum_{j=1}^{n} H(V_j) = r_i \left(H(V_i) - \left(log(m) - log(k)\right) \right) \quad (29)$$

By simplifying, we obtain:

$$log(k) = log(m) - H(V_i) + \beta \frac{r_i}{100} \sum_{j=1}^{n} H(V_j) \quad (30)$$

Therefore, the optimum value k with which the variable V_i should be microaggregated (i.e. the least number of records that should exist in each cluster of the anonymised V_i) in order to achieve the optimum level of Privacy is:

$$k = 10^{\left(log(m) - H(V_i) + \beta \frac{r_i}{100} \sum_{j=1}^{n} H(V_j) \right)} \quad (31)$$

where:

- m is the number of records that exist in the microdata set,
- n is the number of identifying variables in the (non-anonymised and anonymised) microdata set,
- β is the preference for privacy,
- r_i is the number of Intruder Preference Points allocated to the variable V_i, and
- $H(V_i)$ and $H(V_j)$ refer to the information entropy of the non-anonymised variables V_i and V_j, respectively.

As in Step 1, we must ensure that the k value obtained in Equation 31 is not greater than the number of records that exist in the microdata set. Therefore, if $k > m$, we set $k = m$. We also round off the k value obtained in Equation 31 to the nearest non-negative integer value.

Step 3: Select the *k* value that should be used to microaggregate each variable or group of variables

In Steps 1 and 2, we determined two k values with which each variable V_i could be microaggregated. The k value obtained in Step 1 is the one with which V_i should be microaggregated if we only take Information Utility into account. Similarly, the k value obtained in Step 2 only takes Privacy into account. In this step, we need to select one k value that will satisfy the requirements of both information utility and privacy, with minimal deviation from the optimum levels.

Since this step depends on the type of microaggregation used, we continue the discussion of the remainder of this step by considering univariate and multivariate microaggregation seperately.

Univariate microaggregation

In univariate microaggregation, we have only one variable V_1 that needs to be microaggregated. Let k^{iu} refer to the k value obtained in Step 1. That is, k^{iu} is the least number of records that should be present in each cluster of the microaggregated V_1 such that the optimum level of Information Utility will be achieved. Let k^{p} refer to the k value obtained in Step 2. Therefore, k^{p} is the least number of records that should be present in each cluster of the microaggregated V_1 such that the optimum level of Privacy is achieved.

If $k^{iu} = k^{p}$, then we can microaggregate V_1 with this value. However, since the preference for information utility and privacy may not be the same, it is likely that k^{iu} does not equal k^{p}. In that case, we need to find a k value, where $k^{iu} \leq k \leq k^{p}$ or $k^{p} \leq k \leq k^{iu}$, with which V_1 can be microaggregated and that will have the least effect on the calculated optimum levels of Information Utility and Privacy of the variable.

To select the k value that will have the least effect on the optimum levels of Information Utility and Privacy, we need to determine the k value at which there will be the least deviation from the optimum solution obtained in the OPI function.

The optimum solution to the optimisation problem was obtained when the Economic Utility value U was at its maximum, taking into account the constraints of the optimisation problem. Therefore, to determine the least effect on the optimum levels of Information Utility and Privacy of each variable, we propose to take into account the least effect on the value of the Economic Utility (U) that was used to derive the optimum levels. This will ensure that we deviate as little as possible from the calculated value for U.

We determine the value for U at each k value between $min\left(k^{iu}, k^{p}\right) \leq k \leq max\left(k^{iu}, k^{p}\right)$. The optimum value k selected will be the one at which the difference between the resulting value for U and the original value for U is minimal.

Multivariate microaggregation, where each identifying variable is microaggregated seperately

When we need to microaggregate a microdata set with more than one identifying variable, but where each identifying variable is to be microaggregated seperately, then we apply univariate microaggregation to each identifying variable in parallel. Therefore, the k value chosen for one identifying variable does not affect the k value chosen for another identifying variable. Hence, we proceed as in the case of univariate

microaggregation, except that we repeat this step for every identifying variable of the given microdata set.

Multivariate microaggregation, where variables are grouped into blocks

We now address the need to microaggregate a microdata set with more than one identifying variable and where identifying variables can be grouped into blocks of one or more variables. Each block of variables is then microaggregated independently of the other blocks (Nin et al., 2008a).

We are given a microdata set with *n* identifying variables $V_1, \dots, V_n$. The variables can be split into blocks of one or more variables and each variable can be part of only one block.

For each block of variables $V_i, \dots, V_j$, where $1 \leq i \leq j \leq n$, we need to determine the minimum number of records *k* that should exist in each cluster of that block. Since each block of variables will be microaggregated independently of the other blocks, the value *k* does not need to be the same for each block. Therefore, we determine the optimum *k* value for each block independently of the other blocks. Hence, this step is repeated for each block of variables.

For each variable V_t of the block, we need to take into account the *k* value that was calculated as the optimum value to achieve the optimum level of Information Utility x_t, and also the *k* value that was calculated to achieve the optimum level of Privacy y_t.

Let k_t^{iu} refer to the least number of records that should be present in each cluster of the microaggregated variable V_t such that the optimum level of Information Utility x_t will be achieved. Let k_t^{p} refer to the least number of records that should be present in each cluster of the microaggregated variable V_t such that the optimum level of Privacy y_t

will be achieved. Since we have *j* variables in the block, the number of *k* values derived in Steps 1 and 2 will be 2*j*: $k_i^{iu},\ldots,k_j^{iu}$ and $k_i^{p},\ldots,k_j^{p}$.

If $k_i^{iu} = \ldots = k_j^{iu} = k_i^{p} = \ldots = k_j^{p}$, then we can microaggregate the block of variables with this value. However, since the preferences for each variable and for information utility and privacy may not be the same, it is unlikely that the *k* values obtained in Steps 1 and 2 are all equal. In that case, we need to find a *k* value that is between the minimum and maximum possible 2*j* values. That is, we need to find a *k* value such that $min\left\{k_i^{iu},\ldots,k_j^{iu},k_i^{p},\ldots,k_j^{p}\right\} \leq k \leq max\left\{k_i^{iu},\ldots,k_j^{iu},k_i^{p},\ldots,k_j^{p}\right\}$.

Additionally, the *k* value chosen should be one at which the deviation from the maximum value for *U* is minimal, similar to what was proposed in the case of univariate microaggregation.

7.3.1 Implication for *k*-anonymity

Another way in which a microdata set can be anonymised is by altering the original microdata such that it will satisfy the property of *k*-anonymity (Samarati, 2001; Sweeney, 2002a, 2002b). When a microdata set is *k*-anonymised with a certain value for *k* (where $k > 1$), every record is indistinguishable from at least $k - 1$ other records in that microdata set. Therefore, it implies that the anonymised microdata set has a certain number of record groups, or clusters, where each group has at least *k* records. Therefore, the problem of selecting the optimum *k* value in *k*-anonymisation is similar to the problem of selecting the optimum least number of records per cluster in microaggregation.

The research problem of optimal *k*-anonymisation aims to find an anonymisation that will produce the "best" *k*-transformed dataset, as determined by some cost metric (Bayardo & R Agrawal, 2005). For example, if the cost metric is the information loss that occurs as a result of the generalization and suppression applied, then an optimal *k*-anonymisation is an anonymisation that achieves *k*-anonymity with the least number of generalization and suppression combinations, so as to minimise information loss.

Finding an optimal *k*-anonymisation has been proved to be NP-hard (Meyerson & Williams, 2004), although polynomial time approximate algorithms have been developed (such as those of Aggarwal et al. (2005), LeFevre, DeWitt, and Ramakrishnan (2005, 2006), as well as Meyerson and Williams (2004)). Nevertheless, as in the case of microaggregation, it is unclear from the literature what should be the optimum *k* value with which a microdata set should be *k*-anonymised

Domingo-Ferrer and Torra (2005) described how *k*-anonymity can be achieved with microaggregation of continuous, ordinal and nominal data. We can achieve *k*-anonymity by microaggregating all the identifying variables of a microdata set as one group. Therefore, we can use the steps presented in this Section to determine the optimum *k* value that should be used in *k*-anonymisation, if we assume that *k*-anonymity will be achieved by multivariate microaggregation, where all identifying variables are grouped into one block.

7.3.2 Examples of applying the Anonymising function with Microaggregation

To present examples of applying the Anonymising function with microaggregation, we will show how the calculated *k* value changes as the values for the ANOPI input parameters change. For these examples, we used the Wine Data Set, which is available at the UCI Machine Learning Repository (Asuncion & Newman, 2007). We used two variables of the data set, namely "Alcohol" and "Malic acid", which will be referred to as V_1 and V_2, respectively. Table 7.3 shows how the *k* values change as the values for the input parameters change.

As the value for α increases and the value for β decreases, the overall *k* values with which variables are microaggregated decrease. This is due to the fact that the preference for information utility increases as the value of α increases. Therefore, there are fewer records per cluster in the microaggregated variables (although the number of clusters is greater). The opposite is true when the value for α decreases and the value for β increases.

Moreover, as the number of User Preference Points allocated to a variable increases, the *k* value with which the particular variable is microaggregated decreases. This can be attributed to the fact that as the data user's preference for a variable increases, a greater level of Information Utility should be derived from that variable. Hence, there should be a greater amount of information entropy present in the microaggregated variable. This implies that there should be a greater number of clusters in the microaggregated variable, and consequently there are fewer records per cluster (i.e. a lower *k* value).

In a similar way, as the number of Intruder Preference Points allocated to a variable increases, the *k* value with which the variable is microaggregated increases. A higher number of Intruder Preference Points allocated to a variable implies that more Privacy should be derived from the variable. Therefore, more information entropy should be lost from this variable. Hence, there should be fewer clusters in the microaggregated variable, implying that the number of records per cluster (*k* value) should be greater.

Lastly, the way in which variables are microaggregated also impacts the final *k* value. When variables are grouped for multivariate microaggregation, the overall quality of the optimal solution decreases in comparison to performing univariate microaggregation in parallel to each variable.

As an example, consider the first line of Table 7.3. We see that the *k* value for V_1 can range between 10 and 98. However, the final *k* value for V_1 and V_2 as a group is 153, which falls outside of the range of *k* values for V_1. This is due to the fact that the possible range of *k* values for V_2 must also be considered (which in this case is between 30 and 178), since V_1 and V_2 are microaggregated as a group. Hence, the range of possible *k* values with which V_1 and V_2 can be microaggregated as a group is between 10 and 178. The value of 153 is selected since it best reflects the balance between privacy and information utility given the input values. However, if V_1 and V_2 were microaggregated seperately, then the *k* values would be 10 and 158, respectively. Therefore, grouping variables for multivariate microaggregation reduces the quality of the obtained solution, since the *k* value selected for the group of variables may fall outside of the range of the *k* values for a single variable.

Information utility and privacy preference		Allocation of Points				Optimum *k*						
		User Preference Points		Intruder Preference Points		Step 1		Step 2		Step 3		
										Univariate micro-aggregation in parallel		Multivariate micro-aggregation
α	β	q_1	q_2	r_1	r_2	V_1	V_2	V_1	V_2	V_1	V_2	V_1 and V_2 in one group
0.25	0.75	25	75	25	75	98	30	10	178	**10**	**158**	**153**
0.25	0.75	25	75	50	50	98	30	56	53	**56**	**52**	**64**
0.25	0.75	25	75	75	25	98	30	178	9	**173**	**30**	**154**
0.25	0.75	50	50	25	75	54	54	10	178	**10**	**167**	**156**
0.25	0.75	50	50	50	50	54	54	56	53	**56**	**53**	**54**
0.25	0.75	50	50	75	25	54	54	178	9	**167**	**9**	**156**
0.25	0.75	75	25	25	75	30	98	10	178	**30**	**173**	**153**
0.25	0.75	75	25	50	50	30	98	56	53	**56**	**53**	**64**
0.25	0.75	75	25	75	25	30	98	178	9	**158**	**9**	**154**
0.5	0.5	25	75	25	75	54	5	5	53	**54**	**16**	**16**
0.5	0.5	25	75	50	50	54	5	17	16	**17**	**5**	**8**
0.5	0.5	25	75	75	25	54	5	56	5	**56**	**5**	**17**
0.5	0.5	50	50	25	75	17	17	5	53	**17**	**52**	**34**
0.5	0.5	50	50	50	50	17	17	17	16	**17**	**16**	**17**
0.5	0.5	50	50	75	25	17	17	56	5	**56**	**17**	**35**
0.5	0.5	75	25	25	75	5	54	5	53	**5**	**53**	**16**
0.5	0.5	75	25	50	50	5	54	17	16	**5**	**16**	**8**
0.5	0.5	75	25	75	25	5	54	56	5	**17**	**54**	**17**
0.75	0.25	25	75	25	75	30	1	3	9	**30**	**2**	**2**
0.75	0.25	25	75	50	50	30	1	5	5	**30**	**2**	**2**
0.75	0.25	25	75	75	25	30	1	10	3	**10**	**2**	**2**
0.75	0.25	50	50	25	75	5	5	3	9	**5**	**5**	**4**
0.75	0.25	50	50	50	50	5	5	5	5	**5**	**5**	**5**
0.75	0.25	50	50	75	25	5	5	10	3	**5**	**5**	**4**
0.75	0.25	75	25	25	75	1	30	3	9	**2**	**9**	**2**
0.75	0.25	75	25	50	50	1	30	5	5	**2**	**30**	**2**
0.75	0.25	75	25	75	25	1	30	10	3	**2**	**30**	**2**

Table 7.3 Changes to the *k* values as the input parameters change

7.4 Conclusion

In this Chapter we discussed the Anonymising function of ANOPI, with a specific focus on how to use global recoding and microaggregation to anonymise microdata such that the optimum levels of privacy and information utility (obtained in the OPI function) are achieved. We have therefore completed the presentation of the ANOPI microdata anonymisation process in this Chapter. By proposing ANOPI, we have achieved the goal of this study and have also answered our research question. In the next Chapter, we conclude this thesis by discussing how the goal of our study was achieved. We will also discuss the main contributions of this study as well as recommendations for future work.

CHAPTER 8

CONCLUSION

8.1 The research problem addressed by this study

When statistical data, such as in the form of microdata, is released, it is necessary to protect the privacy of individuals whose data is released. In order to protect privacy, microdata needs to be anonymised. However, anonymisation reduces the level of information utility, since data is removed (to some extent) from the identifying variables of the microdata set. Therefore, although anonymisation increases the level of privacy in the microdata, it also reduces the level of information utility. Hence, a conflict between privacy and information utility exists. This conflict between privacy and information utility was the research problem that this study addressed.

This study addressed the above problem by answering the research question *"How can the process of microdata anonymisation be guided such that there will exist an optimum balance between privacy and information utility in the anonymised microdata?"* We answered this research question through two research sub-questions. Firstly, we sought to establish how the optimum levels of information utility and privacy should be determined. Secondly, we sought to determine how microdata should be anonymised such that the determined optimum levels of privacy and information utility are achieved. Hence, the goal of this study was to propose a microdata anonymisation process that anonymises microdata such that it will have an optimum balance between privacy and information utility.

In this Chapter, we describe how the research question was answered and how the goal and objectives of this study were achieved. We also discuss the main contributions our study made towards advancing the state of the art (related to privacy protection and microdata anonymisation), and also provide recommendations for future work.

8.2 How did this study solve the research problem?

To solve the research problem, we proposed a microdata anonymisation process called ANOPI (ANonymisation with Optimum Privacy and Information utility). It anonymised microdata by guiding the anonymisation process such that an optimum balance between privacy and information utility will exist in the microdata. ANOPI had two functions: the OPI function (Optimum Privacy and Information utility function) and the Anonymising function. For both functions, algorithms were proposed (in terms of high-level steps). The ANOPI microdata anonymisation process was proposed in Chapter 6. The OPI function and the Anonymising function were presented in Chapters 6 and 7, respectively.

Given a non-anonymised microdata set, the OPI function determined the optimum levels of privacy and information utility. Microdata anonymisation occurred in the Anonymising Function, which ensured that microdata is anonymised such that the optimum levels of privacy and information utility are achieved.

In the algorithm used by the OPI function, the optimum levels of privacy and information utility were determined by applying concepts from Economic Price Theory. In particular, we applied the concepts and techniques used for solving the problem of *utility maximisation of a consumer*. In the problem of *utility maximisation of a consumer*, the consumer's optimum balance between the consumption of goods is determined, when given constraints in terms of prices for the goods and the consumer's budget available for purchasing the goods. This approach was chosen since the objective and the constraints under which the optimum solution (to the problem of utility maximisation of a consumer) is determined can be used to naturally capture the optimisation problem of balancing privacy and information utility. In our case, privacy and information utility were our "goods". The preferences between different identifying variables were used to set the "prices", while the total amount of information available in the non-anonymised microdata represented the "budget". The proposed algorithm in the OPI function was evaluated through a simulation, which showed how the constraints of the optimisation problem impact the optimum levels of privacy and information utility.

The OPI function was proposed independently of any anonymisation technique that may be used to anonymise microdata. It only determined the optimum levels of privacy and information utility that a microdata set should possess, but it did not specify how the microdata should be anonymised to achieve the optimum levels. To specify how microdata should be anonymised, the Anonymising function was proposed. The Anonymising function is dependant on the type of anonymisation technique used, since it uses the optimum levels of privacy and information utility as inputs to determine how the microdata should be anonymised. Since the Anonymising function is dependant on the type of anonymisation technique used, we limited the specification of the Anonymising function to only two anonymisation techniques, namely global recoding and microaggregation. (These techniques were chosen since they are typically used to achieve *k*-anonymity.) For both anonymisation techniques, we proposed an algorithm (in terms of high-level steps) that should be followed in the Anonymising function to anonymise microdata such that the optimum levels of privacy and information utility are achieved. Both algorithms were also evaluated through a simulation that showed changes in the way in which a microdata set is anonymised based on different optimum levels of privacy and information utility.

By proposing the OPI function, we answered the first research sub-question and also achieved our first objective. By proposing the Anonymising function, we answered the second research sub-question and achieved our second objective. Hence, by proposing the ANOPI microdata anonymisation process, we answered our research question and also achieved our goal stated in Chapter 1.

8.3 Main contributions of this study

8.3.1 Advancement of the state of the art

Most of the existing approaches that address the conflict between privacy and information utility in microdata anonymisation only consider the problem from one angle (Domingo-Ferrer & Torra, 2005; Zhang et al., 2007; Xu et al., 2006; LeFevre et al., 2006b; Stark et al., 2006; Ghinita et al., 2007; B. C. M. Fung, K. Wang, L. Wang, &

Hung, 2009; Gionis & Tassa, 2009; Mohammed, B. C. M. Fung, Hung, & Lee, 2009; D. W. Wang, Liau, & Hsu, 2007). That is, they maximise information utility subject to a given level of privacy. In such cases, privacy is only a constraint of the optimisation problem and does not form part of the objective function. Hence, a truly optimum balance between privacy and information utility is not necessarily achieved, since privacy and information utility are not both maximised.

Moreover, other approaches that aim to maximise both privacy and information utility (for example, the approach proposed by Loukides and Shao (2008)), do not take into account the purpose for which the data user requires the data. Therefore, the utility preferences of a specific data user are not taken into account. This may also lead to a solution that does not necessarily provide the optimum level of information utility for a specific user and the purpose for which the data is released.

The main contribution our study made towards advancing the state of the art (related to privacy protection and microdata anonymisation), is the ANOPI microdata anonymisation process. By proposing ANOPI in this study, we proposed a microdata anonymisation process that finds an optimum balance between privacy and information utility such that *both* information utility and privacy are maximised. In addition, the constraints (of the optimisation problem) used by ANOPI are able to capture the data owner's and the data user's preferences. That is, ANOPI is able to take into account the preferences that exist between each identifying variable in the microdata set, as well as the preference between the resulting levels of privacy and information utility. Therefore, ANOPI is also able to take into account the environment in which the anonymised microdata will be used as well as the purpose for which the microdata is released. Hence, we believe that the use of the proposed ANOPI microdata anonymisation process leads to a truly optimum balance between privacy and information utility. Therefore, by using ANOPI, we are able to anonymise microdata without unnecessary loss in privacy or information utility, ensuring higher quality of the released microdata.

In addition to proposing ANOPI, another main contribution this study made to the advancement of the state of the art is a new way for quantifying information utility and privacy. The measures proposed in this study are able to take into account the purpose for which the data is released, as well as the environment in which the data is used.

Lastly, we discussed how global recoding and microaggregation can be applied in the Anonymising function of ANOPI. Hence, we showed how to choose an optimum coding for global recoding such that the optimum balance between privacy and information utility occurs. In addition, we also showed how to choose the optimum least number of records per cluster in microaggregation such that the optimum levels of privacy and information utility are achieved. Moreover, since *k*-anonymity can be achieved through microaggregation, we also showed how to choose the optimum *k* value with which a microdata set should be *k*-anonymised.

8.3.2 Publications produced

Throughout this study, a number of publications were presented at conferences and published in journals. These publications are listed below.

- We used Chapter 5 as a basis for a conference paper (Zielinski & Olivier, 2009a), in which we discussed the extent to which *k*-anonymity is appropriate for addressing the conflict between privacy and information utility in microdata anonymisation.
- We used Chapter 6 as a basis for a journal paper (Zielinski & Olivier, 2010), in which the ANOPI microdata anonymisation process was presented, specifically applied with global recoding as the anonymisation technique.
- We used Chapter 7 as a basis for another journal paper (Zielinski & Olivier, 2009b), which is currently under review, in which we described how to determine the optimum number of records per cluster in microaggregation. Therefore, this paper mainly discussed the use of microaggregation as the anonymisation technique applied in the Anonymising function of ANOPI.
- A number of other supporting papers were developed in the initial stages of this study. These include:
    - Initial presentations of the research problem were made at two conferences (Zielinski, 2006, 2007b), and also in an additional publication in the context of eParticipation (Zielinski, 2007a).
    - A solution for overcoming shortcomings of *k*-anonymity (as discussed in Chapter 5) was also presented at a conference (Zielinski, 2007c).

8.4 Recommendations for future work

To the best of our knowledge, this is the first attempt at using Economic Price Theory for the purpose of determining the optimum levels of privacy and information utility in microdata anonymisation. However, to date, we have not established whether the ANOPI microdata anonymisation process would be useful in practice. Therefore, one recommendation we make for future work is to evaluate the practicality of ANOPI.

In addition, ANOPI has a number of limitations that still need to be addressed. The way in which privacy has been defined is a limitation of this approach. The definition of privacy is based on the overall amount of information that should be removed from a particular variable in all records, rather than being based on the information loss per record. Moreover, this definition does not address the way in which the information should be removed, but only how much information should be removed from a variable. This may therefore lead to cases where, once the microdata has been anonymised through global recoding, it is still possible to have records in the microdata set that are relatively rare. So, although a variable would have the optimum amount of information loss (privacy) as a group of records, it may still have records that are relatively rare.

To address this limitation, it may be necessary to combine global recoding with suppression to ensure that there are no unique or relatively rare records in the microdata set. Alternatively, the definition of privacy may need to be revised, to take into account information loss per record, rather than the information loss per group of records. We leave these aspects as avenues for future work.

Another limitation of ANOPI is that the Anonymising function can currently be applied only with global recoding and microaggregation. Therefore, as avenues for future work, we recommend creating algorithms for determining how to anonymise microdata with other perturbative and non-perturbative microdata anonymisation techniques such that the optimum levels of privacy and information utility are achieved.

As another avenue for future work, we also propose to expand ANOPI such that it will also use additional concepts from Economic Price Theory. For example, the concept of (economic) rationing may be considered. In Economic Price Theory, consumers may be subjected to rationing, which limits the amount of goods that a consumer may purchase (J. Hirshleifer et al., 2005). In such cases, the consumer's optimum point may be different from the optimum point that would be possible if the rationing constraints would not need to be met.

For example, the data owner may set a minimum Privacy level for a particular identifying variable V_i by requiring that it should be anonymised at least to the *j*-th degree. This requirement provides us with an additional constraint that the optimum solution must satisfy, namely that the minimum Privacy level of the released variable V_i should be at least as large as $Priv\left((V_i)^j\right)$. That is, the optimum solution will need to satisfy an additional constraint: $y_i \geq Priv\left((V_i)^j\right)$. In a similar way, the data user may require a minimum Information Utility level for an identifying variable V_i by requiring that it should be anonymised no more than the *j*-th degree. In this case, we will need to ensure that the optimum solution satisfies an additional constraint, namely that the minimum Information Utility level of the released variable V_i should be at least as large as $Iu\left((V_i)^j\right)$. That is, the additional constraint that the optimum solution will need to satisfy is $x_i \geq Iu\left((V_i)^j\right)$.

One can also consider adapting the optimisation problem discussed in this thesis such that it also considers the (monetary) value of information, in addition considering to privacy and information utility. The use of incentives for both sharing and protecting the information may be useful in this respect.

ANOPI was proposed to determine the optimum levels of privacy and information utility and to anonymise microdata such that these levels are achieved. However, it is still unclear which microdata anonymisation technique is the best one to use in a specific set of circumstances. There are clear differences between perturbative and non-perturbative techniques in the way in which they anonymise microdata and in the way

in which the optimum levels of privacy and information utility can be achieved. Hence, it is unclear how to select the best (or optimum) anonymisation technique to achieve the optimum levels of privacy and information utility. Therefore, as another recommendation for future work, we propose a study to determine how to choose a technique (and also how to optimally combine different techniques between identifying variables), in order to anonymise microdata such that the optimum levels of privacy and information utility are achieved.

In this study, we focused on balancing privacy and information utility in microdata anonymisation. The conflict between privacy and information utility is also present when statistical data is disseminated in other forms, such as dynamically queryable databases and tabular data. The approaches for protecting dynamically queryable databases and tabular data have been briefly discussed in Chapter 4. As our last recommendation for future work, we propose to apply Economic Price Theory to balance privacy and information utility when dynamically queryable databases and tabular data are protected.

8.5 Conclusion

Ideally, we would like to release a microdata set with high levels of privacy and information utility. However, privacy and information utility are conflicting requirements – as microdata is anonymised, its level of privacy increases while its level of information utility decreases. It is therefore difficult to determine how to anonymise a microdata set such that it can be released with an optimum balance between privacy and information utility.

The objective and constraints of this optimisation problem can be captured naturally with concepts from Economic Price Theory. Therefore, in this study, we used Economic Price Theory as a basis for proposing a microdata anonymisation process for guiding the process of microdata anonymisation. The microdata anonymisation process is able to anonymise a microdata set such that it will have an optimum balance between privacy and information utility. The proposed microdata anonymisation process first determines the optimum levels of privacy and information utility. Thereafter, it determines how to

anonymise the microdata set such that the optimum levels of privacy and information utility are achieved.

Although only a small subset of Economic Price Theory was used in this study, it nevertheless provided a new perspective on solving the problem of balancing privacy and information utility. This suggests that other concepts and techniques available in Economic Price Theory may provide further insight into solving optimisation problems that exist in information security.

APPENDIX

The Appendix contains the data tables (Tables A1 to A4) used in the examples in Chapters 6 and 7. The data tables contain artificial data. They represent the microdata of patients (with different diseases) who were admitted to a hospital.

Year of Birth	Disease
1967	Cancer
1967	Hypertension
1961	Cancer
1962	Heart disease
1965	Heart disease
1977	Heart disease
1984	Cancer
1978	Hypertension
1977	Hypertension
1965	Heart disease
1990	Cancer
1988	Hypertension
1988	Cancer
1974	Hypertension
1981	Cancer
1961	Heart disease
1983	Heart disease
1963	Heart disease
1982	Cancer
1983	Hypertension
1984	Hypertension
1984	Heart disease
1985	Cancer
1976	Hypertension
1963	Cancer
1965	Hypertension
1977	Cancer
1966	Heart disease
1988	Heart disease
1975	Heart disease
- data continues on right -	

Year of Birth	Disease
- data continued from left -	
1978	Cancer
1989	Hypertension
1978	Hypertension
1962	Heart disease
1984	Cancer
1973	Hypertension
1965	Cancer
1970	Hypertension
1990	Cancer
1963	Heart disease
1983	Heart disease
1974	Heart disease
1964	Cancer
1971	Hypertension
1981	Hypertension
1980	Heart disease
1970	Cancer
1963	Hypertension
1962	Cancer
1976	Hypertension
1984	Cancer
1976	Heart disease
1976	Heart disease
1982	Heart disease
1972	Cancer
1962	Hypertension
1967	Hypertension
1986	Heart disease
1984	Cancer
1973	Hypertension

Table A1 Non-anonymised microdata used as input in Example 1

Year of Birth	Marital Status	Disease
1967	Married	Cancer
1967	Divorced	Hypertension
1961	Widowed	Cancer
1962	Married	Heart disease
1965	Married	Heart disease
1977	Widowed	Heart disease
1984	Divorced	Cancer
1978	Widowed	Hypertension
1977	Divorced	Hypertension
1965	Married	Heart disease
1990	Single	Cancer
1988	Single	Hypertension
1988	Single	Cancer
1974	Divorced	Hypertension
1981	Married	Cancer
1961	Married	Heart disease
1983	Married	Heart disease
1963	Married	Heart disease
1982	Married	Cancer
1983	Married	Hypertension
1984	Single	Hypertension
1984	Single	Heart disease
1985	Single	Cancer
1976	Married	Hypertension
1963	Widowed	Cancer
1965	Divorced	Hypertension
1977	Married	Cancer
1966	Married	Heart disease
1988	Single	Heart disease
1975	Married	Heart disease
- data continues on right -		

Year of Birth	Marital Status	Disease
- data continued from left -		
1978	Divorced	Cancer
1989	Single	Hypertension
1978	Divorced	Hypertension
1962	Married	Heart disease
1984	Single	Cancer
1973	Married	Hypertension
1965	Divorced	Cancer
1970	Divorced	Hypertension
1990	Single	Cancer
1963	Married	Heart disease
1983	Divorced	Heart disease
1974	Married	Heart disease
1964	Divorced	Cancer
1971	Divorced	Hypertension
1981	Married	Hypertension
1980	Married	Heart disease
1970	Married	Cancer
1963	Married	Hypertension
1962	Married	Cancer
1976	Divorced	Hypertension
1984	Single	Cancer
1976	Married	Heart disease
1976	Married	Heart disease
1982	Married	Heart disease
1972	Married	Cancer
1962	Divorced	Hypertension
1967	Divorced	Hypertension
1986	Single	Heart disease
1984	Married	Cancer
1973	Married	Hypertension

Table A2 Non-anonymised microdata used as input in Example 2

Year of Birth	Disease
1966 - 1970	Cancer
1966 - 1970	Hypertension
1961 - 1965	Cancer
1961 - 1965	Heart disease
1961 - 1965	Heart disease
1976 - 1980	Heart disease
1981 - 1985	Cancer
1976 - 1980	Hypertension
1976 - 1980	Hypertension
1961 - 1965	Heart disease
1986 - 1990	Cancer
1986 - 1990	Hypertension
1986 - 1990	Cancer
1971 - 1975	Hypertension
1981 - 1985	Cancer
1961 - 1965	Heart disease
1981 - 1985	Heart disease
1961 - 1965	Heart disease
1981 - 1985	Cancer
1981 - 1985	Hypertension
1981 - 1985	Hypertension
1981 - 1985	Heart disease
1981 - 1985	Cancer
1976 - 1980	Hypertension
1961 - 1965	Cancer
1961 - 1965	Hypertension
1976 - 1980	Cancer
1966 - 1970	Heart disease
1986 - 1990	Heart disease
1971 - 1975	Heart disease
- data continues on right -	

Year of Birth	Disease
- data continued from left -	
1976 - 1980	Cancer
1986 - 1990	Hypertension
1976 - 1980	Hypertension
1961 - 1965	Heart disease
1981 - 1985	Cancer
1971 - 1975	Hypertension
1961 - 1965	Cancer
1966 - 1970	Hypertension
1986 - 1990	Cancer
1961 - 1965	Heart disease
1981 - 1985	Heart disease
1971 - 1975	Heart disease
1961 - 1965	Cancer
1971 - 1975	Hypertension
1981 - 1985	Hypertension
1976 - 1980	Heart disease
1966 - 1970	Cancer
1961 - 1965	Hypertension
1961 - 1965	Cancer
1976 - 1980	Hypertension
1981 - 1985	Cancer
1976 - 1980	Heart disease
1976 - 1980	Heart disease
1981 - 1985	Heart disease
1971 - 1975	Cancer
1961 - 1965	Hypertension
1966 - 1970	Hypertension
1986 - 1990	Heart disease
1981 - 1985	Cancer
1971 - 1975	Hypertension

Table A3 Anonymised microdata output in Example 1

Year of Birth	Marital Status	Disease
1966 - 1970	Been_Married	Cancer
1966 - 1970	Been_Married	Hypertension
1961 - 1965	Been_Married	Cancer
1961 - 1965	Been_Married	Heart disease
1961 - 1965	Been_Married	Heart disease
1976 - 1980	Been_Married	Heart disease
1981 - 1985	Been_Married	Cancer
1976 - 1980	Been_Married	Hypertension
1976 - 1980	Been_Married	Hypertension
1961 - 1965	Been_Married	Heart disease
1986 - 1990	Never_Married	Cancer
1986 - 1990	Never_Married	Hypertension
1986 - 1990	Never_Married	Cancer
1971 - 1975	Been_Married	Hypertension
1981 - 1985	Been_Married	Cancer
1961 - 1965	Been_Married	Heart disease
1981 - 1985	Been_Married	Heart disease
1961 - 1965	Been_Married	Heart disease
1981 - 1985	Been_Married	Cancer
1981 - 1985	Been_Married	Hypertension
1981 - 1985	Never_Married	Hypertension
1981 - 1985	Never_Married	Heart disease
1981 - 1985	Never_Married	Cancer
1976 - 1980	Been_Married	Hypertension
1961 - 1965	Been_Married	Cancer
1961 - 1965	Been_Married	Hypertension
1976 - 1980	Been_Married	Cancer
1966 - 1970	Been_Married	Heart disease
1986 - 1990	Never_Married	Heart disease
1971 - 1975	Been_Married	Heart disease
- data continues on right -		

Year of Birth	Marital Status	Disease
- data continued from left -		
1976 - 1980	Been_Married	Cancer
1986 - 1990	Never_Married	Hypertension
1976 - 1980	Been_Married	Hypertension
1961 - 1965	Been_Married	Heart disease
1981 - 1985	Never_Married	Cancer
1971 - 1975	Been_Married	Hypertension
1961 - 1965	Been_Married	Cancer
1966 - 1970	Been_Married	Hypertension
1986 - 1990	Never_Married	Cancer
1961 - 1965	Been_Married	Heart disease
1981 - 1985	Been_Married	Heart disease
1971 - 1975	Been_Married	Heart disease
1961 - 1965	Been_Married	Cancer
1971 - 1975	Been_Married	Hypertension
1981 - 1985	Been_Married	Hypertension
1976 - 1980	Been_Married	Heart disease
1966 - 1970	Been_Married	Cancer
1961 - 1965	Been_Married	Hypertension
1961 - 1965	Been_Married	Cancer
1976 - 1980	Been_Married	Hypertension
1981 - 1985	Never_Married	Cancer
1976 - 1980	Been_Married	Heart disease
1976 - 1980	Been_Married	Heart disease
1981 - 1985	Been_Married	Heart disease
1971 - 1975	Been_Married	Cancer
1961 - 1965	Been_Married	Hypertension
1966 - 1970	Been_Married	Hypertension
1986 - 1990	Never_Married	Heart disease
1981 - 1985	Been_Married	Cancer
1971 - 1975	Been_Married	Hypertension

Table A4 Anonymised microdata output in Example 2

BIBLIOGRAPHY

Adam, N. R., & Wortmann, J. C. (1989). Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, *21*(4), 515 - 556.

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigraphy, R., Thomas, D., & Zhu, A. (2005). Achieving anonymity via clustering. In *Proceedings of the 10th International Conference on Database Theory*. Chicago, USA.

Agrawal, D., & Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 247 - 255). Santa Barbara, California, United States.

Asuncion, A., & Newman, D. J. (2007). UCI Machine Learning Repository.

Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering*. Tokyo, Japan.

Bertsekas, D. P. (1982). *Constrained optimization and Lagrange multiplier methods*. Academic Press, Inc.

Besanko, D. A., & Braeutigam, R. R. (2005). *Microeconomics* (2nd ed.). John Wiley & Sons, Inc.

Blum, A., Dwork, C., McSherry, F., & Nissim, K. (2005). Practical privacy: the sulq framework. In *Proceedings of the 24th ACM Symposium of the Principles of Data Systems* (pp. 128 - 138). Baltimore, USA.

Bonchi, F., Malin, B., & Saygin, Y. (2008). Recent advances in preserving privacy when mining data. *Data and Knowledge Engineering*, *65*(1), 1 - 4.

Chen, G., & Keller-McNulty, S. A. (1998). Estimation of deidentification disclosure risk in microdata. *Journal of Official Statistics*, *14*(1), 79 - 95.

Chor, B., Kushilevitz, E., Goldreich, O., & Sudan, M. (1998). Private information retrieval. *Journal of the ACM*, *45*(6), 965-981.

Ciriani, V., De Capitani di Vimercati, S., Foresti, S., & Samarati, P. (2007). Microdata protection. In *Yu, T., Jajodia, S. (editors) Secure Data Management in Decentralized Systems* (pp. 291 - 321). Springer-Verlag.

Clifton, C., Kantarcioglu, M., & Vaidya, J. (2005). Privacy-preserving data mining. In *Chu, W.W., Lin, T.Y. (editors), Foundations and Advances in Data Mining* (pp. 313 - 344). Springer-Verlag.

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. (2002). Tools for privacy preserving distributed data mining. *SIGKDD Explorations, 4*(2), 28 - 34. doi:10.1145/772862.772867

Defays, D., & Anwar, N. (1995). Micro-aggregation: a generic method. In *Proceedings of the 2nd International Symposium on Statistical Confidentiality* (pp. 69 - 78). Luxembourg.

Defays, D., & Nanopoulos, P. (1993). Panels of enterprises and confidentiality: the small aggregates method. In *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys* (pp. 195 - 204). Ottowa, Canada.

Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the 22nd ACM Symposium of the Principles of Data Systems* (pp. 202 - 210). San Diego, USA.

Dixit, A. K. (1990). *Optimization in economic theory* (2nd ed.). Oxford University Press.

Domingo-Ferrer, J. (2007). A three-dimensional conceptual framework for database privacy. In *In: Jonker, W., Petkovic, M. (eds.) Secure Data Management*, Lecture Notes in Computer Science (Vol. 4721, pp. 193 - 202). Springer-Verlag.

Domingo-Ferrer, J., Bras-Amoros, M., Wu, Q., & Manjon, J. (2009). User-private information retrieval based on a peer-to-peer community. *Data and Knowledge Engineering, 68*(11), 1237 - 1252.

Domingo-Ferrer, J., Martinez-Balleste, A., Mateo-Sanz, J. M., & Sebe, F. (2006). Efficient multivariate data-oriented microaggregation. *The VLDB Journal, 15*, 355 - 369.

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

Domingo-Ferrer, J., Oganian, A., & Torres, A. (2002). On the security of microaggregation with individual ranking: analytical attacks. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 477 - 491.

Domingo-Ferrer, J., & Saygin, Y. (2009). Recent progress in database privacy. *Data and Knowledge Engineering*, *68*(11), 1157 - 1159.

Domingo-Ferrer, J., Sebe, F., & Solanas, A. (2008). A polynomial-time approximation to optimal multivariate microaggregation. *Computers and Mathematics with Applications*, *55*, 714 - 732.

Domingo-Ferrer, J., & Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. In *Doyle, P., Lane J.I., Theeuwes, J.J., Zayatz, L. (editors) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (pp. 111 - 134). North-Holland, Amsterdam.

Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, *11*(2), 195 - 212.

Domingo-Ferrer, J., & Torra, V. (2008). A critique of k-anonymity and some of its enhancements. In *Proceedings of the 2008 Third International Conference on Availability, Reliability and Security*. Barcelona, Spain.

Duncan, G. T., Feinberg, S. E., Krishnan, R., Padman, R., & Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data. In *Doyle, P., Lane J.I., Theeuwes, J.J., Zayatz, L. (editors) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (pp. 135 - 166). North-Holland, Amsterdam.

Duncan, G. T., Keller-McNulty, S. A., & Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Technical Report LA-UR-01-6428, Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, USA.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference (TCC)* (pp. 265 - 284). New York, USA.

Emekci, F., Sahin, O. D., Agrawal, D., & Abbadi, A. E. (2007). Privacy preserving decision tree learning over multiple parties. *Data and Knowledge Engineering*, *63*(2), 348 - 361.

FCSM. (1994). Federal Committee on Statistical Methodology. Statistical policy working paper 22: report on statistical disclosure limitation methodology. Office of Management and Budget, USA.

Fung, B. C. M., Wang, K., Wang, L., & Hung, P. C. K. (2009). Privacy-preserving data publishing for cluster analysis. *Data and Knowledge Engineering*, *68*(6), 552 - 575.

Fung, R. J., Wang, K., & Yu, P. (2005). Top-down specialization for information and privacy preservation. In *Proceedings of the 21st International Conference on Data Engineering*. Tokyo, Japan.

Gavison, R. (1980). Privacy and the limits of the law. *Yale Law Journal*, *89*, 421 - 471.

Ghinita, G., Karras, P., Kalnis, P., & Mamoulis, N. (2007). Fast Data Anonymization with Low Information Loss. In *Proceedings of the 33rd International Conference on Very Large Data Bases* (pp. 758 - 769). Vienna, Austria.

Gionis, A., & Tassa, T. (2009). k-anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering*, *21*(2), 206 - 219.

Gostin, L., & Turek-Brezina, J. (1995). Privacy and security of health information in the emerging health care system. *Journal of Law Medicine*, *5*(1), 1 - 36.

Gross, H. (1971). Privacy and autonomy. *Nomos*, *13*, 169 - 180.

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

Hirshleifer, J., Glazer, A., & Hirshleifer, D. (2005). *Price theory and applications: Decisions, markets and information* (7th ed.). Cambridge University Press.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E., et al. (2007). Handbook on statistical disclosure control, Version 1.01.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., De Wolf, P., Domingo-Ferrer, J., et al. (2005). *u-Argus version 4.0.2 Software and User's Manual*. Statistics Netherlands.

Jaro, M. A. (1989). Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association, 84*(406), 414 - 420.

Kantarcioglu, M., & Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Trans. on Knowl. and Data Eng., 16*(9), 1026 - 1037.

Kenthapadi, K., Mishra, N., & Nissim, K. (2005). Simulatable auditing. In *Proceedings of the 24th ACM Symposium of the Principles of Data Systems* (pp. 118 - 127). Baltimore, USA.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics, 9*, 313 - 331.

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2005). Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on management of data* (pp. 46 - 60). Baltimore, USA.

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006a). Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering*. Atlanta, USA.

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006b). Workload-aware anonymization. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 277 - 286). Philadelphia, USA.

Li, J., Wong, R. C., Fu, A. W., & Pei, J. (2006). Achieving k-anonymity by clustering in attribute hierarchical structures. In *Proceedings of the 8th International Data Warehousing and Knowledge Discovery Conference* (pp. 405 - 416). Krakow, Poland.

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd IEEE International Conference on Data Engineering* (pp. 106 - 115). Washington D.C., USA,

Lindell, Y., & Pinkas, B. (2002). Privacy preserving data mining. *Journal of Cryptology*, *15*(3), 177 - 206.

Liu, L., Kantarcioglu, M., & Thuraisingham, B. (2008). The applicability of the perturbation based privacy preserving data mining for real-world data. *Data and Knowledge Engineering*, *65*(1), 5 - 21.

Loukides, G., & Shao, J. (2008). Data utility and privacy protection trade-off in k-anonymisation. In *Proceedings of the 2008 International workshop on Privacy and Anonymity in Information Society* (pp. 36 - 45). Nantes, France.

Machanavajjhala, A., Gehrke, J., Kiefer, D., & Venkitasubramanian, M. (2006). l-diversity: privacy beyond k-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering*. Atlanta, USA.

Machanavajjhala, A., Kiefer, D., Gehrke, J., & Venkitasubramanian, M. (2007). l-diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1).

Magkos, E., Maragoudakis, M., Chrissikopoulos, V., & Gritzalis, S. (2009). Accurate and large-scale privacy-preserving data mining using the election paradigm. *Data and Knowledge Engineering*, *68*(11), 1224 - 1236.

Mansfield, E. (1985). *Microeconomics: theory and applications* (5th ed.). W. W. Norton & Company, Inc.

Medrano-Gracia, P., Pont-Tuset, J., Nin, J., & Muntes-Mulero, V. (2007). Ordered dataset vectorization for linear regression on data privacy. In *Proceedings of the 4th international conference on Modeling Decisions for Artificial Intelligence* (pp. 361 - 372). Kitakyushu, Japan.

Meyerson, A., & Williams, R. (2004). On the complexity of optimal k-anonymity. In *Proceedings of the 23rd ACM Symposium of the Principles of Data Systems* (pp. 223 - 228). Paris, France.

Miller, F. G. (2008). Research on medical records without informed consent. *Journal of Law, Medicine and Ethics, 36*(3), 560 - 566.

Mohammed, N., Fung, B. C. M., Hung, P. C. K., & Lee, C. (2009). Anonymizing healthcare data: a case study on the blood transfusion service. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining* (pp. 1285 - 1294). Paris, France.

Nabar, S. U., Marthi, B., Kenthapadi, K., Mishra, N., & Motwani, R. (2006). Towards robustness in query auditing. In *Proceedings of the 32nd International Conference on Very Large Data Bases* (pp. 151 - 162). Seoul, Korea.

Nin, J., Herranz, J., & Torra, V. (2008a). How to group attributes in multivariate microaggregation. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 161*(1), 121 - 138.

Nin, J., Herranz, J., & Torra, V. (2008b). On the disclosure risk of multivariate microaggregation. *Data and Knowledge Engineering, 67*(3), 399 - 412.

OCR. (2003). Summary of the HIPAA Privacy Rule. Office for Civil Rights, Department of Health and Human Services, USA. Retrieved March 2, 2009, from http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary.pdf.

OECD. (2007). Glossary of statistical terms. *Organisation for Economic Co-operation and Development*.

Oganian, A., & Domingo-Ferrer, J. (2001). On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Commission for Europe, 18*(4), 345 - 354.

Ostrovsky, R., & Skeith, W. (2007). A Survey of Single-Database Private Information Retrieval: Techniques and Applications. In *Public Key Cryptography – PKC 2007* (pp. 393 - 411).

Pagliuca, D., & Seri, G. (1999). Some results of individual ranking method on the system of enterprise accounts annual survey. *Esprit SDC Project, Deliverable MI-3/D2*.

Paul, P. (2001). Mixed signals: when it comes to issues of privacy, consumers are fraught with contradictions. *American Demographics*, *23*, 45 - 49.

Pfleeger, C. P. (1997). *Security in computing* (2nd ed.). Prentice Hall, Inc.

Pongas, G., & Vernadat, F. (2003). Data life-cycle object model for statistical information systems. In *Proceedings of the Joint ECE/Eurostat/OECD meeting on the management of statistical information systems*. Geneva, Switzerland.

Rachels, J. (1984). Why is privacy important? In *Schoeman, F.D. (editor) Philosophical Dimensions of Privacy: An Anthology* (pp. 290 - 299). Cambridge University Press.

Räikkä, J. (2008). Is privacy relative? *Journal of Social Philosophy*, *39*(4), 534 - 546.

Rauhofer, J. (2008). Privacy is dead, get over it! Information privacy and the dream of a risk-free society. *Information and Communications Technology Law*, *17*(3), 185 - 197.

Reiter, J. P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society*, Series A, *168*, 185 - 205.

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

Schirmacher, W. (1985). Privacy as an ethical problem in the computer society. In *Mitcham, C., Huning, A. (editors) Philosophy and Technology II: Information Technology and Computers in Theory and Practice* (pp. 257 - 268). D. Reidel Publishing Company.

Sebe, F., Domingo-Ferrer, J., Mateo-Sanz, J. M., & Torra, V. (2002). Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets. In *Domingo-Ferrer, J. (editor) Inference Control in Statistical Databases, From Theory to Practice*, Lecture Notes in Computer Science (Vol. 2316, pp. 163 - 171). Springer-Verlag.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379 - 423, 623 - 656.

Simon, G. E., Unützer, J., Young, B. E., & Pincus, H. A. (2000). Large medical databases, population-based research, and patient confidentiality. *American Journal of Psychiatry*, *157*(11), 1731 - 1737.

Skinner, C. J., & Elliot, M. J. (2001). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society*, Series B, *64*(4), 855 - 867.

Skinner, C. J., Marsh, C., Openshaw, S., & Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, *10*(1), 31 - 51.

Stark, K., Eder, J., & Zatloukal, K. (2006). Priority-based k-anonymity accomplished by weighted generalisation structures. In *Proceedings of the 8th International Data Warehousing and Knowledge Discovery Conference* (pp. 394 - 404). Krakow, Poland.

Starr, P. (1999). Health and the right to privacy. *American Journal of Law and Medicine*, *25*(2 and 3), 193 - 201.

Sweeney, L. (2002a). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 571 - 588.

Sweeney, L. (2002b). k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557 - 570.

Tavani, H. T. (2007). Philosophical theories of privacy: implications for an adequate online privacy policy. *Metaphilosophy*, *38*(1), 1 - 22.

Torra, V. (2004). Microaggregation for categorical variables: a median based approach. In *Domingo-Ferrer, J., Torra, V. (editors) Privacy in Statistical Databases*, Lecture Notes in Computer Science (Vol. 3050, pp. 162 - 174). Springer-Verlag.

Truta, T. M., & Vinay, B. (2006). Privacy protection: p-sensitive k-anonymity property. In *Proceedings of the 2nd International Workshop on Privacy Data Management* (p. 94). Atlanta, USA.

Vaidya, J., & Clifton, C. (2002). Privacy preserving association rule mining in vertically partitioned data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 639 - 644). Edmonton, Alberta, Canada: ACM.

Von Staden, H. (1996). In a pure and holy way: personal and professional conduct in the Hippocratic oath. *Journal of the History of Medicine and Allied Sciences, 51*, 404 - 437.

Wald, N., Law, M., Meade, T., Miller, G., Alberman, E., & Dickinson, J. (1994). Use of personal medical records for research purposes. *BMJ, 309*(6966), 1422 - 1424.

Wang, D. W., Liau, C. J., & Hsu, T. S. (2007). An epistemic framework for privacy protection in database linking. *Data and Knowledge Engineering, 61*(1), 176 - 205.

Warren, S., & Brandeis, L. (1890). The right to privacy. *Harvard Law Review, 14*(5), 193 - 220.

Willenborg, L., & De Waal, T. (2001). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics. Springer-Verlag.

Winkler, W. E. (2004). Re-identification methods for masked microdata. In *Domingo-Ferrer, J., Torra, V. (editors) Privacy in Statistical Databases*, Lecture Notes in Computer Science (Vol. 3050, pp. 216 - 230). Springer-Verlag.

Wong, R. C., Li, J., Fu, W., & Wang, K. (2006). (a, k) anonymity: an enhanced k-anonymity model for privacy-preserving data publishing (pp. 754 - 759). Presented at the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA.

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W. (2006). Utility-based anonymization for privacy preservation with less information loss. *ACM SIGKDD Explorations, 8*(2), 21 - 30.

Yancey, W. E., Winkler, W. E., & Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In *Domingo-Ferrer, J. (editor) Inference Control in Statistical Databases, From Theory to Practice*, Lecture Notes in Computer Science (Vol. 2316, pp. 135 - 152).

Zhang, L., Jajodia, S., & Brodsky, A. (2007). Information disclosure under realistic assumptions: privacy versus optimality. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*. Alexandria, USA.

Zielinski, M. P. (2006). Guiding the release of microdata by using support and confidence of association rules. In *Proceedings of the EGOV'06 PhD Colloquium*, Krakow, Poland.

Zielinski, M. P. (2007a). Privacy protection in eParticipation: guiding the anonymisation of microdata. In *Avdic, A., Hedstrom, K., Rose, J., Gronuld, A. (editors) Understanding eParticipation - Contemporary PhD eParticipation studies in Europe* (pp. 57 - 69). Örebro University Library, Sweden.

Zielinski, M. P. (2007b). Balancing privacy and information utility in microdata anonymisation. In *Proceedings of the 2007 Digital Identity and Privacy Conference*. Maastricht, The Netherlands.

Zielinski, M. P. (2007c). Overcoming the limitations of k-anonymity through association rule hiding. In *Proceedings of the 2007 Digital Identity and Privacy Conference*. Maastricht, The Netherlands.

Zielinski, M. P., & Olivier, M. S. (2009a). How appropriate is k-anonymity for addressing the conflict between privacy and information utility in microdata anonymisation. In *Proceedings of the 2009 Information Security South Africa (ISSA) Conference*. Johannesburg, South Africa.

Zielinski, M. P., & Olivier, M. S. (2009b). How to determine the optimum number of records per cluster in microaggregation. *(Submitted for publication).*

Zielinski, M. P., & Olivier, M. S. (2010). On the use of Economic Price Theory to find the optimum levels of privacy and information utility in non-perturbative microdata anonymization. *Data and Knowledge Engineering*, *69*(5), 399 - 423.