Privacy and utility assessment within statistical data bases

Louis-Philippe Sondeck

THESE DE DOCTORAT
CONJOINT TELECOM SUDPARIS et L' UNIVERSITE PIERRE ET MARIE CURIE

Spécialité: Informatique et Réseaux

École doctorale: Informatique, Télécommunications et Electronique de Paris

Présentée par

Louis Philippe SONDECK

Pour obtenir le grade de

DOCTEUR DE TELECOM SUDPARIS

Mesure de la vie privée et de l'utilité des données dans les bases de données statistiques

---

Soutenue le 15 décembre 2017 devant le jury composé de :

Sara FORESTI
Professeur associé, Università degli Studi di Milano, Italie - *Rapporteur*
Benjamin NGUYEN
Professeur HDR, INSA Val de Loire, France - *Rapporteur*
Sébastien TIXEUIL
Professeur HDR, Université Pierre et Marie Curie, France - *Examinateur*
Christophe ROSENBERGER
Professeur HDR, ENSICAEN, France - *Examinateur*
Vincent FREY
Ingénieur de recherche, Orange Labs, France - *Encadrant*
Maryline LAURENT
Professeur HDR, Télécom SudParis, France - *Directrice de thèse*

Thèse No : 2017TELE0023

PHD THESIS TELECOM SUDPARIS IN PARTNERSHIP WITH PIERRE ET MARIE CURIE UNIVERSITY

Speciality: Computer Science and Networks

Doctoral School: Informatique, Télécommunications et Electronique de Paris

Presented by

Louis Philippe SONDECK

To obtain the degree of

DOCTOR OF TELECOM SUDPARIS

Privacy and Utility Assessment within Statistical Data Bases

Presented on December 15th, 2017 with the Jury composed by:

Sara FORESTI
Associate Professor, Università degli Studi di Milano, Italy - *Reviewer*
Benjamin NGUYEN
Professor, INSA Val de Loire, France - *Reviewer*
Sébastien TIXEUIL
Professor, Université Pierre et Marie Curie, France - *Examiner*
Christophe ROSENBERGER
Professor, ENSICAEN, France - *Examiner*
Vincent FREY
Research Engeneer, Orange Labs, France - *Supervisor*
Maryline LAURENT
Professor, Télécom SudParis, France - *Thesis Director*

Thesis No : 2017TELE0023

To my mother.

Abstract

Personal data promise relevant improvements in almost every economy sectors thanks to all the knowledge that can be extracted from it. As a proof of it, some of the biggest companies in the world, Google, Amazon, Facebook and Apple (GAFA) rely on this resource for providing their services. However, although personal data can be very useful for improvement and development of services, they can also, intentionally or not, harm data respondent's privacy. Indeed, many studies have shown how data that were intended to protect respondents' personal data were finally used to leak private information. Therefore, it becomes necessary to provide methods for protecting respondent's privacy while ensuring utility of data for services. For this purpose, Europe has established a new regulation (The General Data Protection Regulation) (EU, 2016) that aims to protect European citizens' personal data. However, the regulation only targets one side of the main goal as it focuses on privacy of citizens while the goal is about the best trade-off between privacy and utility. Indeed, privacy and utility are usually inversely proportional and the greater the privacy, the lower the data utility. One of the main approaches for addressing the trade-off between privacy and utility is data anonymization. In the literature, anonymization refers either to anonymization mechanisms or anonymization metrics. While the mechanisms are useful for anonymizing data, metrics are necessary to validate whether or not the best trade-off has been reached. However, existing metrics have several flaws including the lack of accuracy and the complexity of implementation. Moreover existing metrics are intended to assess either privacy or utility, this adds difficulties when assessing the trade-off between privacy and utility. In this thesis, we propose a novel approach for assessing both utility and privacy called Discrimination Rate (DR). The DR is an information theoretical approach which provides practical and fine grained measurements. The DR measures the capability of attributes to refine a set of respondents with measurements scaled between 0 and 1, the best refinement leading to single respondents. For example an identifier has a DR equals to 1 as it completely refines a set of respondents. We are therefore able to provide fine grained assessments and comparison of anonymization mechanisms (whether different instantiations of the same mechanism or different anonymization mechanisms) in terms of utility and privacy. Moreover, thanks to the DR, we provide formal definitions of identifiers (Personally Identifying Information) which has been recognized as one of the main concern of privacy regulations. The DR can therefore be used both by companies and regulators for tackling the personal data protection issues.

Résumé

Les données personnelles sont d'une importance avérée pour presque tous les secteurs d'activité économiques grâce a toute la connaissance qu'on peut en extraire. Pour preuve, les plus grandes entreprises du monde que sont: Google, Amazon, Facebook et Apple s'en servent principalement pour fournir de leurs services. Cependant, bien que les données personnelles soient d'une grande utilité pour l'amélioration et le développement de nouveaux services, elles peuvent aussi, de manière intentionnelle ou non, nuire à la vie privée des personnes concernées. En effet, plusieurs études font état d'attaques réalisées à partir de données d'entreprises, et ceci, bien qu'ayant été anonymisées. Il devient donc nécessaire de définir des techniques fiables, pour la protection de la vie privée des personnes tout en garantissant l'utilité de ces données pour les services. Dans cette optique, l'Europe a adopté un nouveau règlement (Le Règlement Général sur la Protection des Données) (EU, 2016) qui a pour but de protéger les données personnelles des citoyens européens. Cependant, ce règlement ne concerne qu'une partie du problème puisqu'il s'intéresse uniquement à la protection de la vie privée, alors que l'objectif serait de trouver le meilleur compromis entre vie privée et utilité des données. En effet, vie privée et utilité des données sont très souvent inversement proportionnelles, c'est ainsi que plus les données garantissent la vie privée, moins il y reste d'information utile. Pour répondre à ce problème de compromis entre vie privée et utilité des données, la technique la plus utilisée est l'anonymisation des données. Dans la littérature scientifique, l'anonymisation fait référence soit aux mécanismes d'anonymisation, soit aux métriques d'anonymisation. Si les mécanismes d'anonymisation sont utiles pour anonymiser les données, les métriques d'anonymisation sont elles, nécessaires pour valider ou non si le compromis entre vie privée et utilité des données a été atteint. Cependant, les métriques existantes ont plusieurs défauts parmi lesquels, le manque de précision des mesures et la difficulté d'implémentation. De plus, les métriques existantes permettent de mesurer soit la vie privée, soit l'utilité des données, mais pas les deux simultanément; ce qui rend plus complexe l'évaluation du compromis entre vie privée et utilité des données. Dans cette thèse, nous proposons une approche nouvelle, permettant de mesurer à la fois la vie privée et l'utilité des données, dénommée Discrimination Rate (DR). Le DR est une métrique basée sur la théorie de l'information, qui est pratique et permet des mesures d'une grande finesse. Le DR mesure la capacité des attributs à raffiner un ensemble d'individus, avec des valeurs comprises entre 0 et 1; le meilleur raffinement conduisant à un DR de 1. Par exemple, un identifiant a un DR égale à 1 étant donné qu'il permet de raffiner complètement un ensemble d'individus. Grâce au DR nous évaluons de manière précise et comparons les mécanismes d'anonymisation en termes d'utilité et de vie privée (aussi bien différentes instanciations d'un même mécanisme, que différents mécanismes). De plus, grâce au DR, nous proposons des définitions formelles des identifiants encore appelés informations d'identification personnelle. Ce dernier point est reconnu comme l'un des problèmes cruciaux des textes juridiques qui traitent de la protection de la vie privée. Le DR apporte donc une réponse aussi bien aux entreprises qu'aux régulateurs, par rapport aux enjeux que soulève la protection des données personnelles.

Acknowledgements

There are many people who have participated in one way or another in the accomplishment of this work, whether by advices, encouragements or rewarding discussions. I would like to thank all of you for all your support.

I would like to express my deepest gratitude to **Maryline LAURENT**, my thesis director for her support, advices and trust. This thesis has been a very particular experience, with highs and lows, and she always had trust in me even when all was very low. Maryline taught me to never stop going furhter in the work, as anything is still perfectible and can be improved; this mindset has positively influenced various aspects of my life, whether professional or personal. I can hardly express all the gratitude I have for her and how her support has been crucial not only for the accomplishment of this work, but also for the accomplishment of the person I am today.

I am very much thankful to **Vincent FREY**, my senior project supervisor for his commitment and guidance throughout this thesis. Vincent helped me with new ideas and advices that have been useful for the refinement of this work. He has always made himself available for resolving technical as well as administrative difficulties, and has greatly facilitated my integration and collaboration within the various projects in which I participated.

A special thank to the "PhD Dpt." (**Dr. Yanhuang LI**, **Kevin CORRE**, **Marco LOBE KOME**, **Youssou NDIAYE**, **Julien HATIN**) who have been my nearest collaborators during these three years and with whom I shared the same office. It has been a great pleasure to meet you and share with you during these years. I am very thankful to all the support you gave me, this experience would have been very difficult without your support.

I would like to thank **Prof. Benjamin NGUYEN** and **Dr. Sara FORESTI** who are the reporters of my thesis and who provided me with recommandations which greatly improve the quality of this work. Thanks a lot to **Prof. Sébastien TIXEUIL** and **Prof. Christophe ROSENBERGER** for their interest and for being part of the jury of my thesis.

A special thank to the projects I participated to: **ARDECO**, **ADAGE** which gave me the opportunity to explore realistic use cases and allowed me to approach the problematic from different angles. A special thank to **Stéphane PETIT**, **Anne-Sophie PIGNOL**, **Emilie SIRVENT-HIEN**, **Baptiste OLIVIER** for the rewarding discussions.

I would like to thank **Dr. Jacques TRAORE** for his advices, interest and availability for discussing new ideas and approaches. His experience and recommendations have been very usefull for helping me refining my contributions.

I am very grateful to my parents **Gabriel SONDECK** and **Gisèle SONDECK** without whom all this would never have happened, to my little brothers **Maxime**, **Ludovic**, **Dominique** and **Ferdinand** for their support and trust, and for whom I am

a model.

I can not conclude without thanking **Joel EVAIN**, my team manager, for facilitating my integration into the research team and his availability to respond promptly to all administrative requests, my colleagues and time spent together, the meetings and discussions; it has been a great pleasure to share all this time with you.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

1.1 Data Privacy Issues with Identified Limitations of the Current GDPR Regulation

It is difficult to estimate all the benefits that personal data can provide both to users and to companies. Personal data are used in every sector for improvements and development of new services including: consumers' risk analysis, reduction of transaction costs, increase of advertising returns. In 2009, the European Commissioner for Consumer Protection compared personal data to oil (Spiekermann et al., 2015), in order to illustrate its implication for creating added value for companies. A report by the Boston Consulting Group (Global, 2012), projects that the personal data sectors will produce up to 1 trillion euro in corporate profits in Europe by 2020.

On the other hand, personal data can represent an important burden for companies due to the inherent risk of privacy violation. Indeed, the data may reveal more information to the data processors than the respondent (Domingo-Ferrer, 2007) desire, leading to privacy violation which could affect the company in terms of reputation and legal penalties. For example, in 2006, America Online's (AOL) released twenty million anonymized search queries (Barbaro, Zeller, and Hansell, 2006) for the benefit of researchers; thereafter, the data have been used to re-identify Thelma Arnold, a 62-year old widow living in Lilburn. To adress such issues, Europe has defined a new regulation for privacy management (GDPR: General Data Protection Regulation) which will take effect on May 2018 and which aims protecting European citizens' privacy. One of the main changes of this regulation according to previous ones is penalty: offending companies could be fined up to 4% of annual global turnover or 20 million euro.

However, according to our analysis, two points of concern still need clarification to support the implementation of the GDPR, as discussed in Section 1.3.

1.2 GDPR the New Regulation for Protecting Personal Data

The General Data Protection Regulation (GDPR) is the new regulation for protecting users' privacy which has been approved by the European Union (EU) on 14 April 2016 and will be directly applied in all members states on 25 May 2018. GDPR replaces the previous regulation which was the Data Protection Directive 95/46/EC adopted in 1995. The GDPR provides 3 main key changes:

- **Increased Territorial Scope (extra-territorial applicability):** GDPR applies to all companies processing personal data of data subjects residing in the Union,

regardless of the company's location. Indeed, the previous data protection regulation was intended to take into account the context of each member state which was not clearly defined and has arisen in a number of high profile court cases.

- **Penalties:** under GDPR, offending companies can be fined up to 4% of annual global turnover or 20 Million euro (whichever is greater).
- **Consent**: consent should be clearly asked by service providers and long and illegible terms are no longer allowed. Moreover, the request for consent must clearly describe the purpose for data processing.

Other changes include: **breach notification** (breach notification is mandatory under GDPR and must be done within 72 hours), **right to access** (the data subject has the right to obtain from the data processor a copy of its personal data which should be free of charge), **right to be forgotten** (the data subject has the right to ask for erasing of all his personal data from the data processor data bases).

The first reason companies are encouraged to be compliant to GDPR is the fines. Indeed, fines are relatively high (4% of annual global turnover or 20 Million euro, whichever is greater) and this could be persuasive enough.

Another reason is trust. Indeed the main claimed goal of GDPR is to reinforce the trust between companies and customers as without trust, innovation and development of new services can not take place. Moreover, data breaches can damage the reputation of a company with a great impact on incomes.

1.3 Lack of Clarity of GDPR for its Implementation

1.3.1 The Need to Characterize Identifiers, as not Clearly Addressed in the GDPR

The main difficulty for privacy management comes from its legal definition and especially from the difficulty to characterize Identifiers (Schwartz and Solove, 2011). Indeed the GDPR, as other regulations before (Schwartz and Solove, 2011), lacks a proper definition of identifiers which may be biased. Indeed, in its Article 32, the GDPR defines the means to ensure the security of processing, which can be separated into two groups:

- Methods for ensuring and evaluating data privacy: pseudonymization (defined below) and encryption.
- Methods for ensuring confidentiality, integrity, availability and resilience.

In addition to these measures, the GDPR encourages in its Article 40 : " *the drawing up of codes of conduct intended to contribute to the proper application of the Regulation, taking account of the specific features of the various processing sectors and the specific needs of micro, small and medium-sized enterprises.*"

Therefore, the GDPR proposes *pseudonymization* as a mean to ensure data privacy. Pseudonymization is defined in Article 4 as:

"*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that*

such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or Identifiable Natural Person."

In other words, pseudonymization consists in transforming personal data such that they can no longer be linkable to a specific data subject. .

Personal data are defined as :

"any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."

However, the term *identifier* is not defined, this definition only provides an enumeration of specific cases which are obviously not exhaustive taking into account the diversity of applications. Furthermore, considering the proposed set of identifiers, could they be considered as such in any context ?

To illustrate the complexity behind this terminology let us consider the following example which underlines the importance of the context when defining identifiers:

Suppose there is a badly parked car on the street and we need to identify the owner of this car. Suppose we are able to find that the owner, M. John, is in a bar in the same street. Now suppose that in the same bar there are many people named John. Can we consider in this case that the name is an identifier as it does not refer anymore to a single person ? On the other hand, suppose we are able to know that the owner is wearing a white shirt and we find that in the same bar, there is only one person wearing a white shirt. Then the white shirt becomes an identifier in this case since it is enough to identify the car's owner.

This example shows the complexity of the identifiers terminology and the importance of context which is not taken into account in the current legal text.

A proper definition of *identifiers* is an important concern as *personal data* are defined according to *identifiers* and *pseudonymization* (the recommended mechanism) is about the protection of *personal data*. Hence, without a clear definition of *identifiers*, data processors would not be able to identify them (*identifiers*) and therefore, would not be able to protect *personal data*.

1.3.2 The Need for using Anonymization instead of Pseudonymization (a Lacking Point of the GDPR)

Pseudonymization is not enough for protecting data as it only prevents a direct link with a subject (cf. Section 1.3); however, the remaining knowledge within *pseudonymized* data can still be used for re-identifying a specific subject (Hansell, 2006) (Barbaro, Zeller, and Hansell, 2006). In the literature there is a more general mechanism to protect personal data which is *anonymization*. *Anonymization* consists in transforming data such that the data can no longer be linked to a particular data subject and no information can be learned from him while still enabling utility of data. This definition includes both the inability to link an information to a subject and the inability to learn any information on specific subjects.

TABLE 1.1: Original Data Table.

	Age	Disease
1	22	lung cancer
2	22	lung cancer
3	22	lung cancer
4	45	stomach cancer
5	63	diabetes
6	40	flu
7	35	aids
8	35	aids
9	32	diabetes

TABLE 1.2: Anonymized Data Table.

	Age*	Disease
1	2*	lung cancer
2	2*	lung cancer
3	2*	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	3*	aids
8	3*	aids
9	3*	diabetes

The following example describes the anonymization process. Let us consider the two data tables (Table 1.1 and Table 1.2); an Original Data Table (Table 1.1) which represents the raw data (not anonymized) and an Anonymized Data Table (Table 1.2) which is the corresponding anonymized table. This is a specific instantiation of the k-anonymity anonymization model (Samarati and Sweeney, 1998) (Sweeney, 2002) (Samarati, 2001).

The idea of k-anonymity is to transform some attribute's values (key attribute, e.g. *Age*) in order to reduce their identifying capabilities according to another type of attribute (sensitive attribute, e.g. *Disease*) by forming subsets of k records. In our example, the key attribute *Age* is transformed into *Age** within the anonymized table, the sensitive attribute *Disease* is not transformed and subsets of 3 records are formed (table 1.2 is therefore a 3-anonymity table of Table 1.1).

The main concern about anonymization is the *trade off between data privacy and data utility* (Li and Li, 2009) (Loukides and Shao, 2008a). Indeed, while data privacy is recommended by GDPR, service providers need to use the data for improvement of their services and since anonymization is about reducing the attributes' inherent information, there should be a trade off between what is recommended by authorities and what is needed by service providers. Anonymization should therefore be precisely evaluated in order to provide *the good balance between data privacy and data utility*.

The previous anonymization instantiation is therefore intended to respond to both privacy and data utility. Privacy refers to the ability of an attacker to re-identify a given respondent, while utility refers to the capability of a data processor to extract useful information from data. Both issues can be formulated in terms of questions. We provide here an example of possible questions:

- **Privacy:** how much is it possible to go from the transformed *Age** values to its corresponding raw values (*Age*) ?
- **Utility:** how much from the defined Age groups, is it possible to identify the corresponding Disease ?

Indeed, the privacy violation comes from the ability to re-identify a subject and therefore in our example, to go from the transformed attribute *Age** to the raw attribute *Age* (within the Original Data table); and measuring privacy would consist in measuring this ability.

On the other hand, utility refers to the capacity to extract useful information from the data. In this case, measuring utility would consist in measuring how, from Age, we can guess the disease a subject suffers for.

However, as we will see in Chapter 3, utility extraction can also be considered as a privacy violation and the difficulty of privacy evaluation would therefore consist in delimiting the border between privacy and utility.

This example shows how complex anonymization can be, since it refers to specific questions which depend on the considered case.

Alongside the previous concern, the regulation encourages in its article 42: "*the establishment of data protection certification mechanisms and of data protection seals and marks, for the purpose of demonstrating compliance with this Regulation of processing operations by controllers and processors.*"

However, in the literature, although there are many propositions of anonymization mechanisms, there are few practical metrics for quantifying privacy vs. utility. Moreover, there is no uniform approach for comparison of different anonymization techniques/instantiations. This last concern is necessary for choosing the most suitable technique and therefore to assess the compliance with the regulation.

1.4 The Relevant Issues

We can therefore consider the following issues:

1. **How can we characterize identifiers with respect to the context ?** Identifiers are at the centre of the regulation around personal data as personal data are defined according to identifiers; however, the current regulation does not consider the context in its definition and a lack of a proper characterization can lead to misunderstandings.
2. **Which approach is to be considered for privacy assessment ?** Many approaches exist for privacy measurements but for defining consistent regulation rules, there is the need for a global approach.

3. **How can we define a scale for comparing various anonymity mechanisms ?** Many mechanisms exist for protecting privacy but few are pratical. Moreover, there is no uniform scale that can be used for comparison.

4. **How can we measure utility in terms of specific needs (questions) ?** As utility refers to a specific data use and therefore to a specific question, an accurate assessment would consist in measuring data with respect to a need; which is difficult to implement because of the subjectivity of the need.

1.5 Contributions of this Thesis

In this thesis we propose a metric called Discrimination Rate (DR) that enables:

- **Goal 1:** A fine grained characterization of identifiers with respect to the context.
- **Goal 2:** A measurement of anonymity degree in terms of attribute identification capabilities which provide a fine granularity and therefore is usable for different domains.
- **Goal 3:** An accurate evaluation and comparison of the existing anonymization techniques in terms of disclosure risk.
- **Goal 4:** An accurate measurement of utility in terms of specific utility needs.

1.6 Thesis Organization

The rest of the manuscript is organized as follows:

- **Part II: State of the Art**

  Chapter 2 - Statistical Disclosure Control: Goal and Mechanisms. This chapter presents the Statistical Disclosure Control (SDC) goals in terms of the trade-off between data utility and privacy and it describes the related mechanisms (generalization and suppression, microaggregation, PRAM, synthetic data generation...) with their capability to achieve these goals.

  Chapter 3 - Statistical Disclosure Control Metrics. This chapter presents anonymity metrics as the main concern in data privacy and we can distinguish two types of metrics: disclosure risk metrics and utility metrics. Disclosure risk metrics (Uniqueness, Record linkage) assess the re-identification capability of anonymized data while the utility metrics (information loss, machine learning metrics) assess the data utility.

- **Part III: Contributions**

  - **Goal 1, Goal 2 and Goal 3**

    Chapter 4 - Discrimination Rate: an attribute-centric metric to measure privacy. This chapter presents the Discrimination Rate (DR), a new metric that provides an attribute-centric approach for privacy measurement and that is practical and flexible enough to fit with various application

domains. The DR computes the attribute identifying capability (scaled between 0 and 1) by measuring how it refines an anonymity set; the more an attribute can refine an anonymity set, the higher its DR. For example, an identifier has a DR equal to 1 as it refines the anonymity set to single users. Thanks to the DR, we provide the first fine grained evaluation and comparison of two of the most used anonymization techniques which are k-anonymity and l-diversity. This work gives a solution for Goal 1, Goal 2 and Goal 3 of Section 1.5. It was published in Journal of Annals of Telecommunications, 2017 (Sondeck, Laurent, and Frey, 2017c).

– **Goal 3**

Chapter 5 - The Semantic Discrimination Rate. This chapter presents the Semantic Discrimination Rate (SeDR), an improvement of the DR which takes into account semantic considerations. The SeDR enables more flexibility for anonymity measurements and is used to compare l-diversity vs. t-closeness which are two of the best k-anonymity-like anonymization techniques. Moreover, as t-closeness is considered better than l-diversity, the SeDR shows that, depending on the semantic considerations, t-closeness can be worse than l-diversity. This work is an approach for Goal 3. It was published in the Security and Cryptography conference (SECRYPT) in 2017 (Sondeck, Laurent, and Frey, 2017b).

– **Goal 4**

Chapter 6 - A posteriori utility assessment of sanitized data with the Discrimination Rate metric. After using our metric (SeDR) for anonymity measurement, we show how it can be used to provide an accurate *a posteriori* utility assessment for any type of sanitized (anonymized) data. A posteriori assessment is the most practical approach as it is performed only on the basis of the sanitized data and a predefined utility need, while the *a priori* assessment aims to assess the extent to which sanitized data reflect the original data; and is therefore based on anonymized data and original data (which is not accessible). This contribution satisfies Goal 4.

- **Part IV: Conclusion**

Chapter 7 - Conclusion and Perspectives. This chapter concludes the dissertation and provides a summary of contributions together with the perspectives for future works.

Part II

State of the Art

Chapter 2

Statistical Disclosure Control: Goal and Mechanisms

2.1 Introduction

This chapter presents the main anonymization mechanisms used for achieving both data utility and data privacy within databases. We first explain the goal of anonymization and the related challenges, then we introduce some background including attacks targeting respondents' privacy and finally we describe the different mechanisms used to counteract those attacks. We focus on Statistical Disclosure Control (SDC) mechanisms which are about protecting respondents' privacy i.e. privacy of persons to whom the data refer.

SDC is a large field including **tabular data** protection, **queryable databases** and **microdata protection**. The related protection mechanisms are diverse with respect to each domain. However, **tabular data** and **queryable databases** formats can be obtained from microdata format (cf. Section 2.4.3). We provide a quick overview of tabular data protection and queryable databases protection; then we focus on microdata which the protection mechanisms can be splitted into: **deterministic mechanisms** which do not consider noise addition, and **non-deterministic mechanisms** that are based on noise addition and synthetic data generation. No approach is strictly better than the other one as the goal is to reach the good trade-off between privacy and utility and depending on the needs, either deterministic or non-deterministic mechanism can be used.

This chapter is organized as follows. Section 2.2 presents the goals of SDC mechanisms in terms of trade-off between privacy and utility. Section 2.3 gives some background on SDC including related definitions and description of the main SDC file formats. Sections 2.5 and 2.6 focus on the microdata file format and present the related protection mechanisms that are either deterministic or non deterministic. Section 2.7 gives our conclusion.

2.2 SDC Objectives and Assessment Considerations

The goal of data anonymization (SDC) is twofold (Li and Li, 2009) (Makhdoumi and Fawaz, 2013):

- **Privacy**: it should protect against respondents identification
- **Utility**: it should guarantee the usefulness of data for different processes.

anonymization goal is to guarantee the good trade-off between privacy and utility; data are used to carry out various applications among which: development of

new services by companies, scientific research, government studies... and these applications should also guarantee privacy. These specificities have led to another expression which is *sanitization* and which better underlines both of the goals.

However, SDC mechanisms are usually built to ensure privacy as they aim to counteract re-identification attacks (Samarati and Sweeney, 1998), (Machanavajjhala et al., 2007), (Li, Li, and Venkatasubramanian, 2007), (Dwork, 2011) and utility is measured thereafter depending on the case. This approach has led to evaluate SDC mechanisms in terms of their capability to resist to attacks (Ganta, Kasiviswanathan, and Smith, 2008) instead of their capability to respond to the fundamental need which is: the trade-off between privacy and utility. Indeed, privacy and utility are usually considered as inversely proportional (Li and Li, 2009) (Xu et al., 2015) (Brickell and Shmatikov, 2008) as such, comparing SDC mechanisms considering only the privacy dimension could lead to a biased assessment.

2.3 SDC Terminology and Formal Description

Statistical Disclosure Control techniques aim to protect data within statistical databases such that they can be published without harming the privacy of individuals to whom the data correspond and, at the same time they can ensure data usability. The earliest works on SDC date back to 1970s with the contribution of Dalenius (Dalenius, 1974) and the works by Schlörer and Denning (Denning, Denning, and Schwartz, 1979) (Schlörer, 1975). A good survey of more recent SDC technologies is given by (Domingo-Ferrer, Sánchez, and Hajian, 2015a).

This section introduces the terminology of SDC that will be used throughout this chapter.

2.3.1 SDC Terminology and Formal Description

The statistical databases can take one of the following formats: *microdata* (cf. Section 2.4.3), *tabular data* (cf. Section 2.4.1) and *queryable databases* (cf. Section 2.4.2). The two latter can be seen as macro data (Ciriani et al., 2007) as they only refer to aggregated data about respondents, while microdata contain data about single respondents. As macro data formats can be obtained from a micro data format by aggregating values (cf. Section 2.4.3), the terminology presented here is only about the micro data format.

A *microdata* file is generally depicted by a table where each row (record) contains individual data and each column is an attribute shared by every respondents within the table. For example Table 2.1 is a microdata with 4 attributes (*ZIP Code*, *Age*, *Salary* and *Disease*) and 9 records.

Formally, a microdata file M referring to r respondents with s attributes (variables) is a $r \times s$ matrix where M_{ij} is the value of attribute j for respondent i. Attributes in a microdata can be of three categories which are not necessarily disjoint:

- **Identifiers**: attributes that can be used on their own to characterize a single respondent among others. Examples of such attributes are: social security numbers, names, fingerprints.

TABLE 2.1: Original Data Table (Salary/Disease).

	ZIP Code	Age	Salary	Disease
1	35567	22	4K	colon cancer
2	35502	22	5K	stomach cancer
3	35560	22	6K	lung cancer
4	35817	45	7K	diabetes
5	35810	63	12K	diabetes
6	35812	40	9K	aids
7	35502	35	8K	aids
8	35568	35	10K	flu
9	35505	32	11K	lung cancer

- Quasi-Identifiers/Key Attributes: attributes that do not completely characterize a respondent but can be combined with others for complete characterization. Examples of such attributes are: zip code, age, gender...
- Confidential/Sensitive attributes: attributes which contain sensitive information on the respondent. Examples are: salary, religion, health.

Attributes within a microdata file format can be of different types (Domingo-Ferrer, Sánchez, and Hajian, 2015a):

- **Continuous**: attributes on which numerical and arithmetical operations can be performed. Examples: Age, Salary... The main disadvantage of continuous attributes is that their values are usually completely different from each other and so, can be used to characterize a respondent within a dataset.
- **Categorical**: attributes that take values over a finite set and over which arithmetical operations can not be performed. We can distinguish two types of such attributes:
  - **Ordinal**: when the values are ordered. We can therefore apply operators like $\leq$, max and min.
  - **Nominal**: when the values are not ordered. The only possible operator is equality (=). Examples are the colors.

Moreover, depending on the context, any attribute can be used for re-identification. This last observation is underlined by (Schwartz and Solove, 2011) which presents identifiers as one of the most important concerns for privacy regulation as personal data are inherently linked to identifiers (Personally Identifiable Information). In Chapter 4 we propose a more formal and quantifiable definition of the identifiers' terminology which takes into account the context.

2.4 SDC Application Domains

This section presents a description of the three SDC domains which are: *tabular data protection*, *queryable databases protection* and *micro data protection*. While the first two protection mechanisms have been studied from a while (since 1970s (Denning, Denning, and Schwartz, 1979) (Dalenius, 1974)), *microdata protection* is a relatively new field.

2.4.1 Tabular Data

Tabular data is a specific SDC format which the goal is to publish static aggregate information over data (e.g., sums, averages...) rather than original data in order to limit information leakage. Tabular data can be described as follows:

$$T : D(M_{i1}) \times D(M_{i2}) \times ... \times D(M_{ik}) \rightarrow \mathbb{R} \; or \; \mathbb{N} \tag{2.1}$$

Where $k \leq s$ and D refers to the domain where attributes M_{ij} takes its values. An interesting survey on tabular data is provided by (Willenborg and De Waal, 2012a). Tabular data tables can be of two types (Domingo-Ferrer, Sánchez, and Hajian, 2015a):

- **Frequency tables**: that display the count of respondents at the crossing of categorical attributes (in $\mathbb{N}$). For example given a census microdata containing attributes "Marital Status" and "Zip Code", a frequency table can display the count of respondents for each marital status in each Zip Code region.
- **Magnitude tables**: that display a numerical value at the crossing of categorical attributes (in $\mathbb{R}$). For example if the census data also contain the "Salary", a magnitude table could display the average salary for each marital status in each zip code region.

Tables are called *linked* if they share some of the crossed categorical attributes. For example "Marital Status" × "Zip Code" is linked to "Marital Status" × "Age".

While the data displayed by tabular data seem to be constrained they can be subject to attacks among which:

- **External attack**: refers to attacks made by an attacker who is not a respondent. For example suppose an attack targeting a magnitude table displaying average salary and that, there is a single respondent for a given marital status MS_i and living in a given zip code region Z_j; the average salary of respondents would then disclose the actual salary of the only respondent.
- **Internal attack**: refers to attacks made by an attacker who is a respondent. Suppose now that there are only 2 respondents with marital status MS_i and living in Z_j their salary would be displayed to each other.
- **Dominance attack**: is a specific case of the internal attack where the attacker is a respondent who dominates in the contribution to a cell of magnitude table and can therefore upper-bound the contributions of the other respondents. For example if the magnitude table, instead of displaying an average salary, displays the total salary of respondent according to their marital status and Zip Code regions and if a respondent contributes to 90% he can then infer that the other respondents have low incomes.

The methods to counteract these attacks are either perturbative which modify some values of the table (e.g., Controlled Tabular Adjustment (Dandekar and Cox, 2002)) or non-perturbative which do not modify the table values (e.g., Cell Suppression (Fischetti and Salazar, 2000)).

2.4.2 Queryable Databases

Queryable databases are a SDC format depicted by on-line databases to which a user can submit statistical queries (e.g., sums, averages...). These restrictions are made to prevent a user to infer information on a specific respondent. The main approaches to implement queryable data protection can be of three types (Domingo-Ferrer, Sánchez, and Hajian, 2015a):

- **Data perturbation**. This is only possible when randomized answers can be enough for the user's need. Perturbation can either apply on the records to which the queries refer or to the query result after computation over original data.
- **Query restriction**. This is the mechanism used when the user does not want randomized answers (e.g., a number). The data are simply restricted i.e. no answer is provided since they may carry enough information for characterization and therefore for re-identification. Many criteria can be used to restrict the access to a given request, one of them is the set size control i.e. refusing the access to request referring to a set of records which is too small. (Chin and Ozsoyoglu, 1982) propose some examples of query restriction.
- **Camouflage**. This mechanism is used when the user needs approximate but not randomized answers, for example, when small interval answers can be enough. The idea here is to provide an interval answer which includes the exact answer.

2.4.3 Microdata

There are different approaches for describing Microdata: *masking methods* and *synthetic data generation*. *Masking methods* consist in modifying the original data by reducing their inner amount of identification information and can be either deterministic or non-deterministic (noise addition). *Synthetic data generation* refers to methods that generate synthetic data that reflect original data to ensure the respondents confidentiality and is a non-deterministic approach. Note that the goal of both approaches is to preserve some statistical properties of the original data within the sanitized data while preventing re-identification of the respondents.

Another way to characterize masking methods is by considering whether they alter original data or not, they can be either *perturbative* or *non-perturbative* (Willenborg and De Waal, 2012b). Perturbative masking methods transform the original data which may disappear in the anonymized version and can use one of the following mechanisms: *noise addition*, *microaggregation*, *data/rank swapping*, *microdata rounding*, *resampling* and the *Post-RAndomization Method (PRAM)*. Examples of perturbative masking methods are described in (Hundepool et al., 2012). Non-perturbative masking methods do not alter original data but produce partial suppression or reduction. Some examples are: *sampling*, *generalization and suppression*, *top and bottom coding* and *local suppression*.

In this chapter we adopt another classification which is about whether microdata protection mechanisms are *deterministic* or *non-deterministic* (Sections 2.5 and 2.6) as it better underlines their capacity to provide privacy vs. utility.

2.4.4 Conclusion

Queryable database protection and the *tabular data protection* can be obtained from a *microdata* file by first performing a microdata protection mechanism over the data before performing aggregate information in case of *tabular data* (cf. Section 2.4.1) or a query in case of *queryable databases* (cf. Section 2.4.2).

2.5 Deterministic mechanisms

Deterministic mechanisms do not consider noise addition for protection. Different mechanisms have been proposed in the literature including Generalization and Suppression, Anatomy and Microaggregation.

2.5.1 Generalization and Suppression

Generalization and Suppression aim to protect respondents' privacy by replacing quasi-identifier values by more general values while suppressing identifying attributes. Generalization and Suppression aims to counteract *identity disclosure* (cf. Section 2.4.3). The idea is to reduce the identifying capability of quasi-identifiers by preventing the uniqueness of some specific value or combination of values. The quasi-identifier transformation can apply in different ways depending on the quasi-identifier type. For a categorical attribute a specific value can be replaced by a more general value according to a given hierarchy. For a continuous attribute, intervals containing the exact values can be used instead; discretization can also apply (Hundepool et al., 2005). However, applying generalization on continuous attributes is more tricky as the arithmetic operations that were simple to apply on exact data could become less intuitive.

For example, let consider Table 2.2 which is a generalized instantiation of Table 2.1 where the categorical attributes Age and ZIP Code are transformed into Age* and ZIP Code*. As we can observe, unlike in the original data (Table 2.1), where the Age value 45 can be used to characterize a respondent, no value or combination of values of the transformed attributes can be used to characterize a single respondent.

Various schemes of Generalization and Suppression have been proposed in the literature including: *full-domain generalization* (LeFevre, DeWitt, and Ramakrishnan, 2005) (Sweeney, 2002); *Subtree Generalization* (Bayardo and Agrawal, 2005), (LeFevre, DeWitt, and Ramakrishnan, 2005); *Cell Generalization* (Wong et al., 2006). They propose different generalization approaches considering different levels of a given taxonomy tree.

2.5.2 Local Suppression

Local suppression aims to suppress some of the identifying attributes values instead of suppressing the entire attributes in order to increase the usable set of records by building a cluster of values (referring to the Generalization and Suppression mechanism). For example in Table 2.1 we can suppress the subsets of Age values {63, 45, 40}, thus, we would be able to build a new cluster indexed by "≥ 40". As continuous values are all different from each other, this mechanism is only suitable for categorical values. (Hundepool et al., 2008) proposes an implementation of local suppression in combination with the Generalization mechanism.

TABLE 2.2: Generalized Table.

	ZIP Code*	Age*	Disease
1	35***	≤ 32	colon cancer
2	35***	≤ 32	stomach cancer
3	35***	≤ 32	lung cancer
9	35***	≤ 32	lung cancer
4	35***	> 32	diabetes
5	35***	> 32	diabetes
6	35***	> 32	aids
7	35***	> 32	aids
8	35***	> 32	flu

2.5.3 Top and Bottom Coding

Top and Bottom Coding (Domingo-Ferrer and Torra, 2001a) is a specific application of the Generalization and Suppression mechanism. The idea is to rank the data values to form two sets, the top set and the bottom set, and then apply the same generalization for all the values in each set. For this mechanism the attribute should be either continuous or categorical ordinal due to the need of an order relation. Other variants of this mechanism (Domingo-Ferrer and Torra, 2001a) (Domingo-Ferrer and Torra, 2002) use either top coding (only values greater than a given value are replaced), bottom coding (only values lower than a given value are replaced). For example, in Table 2.2, two subsets of values are built for attribute Age, the top values referred by "≤ 32" and the bottom values referred by "> 32".

2.5.4 Anatomy

Anatomy (Xiao and Tao, 2006) releases all the quasi-identifiers and confidential attributes in two different tables and targets *attribute disclosure* (cf. Section 2.4.3). Indeed, splitting confidential attributes and quasi-identifiers in different tables enable breaking their correspondences with the benefit of no transformation on their values, which provides therefore more granularity for accurate analysis. The intuition is that with the generalized Table, the domain values are lost and the uniform distribution assumption is the best to consider since there is no additional knowledge, whereas with the anatomyzed tables, domain values are kept and more accurate analysis can apply. Anatomy increases therefore data utility with respect to the Generalization and Suppression mechanism.

As an example, let us consider Table 2.3 which is the Original Data Table (Table 2.1) where we added attribute ID for specifying 2 groups of records (1 and 2). We consider 3 quasi-identifiers (ZIP Code, Age and Salary) and 1 confidential attribute (Disease). When we compute anatomy on this table, we obtain Tables 2.4 and 2.5. As we can observe, there is no more direct links between quasi-identifiers and confidential attributes in the anatomy Tables, the correspondences are rather made through attribute Group-ID.

Let us now compare this instantiation to the Generalization and Suppression instantiation according to the following request: "*count the number of respondents of age "*≥ 35*" having aids*". The correct number of respondents is 2; if we compute this number using the Generalized Table (Table 2.2) we obtain: 2 (the number of aids values

TABLE 2.3: Original Data Table (Anatomy).

ID	ZIP Code	Age	Salary	Disease
1	35567	22	4K	colon cancer
1	35502	22	5K	stomach cancer
1	35560	22	6K	lung cancer
2	35817	45	7K	diabetes
2	35810	63	12K	diabetes
2	35812	40	9K	aids
2	35502	35	8K	aids
2	35568	35	10K	flu
1	35505	32	11K	lung cancer

TABLE 2.4: The Quasi-Identifier Table Obtained with Anatomy (QIT).

ZIP Code	Age	Salary	Group-ID
35567	22	4K	1
35502	22	5K	1
35560	22	6K	1
35817	45	7K	2
35810	63	12K	2
35812	40	9K	2
35502	35	8K	2
35568	35	10K	2
35505	32	11K	1

in the table) $\times \frac{1}{5}$ (the probability of having a subject with age ≥ 35 and who belongs to the subset containing the aids values within Table 2.2) = $\frac{2}{5}$. When computed on the anatomyzed Tables (Tables 2.4 and 2.5) we obtain: 2 (the number of aids values in the table) $\times \frac{2}{5}$ (the probability of having a subject with age ≥ 35 and who belongs to the subset containing the aids values within Tables 2.4 and 2.5) = $\frac{4}{5}$ which is closer to the real value.

However, anatomy is not suitable for continuous attributes as it does not add any benefit for such attributes. Indeed, continuous values are usually completely different from each other and this would provide a uniform distribution assumption as well for the anatomy mechanism. Moreover, as for anatomy, data are published in different tables, it is unclear how standard data mining tools (classification, clustering, association...) would apply; then new tools and algorithms need to be designed.

2.5.5 Microaggregation

Microaggregation is a SDC mechanism usually performed on *continuous attributes* (cf. Section 2.3.1) as the common application requires average computation. However some applications exist for *categorical attributes* (Domingo-Ferrer and Torra, 2005), (Torra, 2004). The idea of microaggregation is to prevent privacy violation by

TABLE 2.5: The Sensitive Table Obtained with Anatomy (ST).

Group-ID	Disease	Count
1	colon cancer	1
1	stomach cancer	1
1	lung cancer	2
2	diabetes	2
2	aids	2
2	flu	1

building groups of at least k subjects where the individual confidential values are replaced by a common value (usually an average over the k subjects, in case of *continuous attributes*). When the original data includes several attributes (variables), there are different approaches for microaggregation: *univariate microaggregation* (Hansen and Mukherjee, 2003), (Nin and Torra, 2009) which is applied to each variable independently, *Multivariate microaggregation* (Domingo-Ferrer, Sebé, and Solanas, 2008) which constructs clusters taking into account all or subsets of variables at a time.

Microaggregation is performed using the two following operations:

- **Partition**: original records are partitioned into several groups where values correspond to at least k subjects and where no individual subject dominates too much (i.e. where one of the value is not too high or too low with respect to the other values)
- **Aggregation**: which is applied on attribute values using a specific computation (e.g., the mean for *continuous attributes*, the median for *categorical attributes*)

To illustrate the microaggregation process, we propose an example of a specific type of microaggregation which is k-anonymity microaggregation (Domingo-Ferrer, 2006). Let us consider Table 2.1 as our original table, Age and Salary as our quasi-identifiers and Disease as our confidential attribute. Table 2.6 is a possible microaggregation instantiation with a partition of 3 groups (1-3, 4-6, 7-9); aggregation consists in average computation for attribute Age and median computation for attribute Salary. Table 2.6 provides therefore protection against the *identity disclosure* as an observer can no longer infer the disease of a specific individual within the table.

However, partitioning should ensure the minimal information loss obtained when we reach the optimal k-partition. The optimal k-partition maximizes the within-group homogeneity and the higher the within-group homogeneity, the lower the information loss. To measure the within-group homogeneity the sum of squares criterion is commonly used (Hansen, Jaumard, and Mladenovic, 1998), (Gordon and Henderson, 1977), (Edwards and Cavalli-Sforza, 1965).

The sum of squares criterion can be described as follows (Domingo-Ferrer and Mateo-Sanz, 2002): let n be the number of records, g the number of groups of size at least k, n_i be the number of records in the i-th group ($n_i \geq k$ and $n = \sum_{i=1}^{g} n_i$). Let x_{ij} be the j-th record in the i-th group and $\bar{x}_i$ the average data vector over the i-th group, let $\bar{x}$ be the average data vector over the whole set of n individuals. The sum of squares criterion is computed as:

TABLE 2.6: Microaggregation Table.

	Age	Salary	Disease
1	22	5K	colon cancer
2	22	5K	stomach cancer
3	22	5K	lung cancer
4	49.33	9K	diabetes
5	49.33	9K	diabetes
6	49.33	9K	aids
7	34	10K	aids
8	34	10K	flu
9	34	10K	lung cancer

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i) \tag{2.2}$$

Where "$(x_{ij} - \bar{x}_i)'$" refers to the micro aggregated instantiation and "$(x_{ij} - \bar{x}_i)$" to the original microdata. Then the optimal k-partition is the one that minimizes the SSE. However there exist different ways to form groups. Several taxonomies are possible including: (1) fixed group size (Defays and Nanopoulos, 1993), (Domingo-Ferrer and Torra, 2005) vs. variable group size (Laszlo and Mukherjee, 2005), (Domingo-Ferrer and Mateo-Sanz, 2002), (2) exact optimal (Hansen and Mukherjee, 2003) vs. heuristic microaggregation, (3) categorical (Domingo-Ferrer and Torra, 2005) vs. continuous.

2.6 Non-Deterministic Mechanisms

Non-deterministic protection mechanisms are based on noise addition and randomization to protect respondents' data. Different methods exist for noise addition. We first describe the noise addition principle and its related algorithms and then we present the existing protection mechanisms.

2.6.1 Noise Addition

Noise addition refer to a set of perturbative masking methods that aim to protect confidential data by mixing them with noise/fake data. This way, an attacker can no longer distinguish which data actually belong to real respondents. The mix is usually performed by adding or multiplying a sensitive attribute with a random variable with a given distribution.

Let X_j be the j_{th} column (sensitive attribute) of the original microdata table with N records. There exist many algorithms for noise addition (Brand, 2002) among which:

- **Masking by uncorrelated noise addition**: The sensitive attribute X_j of the original data set is replaced by a vector:

  $Y_j = X_j + \epsilon_j$

where ϵ_j is a vector of normally distributed noise referring to a random variable $\epsilon_j \sim N(0, \sigma_{\epsilon_j})^2$, such that $Cov(\epsilon_j, \epsilon_i) = 0, \forall i \neq j$. Uncorrelated noise addition preserves the mean and the covariance and does not preserve neither variance nor correlations.

- **Masking by correlated noise addition**: this method unlike the previous does preserve means and correlation coefficients; it ensures that the covariance matrix of noise is proportional to the covariance matrix of the original data set i.e. $\epsilon \sim N(0, \Sigma_\epsilon)$, where $\Sigma_\epsilon = \alpha\Sigma$ where Σ is the covariance matrix of the original data. However, this method provides a weak level of protection Tendick, 1991 Tendick and Matloff, 1994.

- **Masking by noise addition and linear/non linear transformation**: These methods (Hundepool et al., 2012) (Kim, 1986) aims to improve the protection while still providing a good level of correlations; therefore, the microdata obtained after the additive noise is linearly/non linearly transformed before release. However, such transformation relies on a parameter p that should not be revealed, otherwise the released data would have the same protection level as for data with additive noise.

Parameter p can be used for specific adjustments in case of sub populations but with the drawback to reduce the protection level.

Noise addition is well suited for continuous data for many reasons: (i) there is no assumption on the range of possible values (which may be infinite); (ii) the added noise is usually continuous which matches with the data type; (iii) no exact matching can be made with external data, however approximate matching can still apply.

These last properties guarantee resistance to the *identity disclosure* (cf. Section 2.4.3).

2.6.2 Data Swapping

Data swapping exchanges specific confidential values between respondents in order to break the correspondence between key and confidential attributes. Data swapping targets therefore *attribute disclosure* (cf. Section 2.4.3). The first propositions date back to 80's by Dalenius (Dalenius and Reiss, 1982) and Reiss (Reiss, 1984), for continuous and categorical attributes respectively. A variant of data swapping is rank swapping (Carlson and Salabasis, 2000) which first performs a ranking of confidential values and then, swaps them according to their rank instead. The ranked value is swapped with another within a restricted range according to an input parameter p which refers to the number of records. The goal of rank swapping is to enable multivariate analysis which is less practical on data swapping without ranking and therefore to provide more utility for data. Indeed an empirical work on data swapping for continuous attributes (Domingo-Ferrer, Mateo-Sanz, and Torra, 2001) have shown its effectiveness in terms of the trade-off between privacy and utility.

2.6.3 Sampling

Sampling is about publishing a sample of the original set of records (Willenborg and De Waal, 2012a), instead of the whole original records. The idea is to reduce accuracy by the uncertainty of presence or absence of a given respondent within the sample. For example we can decide to publish only the even records of the original microdata set. As such, this mechanism is more suited for categorical attributes as

continuous attributes would highly enhance the re-identification capability. Indeed, as the continuous values are most of the time completely different from each other, they could easily serve for characterizing a given respondent within the sample, and used for linkage with external data.

2.6.4 Rounding

Rounding (Denning, 1982) aims to replace original values with rounded values and is applicable only on continuous values. For this purpose, intervals of sensitive values are built and for each interval i, a corresponding value v_i is chosen within i; therefore, each original value falling in i is replaced by v_i.

2.6.5 Post-RAndomization Method (PRAM)

The Post-RAdomization Method (PRAM) (Gouweleeuw, Kooiman, and De Wolf, 1998), (Kooiman, Willenborg, and Gouweleeuw, 1997), (Wolf, 2006) is a perturbative masking method that applies on categorical attributes and uses data randomization. PRAM relies on a Markov matrix for providing probability for replacing a category with another. Let $P = [p_{ij}]$ be a Markov matrix (i.e. a real $n \times n$ matrix, where each element $p_{ij} \geq 0$ and $\sum_{j=1}^{n} p_{ij} = 1$, $i = 1, ..., n$), which contains the probability to replace a category within the original data with another category. Therefore, p_{ij} is the probability that category c_i in the original data is replaced by category c_j for producing the anonymized data. It is therefore difficult for an attacker to identify the real records. As modifications can apply indistinguishably on key attributes and on confidential attributes, this mechanism addresses both *identity disclosure* and *attribute disclosure* (cf. Section 2.4.3). The Markov matrix, provides both suppression and generalization which make the PRAM method more generic. However PRAM can not apply on *continuous attributes* as the PRAM matrix must contain a row for each possible value of each attribute, and continuous attributes may have an infinite number of values.

2.6.6 MASSC

MASSC (Singh, Yu, and Dunteman, 2003) is a masking method which acronym refers to: Micro Agglomeration, Substitution, Subsampling and Calibration. The corresponding operations are the following:

1. Micro agglomeration refers to the generalization principles (cf. Section 2.5.1) and is applied on key attributes to partition them into groups that have the same disclosure risk.

2. Optimal probabilistic substitution is then used to modify the original data. This substitution uses a markov matrix like the PRAM mechanism (Section 2.6.5)

3. Optimal probabilistic subsampling is used to suppress some values according to predefined probabilities.

4. Optimal sampling weight calibration is used to increase data utility and preserves estimates for the concerned variables.

The main advantage of MASSC is that, it is designed in such a way that disclosure risk can be analytically quantified. Quantification is based on the uniqueness

principle (cf. Section 2.4.3) which applies mainly on categorical attributes. Indeed, as continuous values are usually completely different from each others, they can be used to characterize each of the respondents in the microdata and would always provide the worst security. That is why this method is not well suited for continuous attributes.

In Chapter 3 we provide a quantitative definition of such attributes with different values and we refer to them as identifiers.

2.6.7 Synthetic Data Generation

The difference between the **Synthetic Data Generation** mechanism and the **Noise Addition** mechanisms is that, unlike the latter which mixes fake data with original data, **Synthetic Data Generation** generates completely new data that do not contain original data but some of its characteristics. The idea is to generate random data that contain some statistics about the original data and guarantee utility for specific purposes while protecting respondents. The earliest works on Synthetic Data Generation dates back to 90's with the work of Rubin (Rubin, 1993) and aimed to create an entire synthetic data based on original data. (Dwork, 2008) proposes a survey of a more recent Synthetic Data Generation mechanism called Differential Privacy.

Synthetic Data Generation counteracts by definition all the existing attacks on microdata (cf. Section 2.4.3) as the synthetic data are assumed not belonging to any respondent and are generated randomly. However, the randomness of the process can bring in some issues (Reiter, 2005), (Winkler, 2004). Suppose, by chance, some records match the original records within the sanitized data, confidential data of the corresponding respondent would therefore be revealed even if the data are assumed to be randomly generated. However, this is not the main issue concerning data generation, the main issue is about data utility. Unlike the previous mechanisms, specific use cases should be defined prior to data sanitization using Synthetic Data Generation, and this makes the Synthetic Data Generation mechanism very limited in terms of utility; the sanitized data are only useful for the predefined use cases. Moreover, the statistics that are defined using synthetic generation do not apply to sub domains i.e. to subsets of values; statistics are rather extracted considering the overall data. To counteract this latter issue, some hybrid mechanisms have been proposed in the literature that combine Synthetic Data Generation and deterministic approaches like k-anonymity (Li, Qardaji, and Su, 2011), (Soria-Comas et al., 2014).

2.7 Conclusion

This chapter presents an overview of the main Statistical Disclosure Control (SDC) mechanisms and the SDC goal which is twofold: protecting respondent's privacy within statistical databases while ensuring data utility. We focus on microdata, a specific SDC file format from which the other file formats can be derived. Microdata are subject to various attacks and many protection mechanisms are proposed to counteract these attacks. We propose a classification of microdata protection mechanisms according to whether they are deterministic or non-deterministic, and specify for each mechanism which attack it aims to counteract. This classification underlines the capacity of those mechanisms to provide the good trade-off between utility and privacy. Indeed, while deterministic mechanisms are more subject to flaws, they enable a better control over data which supports accurate analysis and therefore

improves data utility. On the other hand, while non-deterministic mechanisms provide a stronger protection than deterministic mechanisms with respect to attacks, the synthetic data addition prevents accurate analysis and therefore reduces data utility. Finally, no category is strictly better than another as the goal is to reach the good trade-off between privacy and utility. However, we can not achieve this goal without a proper mean to assess the privacy/utility level of data. In the next chapter, we present the main metrics used for evaluating SDC and to measure privacy vs. utility.

Chapter 3

Statistical Disclosure Control Metrics

3.1 Introduction

While the SDC protection mechanisms (Chapter 2) are useful for providing the trade-off between privacy and utility, the data processor still does not know which would be the best mechanism for fulfilling his goals and therefore, needs metrics for choosing the good one. This validation process is even more important as it is needed both by the data processor and the regulator. Indeed, the trade-off can be translated from the data processor's point of view into: how to extract the maximum information from data while ensuring compliance with the regulation ? While the regulator is concerned by: how to define a threshold for validating or not the data processors' implementations ?

In fact, companies and other processors (researchers, governments...) are interested in extracting the maximum information from the data while protecting respondent's privacy; yet, privacy is defined according to the regulation. The regulator on its side needs to define a protection threshold that would serve for validating the compliance of processors' data with regulation. However, data utility and privacy are usually inversely proportional, meaning that the greater the utility, the smaller the privacy and vice versa (Karr et al., 2006)(Li and Li, 2009) (Xu et al., 2015) (Brickell and Shmatikov, 2008). This comes from the fact that to protect against re-identification, data should be transformed in order to reduce their identification capabilities, which implies reducing information, and leads to a weaker data accuracy. Therefore, the data processors' challenge would be to maximize the data utility (corresponding to minimize the privacy) and the regulator's challenge, to define the threshold that would be enough for protecting respondents' privacy. While the needs of each party (processors/regulator) seem to be contradictory, they refer to the same issue which is privacy assessment and hence, metrics.

Metrics are at the centre of data sanitization as they are intended to guide the answers to the previous challenges. Indeed, while the SDC mechanisms can be useful for data sanitization they should be guided to address the previous challenges. In fact, any mechanism is not suitable for every use case and for the same mechanism different instantiations can provide very different results (Lee and Clifton, 2011) (Li and Li, 2009). Metrics are therefore intended to answer the following issues:

1. **Assessing the Data Disclosure risk** level (both for the processors and the regulator).

2. **Assessing the data utility level** (by processors).

However, ensuring the trade-off between privacy and utility is a very challenging concern for 2 main reasons:

- **The diversity of privacy models**: there is no consensus about the privacy model to be considered; various approaches are proposed in the literature which are most of the time difficult to compare and this does not enable a uniform assessment of the effective privacy level.
- **The subjectivity of utility**: data utility refers to the capacity of data to respond to a given need or problem and this is subjective (Domingo-Ferrer and Torra, 2001a)(Karr et al., 2006).

There are two main models in SDC which are: k-anonymity and ϵ-differential privacy.

These models provide both protection mechanisms (cf., Sections 2.5.5 and 2.6.7 respectively) and metrics (cf., Sections 3.3.1 and 3.3.2 respectively). However, while the related metrics can at some extent quantify privacy/disclosure risk, they fail reflecting the data utility, especially for the ϵ-differential privacy metric (Lee and Clifton, 2011). Moreover, the related metrics are not comparable in terms of privacy degree.

These issues have led to various propositions intended to measure on the one hand data utility (Sankar, Rajagopalan, and Poor, 2013) (Xu et al., 2006)(Bindschaedler, Shokri, and Gunter, 2017) and on the other hand the disclosure risk (Domingo-Ferrer and Torra, 2001a). The disclosure risk assessment is computed according to the existing SDC attacks (identity disclosure and attribute disclosure), and the related approaches for measurements can be splitted into: *uniqueness* or Record Linkage (cf., Sections 3.4). Concerning utility, we should first consider if we plan to perform Privacy Preserving Data Mining (PPDM) or Privacy Preserving Data Publishing (PPDP). PPDM is about publishing data for a predefined use (cf., Section 3.5.2), while PPDP is about publishing data that could serve for various uses (cf., Section 3.5.1)(Clifton and Tassa, 2013). The utility assessment for the k-anonymity model refers to information loss i.e. the extent to which the sanitized data are degraded with respect to the original data (Domingo-Ferrer, Sánchez, and Hajian, 2015a) and, for differential privacy, utility assessment refers to how much data mining algorithms (clustering, classification...) are able to accurately process the sanitized data (Bindschaedler, Shokri, and Gunter, 2017) (Abadi et al., 2016).

This chapter is organized as follows. Section 3.2 provides some background on utility and disclosure risk measurements including the definition of disclosure risk and data utility. Section 3.3 presents the two main privacy models in SDC which are k-anonymity and ϵ-differential privacy and their limitations for measuring disclosure risk and utility. Section 3.4 presents the main disclosure risk metrics in terms of uniqueness and record linkage. Section 3.5 presents the common utility metrics for measuring data utility both for PPDM and PPDP, Section 3.6 presents our comparative analysis of disclosure risk metrics, and Section 3.7 gives our conclusion.

3.2 Background

This section gives some background on disclosure risk and utility assessment.

3.2.1 What is Disclosure Risk ?

Disclosure risk can be defined as the capacity of an intruder to use a sanitized microdata to infer confidential information on a respondent among others within the original microdata (Bernardo et al., 2003). In practice, the disclosure risk is performed by matching different micro data files shared by the same respondent or by assessing the correlation between a respondent data within the same microdata file. Disclosure risk refers to disclosure scenarios and is evaluated accordingly. There are two different disclosure scenarios (Domingo-Ferrer, Sánchez, and Hajian, 2015a) which are:

- **Identity disclosure**: which aims to link a respondent to a specific record within the microdata and is usually performed by composing different microdata files. Indeed by using only the sanitized data, an intruder can not identify a respondent due to the transformations performed by the protection mechanisms and will therefore need external data for identification.
- **Attribute disclosure**: which is about the knowledge an attacker could gain on attributes themselves without necessarily relate them to a given respondent. This attack is based on the correspondences between attributes within microdata sets (usually between key attributes and confidential attributes). If an attacker is able to identify such correspondences, thereafter, he will only need to link a respondent to an involved quasi-identifier to infer his confidential attribute. There exist various types of attribute disclosure including: the *homogeneity attack*, the *background knowledge attack*, the *semantic attack* and the *skewness attack* (cf., Section 3.3.1 for details).

Hence, *identity disclosure* ensures complete re-identification of respondents while *attribute disclosure* is about information gain. Nevertheless, it is still useful to protect against the latter attack as it may provide valuable information to an attacker that could be used later in case of correspondence with data containing real identities. We provide examples of *identity disclosure* and *attribute disclosure* in Section 3.3.1.

Another point is that, *attribute disclosure* can either be considered as an attack or as a utility indicator, as discussed in Section 3.3.1

Two approaches can be considered for assessing disclosure risk (Domingo-Ferrer, Sánchez, and Hajian, 2015b):

- **Uniqueness**: the idea of uniqueness is to identify specific combinations of attributes that are unique or rare and can be used to link a record in the sanitized microdata to a record within the original data. This attack is based on key attributes as by definition some of their combinations can lead to unique characterization of respondent (cf., Section 2.3). For identity disclosure, this approach does not apply on *perturbative masking methods* (cf. Section 2.4.3) as they usually transform the original data and there are no longer correspondences with original data; but can still be performed for attribute disclosure as shown in Section 3.3.1.
- **Record linkage**: it provides a more general definition that is about the knowledge a specialized intruder can use to link a record to a respondent. This approach requires a specific attack model for application, and can target both *perturbative masking mechanisms* and *non-perturbative masking mechanisms*.

While in (Domingo-Ferrer, Sánchez, and Hajian, 2015b) *uniqueness* and *record linkage* are presented as targeting *identity disclosure*, they can also be used for assessing *attribute disclosure* as described in Section 3.3.1.

The assessment mechanisms related to *uniqueness* an *record linkage* are presented in Section 3.4.

3.2.2 What is Data Utility ?

Data utility refers to the capacity of data to respond to a given need. Indeed, it is difficult to assess utility without specifying the need as data can be useful for some uses but not for others. As such, defining a general data utility can be a tricky issue (Karr et al., 2006). However, in practice many applications require data publication without specifying a given need (e.g., health data, census data, educational data...). These issues have led to two different approaches for defining utility when publishing data: utility according to a specific use which is about Privacy Preserving Data Mining (PPDM) or utility that can serve for various uses which is about Privacy Preserving Data Publishing (PPDP).

We provide a detailed description of these approaches in Section 3.5.

3.3 Privacy Models (k-anonymity and Differential Privacy)

There are two main privacy models for SDC: *k-anonymity* (Samarati and Sweeney, 1998) (Samarati, 2001) (Sweeney, 2002) and *ϵ-differential privacy* (Dwork et al., 2006) (Dwork, 2008) (Dwork, 2011). *k-anonymity* is a deterministic approach that can be implemented using different protection mechanisms (*Generalization and Suppression*, *Aggregation*...; cf., Sections 2.5.1 and 2.5.5) and various improvements have been proposed to counteract different types of attacks. *Differential privacy* (DP) is a non-deterministic approach and refers to the *Synthetic Data Generation* mechanism (cf., Section 2.6.7).

These models depict both protection mechanisms and metrics. However, while the related metrics can at some extent reflect the disclosure risk, they do not reflect utility of data. Moreover, the related privacy metrics are not comparable in terms of privacy degree which do not allow a uniform assessment of approaches.

3.3.1 k-anonymity Based Models

k-anonymity is a model, and there exist many mechanisms/metrics based on this model including: k-anonymity (itself) (Samarati and Sweeney, 1998) (Samarati, 2001) (Sweeney, 2002) , l-diversity (Machanavajjhala et al., 2007) and t-closeness (Li, Li, and Venkatasubramanian, 2007). In this section, we present this model and related mechanisms/metrics and show how they are used to assess the *disclosure attacks* (*identity disclosure* and *attribute disclosure*), according to the disclosure approaches which are: **uniqueness** and **record linkage**. We then propose a discussion on the data utility assessment using this model.

k-anonymity to Mitigate Identity Disclosure

k-anonymity (Samarati and Sweeney, 1998) (Samarati, 2001) (Sweeney, 2002), aims to protect against *identity disclosure* and proposes to build blocks of k indistinguishable subjects to prevent characterization of single subjects. This prevents at some extent *uniqueness* analysis, as a subject is less characterizable with specific values.

While there exist three types of attributes: identifiers, quasi-identifiers and confidential attributes, k-anonymity considers only two types: quasi-identifiers and confidential attributes. Indeed, identifiers are either suppressed or transformed into quasi-identifiers in order to reduce their identification capability; confidential attributes are not transformed.

k-anonymity (Sweeney, 2002) (Samarati, 2001) is therefore defined as follows:

Definition 1 *(k-anonymity)*
Let T be a table and Q be the quasi-identifier associated with it. T is said to satisfy k-anonymity if and only if each sequence of values in T[Q] appears with at least k occurences in T[Q].

Where T[Q] refers to all the values of the quasi-identifier (cf. Chapter 2) Q within T.

For example, let us consider Table 3.1 and Table 3.2 which are a raw microdata and its corresponding 3-anonymity instantiation respectively.

TABLE 3.1: Original Data Table (Disease.

	email address	Age	Disease
1	marco@orange.com	22	lung cancer
2	simon@orange.com	22	lung cancer
3	kevin@orange.com	22	lung cancer
4	a125@orange.com	45	flu
5	anne@orange.com	63	diabetes
6	youssou@orange.com	40	flu
7	vincent@orange.com	35	aids
8	nicolas@orange.com	35	aids
9	louis@orange.com	32	diabetes

TABLE 3.2: 3-anonymity Table (Disease).

	email address*	Age*	Disease
1	*	2*	lung cancer
2	*	2*	lung cancer
3	*	2*	lung cancer
4	*	≥ 40	flu
5	*	≥ 40	diabetes
6	*	≥ 40	flu
7	*	3*	aids
8	*	3*	aids
9	*	3*	diabetes

This is an instantiation of k-anonymity using the *Generalization and Suppression* protection mechanism (cf., Section 2.5.1). As we can observe attribute *email address* has been completely suppressed as its values are all different and can be used for identification. Attribute *Age* has been generalized into attribute *Age** by building groups of 3 subjects and therefore, an attacker can no longer link a subject to a specific *Age* value.

However, if k-anonymity mitigates *identity disclosure*, it fails mitigating *attributes disclosure*, especially: *homogeneity* and *backgroung knowledge attacks* (Machanavajjhala et al., 2007).

l-diversity to Mitigate Homogeneity and Background Knowledge Attacks

The l-diversity technique (Machanavajjhala et al., 2007) has been introduced to counteract two specific *attribute disclosure* attacks:

- **homogeneity attack** which refers to the knowledge gained by correlating *key attributes* and *sensitive attributes* within the table. For instance, in the 3-anonymity table (Table 3.2), the key attribute value "2*" completely corresponds to the sensitive value "lung cancer"; as such, an attacker only needs to know that the subject is twenties to link him to the disease "lung cancer". This is an application of **uniqueness** between *key attributes* and *confidential attributes*.
- **backgroung knowledge attack** which uses external data to improve subjects identification. For example, in the 3-anonymity table (Table 3.2), if we consider the third equivalence class (with key attribute value "3*"), the correspondence between "Age*" and "Disease" is not complete, and an attacker will therefore need external information (for example that the subject is less likely to have *diabetes*) to link him to *aids*. This is a specific case of **record linkage** where the attacker can link a key attribute value to a sensitive attribute value with a given probability.

To counteract those attacks, l-diversity adds the restriction that all the confidential attributes should have at least l "*well represented*" values.

Definition 2 *(The l-diversity principle)*
An equivalence class is said to have l-diversity if there are at least l "well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity.

In this definition, "*well-represented*" refers to the number of different values of the sensitive attribute within an equivalence class. Indeed, the more the sensitive values are diverse, the better the protection is, as an attacker would not be able to infer a specific sensitive value.

This restriction enables to reduce the correlation between key attribute values and sensitive attribute values and helps to mitigate both *homogeneity* and *background knowledge* attacks. This prevents *uniqueness* analysis between attributes within the microdata. For instance, in the 3-diverse table (Table 3.3), there are 3 different values of the sensitive attributes "Disease" and "Age" in each equivalence class, and this diversity prevents from correlating key attributes and sensitive attributes.

Despite these improvements, l-diversity has been proved (Li, Li, and Venkatasubramanian, 2007) to be inefficient to counteract *attribute disclosure* attacks as it does not take into account the semantic of attributes. This flaw is depicted through two main attacks: *skewness attack* and *similarity attack*.

t-closeness to Mitigate Skewness and Similarity Attacks

The t-closeness technique (Li, Li, and Venkatasubramanian, 2007) has been introduced to counteract:

TABLE 3.3: A 3-diverse Table.

	ZIP Code*	Age*	Salary	Disease
1	355**	2*	4K	colon cancer
2	355**	2*	5K	stomach cancer
3	355**	2*	6K	lung cancer
4	3581*	≥ 40	7K	stomach cancer
5	3581*	≥ 40	12K	diabetes
6	3581*	≥ 40	9K	aids
7	355**	3*	8K	aids
8	355**	3*	10K	flu
9	355**	3*	11K	lung cancer

- the **skewness attack** which is based on the skewness between the distribution of sensitive attribute values within the original table and the distribution within equivalence classes. Let us consider the following example:

  Example 1 *Suppose we have an original skewness table containing data of 1000 patients with and without cancer; the key attributes are "Age", "ZIP Code" and the sensitive attribute is "Cancer"; and "Cancer" can have two values "Yes" or "No". Suppose we have only 10 "Yes" in the table. A 2-diverse table (formed by equivalence class of 2 subjects) would provide 50% probability of having cancer for each subject within classes instead of 10/1000% in the original table and then, an information gain from the anonymized table.*

- the **similarity attack** which relies on similarity between sensitive values. Indeed, when the sensitive attribute values are distinct but semantically similar (e.g. "stomach cancer", "colon cancer", "lung cancer"), the *similarity attack* can occur. For example, let us consider the first class of the 3-diverse table (Table 3.3) with key value "2*". The value "2*" corresponds to the subset of sensitive values: {4K, 5K, 6K}. Even if those values are diversified, they still contain semantic information as an attacker can infer that all subjects who are twenties have low incomes. This is a specific case of **record linkage** where the attacker is not interested in specific values but instead in subsets of values.

To overcome *skewness* and *similarity attacks*, the t-closeness principle was proposed by Li and al (Li, Li, and Venkatasubramanian, 2007) and states that:

Definition 3 *(The t-closeness principle)*
An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the original table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

The t-closeness metric measures therefore the distance between distributions of sensitive values within classes and within the original table to ensure it does not exceed a given threshold. This property is claimed to mitigate both *skewness* and *similarity attacks*.

Table 3.4 is an instantiation of the t-closeness principle and we can observe that *semantic* and the *skewness* attacks can no longer be performed.

TABLE 3.4: An 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease.

	ZIP Code*	Age**	Salary	Disease
1	3556*	≤ 40	4K	colon cancer
3	3556*	≤ 40	6K	lung cancer
8	3556*	≤ 40	10K	flu
4	3581*	≥ 40	7K	stomach cancer
5	3581*	≥ 40	12K	diabetes
6	3581*	≥ 40	9K	aids
2	3550*	≤ 40	5K	stomach cancer
7	3550*	≤ 40	8K	aids
9	3550*	≤ 40	11K	lung cancer

Data Utility for k-anonymity

While it is quite simple to observe the gain in resistance for each k-anonymity improvement, it is not the case for data utility. Indeed, minimizing information loss for k-anonymity requires a prior analysis of the needs and implementations should consider a taxonomy and hierarchies of values (Xu et al., 2006) (Loukides and Shao, 2008b).

Utility is Handy with k-anonymity

An interesting point is about the trade-off between utility and privacy which, unlike ϵ-differential privacy (cf., Section 3.3.2), is particularly palpable for the k-anonymity model. Indeed, the more a given instantiation can resist attacks, the lower the data utility.

To illustrate this observation, let us consider Tables 3.2 and 3.3 which are a 3-anonymity instantiation and a 3-diversity instantiation respectively. Suppose a study which aim to provide a treatment for diseases according to the age; a possible utility assessment would be about the capability to infer from attribute Age a given Disease and be able thereafter, to prescribe the good treatment according to the age. It is straightforward to observe that, for this utility study, the 3-anonymity instantiation provides more utility than the 3-diversity instantiation.

This observation also shows the versatility of *attribute disclosure* that can be interpreted either as an attack or as a utility indicator.

3.3.2 ϵ-Differential Privacy

Unlike k-anonymity, ϵ-differential privacy (Dwork et al., 2006) is a non-deterministic approach which guaranties privacy of respondent by generating synthetic Data (cf., Section 2.6.7). Differential privacy has been introduced for sanitization of queryable data bases which is an interactive approach (cf., Section 2.4.2). The idea is to ensure that a data base user can not guess - at some extent, defined by the parameter (ϵ) - the presence/absence of any single respondent record. Thus, the query answer is transformed in such a way that the influence of presence or absence of any respondent record is limited. This transformation is performed according to the parameter ϵ; the smaller ϵ, the more difficult it is for an attacker to infer the contribution of a respondent on a given query answer. The formal definition of ϵ-differential privacy is as follows (Dwork, 2008):

Definition 4 *ϵ-Differential Privacy*
A randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all datasets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,

$$Pr[(\mathcal{K}(D_1) \in S] \leq exp(\epsilon) \times Pr[(\mathcal{K}(D_2) \in S] \tag{3.1}$$

In practice, ϵ-differential privacy prevents re-identification by adding noise to the query response. Suppose $f(X)$ is the real response to the user query f (e.g., the average of an attribute value, the number of records with a specific attribute value,...), this response is perturbed by adding a random amount of noise $g(X)$, to $f(X)$; the noise depends on ϵ and the *variability of the query response*. Finally, the returned response $r(X)$ has the following form:

$$r(X) = f(X) + g(X) \tag{3.2}$$

To generate the noise $g(X)$ according to ϵ-differential privacy, the Laplace distribution is commonly used with zero mean and $\Delta(f)/\epsilon$ scale parameter where:

- ϵ is the differential parameter
- $\Delta(f)$ is the L_1-sensitivity of f which is the maximum variation of the query function between neighbours data sets i.e. sets differing in at most one record.

The density function of Laplace noise is then:

$$p(x) = \frac{\epsilon}{2\Delta(f)} \exp^{-|x|\epsilon/\Delta(f)}. \tag{3.3}$$

The *variability of the query response* ($\Delta(f)$) refers to the variability of the query function between neighbour data sets i.e. how much the neighbours data sets influence the query response; indeed, as ϵ-differential privacy aims preventing the knowledge of presence or absence of any single respondent within the table, a high variability between neighbours is not desirable. Hence, the higher the variability, the higher the amount of necessary noise for smoothing the response would be.

According to equation 3.3, for fixed ϵ, the higher the sensitivity $\Delta(f)$ of the query function f, the more noise is added which is normal as ϵ-differential privacy aims to prevent observation. Also, for fixed sensitivity $\Delta(f)$, the smaller the ϵ parameter, the more Laplace noise is added. This last property is depicted by equation 3.1 as the closer ϵ is to zero, the closer are the probabilities and therefore the similarity between different data sets differing on at most one element (D_1 and D_2).

Therefore, privacy for ϵ-differential privacy is ensured by the epsilon parameter (ϵ) as the variability depends on the data set.

Differential privacy has been extended to be used for sanitization of microdata sets which refers to a non-interactive approach (Sarathy and Muralidhar, 2011)(Chen et al., 2011)(Hardt, Ligett, and McSherry, 2012).

Data Utility for ϵ-differential privacy

The main drawback of differential privacy is data utility (Lee and Clifton, 2011)((Clifton and Tassa, 2013)). Indeed, the noise addition greatly reduces the control on data and this prevents an accurate utility assessment.

In practice, the noise addition should be added in order to guarantee utility (Dwork et al., 2006), hence, the analysis to be performed should be known in advance in order to add to good amount of noise, taking into account $\Delta(f)$ (the global sensitivity) and ϵ (the differential parameter); this then refers to PPDM. Moreover, as noise addition is usually performed according to the global sensitivity, it prevents analysis over sub domains. Finally, the fact that sanitized data should be generated according to a predefined purpose has raised some criticisms asking: *"why not publish the statistics one wants to preserve rather than release the synthetic data ?"* (Domingo-Ferrer, 2008a). This contrasts with the k-anonymity model that provides more control over data to provide general purpose utility (cf., Section 3.3.1).

3.4 Disclosure Risk Metrics

This section presents the disclosure risk metrics that are: uniqueness and record linkage (cf., Section 3.4). They aim to link records within a data set A to records within another data set B and use different approaches which are mostly empirical. Note that, as stated in the introduction, the two privacy models (k-anonymity and ϵ-differential privacy) intrinsically provide privacy metrics (referred to as k and ϵ). However, they are insufficient for quantifying the disclosure risk as they provide global measurements over data, which the link with the capacity of re-identification is not direct.

3.4.1 Uniqueness

The idea of uniqueness is that unique combination of key variables values have higher risks of re-identification and mainly targets the *sampling protection mechanism* (cf., Section 2.6.3) (Skinner, Marsh, and Wymer, 1994)(Elamir and Skinner, 2006); the goal is to use the sample to re-identify a respondent within the whole population. Uniqueness can also be used for *attribute disclosure* as shown in Section 3.3.1.

We can distinguish **simple uniqueness** and **special uniqueness** (Templ, Meindl, and Kowarik, 2013).

Simple uniqueness

Let us consider Table 3.5 and perform uniqueness according to the sampling mechanism (cf., Section 2.6.3). Let f_k be the frequency counts of records with pattern k in the sample. A record is called a sample unique if it has a pattern k such that $f_k = 1$. Let F_k be the number of units in the population having the same pattern k. A record is called a population unique if $F_k = 1$.

For example record 2 in Table 3.5 has $f_k = 1$ as the combination of ZIP Code, Age and Disease (*75005*, *23* and *lung cancer*) is unique within this table. Records 7 and 8 have $f_k = 2$ as they are the only records with the same pattern.

Suppose now that an intruder observes f_k on the sample and wishes to re-identify using the whole population, the probability to re-identify is therefore computed as: $\frac{f_k}{F_k}$; as F_k is the frequency counts of pattern k within the whole population.

Special uniqueness

Unlike simple uniqueness, special uniqueness (Elliot, Skinner, and Dale, 1998) also takes into account subsets of key variable sets. This enables enhancing the probability of re-identification. For example in Table 3.5, record 2 is a sample unique for

TABLE 3.5: uniqueness Table.

	ZIP Code	Age	Disease
1	75002	22	lung cancer
2	75005	23	lung cancer
3	75002	22	lung cancer
4	75012	45	flu
5	75002	63	diabetes
6	75002	40	flu
7	75012	35	aids
8	75012	35	aids
9	75002	25	diabetes

attributes ZIP Code, Age and Disease (*75005*, *23* and *lung cancer*) but also for ZIP Code and Age (*75005* and *23*). A record is defined as special unique with respect to a variable set K if it is sample unique both on K and on a subset of K (Elliot, Skinner, and Dale, 1998). It has been shown that special uniqueness provides a higher probability of re-identification (Elliot, Manning, and Ford, 2002).

3.4.2 Record linkage

Uniqueness is not applicable on perturbative masking method as data are transformed and it becomes useless to search for matching key values (cf., Section 3.2). A more general approach consist in record linkage.

Record linkage is about linking records together (whether within the same micro data or between two different microdata) and requires defining an attacker model. We can distinguish four general approaches for record linkage:

- **Distance-based record linkage** the idea is to compute the distance between the records we wish to link (Pagliuca and Seri, 1999), (Domingo-Ferrer and Torra, 2001a)(Torra and Miyamoto, 2004); as such we should first define the distance to be computed. (Pagliuca and Seri, 1999) proposes a specific case of this approach using the microaggregation masking method with the Euclidean distance. For each record in the sanitized data set, the distance to every record in the original data set is computed; then the *nearest* and *second nearest* records in the original dataset are considered. A record in the sanitized data set is labelled as *linked* when the nearest record in the original dataset turns out to be the corresponding original record and; labelled as *linked to 2nd nearest* when the second nearest record in the sanitized data set turns out to be the corresponding original record. In all other cases, a record in the masked dataset is labelled as *not linked*. The percentage of *linked* and *linked to 2nd nearest* is a measure of disclosure risk. An empirical work shows that the record based linkage provides better performance than the probabilistic record linkage (Domingo-Ferrer and Torra, 2001a).

- **Interval disclosure** it is a particular case of distance-based record proposed by (Pagliuca and Seri, 1999). Instead of linking records (in original and sanitized data sets) between them, each sanitized value is ranked and intervals are constructed around the value. The width of the interval is based on the rank of the value or on its standard deviation. The proportion of original values that fall within the interval centred around their corresponding sanitized value is a

measure of disclosure risk. This approach allows computation on larger data sets than the simple distance-based record linkage.

- **Probabilistic record linkage** the idea is to compute probabilities according to coincidence between records in the sanitized data set and in the original data set (Fellegi and Sunter, 1969) (Jaro, 1989) (Torra and Domingo-Ferrer, 2003). As such, for each pair of records an index is computed: *linked*, *not linked* and *clerical*. Indexes *linked* and *not linked* are computed according to the following conditional probabilities for each variable in the pair of record: the probability $P(1|M)$ of correspondence of values between attributes for a real match; and $P(0|M)$ the probability of non correspondence between values of attributes for a real match; finally, the index *clerical* is about records that can not be automatically classified as *linked* or *not linked* and requires a human inspection. Disclosure is then computed as the number of matches between sanitized and original data.

- **Other record linkage methods** have also been proposed (Domingo-Ferrer and Torra, 2003)(Torra, 2004). While the previous record linkage approaches require that the data sets file share same variables; (Domingo-Ferrer and Torra, 2003)(Torra, 2004) propose a method which, under appropriate conditions, shows that re-identification is still possible when the files do not share the same variables. The conditions include the fact that both files contain similar structural information and that this structural information can be expressed by means of partitions. Therefore, the relationship between variables is established using the built partitions as common partitions in both files reflect the common structural information.

3.5 Utility Metrics

When talking about utility, we should consider two approaches: utility according to various data uses (PPDP) and utility according to specific data uses (PPDM) (cf., Section 3.2.2).

This section presents the metrics that are used for providing both approaches according to the k-anonymity model and the ϵ-differential privacy model. While for the k-anonymity model, utility measurements refer to the information loss measurements, the ϵ-differential utility assessment is more about machine learning process (classification, clustering ....) (Clifton and Tassa, 2013).

Note that, these metrics use mostly an *a priori* approach for assessment, meaning that the measurements are performed in order to provide anonymized data that reflect at some extent the original data set. On the other hand, an *a posteriori* approach would consist in measurements over anonymized data with respect to a given utility need. However, this latter approach is more complex as it requires a framework that would capture the utility need, which is subjective. In Chapter 6 we propose a framework for performing *a posteriori* assessment.

3.5.1 Utility for PPDP

Privacy Preserving Data Publishing (PPDP) aims to provide sanitized data that can be used for various applications. This approach is suitable to the k-anonymity model (cf., Section 3.2.2). The common way to ensure PPDP is by measuring information loss, more precisely the distance between original and sanitized data according to

specific statistics indicators.

Information loss measurement depends on the data type to be evaluated (continuous or categorical, cf., Section 2.3).

Utility Measurement for Continuous Data

Domingo Ferrer et al. (Domingo-Ferrer, Sánchez, and Hajian, 2015a), propose the following definition for information loss suited for continuous data:

Let us consider a microdata set with n respondents (records), $I_1, I_2, ..., I_n$ and p continuous attributes $Y_1, Y_2, ..., Y_p$. Let M be the matrix representing the original microdata set (rows are records and columns are attributes). Let M' be the matrix representing the sanitized microdata set. They consider the following indicators to characterize the information within sanitized data:

- Covariance matrices V on M and V' on M'.
- Correlation matrices R and R'.
- Correlation matrices RF and RF' between the p attributes and the p factors $PC_1, PC_2, ..., PC_p$ obtained through principal components analysis.
- Communality between each of the p attributes and the first principal component PC_1 (or other principal components PC_i's). Communality representing the percent of each attribute that is explained by PC_1 (or PC_i). Let C be the vector of communalities for M and C' the corresponding vector for M'.
- Factor score coefficient matrices F and F'. Matrix F contains the factors that should multiply each attribute in M to obtain its projection on each principal component. F' is the corresponding matrix for M'.

They propose thereafter to measure the discrepancy of these indicators between original and sanitized data. More precisely, they propose to measure the discrepancies between matrices M, V, R, RF, C, and F obtained on the original data and the corresponding M', V', R', RF', C', and F' obtained on the sanitized data set. For this purpose they propose three specific measures which are:

- **Mean square error** which refers to the sum of squared differences between components of pairs of matrices, divided by the number of cells in either matrix.
- **Mean absolute error** which refers to the sum of absolute differences between components of pairs of matrices, divided by the number of cells in either matrix.
- **Mean variation** which refers to the sum of absolute percent variation of components in the matrix computed on sanitized data with respect to components in the matrix computed on original data, divided by the number of cells in either matrix. This approach is not influenced by the scale changes of attributes.

Utility Measurement for Categorical Data

While this indicators are suited for continuous attributes, for categorical attributes, utility assessment usually refers to (Domingo-Ferrer and Torra, 2001b):

- **Direct comparison of values**
- **Comparison of contingency tables**
- **Entropy-based measures**

Direct comparison of values depends on whether the categorical values are *nominal* (which takes values over an unordered set) or *ordinal* (which takes values over a totally ordered set).

Indeed *nominal* values only accept equality as comparison operation and the comparison result between a sanitized category a' and its original correspondence a is binary i.e. either we have a correspondence or not:

$$d_c(a, a') = \begin{cases} 0 & (\text{if } a \neq a') \\ 1 & (\text{if } a = a') \end{cases} \tag{3.4}$$

For *ordinal* values, we can consider a distance. Let $\leq C$ be the total order operator over the range $R(C)$ of variable C. We define the distance between categories a and a' as the number of categories between the minimum and the maximum of a and a' divided by the cardinality of the range:

$$d_c(a, a') = \frac{|a'' : min(a, a') \leq a'' \leq max(a, a')|}{|R(C)|} \tag{3.5}$$

Comparison of contingency tables measures the distance between contingency tables rather than directly comparing values. The idea is to first build the contingency tables of original and sanitized data sets and then compare them by computing the sum of their differences component by component (Domingo-Ferrer and Torra, 2001b). *Comparison of contingency tables* generalizes some of the information loss measures based on counting (Hundepool et al., 2012).

Entropy-based measures this is one of the most used approaches for measuring utility for categorical attributes. The idea is to model sanitization as a process that removes information and the goal is therefore to assess the remaining information within the data. However, this approach requires to define the assessment model as random variables may be interpreted in different ways (Shin et al., 2012a). Propositions for this approach mainly target the PRAM protection mechanism and refer to Mutual Information which is a specific case of the KL-divergence metric (Rebollo-Monedero, Forne, and Domingo-Ferrer, 2010) (Rodriguez-Carrion et al., 2015).

3.5.2 Utility for PPDM

Privacy Preserving Data Mining (PPDM) aims at providing sanitized data which utility is calibrated according to a specific use or set of uses. The advantage of this approach is that data are optimized and can therefore provide a good amount of utility while protecting respondent's privacy. The main approaches for measuring PPDM are: classification, regression and clustering (Torra, 2017a).

Classification

Classification is one of the most used metrics for PPDP, they are more suited for categorical data and aims to classify a given element with respect to the data to which it belongs referring to different concepts including: *decision trees* and *nearest neighbour* (Torra, 2017b).

Decision trees aims to classify a given element according to a binary tree of questions where the each node is a question except the leaves which are categories. Then for each answer of questions with respect to the element, the element is oriented whether to the left child or to the right child of the node until it reaches its corresponding category. The decision trees is built from the data and the goal is to classify new elements correctly and minimize the eight of the tree (the number of questions to be asked).

Nearest neighbour aims to classify a given element with respect to its nearest elements. When a nearest element is found the category of this element is returned. Another approach (k-nearest neighbour) consists in finding the k-nearest elements and returning the category of the majority of these k elements.

Using these techniques, the utility assessment is performed by comparing the capability of anonymized data to classify a record with respect to original data (Bapna and Gangopadhyay, 2006) (Agrawal and Srikant, 2000) (Xue et al., 2017).

Some results show that sanitized data can even improve the data quality with respect to original data ((Bapna and Gangopadhyay, 2006))(Sakuma and Osame, 2017). The reason can be the fact that aggregation of values reduces the number of categories to be analysed and may therefore improve the capability to identify the category of interest. Note however that this is only true for a predefined need, as the definition of values to be aggregated depends on the use of data (Domingo-Ferrer, Sánchez, and Hajian, 2015a).

While those approaches are intended to target a very narrow set of uses, Fung et al. (Fung, Wang, and Philip, 2007) argue that: "*even if the data processor knows in advance that data will be used for classification, it may not know how the user may analyse the data. Application-specific details often influence the building of classifier. For example, some users prefer accuracy while others prefer interpretability. In many cases, visualization or exploratory analysis are useful for defining the right approach for classification. Therefore, even data sanitized with specific data mining tasks in mind can serve for other data mining tasks as well*".

Regression

Regression provides an estimation of the relationships among variables splitted into a *dependent variable* and one or more *independent variables* (predictors) and is more suited for continuous data. The goal is to measure how much the values of the dependent variable change when a given independent variable varies while the other independent variables do not change. Many derivations of regression exist: linear regression are used to assess anonymized data, for example (Muralidhar and Sarathy, 2005) compares different results of regression models while (Raghunathan, Reiter, and Rubin, 2003) compares the predictions of different regression models using the sum of squared error.

Clustering

Clustering can apply either on categorical or continuous data. The idea is to group elements within groups in such a way that elements in a given group are similar with respect to criteria. The common parameters to be fixed are the number of clusters to be formed and the considered criterion (usually a distance). There are many structures for clusters used for evaluating sanitized data: clusters, fuzzy clusters, dendrograms... (Batet et al., 2013) (Feldman et al., 2009).

3.6 Comparative Analysis of Disclosure Risk Metrics and Limitations

This section compares the previously described disclosure risk metrics and underlines their limitations. We focus on disclosure risk metrics as they are very much critical for enforcing the new regulation (cf. Chapter 1). We first present the criteria used for our comparison and then, our comparative analysis.

3.6.1 Assessment Criteria

We use the following criteria for comparing the disclosure risk metrics:

1. **Link with re-identification**: Disclosure risk metrics measure how much a respondent is about to be re-identified, therefore, the measurements should clearly establish the link with the capability to re-identify a given respondent.

2. **Empirical or analytical**: While an empirical assessment can provide accurate results, an analytical approach enables a better interpretation of the assessment results, which is necessary for a wide adoption and usage.

3. **Granularity**: Granularity enables more flexibility and accuracy for providing assessment with respect to specific use cases. Indeed, different interpretations can fit to the same privacy assessment and granularity enables to take into account specific needs. For example, for assessing the capability to re-identify a respondent with respect to his location, one may want to use specific locations, or subsets of locations instead, and the chosen metric should take this into account.

   Is it possible to perform measurements over many attributes, down to attributes'values, combination of attributes'values...

4. **Generality**. A generic approach of measurements is important for a large scope metric. For example, authors in (Domingo-Ferrer and Torra, 2003) underline the fact that few record linkage metrics are able to link records within different tables that do not share similar attributes'values, and propose a new record linkage approach for tackling this issue. This is an important concern as record linkage is usually performed over data that do not contain same values (Domingo-Ferrer and Torra, 2003). This feature is especially relevant for defining a large scope metric especially interesting for regulation definition (cf. Chapter 1)

   Can the metric be used with different anonymization mechanisms ? Does the metric take into account different types of attributes (categorical and continuous) ? Can the metric be used to link records within tables that do not contain same or similar attributes'values ?

5. **Applicability and scalability**. As they are intended to be used over large sets of data, record linkage metrics should be easily applicable (in terms of time consumption) and scalable.

3.6.2 Comparative Analysis of Existing Metrics

Let us now compare the existing disclosure risk metrics according to our criteria (cf. Section 3.6.1). Table 3.6 summarizes the comparison between these metrics.

k-anonymity-like metrics and the epsilon parameter

The k-anonymity-like metrics (k-anonymity, l-diversity and t-closeness) and the ϵ parameter of ϵ-differential privacy provide measurements that are **difficult to link to the re-identification capability** (Li, Li, and Venkatasubramanian, 2007) (Machanavajjhala et al., 2007) (Lee and Clifton, 2011). Moreover, the epsilon parameter suffers from lack of granularity as it does not provide measurements with respect to specific attributes or attributes'values but rather over tables as a whole. While epsilon **provides an analytical approach** for assessment, this is not validated in practice due to the difficulty to link the results to the re-identification capability. Both models **provide few generic** assessments as they can only apply to their corresponding application domain; moreover, epsilon is more suitable for continuous data. Finally, unlike the **k-anonymity model which is easily applicable over large sets** of data and can be easily assessed, **the epsilon parameter is less assessable** as it should be fixed before the anonymization and it is not possible to extract from an anonymized data set, the corresponding epsilon value.

Uniqueness metrics

The uniqueness metrics (simple uniqueness and special uniqueness cf. Section 3.4) **depict an empirical approach** for assessment (Domingo-Ferrer and Torra, 2001a) which requires specific implementation for each study case. **The simple uniqueness method is not granular** as we can only compare complete records, unlike **the special uniqueness approach** which compares subsets of key variable sets. However, **the approach is not generic** as it is suitable only for non-perturbative masking methods (cf. Section 3.4). However, as they do not provide an analytical approach, those metrics require, for each use case, specific considerations which restrict the definition of a global approach for regulation. **Uniqueness metrics are easily applicable** as they act over non-masked values.

Record Linkage

Record linkage as uniqueness **provides empirical approaches** (Domingo-Ferrer and Torra, 2001a) for assessment which **the link with re-identification varies with respect to the considered model and the specific parameters**. For instance, Distance-based record linkage depends on the considered distance which also depends on the use case. Also, once the distance is chosen, *distance-based record linkage* provides less accuracy than *interval disclosure* in terms of link with re-identification, as there are only two possibilities with *distance-based record linkage* (*nearest* and *second nearest* records cf. Section 5.4.5) while there is a wider range of values for *interval disclosure*. Record linkage provides **granular assessments down to attributes'values**. As

record linkage assumes the definition of a model that can fit to the case (e.g. specific distance in case of distance-based linkage), **record linkage is somehow generic**. However, **record linkage is less easily applicable** than uniqueness metrics as we should consider specific parameters which could add complexity over assessments (Domingo-Ferrer and Torra, 2003).

Other record linkage methods

The main difference of those metrics with uniqueness and record linkage is the capacity to link records that do not contain similar variables (Domingo-Ferrer and Torra, 2003)(Torra, 2004). **These metrics are therefore granular** and provide a **link with re-identification which is variable**. They are also **somehow generic** but **do not provide an analytical approach** for assessment. **Application of those metrics is as complex as for record linkage**.

TABLE 3.6: Table comparing existing disclosure risk metrics.

Metrics	Link with re-id	Granul	Analyti	Gener	Appl/Scal
k-anonymity	★ ☆	★ ☆ ☆	★ ☆	☆ ☆ ☆	★
epsilon	☆ ☆	☆ ☆ ☆	★ ★	☆ ☆ ☆	☆
Simple U	★ ★	★ ☆ ☆	☆ ☆	☆ ☆ ☆	★
Special U	★ ★	★ ★ ☆	☆ ☆	★ ☆ ☆	★
D-b linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ☆	☆
Itv linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ☆	☆
Prob linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ☆	☆
O record linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ★	☆

★: fulfilled criterion
☆: unfulfilled criterion

As depicted by the comparison table (Table 3.6), the "Other record linkage methods" is the one which fulfills most of the criteria. We also observe that, while the "epsilon" parameter provides the most analytical approach is lacks most of the described criteria. In the rest of this thesis we propose a metric which both includes the uniqueness and the record linkage approaches.

3.7 Conclusion

This chapter presents the SDC metrics for measuring both the disclosure risk of sanitized data and the remaining utility within sanitized data. We first present the two main privacy models in SDC which are k-anonymity and ϵ-differential privacy, and show their limitations to capture utility and disclosure risk. We then present existing metrics for disclosure risk and utility. The disclosure risk metrics assess either unique correspondences between attributes' values (uniqueness), or rely on more general approaches based on linkage models (record linkage). On the other hand, utility metrics can be classified according to whether they target sanitized data intended to respond to various uses (PPDP) or sanitized data calibrated for responding specific uses (PPDM). The disclosure risk metrics are then assessed and compared according to different criteria: link with re-identification, empirical/analytical, granularity, generality, applicability. As a result, none of the existing metrics fulfil all the criteria. However, disclosure risk metrics are especially relevant for regulation as it

aims to define rules for protecting respondent's privacy. Moreover, even for the k-anonymity model, which is the most well studied model (as it is the oldest), there are very few metrics enabling accurate comparison of its related derivations. In Chapter 4, we present our first contribution (The Discrimination Rate Metric) which fulfil almost all the previously listed criteria and provides an analytical approach for assessing disclosure risk with application over the k-anonymity model.

Part III

Contributions

Chapter 4

Discrimination Rate: An Attribute-Centric Metric to Measure Privacy

4.1 Introduction

This chapter is about our first contribution, the Discrimination Rate (DR) metric which is an attribute-centric metric as it measures the capability of attributes to refine an anonymity set. The DR enables tackling the limitations of the existing metrics (cf. Chapter 3).

Moreover, the DR enables accurate and formal definition of identifiers (more generally Personally Identifying Information), which has been recognized as one of the main concerns of privacy regulation (Schwartz and Solove, 2011). Indeed, as underlined in Chapter 1 identifiers are at the centre of personal data protection as personal data are defined with respect to identifiers. However, the current regulation does not provide a characterization of identifiers, and without a proper definition, one can not identify an identifier and therefore, can not protect personal data. The main concern with identifiers comes from the difficulty to provide a general definition of identifiers as they depend on the context.

The DR enables to address all these issues through fine grained assessments which provide: a direct link with re-identification, granularity, generality, an analytical approach and applicability. The DR relies on a largely adopted anonymity definition provided by Pfitzmann et al. (Pfitzmann and Hansen, 2010) and which states that: "*the anonymity of a subject from an attacker's perspective means that the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set*", where the anonymity set is defined as "*the set of all possible subjects*". We then use this definition to define our DR metric which measures the attacker's capability by evaluating the capability of the attributes the attacker owns, to refine the anonymity set. The maximum refinement leading to a single subject.

The following example drawn from the Location Based Systems (LBS) (Wernke et al., 2014) (Shin et al., 2012b), underlines the flexibility of this approach as it is attribute-centric. Assume Alice uses a *location service*[1] to find a cardiology clinic but wishes to hide her location to the service provider. The attacker is the service provider in this case and his goal is therefore to link "Alice" to "Cardiology clinic" and infer that Alice suffers from heart disease. We call *anonymity set* the set of all users in the same location who send a request to the service provider - at the same

[1] Location service: service delivered through mobile platforms and based on location data.

period of time - to find a Cardiology clinic. If Alice is the only user in the *anonymity set*, the identification is direct as the information into the request is enough to identify Alice (link "Alice" to "Cardiology Clinic"). Alice's request carries the maximum amount of information to identify Alice. Now assume there are $k-1$ other users in the *anonymity set* and that their requests are indistiguishable, the attacker then needs an extra amount of information (e.g. from some context data) to identify Alice - i.e. reduce the anonymity set to 1 user - as each request refers to k subjects. Alice's request does not carry the maximum amount of information to identify Alice and the attacker needs extra identifying information. The attacker's knowledge can therefore be measured by the identification capability of the information he owns.

The DR provides many features that can be configured to fit with different applications including measurement of single and combined attributes. The DR is very much practical with some algorithms provided. As the DR measures the identification capability of attributes, the link with identification is direct. The usefulness of the DR is illustrated through evaluation and comparison of the k-anonymous and l-diverse mechanisms, two of the most popular Statistical Disclosure Control (SDC) techniques. We are therefore able to provide an attack driven assessment by computing how much information is gained by an attacker after applying a given disclosure attack. Finally, the formalism introduced by the DR enables to formalize well known definitions like identifiers and quasi-identifiers, and to propose new definitions of zero-identifiers and partial-identifiers.

The rest of this chapter is organized as follows. Section 4.2 identifies key features for a good privacy metric. After giving first informal definitions about identifiers in Section 4.4, Section 4.3 introduces the Discrimination Rate metric with useful theoretical background and definitions for Sensitive vs Key Attributes. Section 4.5 revisits the identifiers' definitions with the formalism of the DR. Section 4.6 illustrates the DR relevancy through k-anonymity and l-diversity evaluation, and information loss computation. Section 4.7 illustrates the practical dimension and the information loss computation over a real dataset (the Adult data set). Section 4.8 provides a comparison of the DR with the existing metrics and Section 5.7 gives our conclusions.

4.2 Key Features For a Good Privacy Metric

We identify two key features for getting a relevant general privacy metric. This metric should :

1. quantify how much an attacker can refine an anonymity set from a given information, with fine granularity support in that measurement.

2. enable quantifying the amount of knowledge gained after applying a given attack on a given system.

The idea of feature (1) comes from Pfitzman et al.'s works (Pfitzmann and Hansen, 2010). For them, "*the anonymity of a subject from an attacker's perspective means that the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set*", where the anonymity set is defined as "*the set of all possible subjects*". They introduce the requirement for fine granularity, providing a full useful range of intermediary scores between "identified" and "not identified".

TABLE 4.1: Example of data table

Subjects	ZIP Code	Age	Salary	Disease
subject 1	35000	22	4K	cancer
subject 2	35000	35	5K	diabetes
subject 3	35000	63	3K	malaria
subject 4	35000	22	13K	cancer
subject 5	35000	22	8K	cancer
subject 6	35000	35	15K	malaria
subject 7	35000	45	9K	malaria
subject 8	35000	35	7K	diabetes
subject 9	35000	40	11K	diabetes

Feature (2) comes from the practical consideration that anonymization mechanisms are introduced to counteract some identified attacks (Li, Li, and Venkatasubramanian, 2007) (Singh, Bansal, and Sofat, 2014). Therefore, assessing anonymization mechanisms in terms of attacks, provides the most pragmatic approach for assessment. K-anonymity (Samarati and Sweeney, 1998) was designed against the *identity attack* targeting the disclosure of the identity, l-diversity (Machanavajjhala et al., 2007) improved k-anonymity by mitigating the *homogeneity* and *background knowledge attacks* (refer to Chapter 3 for more details). As such, the idea of the general privacy metric is to give a clear measured evaluation and comparison of some anonymization mechanisms, based on identified attacks. This analysis is provided in Section 4.6.1.

4.3 Our Informal Definitions Related to Identifiers

Hereafter are several informal definitions or explanations related to identifiers. Note that, for these definitions, an attribute is considered as a variable which can take different values. For example, in Table 4.1, Age can take values: 40, 35, 63, 22 and 45.

Definition 5 *An **Identifier** is an attribute or set of attributes whose knowledge helps, for each of its (combination of) values, to reduce an anonymity set of more than one subject to exactly one subject.*

Definition 6 *A **Zero-Identifier** is an attribute or set of attributes whose knowledge does not help, for each of its (combination of) values, to reduce an anonymity set of more than one subject.*

Definition 7 *A **Sketchy-Identifier** is an attribute or set of attributes whose knowledge helps, for at least one of its (combination of) values, to reduce an anonymity set of more than 2 subjects to at least 2 subjects.*

In the light of these definitions, we consider that the identification process refers to reducing the set of subjects to refine the target. As illustrated in Figures 4.1 and 4.2, the knowledge of an **identifier** enables - for each of its values - to reduce the anonymity set to a single subject whereas the knowledge of a **sketchy-identifier** permits - for at least one of its values - to reduce a subset to at least 2 subjects. Therefore,

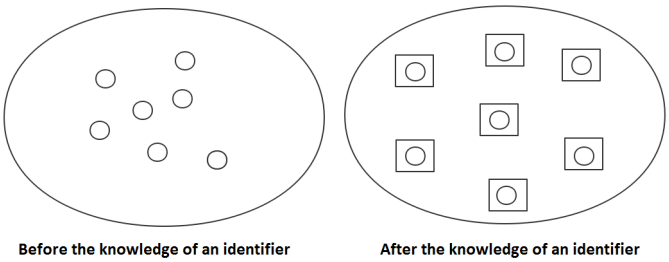

FIGURE 4.1: Anonymity set before and after the knowledge of an Identifier

an attribute that can not help to reduce a set of subjects can not help in identification and is considered, neither as an identifier nor as a sketchy-identifier and according to our definition, is a **zero-identifier**.

We consider identifiers, sketchy-identifiers and zero-identifiers as **context dependent**. An identifier (sketchy-identifier/zero-identifier) for a particular anonymity set, is not necessarily an identifier (sketchy-identifier/zero-identifier) for another.

For illustration, let us consider the following example: suppose we have a set of 9 subjects (Table 4.1) and an attacker only knows that there are nine indistinguishable subjects in the Table. Our anonymity set then refers to the sensitive attribute *Subjects*; the attacker's goal is therefore to refine the set of the *Subjects*'s values. In this context, attribute *ZIP Code* is a **zero-identifier** as everybody shares the same attribute value. Therefore, it does not help to reduce the anonymity set. Attribute *Salary* is an **identifier** as each of its values reduces the anonymity set to a subset of exactly 1 subject. Attribute *Disease* is a **sketchy-identifier** as it enables refinement of the anonymity set to 3 subsets of 3 subjects with respect to its values. Finally attribute *Age* is also a **sketchy-identifier**.

Also, we should distinguish different types of identifiers, **global** and **partial**. **Global identifiers** refine the anonymity set to subsets of exactly one subject whereas **partial identifiers** refine the main set to subsets among which there is at least one subset of one subject and one subset of at least two subjects. For example, attribute *Age* is a partial identifier; it refines the anonymity set to 2 subsets of 3 subjects (values 22 and 35) and 3 subsets of 1 subject (values 63, 45 and 40). Attribute *Salary* is a global identifier; it refines the main set to subsets of exactly 1 subject. In the following, **global identifiers** are referred to as **identifiers**.

As explained previously, identifiers are context dependent. For example, attribute *ZIP Code*, although a **zero-identifier** for this anonymity set, then becomes a **sketchy-identifier** if we introduce another subject with a different *ZIP Code* value.

4.4 Discrimination Rate (DR)

This section presents our general privacy metric, **the Discrimination Rate** (DR) which is based on information theory.
After giving some short introduction to entropy metrics, we introduce the Simple

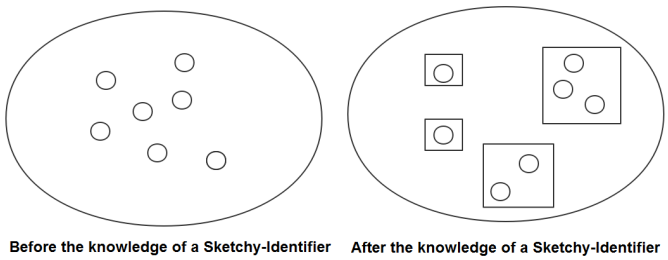

FIGURE 4.2: Anonymity set before and after the knowledge of a Sketchy-Identifier

Discrimination Rate (SDR) for single attribute measurements and the Combined Discrimination Rate (CDR) for multiple attributes measurements.

4.4.1 Background on Entropy

We give here short definitions of **entropy**, **conditional entropy** and **joint entropy** that are used to define our metric. For complete definitions, the reader can refer to (Kolmogorov, 1956).

- The **Entropy** "**H(X)**" of a d.r.v.[2] **X**, taking its values in $\mathcal{X}$, with a probability mass function **P** is a measure of its uncertainty and is defined as follows:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x)) \tag{4.1}$$

- The **Conditional Entropy** "$H(X|Y)$" of a d.r.v. **X** given a d.r.v. **Y**, is the entropy of **X** conditioned on each value y of **Y** averaged over all values y and defined as follows:

$$\begin{aligned} H(X|Y) &= -\sum_{y} p(y) \sum_{x} p(x|y) \log(p(x|y)) \\ &= -\sum_{x,y} p(x,y) \log(p(x|y)) \end{aligned} \tag{4.2}$$

- The **joint entropy** "**H(X,Y)**", of two d.r.v. taking their values within $\mathcal{X}$ for **X** and within $\mathcal{Y}$ for **Y** with a probability mass function **P** is a measure of uncertainty defined as follows:

$$H(X,Y) = -\sum_{x \in \mathcal{X}} p(x,y) log(p(x,y)) \tag{4.3}$$

  For more than two variables $X_1, ..., X_n$ the joint entropy is computed as:

$$H(X_1, .., X_n) = -\sum_{x_1} .. \sum_{x_n} p(x_1, .., x_n) log(p(x_1, .., x_n)) \tag{4.4}$$

Entropy is informally considered as the average amount of information contained in a source X that can take different values within $\mathcal{X}$.

Referring to Information theory, uncertainty is computed based on the fact that, the more likely an event is to happen, the less information it contains.

[2] d.r.v.: discrete random variable

4.4.2 Simple Discrimination Rate (SDR), and Sensitive vs Key attributes

The goal of our metric is to compute the identification capacity of an attribute into a given anonymity set, more precisely, to compute the amount of identification information carried by an attribute into that anonymity set. To remain far more general, we consider attributes as d.r.v. and the anonymity set as the set of outcomes (values and occurrences) of another d.r.v. To clarify our idea, let us consider 2 d.r.v. X and Y; Y the attribute we wish to measure the identification capacity and X, the attributes which the set of values is our anonymity set. As underlined in (Shin et al., 2012a), one of the main concern about entropy based anonymity metrics is the considered random variable. In our case, we want to compute the amount of information carried by a d.r.v. according to the refinement of the set of outcomes of another d.r.v. For that purpose, we consider $H(X)$ the amount of information (uncertainty) carried by X as our initial state. We then compute entropy of X conditioned on Y ($H(X|Y)$) as we wish to measure the effect of Y on X. This quantity represents the remaining uncertainty within X, after Y is divulged. In order to compute the amount of information carried by Y according to X, we need to subtract that quantity from $H(X)$ and thus we obtain $H(X) - H(X|Y)$, which is the effective amount of identification information carried by attribute Y according to the anonymity set. Finally, we divide that quantity by $H(X)$ to normalize the value.

Let us propose the following definition for the *Simple Discrimination Rate*:

Definition 8 ***(Simple Discrimination Rate)***
Let $\mathbf{X}$ *and* $\mathbf{Y}$ *be two d.r.v. The* ***Simple Discrimination Rate*** *of* $\mathbf{Y}$ *relatively to* $\mathbf{X}$ *is the capacity of* $\mathbf{Y}$ *to refine the set of outcomes of* $\mathbf{X}$ *and is computed as follows:*

$$DR_X(Y) = \frac{H(X) - H(X|Y)}{H(X)} = 1 - \frac{H(X|Y)}{H(X)} \tag{4.5}$$

It is easy to observe that $0 \leq DR_X(Y) \leq 1$ and:

- $DR_X(Y) = 0$ when Y is a **zero-identifier** as we have $H(X|Y) = H(X)$; the rest of information is maximal.
- $DR_X(Y) = 1$ when Y is an **identifier** as we have $H(X|Y) = 0$, the rest of information is null.

In the following, **X** is referred to as the **Sensitive Attribute** and **Y** is the **Key Attribute**.

SDR Computation illustration

This section describes how the Simple Discrimination Rate can be computed. We also provide an algorithm to describe the computation steps.

Example 2 *Let us consider Table 4.1 and compute the SDR of attribute Age (our key attribute) over Subjects (our sensitive attribute, cf. Section 4.4.2); the discrete random variables are therefore:*
X: Subjects and Y: Age.

$$
\begin{aligned}
SDR_X(Y) &= 1 - \frac{H(X|Y)}{H(X)} \\
&= 1 - \frac{-1/3\log_2(1/3) - 1/3\log_2(1/3)}{-\sum_{s=1}^{9} 1/9\log_2(1/9)} \\
&= 1 - \frac{1/3\log_2(3) + 1/3\log_2(3)}{\log_2(9)} \\
&= 0.66
\end{aligned}
$$

For $H(X|Y)$, the distribution is computed according to the definition of the Conditional Entropy in Section 4.4: the attribute Age can take 5 values 22, 35, 40, 45, 63. This helps to reduce the main set to subsets of 3, 3, 1, 1 and 1 subject(s) respectively, corresponding to 1/3, 1/3, 1/9, 1/9 and 1/9 of the whole set respectively. The conditional entropies are respectively: $H(X|Y = 22) = -log_2(1/3)$, $H(X|Y = 35) = -log_2(1/3)$, $H(X|Y = 40) = 0$, $H(X|Y = 45) = 0$ and $H(X|Y = 63) = 0$. $H(X|Y)$ is therefore the sum of $-1/3\ log_2(1/3)$ and $-1/3\ log_2(1/3)$.

Similarly, we can compute the DR of attributes ZIP Code, Disease and Salary which are respectively 0 (zero-identifier), 1/2 (sketchy-identifier) and 1 (identifier).

The SDR computation is depicted in Algorithm 1 where H(X) represents a precomputed value of the entropy of X.

Algorithm 1 Simple DR Computation of key attribute Y according to the sensitive attribute X. **Input**: The set $\mathcal{X}$ of values of X, the set $\mathcal{Y}$ of values Y and the connection between each Y outcome and X outcome. **Output**: $DR_X(Y)$.

```
Part2 ← 0
Sum ← total number of subjects
for each value y in (𝒴) do
    Correlate-Sum ← 0
    Sum-y ← number of subjects who share y
    i ← 0
    while ((x in 𝒳) and (number of subjects sharing x and y != 0)) do
        Tab[i] = ← number of subjects who share x and y
        Correlate-Sum ← Correlate-Sum + Tab[i]
        i ← i + 1
    end while
    Cond-Entropy ← 0
    if |𝒳| = 1 then
        Cond-Entropy ← log2(Correlate-Sum/Sum)
    else
        for each value t in Tab do
            Cond-Entropy ← Cond-Entropy + (t/Correlate-Sum) * log2(t/Correlate-Sum)
        end for
    end if
    Part2 ← Part2 - (Correlate-Sum/Sum) * Cond-Entropy
end for
DR ← 1 - Part2/H(X)
return DR
```

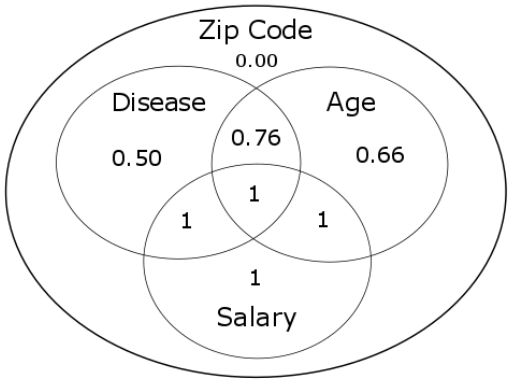

FIGURE 4.3: The Discrimination Rate for Table 4.1

4.4.3 Combined Discrimination Rate (CDR)

The SDR (Section 4.4.2) enables to measure the identification capacity of a single attribute according to a given anonymity set. The Combined Discrimination Rate goes one step further by measuring the identification capacity of a combination of attributes.

From **joint entropy**, we define the **Combined Discrimination Rate**. For convenience, we use the same DR notation for Combined Discrimination Rate and Simple Discrimination Rate, where only the number of parameters differs.

Definition 9 ***(Combined Discrimination Rate)***
*Let $X, Y_1, ..., Y_n$ be d.r.v. The **Combined Discrimination Rate** of $Y_1, Y_2, ..., Y_n$ relatively to X is the capacity of $Y_1, ..., Y_n$ to refine the set of outcomes of $\boldsymbol{X}$ and is computed as follows:*

$$DR_X(Y_1, ..., Y_n) = 1 - \frac{H(X|Y_1, ..., Y_n)}{H(X)}. \tag{4.6}$$

CDR Computation illustration

This section gives some computation examples of CDR and its algorithm.

Example 3 *Referring to Table 4.1 we can compute the Combined DR of the combination of attributes Age and Disease (our key attributes) over the Subjects (our sensitive attribute). The three d.r.v. are in this case:*
X: Subjects and Y_1: Disease and Y_2: Age

$$\begin{aligned} DR_X(Y_1, Y_2) &= 1 - \frac{H(X|Y_1,Y_2)}{H(X)} \\ &= 1 - \frac{-1/3\log_2(1/3) - 1/2\log_2(2/9)}{-\sum_{s=1}^{9} 1/9\log_2(1/9)} \\ &= 1 - \frac{1/3\log_2(3) + 2/9\log_2(2)}{\log_2(9)} \\ &= 0.76 \end{aligned}$$

*Hence, the combination of attributes **Disease** and **Age** gives more identification information than they give individually.*

Thereafter, $DR_X(Age, ZIP\ Code)$ *and* $DR_X(Age, Salary)$ *are respectively* $DR_X(Age) = 0.66$ *and* $DR_X(Salary) = 1$. *This is normal as ZIP Code is a zero-identifier in the current anonymity set (i.e. it does not bring any information) and Salary is an identifier (i.e. thus it brings the maximum of information for the current anonymity set).*

The DR can also be computed on attributes' values and combination of attributes' values. For example: $DR_X(cancer) = 0.83$, $DR_X(cancer, 22) = 0.83$, $DR_X(cancer, 35) = 0.93$, $DR_X(malaria, 35) = 1$.

Finally the values of $DR_X(Age, Disease, ZIP\ Code)$, $DR_X(Age, Disease, Salary)$, *and* $DR_X(Age, Disease, Salary, ZIP\ Code)$ *are respectively* $DR_X(Age, Disease) = 0.76$, $DR_X(Salary) = 1$ *and* $DR_X(Salary) = 1$. *We summarize the results in Figure 4.3.*

The CDR computation is depicted in Algorithm 2 where H(X) represents a pre-computed value of the entropy of X.

Algorithm 2 Combined DR of key attributes $Y_1, ..., Y_n$ relatively to the sensitive attribute X. **Input**: The set $\mathcal{X}$ of values of X, the key attributes $(Y_1, Y_2, ..., Y_n)$ and the connection between each $(Y_1, Y_2, ..., Y_n)$ outcome and X outcome. **Output**: $CDR_X(Y_1, Y_2, ..., Y_n)$.

```
Part2 ← 0
Sum ← total number of subjects
for each value (y_1, ..., y_n) in (Y_1, ..., Y_n) do
    Correlate-Sum ← 0
    Sum-y ← number of subjects who share (y_1, ..., y_n)
    i ← 0
    while ((x in X) and (number of subjects sharing x and (y_1, ..., y_n) != 0)) do
        Tab[i] = ← number of subjects who share x and (y_1, ..., y_n)
        Correlate-Sum ← Correlate-Sum + Tab[i]
        i ← i + 1
    end while
    Cond-Entropy ← 0
    if |X| = 1 then
        Cond-Entropy ← log_2(Correlate-Sum/Sum)
    else
        for each value t in Tab do
            Cond-Entropy ← Cond-Entropy + (t/Correlate-Sum) * log_2(t/Correlate-Sum)
        end for
    end if
    Part2 ← Part2 - (Correlate-Sum/Sum) * Cond-Entropy
end for
CDR ← 1 - Part2/H(X)
return CDR
```

Remark

From the examples, we observe that attributes with $DR = 0$ (the zero-identifiers) are useless for identification (for example ZIP Code). This remark has been underlined by Diaz et al (Diaz, Troncoso, and Danezis, 2007) who shown that, with more information, the attacker's uncertainty does not necessarily decrease.

4.5 Revisited Identifiers Definitions with DR

In the light of the CDR, we revisit the definitions of identifiers given in Section 4.3, to get the more formal definitions below.

Definition 10 ***Identifier***
Let X be a sensitive attribute and $Y_1, Y_2, ...Y_n$ be a set of key attributes, $n \in \mathbb{N} \setminus \{0\}$. $(Y_1, Y_2, ...Y_n)$ is an ***Identifier*** *relatively to X if, and only if:*
$DR_X(Y_1, Y_2, ..., Y_n) = 1$.

Definition 11 ***Sketchy-Identifier***
Let X be a sensitive attribute and $Y_1, Y_2, ...Y_n$ be a set of key attributes, $n \in \mathbb{N} \setminus \{0\}$. $(Y_1, Y_2, ...Y_n)$ is a ***Sketchy-Identifier*** *relatively to X if, and only if:*
$DR_X(Y_1, Y_2, ..., Y_n) \in]0, 1[$.

Definition 12 ***Zero-Identifier***
Let X a sensitive attribute and $Y_1, Y_2, ...Y_n$ be a set of key attributes, $n \in \mathbb{N} \setminus \{0\}$. $(Y_1, Y_2, ...Y_n)$ is a ***Zero-Identifier*** *relatively to X if, and only if:*
$DR_X(Y_1, Y_2, ..., Y_n) = 0$.

Definition 13 ***Partial-Identifier***
Let X be a sensitive attribute and $Y_1, Y_2, ...Y_n$ be a set of key attributes, $n \in \mathbb{N} \setminus \{0\}$, and $\mathcal{Y}_1, \mathcal{Y}_2, ..., \mathcal{Y}_n$ be the sets of possible values of $Y_1, Y_2, ...Y_n$. $(Y_1, Y_2, .., Y_n)$ is a ***Partial-Identifier*** *relatively to X, if and only if $(Y_1, Y_2, .., Y_n)$ is a* ***Sketchy-identifier*** *relatively to* ***X*** *and if $\exists\ (y_1, y_2, ..., y_n) \in (\mathcal{Y}_1, \mathcal{Y}_2, ..., \mathcal{Y}_n)$/*
$DR_X(y_1, y_2, ..., y_n) = 1$.

4.6 DR application To SDC

In this section, we apply the DR to measure anonymity within SDC systems (cf. Chapter 3) and use an attack driven assessment evaluation which computes how much information is gained by an attacker after applying a given attack. We perform our measurements over the k-anonymous and l-diverse techniques.

TABLE 4.2: Generalization Table

ZIP Code	ZIP Code*	Age	Age*
35510	355**	22	2*
35512	355**	22	2*
35517	355**	22	2*
35877	358**	35	3*
35830	358**	39	3*
35618	356**	35	3*
35620	356**	45	≥ 40
35842	358**	40	≥ 40
35655	356**	63	≥ 40

TABLE 4.3: A 3-anonymous Table

	ZIP Code*	Age*	location 1	location 2
1	*****	2*	diabetes-clinic	church
4	*****	2*	cardiology-clinic	church
5	*****	2*	cancer-clinic	church
2	*****	3*	hiv-clinic	synagogue
6	*****	3*	cancer-clinic	mosque
8	*****	3*	cardiology-clinic	synagogue
3	*****	≥ 40	diabetes-clinic	mosque
7	*****	≥ 40	hiv-clinic	mosque
9	*****	≥ 40	hiv-clinic	synagogue

TABLE 4.4: A 3-diverse Table

	ZIP Code*	Age*	location 1	location 2
5	355**	*	cancer-clinic	church
2	355**	*	hiv-clinic	synagogue
4	355**	*	cardiology-clinic	mosque
1	358**	*	diabetes-clinic	church
6	358**	*	cancer-clinic	synagogue
7	358**	*	hiv-clinic	mosque
3	356**	*	diabetes-clinic	church
8	356**	*	cardiology-clinic	mosque
9	356**	*	hiv-clinic	synagogue

4.6.1 Measuring SDC anonymization mechanisms with the DR

Let us now evaluate k-anonymity and l-diversity mechanisms, based on the three attacks listed in Chapter 3 and the DR metric.

Identity disclosure: The protection for k-anonymity and l-diversity against this attack comes directly from the generalization process (Table 4.2). The three following cases are usually considered:

- The black box approach: it refers to an attacker who only has the transformed data (public data) and wishes to re-identify.
- The white box approach: it refers to an attacker who has both the original and the generalized non-sensitive data but not the exact correspondences; for example {355**, 355**, 355**} as transformed data and {35510, 35512, 35517} as original data. The attacker can be an intruder within a company.
- The intermediary (grey box) approach: it refers to an attacker who has the generalized data and some external data that help him guess the original data. This approach is the hardest one as it is difficult to speculate on the external data.

In this chapter we only consider the white box case and we evaluate this attack by computing the amount of information an attacker can gain from the generalized key attributes.

To evaluate this attack, we therefore compute the amount of information that an attacker could gain from the generalized key attributes in the white box approach.

TABLE 4.5: Risk measurements for the *Identity disclosure*

X	Y	$DR_X(Y)$
ZIP Code	355**	0.83
ZIP Code	356**	0.83
ZIP Code	358**	0.83
ZIP Code	ZIP Code*	0.5
Age	2*	1
Age	3*	0.87
Age	≥40	0.78
Age	Age*	0.66

TABLE 4.6: Risk measurements for the *Homogeneity attack*

X	Y	$DR_X(Y)$
3-anonymity Table		
location 1	2*	0.85
location 1	3*	0.73
location 1	≥40	0.73
location 1	Age*	0.31
location 2	2*	1
location 2	3*	0.81
location 2	≥40	0.81
location 2	Age*	0.61
3-diversity Table		
location 1	355**	0.73
location 1	356**	0.73
location 1	358**	0.73
location 1	ZIP Code*	0.2
location 2	355**	0.67
location 2	356**	0.67
location 2	358**	0.67
location 2	ZIP Code*	0

The DR is computed from Table 4.2 over each key attribute using the original key attributes as *sensitive attributes*; Table 4.5 gives the risk measurements to *identity disclosure*.

Homogeneity attack: As explained in Chapter 3, this attack refers to the relative distribution between key and sensitive attributes. To evaluate resistance of k-anonymity and l-diversity to this attack, we measure how much the key attributes can serve to refine the sensitive attributes set. Thus the DR is computed over Tables 4.3 and 4.4 over the key attributes using each sensitive attribute as *sensitive attribute*. Table 4.6 gives the risk measurements to the *homogeneity attack*.

Background knowledge attack: refers to the external knowledge that can be used to link a quasi-identifier to a sensitive attribute within the table. However, using the attributes within the table, an attacker can already extract some information for linking a quasi-identifier to a sensitive attribute, and this refers to the **homogeneity attack**. Therefore, for computing the external information needed, we should

TABLE 4.7: Resistance measurement to combine *Homogeneity and Background knowledge attacks (computed from the results of Table 4.6)*

X	Y	$1 - DR_X(Y)$
3-anonymity Table		
location 1	2*	0.15
location 1	3*	0.27
location 1	≥40	0.27
location 1	Age*	0.69
location 2	2*	0
location 2	3*	0.19
location 2	≥40	0.19
location 2	Age*	0.39
3-diversity Table		
location 1	355**	0.27
location 1	356**	0.27
location 1	358**	0.27
location 1	ZIP Code*	0.8
location 2	355**	0.33
location 2	356**	0.33
location 2	358**	0.33
location 2	ZIP Code*	1

subtract the internal information (referring the the **homogeneity attack**) from the maximum amount of information needed for complete linkage. Hence, using the DR metric for computing this background knowledge, we should subtract the knowledge about the homogeneity attack (the internal knowledge) from the maximum DR which is 1. For example, Table 4.6 reports that the knowledge about the homogeneity attack for quasi-identifier $Age*$ is 0.61, meaning that for re-identifying a subject, the attacker needs extra data with accuracy evaluated with a DR of $1 - 0.61 = 0.39$, which represents the background knowledge. Table 4.7 summarizes the capacity of an attacker for the *background knowledge attack* within the 3-anonymous Table (4.3) and the 3-diverse Table (4.4).

Interpretation of results

The measurements in Tables 4.5, 4.6 and 4.7 are computed on attributes as a whole (ex: Age*, Zip Code*) and on attributes' values (ex: 2*, 355**).

Identity disclosure: From Table 4.5, we note that, although attribute *ZIP Code* has a global low risk to be used for re-identification ($DR = 0.5$), the risk related to its values remains relatively high ($DR = 0.83$). The DR permits an accurate evaluation that can be used during the generalization process to balance between privacy and information loss.

Homogeneity attack: From global measurements of attributes Age* and ZIP Code* in Table 4.6, we observe that the 3-diversity table provides a better security than the 3-anonymity table with lower DR values. However, according to some attributes' values, both mechanisms provide the same security level. For example, attribute's values ≥ 40 (3-anonymity table) and 356** (3-diversity table) have the same security

level according to *location 1* (DR = 0.73). We also observe that, for the 3-diversity table, the DR of *ZIP Code**'s values is lower for *location 2* (DR = 0) than for *location 1* (DR = 0.2); this can be explained as *location 1* has 4 different values while *location 2* has only 3. An attacker is therefore more able to relate a user to *location 1* from *ZIP Code**.

Background knowledge attack: Table 4.7 reports statistics about the amount of information needed to completely relate a sensitive attribute to a key attribute using a *backgroung knowledge attack*. To re-identify a user, an attacker can therefore focus on attributes with the lowest values ($1 - DR$) since they minimize his efforts. For example, for the 3-anonymity table, the attribute's value $2*$ is the most vulnerable one (DR = 0).

Global information loss: *information loss* is inversely proportional to the DR. That is, the less the DR, the more the information loss, and the less an attacker is able to re-identify a user. This comes directly from the definition of the DR: it quantifies the identification capacity of attributes. Therefore, the less an attacker can identify a user from attributes, the less he is able to re-identify.
The *overall information* loss can therefore be computed according to attacks by computing a mean over DRs. We then obtain:

- Global risk measurement according to Identity disclosure:
  $DR_{id} = (DR_{ZipCode}(ZipCode*) + DR_{Age}(Age*))/2 = (0.5 + 0.66)/2 = 0.58$
- Information loss related to *Identity disclosure* : $IL_{id} = 1 - DR_{id} = 0.42$

- Global risk measurement according to Homogeneity attack (3-anonymization):
  $DR_{haa} = (DR_{location1}(Age*) + DR_{location2}(Age*))/2 = (0.31 + 0.61)/2 = 0.46$
- Information loss related to *Homogeneity attack* (3-anonymization):
  $IL_{haa} = 1 - DR_{haa} = 0.54$

- Global risk measurement according to Homogeneity attack (3-diversification):
  $DR_{had} = (DR_{location1}(ZipCode*) + DR_{location2}(ZipCode*))/2 = (0.2 + 0)/2 = 0.1$
- Information loss related to *Homogeneity attack* (3-diversification):
  $IL_{had} = 1 - DR_{had} = 0.90$

- Overall *information loss* measurement (3-anonymization): $(IL_{id} + IL_{haa})/2 = (0.42 + 0.54)/2 = 0.46$

- Overall *information loss* measurement (3-diversification): $(IL_{id} + IL_{had})/2 = (0.42 + 0.90)/2 = 0.66$

Note that, all the above measurements can also apply on attributes'values, for specific interpretations.

TABLE 4.8: Attributes of the Adult dataset used in the experiment

	Attribute	Type	#Values
1	Marital Status	key attribute	7
2	Native Country	key attribute	41
3	Race	key attribute	5
4	Work Class	key attribute	8
5	Occupation	Sensitive	14

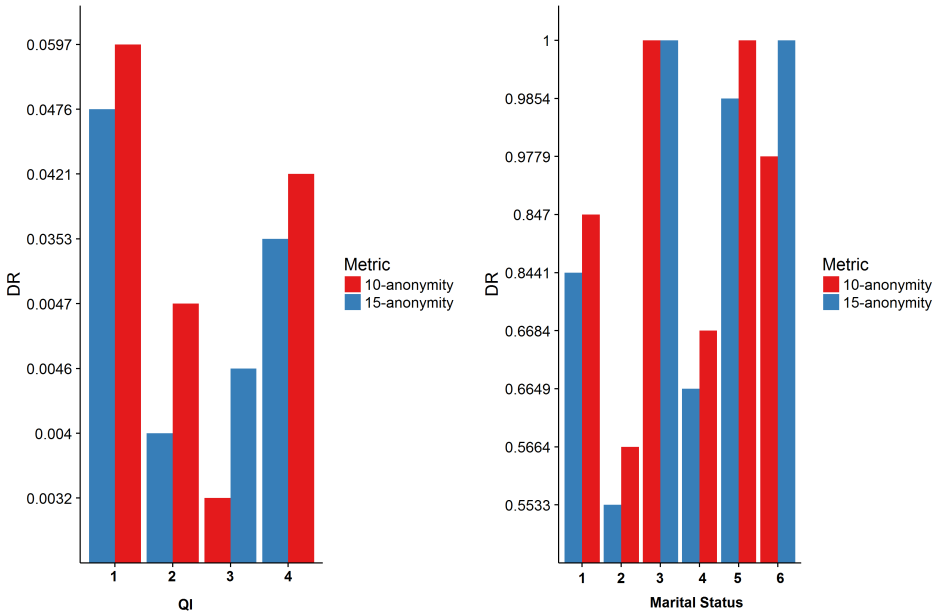

FIGURE 4.4: Identity Attack measurements in the Adult dataset

4.7 Experiments (k-anonymity and l-diversity assessment and comparison)

The goal of this experiment is to compare k-anonymity and l-diversity using a realistic dataset and various levels of anonymization.

The data used in the experiment is the "Adult dataset" from the UC Irvine Machine Learning Repository which contains data collected from a US census. After removing missing values, we obtain a total of 30161 valid records. We use five attributes of the dataset as depicted in Table 4.8 (4 key attributes and 1 sensitive attribute). We use the ARX tool (Polonetsky, Tene, and Jerome, 2014) (version 3.5.1) to compute the anonymization techniques (k-anonymity and l-diversity) and the R tool (Chokkathukalam et al., 2013) (version 3.3.1) to compute our DR metric. Our measurements are computed according to the examples given in Section 4.6. The experiments are running on a i5-4300U CPU with 1.90GHz - 2.50GHz and 4GB memory.

For k-anonymity, we generated 2 instantiations (10-anonymity, 15-anonymity) based on "Generalization and Suppression". For l-diversity, we generated 3 instantiations: a distinct-10-diversity, a shannon-entropy-10-diversity and a recursive-(5-10)-diversity (cf. Chapter 3).

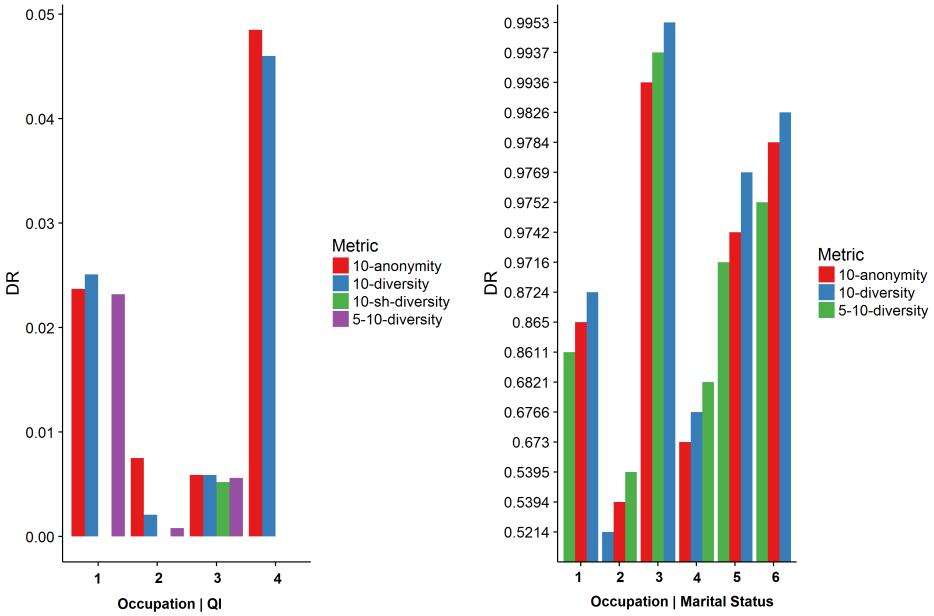

FIGURE 4.5: Homogeneity Attack measurements in the Adult dataset

TABLE 4.9: Values of attribute "Marital Status" used in the experiment

	Marital Status
1	Divorced
2	Married-civ-spouse
3	Married spouse absent
4	Never Married
5	Seperated
6	Widowed

4.7.1 Identity Attack

Figure 4.4 depicts the identity attack measurements. The measurements are made on key attributes as a whole (QI: see Table 4.8 for the correspondence between numbers and attributes) and on the specific values of the key attribute "Marital Status" (see Table 4.9 for the correspondence between numbers and values). The measurements describe the identification capacity of generalized attributes according to original attributes (cf. Section 4.6.1). We can observe from Figure 4.4 that the 15-anonymity instantiation is globally better than the 10-anonymity instantiation as it has fewer identifying attributes. However, for the key attribute "Native Country" the 15-anonymity instantiation provides weaker resistance (0.0047 vs 0.004). For the specific measurements over the key attribute "Marital Status", the 15-anonymity instantiation also provides a global better resistance even if for some values ("Widowed") the 15-anonymity instantiation is weaker (1 vs 0.9779).

4.7.2 Homogeneity attack

Figure 4.5 depicts the homogeneity attack measurements. We consider measurements of identification capacity of the key attributes according to the sensitive attribute "Occupation" (see Section 4.6.1 for more details). The measurements are

made on key attributes as a whole (Occupation|QI: see Table 4.8 for the correspondence between numbers and attributes) and on the specific attribute values of attribute "Marital Status" (Occupation|Marital Status: see Table 4.9 for the correspondence between numbers and values). For key attributes as a whole, we observe that the shannon-entropy-10-diversity is globally better than the recursive-(5-10)-diversity which in its turn is better than the distinct-10-diversity. With no surprise, the 10-anonymity instantiation provides the global lowest resistance.

However, for some key attributes ("Marital Status"), the 10-anonymity instantiation is more resistant than the distinct-10-diversity instantiation (0.237 vs 0.251). This is a non-expected result as 10-anonymity is supposed to be weaker than the distinct-10-anonymity. This is due to the anonymization process which uses an heuristic approach (Prasser et al., 2014) and tries to do the best; this enables a global better resistance for the distinct-10-diversity but which can be weaker for specific values. As such, the DR underlines inconsistencies which can occur during practical implementations.

For key attribute "Marital Status", we did not consider the shannon-entropy-10-diversity as for that model, the "Marital Status" values were completely destroyed during the anonymization process (see left graph on Figure 4.5). We observe that for this particular attribute, the distinct-10-diversity instantiation is globally less resistant than the 10-anonymity instantiation, as depicted by the "occupation|QI" graph.

4.8 Comparison of the DR with the existing disclosure metrics

The Discrimination Rate has the following interesting features:

1. **Context awareness**. The DR is context aware as it is computed relatively to a *sensitive attribute* and it has the flexibility to change the context by replacing a *sensitive attribute* with another one. This feature is not considered by the existing disclosure risk metrics as they only aim to link complete records within different tables.

2. **Granularity**. With its attribute-centric approach the DR offers a good granularity and can be computed with different parameters (over an attribute as a whole, over a combination of attributes, over specific attribute's value, over combination of attributes'values).

3. **Link with re-identification**: the link between DR measurements and re-identification is direct as the DR computes the capability of attributes to identify a subject.

4. **Generality**. The DR can apply on different anonymization mechanisms, to different types of attribute and even when the attributes to be linked do not share same or similar values. This property is useful for addressing different application domains.

5. **Analytical**: the DR provides an analytical approach

6. **Applicability and scalability**: applicability of DR is simple with an algorithm provided which is easily scalable to large data sets.

Note that context awareness is not new as it was identified as a lacking property in anonymity measures in (Tóth, Hornák, and Vajda, 2004).

The comparison is summarized in table 4.10. As we can observe, the DR addresses almost all the considered features except a specific generality feature which is the capability to compare tables that do not contain similar values. Indeed, as underlined by (Domingo-Ferrer and Torra, 2003) this type of assessment should take into account a mean to express the relationship between non-similar values within the different tables. This issue can be addressed by means of partitions (Domingo-Ferrer and Torra, 2003). This last issue is addressed in Chapter 5.

Table 4.10 provides a comparison of existing disclosure risk metrics with the DR metric. We use full and empty stars to reflect the capability to fulfil or not a requirement respectively.

TABLE 4.10: Table comparing existing disclosure risk metrics with DR.

Metrics	Link with re-id	Granul	Analyti	Gener	Appl/Scal
k-anonymity	★ ☆	☆ ☆ ☆	★ ☆	★ ☆ ☆	★
epsilon	☆ ☆	☆ ☆ ☆	★ ★	☆ ☆ ☆	☆
Simple U	★ ★	★ ☆ ☆	☆ ☆	☆ ☆ ☆	★
Special U	★ ★	★ ★ ☆	☆ ☆	★ ☆ ☆	★
D-b linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ☆	☆
Itv linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ☆	☆
Prob linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ☆	☆
O record linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ★	☆
DR	★ ★	★ ★ ★	★ ★	★ ★ ☆	★

★: fulfilled criterion
☆: unfulfilled criterion

4.9 Conclusion and Future Work

This chapter presents ou first contribution, the Discrimination Rate (DR) metric, for measuring disclosure risk within data bases. The DR uses an attribute-centric approach for providing a flexible and practical metric for disclosure risk assessment. Thanks to its attribute-centric approach, the DR enables to be more general than other proposals and tackles the limitations of the existing disclosure risk metrics in terms of: link with re-identification, granularity, generality, applicability and scalability. We are therefore able to provide an attack-driven privacy assessment by measuring how much information is gained by an attacker after applying a given attack; this allows to evaluate and compare k-anonymity and l-diversity, two of the most popular Statistical Disclosure Control (SDC) techniques. Finally, the DR is used for providing an accurate and quantified definition of fundamental privacy notions that are identifiers, and which are recognized as one of the critical concerns for definition of data protection regulations.

Improvements of the DR are still possible to take into account disclosure assessment between different tables which do not contain similar values. This is depicted by the capability to capture the proximity of some sensitive values for expressing their semantic similarity. Indeed, sensitive values can be grouped into subsets for

expressing meaning, and it might be of interest to discriminate those particular subsets of values instead of single values; for instance for getting the information that a person suffers from a cancer, whatever the cancer type. Chapter 5 describes the Semantic Discrimination Rate (SeDR) as an improvement of the DR; it provides a more generic disclosure risk assessment for improving record linkage.

Chapter 5

The Semantic Discrimination Rate

5.1 Introduction

To increase the DR with the semantic dimension, we define the Semantic Discrimination Rate (SeDR). The term semantic refers to the possibility for a data processor to perform specific measurements according to the meaning he gives to some attributes'values. Indeed, using the SeDR, a data processor can define subsets of values of interest (which reflect the meaning) and perform measurements with respect to these subsets of interest. This enables generic assessment for improving disclosure risk assessments. For example, a data processor could be interested in assessing the identification level of subjects suffering from cancer, whichever the cancer type; or he may be interested in identifying subjects living in a subset of locations, instead of specific locations. This new property improves assessment in different ways among which:

1. **A a more generic disclosure risk evaluation than the DR metric for measuring record linkage**
2. **Anonymity measurements from the attacker's perspective, by computing how much information is gained after applying a given attack. Especially for the *semantic attack*.**
3. **An illustrative experiment leading to the comparison of t-closeness and l-diversity, the proof that t-closeness as a metric, is not as protective as claimed by the authors and that, depending on the semantic considerations, t-closeness can be worse than l-diversity**.

This new feature enables disclosure risk assessment between tables that do not share similar values but which the values share semantic similarity and belong to the same respondents (Domingo-Ferrer and Torra, 2003). For that purpose, we introduce a new notion, the semantic partition which depicts a specific partition of an attribute's values for expressing the meaning they carry. As a result, and using an attack driven assessment approach, the SeDR enables to show that t-closeness, which is considered better than l-diversity (cf. Chapter 3), can be worse than l-diversity depending on the case.

The rest of the chapter is organized as follows: Section 5.2 presents our critical analysis on t-closeness technique, and points out its irrelevance to quantify privacy. Section 5.4 describes the semantic empowered Discrimination Rate together with our *semantic partition* definition; we then show how it can be used for improving *record linkage* assessment. Section 5.5 presents our measurements and comparison of l-diversity vs t-closeness. Section 5.6 provides our experiment on a real data set. Finally Section 5.7 gives our conclusions.

5.2 t-closeness Limitations and Inability to Quantify Privacy

(Domingo-Ferrer and Torra, 2008) identified the following limitations on the t-closeness metric:

- t-closeness does not provide a computational procedure;
- If such a procedure was available, it would greatly damage the utility of data. Indeed, by definition, t-closeness aims to destroy the correlations between key attributes and sensitive attributes and this, for any combination of key attribute values.

Additionally, we identified another criticism as t-closeness, taken as a metric, does not measure the effective disclosure risk but instead the accomplishment of the anonymization process. Indeed, two attributes with the same t-closeness measurement can have different privacy levels. This comes directly from the definition (cf. Chapter 3): *t-closeness computes the distance between the distribution of a sensitive attribute within classes and the distribution of attributes in the original table*. That is, a t-closeness measurement relies on the distribution of attributes in the original table; hence, if two attributes have different distributions in the original table, they can have the same t-closeness measurement, but not the same disclosure risk.

Another concern is that the t-closeness measurement relies on a pre-built hierarchy of attribute values that can differ according to the attacker's model. Indeed, in order to compute a t-closeness measurement, the attribute values should be classified and the measurement relies on this classification (Li, Li, and Venkatasubramanian, 2007) which is subjective. We give more details about semantical subjectivity and attacker's model in Section 5.4.1.

Finally, there is no direct link between t-closeness measurements and the re-identification process. Indeed, t-closeness computes a distance between distribution sets and the relationship with information gain or loss is unclear as acknowledged by the authors (Li, Li, and Venkatasubramanian, 2007): *"...the relationship between the value t and information gain is unclear"*.

TABLE 5.1: An 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease.

	ZIP Code*	Age**	Salary	Disease
1	3556*	≤ 40	4K	colon cancer
3	3556*	≤ 40	6K	lung cancer
8	3556*	≤ 40	10K	flu
4	3581*	≥ 40	7K	stomach cancer
5	3581*	≥ 40	12K	diabetes
6	3581*	≥ 40	9K	aids
2	3550*	≤ 40	5K	stomach cancer
7	3550*	≤ 40	8K	aids
9	3550*	≤ 40	11K	lung cancer

5.3 Inability for Basic DR to Measure Semantic

The DR does not take into account the semantic behind attribute values. For example, the t-closeness instantiation (Table 5.1) can provide more semantic privacy than the l-diverse instantiation (Table 5.2). Indeed, from key attribute value 355** in the l-diverse instantiation, an attacker can infer with 50% success that the user's salary is low (between 4K and 6K). This reflects the semantic aspect of attributes which is not taken into account by l-diversity, but is included in the t-closeness approach (Table 5.1).

TABLE 5.2: A 3-diverse Table.

	ZIP Code*	Age*	Salary	Disease
1	355**	2*	4K	colon cancer
2	355**	2*	5K	stomach cancer
3	355**	2*	6K	lung cancer
4	3581*	≥ 40	7K	stomach cancer
5	3581*	≥ 40	12K	diabetes
6	3581*	≥ 40	9K	aids
7	355**	3*	8K	aids
8	355**	3*	10K	flu
9	355**	3*	11K	lung cancer

5.4 Semantic Empowered Discrimination Rate

This section presents the semantic DR (SeDR) that supports semantic measurements. After arguing that the semantic measurement is a subjective measurement, we define our *semantic domains* that permit to capture this subjectivity. Then, we present our semantic Discrimination Rate (SeDR), along with illustration of SeDR computation.

5.4.1 Semantic as a Subjective Measurement with Regard to Attacker's Model

The term *semantic* refers to the meaning of attributes or attributes values, which is fully subjective. Indeed, an attribute's value can have different meanings according to the **attacker's model**. The attacker's model here refers to **the attacker's goal** and **previous knowledge to achieve this goal**. The attacker's model can be specified according to the categories an attacker is classifying the sensitive values.

For instance, let us consider the following three attacker's models over Tables 5.2 and 5.1 where the **attacker's knowledge** is made of the *key attributes* Age* and ZIP Code*:

1. The attacker wants to know the exact Salary's value of a subject;
2. The attacker wants to know which Salary category the subject belongs to: low (4K-6K), medium (7K-9K) or high (10K-12K);
3. The attacker wants to link the subject to one of the following Salary's subsets: {4K, 6K, 10K}, {7K, 12K, 9K} and {5K, 8K, 11K}.

For the **attacker's model 1**, the attacker needs to know the exact value. The set of categories contains therefore single values: {4K}, {5K},...,{12K}. Hence, the similarity between values is not taken into account for this model as the attacker is interested in single values. Therefore, the t-closeness instantiation provides the same semantic security than the l-diverse instantiation and the current DR is enough to compute the disclosure risk for both techniques.

For the **attacker's model 2**, the attacker's needs are not as restrictive as for attacker's model 1 as the attacker only wants to know the average salary. For this attacker's model, the similarity between values is worthwhile and is a privacy risk. As such, the t-closeness metric and the l-diversity metric do not provide the same semantic security, and adaptation of the Discrimination Rate is therefore necessary to measure that disclosure risk.

The **attacker's model 3** is somewhat interesting as it refers to the subsets of Salaries within classes of the t-closeness table (Table 5.1). As shown in Section 5.5, for this model, the t-closeness instantiation (Table 5.1) is proved to be worse than the l-diverse instantiation (Table 5.2).

5.4.2 Semantic Domain Definitions

This section gives our definitions about *semantic partition* and *semantic domains*, which help to capture the subjectivity of semantic based on attacker's models of section 5.4.1. These definitions are illustrated through an example.

Definition 14 *(Semantic Partition)*
*Let X be an attribute and $\mathcal{X}$ be the set of all possible values of X. A **Semantic Partition** of X is a partition of $\mathcal{X}$ according to a given attacker's model.*

Definition 15 *(Semantic Domain)*
*A **Semantic Domain** is an element of a Semantic Partition.*

The semantic domains refer to the classification of sensitive values with respect to their sensitivity similarity as identified by the attacker's model (Section 5.4.1). We refer to the set of *semantic domains* as the *semantic partition*. Indeed, for the purpose of this work, we suppose the semantic domains to be disjoint and the *semantic partition* to be a partition[1] of the set of sensitive values.

We use partitions instead of more sophisticated structures as dendrograms or ontologies because as shown by (Neumann and Norton, 1986), partitions are more robust to changes in data.

The corresponding *semantic partitions* of attacker's models in Section 5.4.1 are:

- Attacker's model 1: SP_1 = {{4K}, {5K},..., {12K}}.
- Attacker's model 2: SP_2 = {{4K, 5K, 6K}, {7K, 8K, 9K}, {10K, 11K, 12K}}.
- Attacker's model 3: SP_3 = {{4K, 6K, 10K}, {7K, 12K, 9K} {5K, 8K, 11K}}.

Note that, the methodology for getting a *semantic partition* is out of scope of this chapter. Our objective is only to show how subjective are the anonymity measurements and how semantic can be introduced in our DR metric. There are however some works (Erola et al., 2010) (Abril, Navarro-Arribas, and Torra, 2010) proposing a way to cluster values according to their semantic similarity, and therefore, a way to build semantic partitions.

[1] Partition of a set A: is a subdivision of A into subsets that are disjoint, non-empty and which the union equals to A.

5.4.3 SeDR as DR with Semantic Measurement

To cope with the DR's inability to handle semantic dimension as explained in Section 5.3, this section defines the Semantic DR (SeDR) which supports semantic measurements based on the *semantic domains* (Section 5.4.2).

Thanks to the *semantic domains*, the SeDR has the objective to measure how much an attacker provided with key attributes, is able to refine the set of semantic domains (the *semantic partition*) instead of the set of single values. As such, with SeDR, it is possible to know the attacker's capacity to infer **subsets** of user's sensitive values from a key attribute value.

Before applying the SeDR, we should first transform the sensitive attribute X according to a given semantic partition SP. Let sX be the result of the transformation. We define the **semantic partition transformation** f_{SP} as follows:

$$f_{SP} : X \rightarrow sX. \tag{5.1}$$

The SeDR is then defined as follows:

Definition 16 ***(Semantic Discrimination Rate)***
*Let X be a sensitive attribute and SP a semantic partition of X. Let $sX = f_{SP}(X)$ and $Y_1, ..., Y_n$ be a set of key attributes. The **Semantic Discrimination Rate** (SeDR) of $Y_1, ..., Y_n$ relatively to X is the DR of $Y_1, ..., Y_n$ relatively to sX and is computed as follows:*

$$SeDR_X(Y_1, Y_2, ..., Y_n) = DR_{sX}(Y_1, Y_2, ..., Y_n) \tag{5.2}$$

Therefore, the original DR is a particular case of the SeDR with a *semantic partition* composed of single sensitive values.

5.4.4 Illustration of the SeDR Computation and Comparison with the DR

Let us illustrate the SeDR over the original data Table 5.3 with the *semantic partition* SP_4 = {{diabetes, flu, aids}, {colon cancer, lung cancer, stomach cancer}}.

TABLE 5.3: Original Data Table (Salary/Disease).

	ZIP Code	Age	Salary	Disease
1	35567	22	4K	colon cancer
2	35502	22	5K	stomach cancer
3	35560	22	6K	lung cancer
4	35817	45	7K	stomach cancer
5	35810	63	12K	diabetes
6	35812	40	9K	aids
7	35502	35	8K	aids
8	35568	35	10K	flu
9	35505	32	11K	lung cancer

The semantic partition transformation f_{SP} is applied on X by replacing the set of values $\mathcal{X}$ (of X) by the set of values $s\mathcal{X}$ (of sX). For example, for *sensitive attribute* "Disease", we transform $\mathcal{X}$ = *{colon cancer, stomach cancer, lung cancer, stomach cancer, diabetes, aids, aids, flu, lung cancer}* using SP_4 = *{{diabetes, flu, aids}, {colon cancer, lung*

TABLE 5.4: Semantic DR in Table 5.3.

X	Y	$DR_X(Y)$
Disease	22	1
Disease	32	1
Disease	35	1
Disease	40	1
Disease	45	1
Disease	63	1
Disease	Age	1

cancer, stomach cancer}} into $s\mathcal{X}$ = *{cancer, cancer, cancer, cancer, other disease, other disease, other disease, other disease, cancer}*.

When applying the previous transformation on the sensitive attribute Disease in Table 5.3, and computing the SeDR according to the key attribute "Age*", we obtain the results in Table 5.4.

As shown in Table 10 vs Table 5, the SeDR is able to extract more information from the same database than the non-semantic DR, as higher values are obtained in Table 5.4. For instance, for key attribute value 22, the SeDR is 1 compared to 0.79 for the DR, as this key attribute fully corresponds to the semantic domain {colon cancer, lung cancer, stomach cancer} of the original data table (Table 5.3).

5.4.5 Measuring Record Linkage with SeDR

Record linkage aims to link records that belong to the same respondents between different tables. Some existing techniques (Domingo-Ferrer and Torra, 2003) (Torra, 2004) enable comparison of records that do not contain similar values using different models. These approaches require first to express the similarities between records before comparing them.

Using the SeDR, the similarities are reflected by the semantic partitions and measurements are performed with respect to these semantic partitions. For example the previous illustration of SeDR computation in Section 5.4.4 can be considered as a record linkage assessment where we try to link Disease's values to specific groups of Age's values. In the context of record linkage, we can suppose that these attributes belong to different tables. Disease and Age do not contain similar values but we are able to compute the correspondence between Disease's values and a specific partition of Age's values (which reflects the meaning a data processor can give to attribute Age's values) and the computation depicts their linkage. We then observe a perfect match between the Disease's values and this specific partition of Age's values which would provide the maximal capacity for an attacker trying to link these subsets of values (cf. Section 5.5.3). The comparison of the SeDR and the other metrics is depicted in Table 5.5 (cf. Chapter 4).

5.5 Measurement and Comparison of l-diversity vs t-closeness with SeDR

This section shows first how the semantic attacks - skewness attack and the similarity attack (Section 2.3) - can be measured with either the DR or the SeDR. Then it proves through the SeDR, for the similarity attack only, that t-closeness is not as

TABLE 5.5: Table comparing existing disclosure risk metrics with DR and SeDR.

Metrics	Link with re-id	Granul	Analyti	Gener	Appl/Scal
k-anonymity	★ ☆	☆ ☆ ☆	★ ☆	★ ☆ ☆	★
epsilon	☆ ☆	☆ ☆ ☆	★ ★	☆ ☆ ☆	☆
Simple U	★ ★	★ ☆ ☆	☆ ☆	☆ ☆ ☆	★
Special U	★ ★	★ ★ ☆	☆ ☆	★ ☆ ☆	★
D-b linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ☆	☆
Itv linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ☆	☆
Prob linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ☆	☆
O record linkage	★ ☆	★ ★ ★	☆ ☆	★ ★ ★	☆
DR	★ ★	★ ★ ★	★ ★	★ ★ ☆	★
SeDR	★ ★	★ ★ ★	★ ★	★ ★ ★	★

★: fulfilled criterion
☆: unfulfilled criterion

privacy protective as claimed by the authors, and that it can provide lower privacy protection than l-diversity. Both t-closeness and l-diversity techniques are instantiated over the original data Table 5.3 to give Tables 5.2 and 5.1 respectively. Note that these tables are similar to the ones of the original paper related to the t-closeness metric (Li, Li, and Venkatasubramanian, 2007).

5.5.1 Skewness Attack - Measurement with DR

The original DR is enough to evaluate this attack as only the skewness between the original distribution of sensitive values and their distribution within equivalence classes needs to be measured. For explaining this measurement, let us recall the skewness example of Chapter 3:

Example 4 *Suppose we have an original skewness table containing data of 1000 patients with and without cancer; the key attributes are "Age", "ZIP Code" and the sensitive attribute is "Cancer"; and "Cancer" can have two values "Yes" or "No". Suppose we have only 10 "Yes" in the table. A 2-diverse table (formed by equivalence class of 2 subjects) would provide 50% probability of having cancer for each subject within classes instead of 10/1000% in the original table and then, an information gain from the anonymized table.*

The objective of the attack is to improve the attacker's knowledge within the equivalence classes. As such, the DR enables to quantify how much information is gained by an attacker from equivalence classes, according to the original table.

Therefore, for evaluating this attack, we compute the difference between the DRs of the involved key attributes in the original table and in the equivalence classes. Based on the skewness table of Example 4, we compute the DR of key attributes "Age" and "ZIP Code" using "Cancer" as the *sensitive attribute* in the original table ($DR_{Cancer}(Age)$ & $DR_{Cancer}(ZIPCode)$) and the DR of the key attributes "Age*" and "ZIP Code*" within equivalence classes ($DR_{Cancer}(Age*)$ & $DR_{Cancer}(ZIPCode*)$). Finally the actual information gain related to *skewness attack* is:

- $DR_{cancer}(Age) - DR_{cancer}(Age*)$ for key attribute Age.
- $DR_{cancer}(ZIPCode) - DR_{cancer}(ZIPCode*)$ for key attribute ZIP Code.

TABLE 5.6: Risk measurement for Tables 5.2 & 5.1 for the similarity attack using SP_4 as the semantic partition and Age* & ZIP Code* as key attributes.

X	Y	$SeDR_X(Y)$
3-diverse Table		
SP_4	2*	1
SP_4	≥ 40	0.69
SP_4	3*	0.69
SP_4	Age*	0.38
SP_4	355**	0.38
SP_4	3581*	0.69
SP_4	ZIP Code*	0.07
t-closeness Table		
SP_4	≤ 40	0.38
SP_4	≥ 40	0.69
SP_4	Age**	0.07
SP_4	3550*	0.69
SP_4	3581*	0.69
SP_4	3556*	0.69
SP_4	ZIP Code*	0.07

This computation can also be performed on attribute's values instead of the attributes.

This evaluation through DR computation gives far more results than merely computing the ratio between probabilities (50% and 10/1000%), as the DR takes into account the correlation between key attributes and sensitive attributes and since the attacker's knowledge refers to key attributes, the DR quantifies the actual information gain.

5.5.2 Similarity Attack - Measurement with SeDR

The SeDR is computed to evaluate the similarity between values of sensitive attributes. The similarity between values is formalized through some defined *semantic partitions*.

We consider three *semantic partitions*; two partitions of "Salary" (according to the attacker's models 2 and 3, Section 5.4.2) and one partition of "Disease":

- SP_2 = {{4K, 5K, 6K}, {7K, 8K, 9K}, {10K, 11K, 12K}} for "Salary".
- SP_3 = {{4K, 6K, 10K}, {7K, 12K, 9K} {5K, 8K, 11K}}
- SP_4 = {{diabetes, flu, aids}, {colon cancer, lung cancer, stomach cancer}}

We then use these *semantic partitions* and each key attribute ("Age*" and "ZIP Code*") to compute the SeDR for the l-diverse and the t-closeness instantiations (Tables 5.2 and 5.1). The results are depicted in Tables 5.7, 5.8 and 5.6.

TABLE 5.7: Risk measurement for Tables 5.2 & 5.1 for the similarity attack using SP_2 as the semantic partition and Age* & ZIP Code* as key attributes.

X	Y	$SeDR_X(Y)$
3-diverse Table		
SP_2	2*	1
SP_2	≥ 40	0.81
SP_2	3*	0.81
SP_2	Age*	0.61
SP_2	355**	0.39
SP_2	3581*	0.81
SP_2	ZIP Code*	0.19
t-closeness Table		
SP_2	≤ 40	0.39
SP_2	≥ 40	0.81
SP_2	Age**	0.19
SP_2	3550*	0.67
SP_2	3581*	0.81
SP_2	3556*	0.81
SP_2	ZIP Code*	0.28

5.5.3 Results Proving the Lower Privacy Protection of T-closeness vs L-diversity

As reported in Section 5.4.3, the SeDR can compute the refinement capacity of a given *key attribute* over a *semantic partition* of a *sensitive attribute* (Section 5.4.2). The *semantic partition* reflects the subjectivity related to the semantic interpretation. The semantic risk measurement consists therefore in measuring how much from a given *key attribute*, an attacker is able to refine the *semantic partition* of a *sensitive attribute*. The more an attacker is able to refine the *semantic partition*, the higher the related risk.

Each computation is therefore performed according to a given *key attribute* and a given *semantic partition*.

Tables 5.7, 5.8 and 5.6 depict the risk measurements related to the *similarity attack* performed over the l-diverse table (Table 5.2) and the t-closeness table (Table 5.1). The considered key attributes are ZIP Code* and Age* and their SeDR are computed over the semantic partitions SP_2, SP_3 (related to "Salary") and SP_4 (related to "Disease").

Hereafter, we prove that **the assertion that t-closeness is semantically more secure than l-diversity is wrong**:

1. Table 5.7 shows that **an attacker is more able to refine the semantic partition SP_2 within the t-closeness table based on key attribute ZIP Code* than with the l-diversity table**. ZIP Code* gives a SeDR of 0.28 for t-closeness vs 0.19 for l-diversity.

2. Table 5.8 proves that **the t-closeness instantiation is weaker than the l-diverse instantiation against the *similarity attack* for the semantic partition SP_3**. Based on attribute ZIP Code*, an attacker is able to completely refine the *semantic partition* SP_3 (DR = 1), as the ZIP Code*'s values directly refer to the considered *semantic domains*.

TABLE 5.8: Risk measurement for Tables 5.2 & 5.1 for the similarity attack using SP_3 as the semantic partition and Age* & ZIP Code* as key attributes.

X	Y	$SeDR_X(Y)$
3-diverse Table		
SP_3	2*	0.81
SP_3	≥ 40	1
SP_3	3*	0.81
SP_3	Age*	0.61
SP_3	355**	0.58
SP_3	3581*	1
SP_3	ZIP Code*	0.58
t-closeness Table		
SP_3	≤ 40	0.58
SP_3	≥ 40	1
SP_3	Age**	0.58
SP_3	3550*	1
SP_3	3581*	1
SP_3	3556*	1
SP_3	ZIP Code*	1

3. Table 5.6 shows that **an attacker is more able to refine the semantic partition SP_4 with some key attribute ZIP Code* values (3550* and 3556*)**. For these two values, the computed SeDR is higher for the t-closeness instantiation (0.69 vs 0.38).

5.6 Experiment

The goal of this experiment is to demonstrate over a realistic dataset, that depending on the chosen semantic partition, a t-closeness instantiation can be weaker than a l-diverse instantiation.

The data used in the experiment is the "Adult dataset" from the UC Irvine Machine Learning Repository which contains data collected from a US census. After removing missing values within the dataset, we obtain a total of 30161 valid records. We used six attributes of the dataset as depicted in Table 5.9 (4 key attributes and 2 sensitive attributes). We used the ARX tool (Prasser et al., 2014) (version 3.5.1) to compute the anonymization techniques (l-diversity and t-closeness) and the R tool (Chokkathukalam et al., 2013) (version 3.3.1) to compute our DR metric. Our measurements are computed according to the examples given in Section 5.5. The experiments are running on a i5-4300U CPU with 1.90GHz - 2.50GHz and 4GB memory.

The Figure in the experiment refers to values within Table 5.9 according to their corresponding numbers. For example, the number 1 in Figure 5.1 refers to the key attribute Age in Table 5.9.

We compare an l-diversity instantiation to a t-closeness instantiation of the dataset. The considered sensitive attribute in this case is attribute "Occupation". Using the ARX tool we generated 2 instantiations of the "Adult dataset" : a **10-diverse** instantiation and a **0.3-closeness** instantiation. We then compute our measurement both with and without considering the semantic partition.

TABLE 5.9: Attributes of the Adult dataset used in the experiment

	Attribute	Type	#Values
1	Age	key attribute	72
2	Education	key attribute	16
3	Race	key attribute	5
4	Sex	key attribute	2
5	Occupation	Sensitive	14
6	Salary	Sensitive	2

The semantic partition is computed as follows: attribute "occupation" has a total of 14 values:

- {Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-speciality, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces}

We considered the following semantic partition for our measurements that demontrates less resistance for the t-closeness instantiation:

- SP = {{Adm-clerical, Exec-managerial, Prof-specialty, Transport-moving, Other-service}, {Handlers-cleaners, Protective-serv, Tech-support}, {Craft-repair, Protective-serv, Sales}, {Machine-op-inspct, Armed-Forces, Priv-house-serv}}

Results interpretation

Let us first consider the graph on the left-hand side of Figure 5.1. This graph depicts the global identification capacity of attributes Age ("1"), Education ("2"), Race ("3") and Sex ("4") according to the sensitive attribute "Occupation". The measurements are computed both with and without the considered semantic partition (SP). We can first observe that for instantiations without the semantic partition (0.3-closeness and 10-diversity), the 10-diversity instantiation is globally weaker than the 0.3-closeness instantiation as the key attributes are more identifying for the 10-diversity instantiation. However, when we consider the instantiations with semantic partition (0.3-closeness-SP and 10-diversity-SP), the 0.3-closeness instantiation becomes globally weaker than the 10-diversity instantiation. This is especially true for attribute "4" (Sex) as we obtain: DR = 0.025 vs DR = 0.001.

The graph on the right-hand side (Figure 5.1) focuses on attribute Sex ("4") and depicts the identification capacity of its values (Female and Male) according to the sensitive attribute "Occupation". The graph faithfully translates what is depicted by the global measurement on attribute Sex (the graph on left-hand side). That is, the t-closeness instantiation with SP, is worse than the l-diverse instantiation with SP.

This experiment validates the conclusion of Section 5.5 that is: considering specific cases (specific semantic partitions) over the same data set, t-closeness can be worse than l-diversity.

5.7 Conclusion

Data publishing promises significant progress for emergence and improvement of new services. However, to mitigate privacy leakages due to poor anonymization

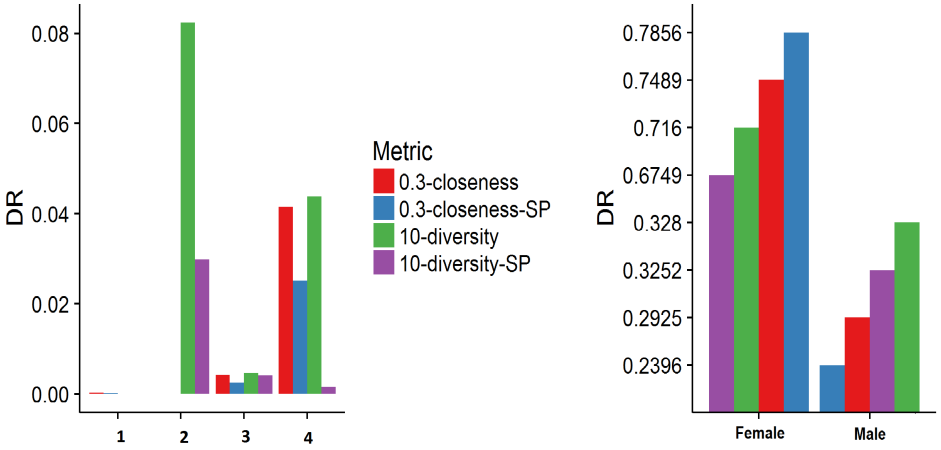

FIGURE 5.1: SeDR measurements for the Experiment

procedures, there is a strong need for publishers to have a practical and precise metric to assess the data anonymity level prior to publishing datasets. However, the anonymity level is reflected by the resistance capacity of anonymized data to disclosure attacks which are uniqueness and record linkage.

This chapter presents the Semantic Discrimination Rate, an improvement of the DR metric described in Chapter 4. The SeDR enables a more generic disclosure risk assessment for addressing both record linkage and uniqueness, and therefore the full range of approaches used to describe disclosure risk. The SeDR enables to tackle the de-anonymization issue from the attacker's perspective, by computing how much information is gained after applying a given disclosure attack, and all the existing attacks are evaluated. Illustration of that metric is given over some classical anonymization techniques (t-closeness and l-diversity), and proves that t-closeness is not as privacy protective as it was originally claimed to be as it can behave worse than l-diversity.

However, preventing disclosure of sensitive data is one of the two main goals of data anonymization as data anonymization aims to ensure both privacy and utility of anonymized data. In the next chapter, we propose a new method for assessing utility of anonymized data based on the SeDR metric.

Chapter 6

A Posteriori Utility Assessment of Sanitized Data with the Discrimination Rate Metric

6.1 Introduction

Chapters 4 and 5 describe our DR and SeDR metrics for assessing disclosure risk (uniqueness and record linkage) within data bases. This chapter provides a method for measuring utility within data bases based on the SeDR metric. We use an *a posteriori* approach (cf. Sections 6.4 and 6.5) i.e. an approach based only on a sanitized data set and a predefined utility formulation. We provide a formalization of the utility need which captures the versatility of utility need and we use our SeDR metric for assessing the utility degree.

With the adoption of the General Data Protection Regulation (the new data protection regulation in Europe), companies and the over all system built around big data will face a very big concern. Indeed, the current data utility assessment performed on raw data, will be performed on sanitized data instead. The main difficulty of this issue is that, sanitization destroys data by reducing their capability to re-identify a respondent and therefore, in most of cases, it reduces their capability to respond accurately to the utility needs. It is therefore necessary to find specific metrics for assessing utility based on sanitized data. In practice, utility assessment is based on a formulated need and evaluated on the basis of sanitized data only (a posteriori assessment); however, the current metrics aim to assess utility (cf. Section **??**) with respect to the capability of sanitized data to reflect the original data (a priori assessment). While this latter approach can at some extent capture the information loss, it does not target the utility need and may not respond to the data processor's requirement. However, the *a posteriori* approach is complex as it requires defining the utility need, which is very diverse and depends on the semantic, which is based on human interpretations. Moreover, even with a framework for capturing that utility need, one should also define a metric, flexible enough to fit any cases. Therefore, the definition of an *a posteriori* approach for utility assessment would require: (a) a framework for capturing the versatility of utility needs, (b) a metric that would be flexible enough to fit any possible needs.

In the literature, due to the complexity of the previous issues, the proposed utility metrics focus on the *a priori* approach and can be classified in two main categories: Privacy Preserving Data Publishing (PPDP) which is about information loss and Privacy Preserving Data Mining (PPDM) which is about specific uses metrics.

Information loss reflects how much data are destroyed between original and sanitized data, and their main goal is to guide anonymization processes for providing data for various data uses (Torra, 2017a) (Murdoch, 2014) (Rebollo-Monedero, Forne, and Domingo-Ferrer, 2010). On the other hand, specific uses metrics aim to provide sanitized data that can be used for specific uses and which include classification metrics (Bapna and Gangopadhyay, 2006) (Agrawal and Srikant, 2000) (Xue et al., 2017), regression based metrics (Muralidhar and Sarathy, 2005) (Regner and Riener, 2017) (Gomatam, Karr, and Sanil, 2005) and clustering (Batet et al., 2013) (Ni, Xie, and Qian, 2017). These latter metrics consider at some extent the utility need (Fung, Wang, and Philip, 2007).

However, concerning the specific uses *a priori* metrics, (Domingo-Ferrer, 2008b) criticized the approach asking : *"why not directly publish the statistics one wants to preserve rather than release a synthetic microdata set"*. A response to this critic, concerning *classification* (one of the widest used method for specific assessment) can be found in the following observation by Fung et al. (Fung, Wang, and Philip, 2007): "*knowing that the data is used for classification does not imply that the data provider knows exactly how the recipient may analyse the data. The recipient often has application-specific bias towards building the classifier. For example, some recipient prefers accuracy, whereas the others prefer interpretability*" and that "*In other cases, the recipient may not know exactly what to do before seeing the data, such as visual data mining, where the human makes decisions based on certain distributions of data records at each step*".

As a global observation, both metrics do not directly respond to the data processor's *a posteriori* need and further investigation is still necessary.

In this chapter, we propose a new approach for providing *a posteriori* utility assessment, which directly targets the data processor's need and applies on sanitized data obtained from any anonymization mechanism. This approach differs from the current approaches which focus on how to optimize the anonymization process in terms of information (*a priori* assessments). Our approach relies on two elements:

- the utility need expressed in terms of attributes of interest which the values are partitioned into partitions of interest
- the utility degree computed using the Semantic Discrimination Rate metric (SeDR) (Sondeck, Laurent, and Frey, 2017b) (cf. Sections 6.4 and 6.5) which is an improvement of the Discrimination Rate metric (Sondeck, Laurent, and Frey, 2017a) (DR), introduced by Sondeck et al. and which computes identification capability of attributes (scaled between 0 and 1) by measuring how they can refine an anonymity set; the maximum refinement leading to a single subject (e.g. an identifier has a DR of 1). The Semantic DR (SeDR) takes into account semantic considerations by computing the capability of attributes to refine subsets of subjects (*semantic partitions*) instead of single subjects. This enables measurements with respect to specific subsets of interest to reflect the need (semantic). Let us take a simple example for illustration. A data processor can be interested in identifying the respondents living in regions (sets of zip code locations) instead of specific zip code locations. Our approach will enable him to build a semantic partition of regions, instead of using single zip code locations, prior to measure the utility degree with the SeDR.

We show that the expressed utility need reflects the 2 aspects of semantic in psychology (Sánchez et al., 2012) which are *semantic similarity* and *semantic relatedness* and that the *a posteriori* utility need can be expressed thanks to *semantic partitions*. We are then able to:

1. **Assess utility while taking into account the versatility of utility needs**
2. **Measure utility accurately (down to specific subsets of values)**.
3. **Assess utility over sanitized data obtained from any anonymization mechanism**.

As a proof of concept, we provide a utility evaluation over the Adult dataset from the UCI machine learning repository.

This chapter is organized as follows. Section 6.2 gives some background on semantic for utility assessment. Section 6.3 underlines the frontier between privacy and utility and shows that the *attribute disclosure* attack, one of the well known privacy attacks, belongs to this frontier. Section 6.4 presents our informal definition of the *a posteriori* utility assessment. Section 6.5 presents the formal definition of the *a posteriori* utility assessment based on the observations of Section 6.4. Section 6.6 provides our experiment on real data. Finally, Section 6.7 gives our conclusion.

6.2 Semantic in Utility Assessment

In this section we introduce the semantic aspect of the utility need as utility is subjective and refers to semantic considerations (Abril, Navarro-Arribas, and Torra, 2010) (Erola et al., 2010).

Utility refers to semantic considerations. Semantic is defined in psychology (Goldstone, 1994) by how humans organize and classify objects and is depicted by two different paradigms (Sánchez et al., 2012): *semantic similarity* and *semantic relatedness*. *Semantic similarity* refers to how different objects are similar with respect to a taxonomy (e.g., a car and a motorcycle are similar as they are both automotive). On the other hand, *semantic relatedness* does not necessary rely on a taxonomy (e.g., antonymy, functionality, cause-effect). While *semantic similarity* is clearly defined (with respect to a taxonomy), relatedness is less clearly defined and depends on the use case. We believe that these two notions can be used to characterize semantic through the existing utility metrics.

In the literature these concepts are used both for information loss and specific uses metrics (Abril, Navarro-Arribas, and Torra, 2010) (Erola et al., 2010) (Batet et al., 2013) (Ni, Xie, and Qian, 2017). Usually, they refer to statistics computation such as: mean, variance, covariance matrices which reflect the *semantic relatedness*. For example, (Abril, Navarro-Arribas, and Torra, 2010) (Erola et al., 2010) measure semantic with respect to the microaggregation anonymization mechanism, which aim to group individual values into small aggregates, and within each aggregate, individual values are replaced by mean values. Utility refers then to how to classify values (which reflects the *semantic similarity*) for maximizing the within-aggregate homogeneity (Domingo-Ferrer, Sánchez, and Hajian, 2015a) i.e. the extent to which the mean value reflects each of the values taken into account within the mean computation (which reflects the *semantic relatedness*).

We propose our definitions of *semantic similarity* and *semantic relatedness* for capturing the *a posteriori* utility need over sanitized microdata in Sections 6.4 and 6.5.

TABLE 6.1: Data table example

Respondent	ZIP Code	Age	Salary	Disease
resp 1	35510	22	4K	cancer
resp 2	35510	35	5K	diabetes
resp 3	35510	63	3K	malaria
resp 4	35510	22	13K	cancer
resp 5	35510	22	8K	cancer
resp 6	35510	35	15K	malaria
resp 7	35510	45	9K	malaria
resp 8	35510	35	7K	diabetes
resp 9	35510	40	11K	diabetes

6.3 On the Frontier Between Utility and Privacy

Privacy and utility are considered as two faces of the same coin. Indeed, both work together and we can not act on one without acting on the other one; the more we have privacy, the less the data are useful and vice versa. As such, some attacks on privacy can also be considered as evaluation of utility as attacks on privacy reduce privacy and therefore enhance utility. As criteria for defining attacks on privacy are the most well established, let us consider the following privacy attacks which are commonly accepted (Domingo-Ferrer, Sánchez, and Hajian, 2015a):

- **attribute disclosure** consists in inferring information about attributes of an individual based on the sanitized data set. This usually occurs through the computation of correspondences between attributes within the sanitized data set (Machanavajjhala et al., 2007) (Li, Li, and Venkatasubramanian, 2007).
- **identity disclosure** consists in linking a record of a respondent within a sanitized data set to his identity. This is also called re-identification. Unlike attribute disclosure which does not imply disclosure of the identity of a respondent, identity disclosure implies disclosure of his identity.

Hence, while identity disclosure refers to complete re-identification and is clearly not acceptable, attribute disclosure is less strictly defined and belongs to the frontier as utility assessment also refers to computing the correspondence between attributes (cf. Section 6.4).

Most of the examples of utility assessment in this work can also be considered as an attribute disclosure assessment (cf. Section 6.4.3) and the distinction between privacy disclosure and utility can be defined through a threshold that can be fixed using our model (cf. Section 6.5).

6.4 Informal Definitions of the A Posteriori Utility and Illustrations

This section shows how the *a posteriori* utility can be expressed in terms of *semantic similarity* (based on *semantic partitions*) and *semantic relatedness* (based on the correlation degree of *semantic partitions*) (cf. Chapter 5). *A posteriori* needs do not consider original data and is only based on a sanitized data set and an expressed need.

Note that, unlike *a priori* utility assessments, the *a posteriori* utility assessment does not have the constraint for considering differently categorical and continuous data. The main reason for considering continuous and categorical data differently for the *a priori* approach is that they do not use the same operations for comparison (cf. Chapter 2). Indeed, the *a priori* approach aims to guarantee the similarity between sanitized and original data and therefore needs to compare them. For our measurements, we only consider *semantic partitions* with SeDR computations over sanitized data without considering original data.

Moreover, our method is completely independent of any anonymization mechanism and can apply to all the existing anonymization mechanism, all we need is a formulated utility need and a sanitized micro data set.

6.4.1 Informal Definitions

As the utility need is intended to be answered by a microdata set, let us first define what is a microdata set.

A *microdata* is a file generally depicted by a table where each row (record) contains individual's information splitted into different columns (attributes). A record refers to a single respondent and an attribute is an information shared by all the respondents within the microdata. For example in Table 6.2, there are 9 respondents and 4 attributes (ZIP Code, Age, Salary and Disease).

From this definition, we propose a characterization of the need as a function taking as input a set of attributes and returning a value between 0 and 1 reflecting the capability of the considered attributes to respond to the need. However, specific conditions should also apply on the attributes' values to reflect the *semantic similarity*. Considering a sanitized microdata, the formulation of the *a posteriori* utility need should therefore take into account:

1. **A set of attributes of interest** (for the data processor) as all attributes are not necessarily useful for responding to the need.
2. **A partition of the attributes' values into subsets of interest** (for the data processor) as a need is reflected by how the attributes' values are organized into sub domains. The sub domains of interest are used to build the *semantic partitions* (Section 6.5).

The term interest refers to the fact that it should be defined by the data processors and this reflects the subjectivity of the need. The first property (1) refers to *semantic relatedness* (cf. Section 6.2) as the attributes of interest are chosen according to specific considerations related to the data processor's need, however, the second property (2) refers to *semantic similarity* as attributes's values are partitioned according to the similarity of some of the subsets of values.

Finally the **quality of utility** is reflected by the **correlation degree** between subsets of values within partitions of different attributes of interest; which also refers to the *semantic relatedness*. We use the SeDR approach for computing this correlation degree.

6.4.2 Illustration (Global Recoding)

Let us consider the following example for illustration. Suppose we want to evaluate utility using the sanitized data Table 6.3, extracted from Table 6.1 (by considering

TABLE 6.2: Original data table (Global Recoding)

	ZIP Code	Age	Disease
1	35510	22	cancer
2	35602	35	diabetes
3	35712	63	malaria
4	35510	22	cancer
5	35510	22	cancer
6	35602	35	malaria
7	35715	45	malaria
8	35602	32	diabetes
9	35703	40	diabetes

only *ZIP Code*, *Age* and *Disease*). The protection mechanism is the global recoding (Domingo-Ferrer, Sánchez, and Hajian, 2015a) and we use the k-anonymity model (Samarati and Sweeney, 1998). We provide a 3-anonymity instantiation as clusters of 3 respondents are built by recoding *Age* into *Age** to prevent re-identification of respondents as shown in table 6.3.

TABLE 6.3: 3-anonymity Table (Disease) of Table 6.2.

	ZIP Code*	Age*	Disease
1	355**	2*	cancer
2	355**	2*	cancer
3	355**	2*	cancer
4	356**	≥ 40	malaria
5	356**	≥ 40	diabetes
6	356**	≥ 40	malaria
7	357**	3*	malaria
8	357**	3*	diabetes
9	357**	3*	diabetes

Suppose now the following scenario for utility measurement.

Consider a study which aims to provide respondent with treatment with respect to their age and which is based on the sanitized data in Table 6.3. Then a possible *utility need* would be the capacity to know from attribute *Age** the corresponding *Disease*. Therefore, the *attributes of interest* are *Age** and *Disease* and we can consider the predefined *semantic partition* (depicted by values 2*, ≥ 40 and 3* as they are defined by the anonymization mechanism) for assessing the utility need. Note that we can also define our own semantic partitions for computation (see Section 6.4.3).

Let us consider the predefined *semantic partition*, and assess the utility need. The utility need in this case refers to the correlation degree between the *semantic partition* of *Age** and each value of attribute *Disease*. We use the SeDR to compute the capability of *Age**'s values to refine the *Disease*'s values for computing this correlation degree. The result is depicted in Table 6.4.

We observe that the global capability (SeDR) to respond to the utility need i.e. to prescribe a treatment according to age is **0.6**. We can also observe that the value "2*" provides the highest utility (SeDR = **1**) as for respondents who have twenties, we

TABLE 6.4: Utility assessment within Table 6.3.

X	Y	$DR_X(Y)$
Disease	$2*$	1
Disease	≥ 40	0.8
Disease	$3*$	0.8
Disease	Age*	0.6

TABLE 6.5: Original data table (Microaggregation)

	ZIP Code	Age	Salary	Disease
1	35510	22	4K	cancer
2	35510	35	5K	diabetes
3	35510	63	3K	malaria
4	35620	22	13K	cancer
5	35620	22	8K	cancer
6	35620	35	15K	malaria
7	35740	45	9K	malaria
8	35740	32	7K	diabetes
9	35740	40	11K	diabetes

can prescribe without ambiguity the cancer treatment.

6.4.3 Illustration (Microaggregation)

Let us consider a second example for illustration. Suppose we want to evaluate utility using the sanitized data Table 6.6, obtained from the original data Table 6.5. The protection mechanism is microaggregation (Domingo-Ferrer, 2006) applied on attribute *Salary* by replacing some of its values by average measurements; the considered set of values are: {3K, 4K, 5K}, {7K, 8K, 9K}, {11K, 13K, 15K} and average values are computed as depicted in Table 6.6.

Let us now evaluate this sanitized microdata in terms of utility.

Suppose a study that aims to subsidize people with respect to their *Salary* based on their *Age* and *ZIP Code* considering different cases:

1. Whether their Salary belongs to the following intervals: $[0K, 5K[$, $[5K, 10K[$ and $[10K, 15K[$ (Table 6.6)
2. Whether their salary is lower or greater than $10K$ (Table 6.7).

Suppose also that attributes *Age* and *ZIP Code* are evaluated using the following considerations:

(a) Attribute *Age* only

(b) Combination of *ZIP Code* and *Age*

(c) Whether *Age* is lower or greater than 35

The attributes of interest are therefore *Age*, *Salary* and *ZIP Code* and the partitions of interest are defined by (1) and (2) (for *Salary*) and (a), (b) and (c) (for *Age* and *ZIP*

TABLE 6.6: Microaggregated data table of Table 6.5

	ZIP Code	Age	Salary*
1	35510	22	4K
2	35510	35	4K
3	35510	63	4K
4	35620	22	13K
5	35620	22	8K
6	35620	35	13K
7	35740	45	8K
8	35740	32	8K
9	35740	40	13K

TABLE 6.7: Microaggregated data table of Table 6.6 with application of the *semantic partition* (2) over attribute Salary

	ZIP Code	Age	Salary**
1	35510	22	$< 10K$
2	35510	35	$< 10K$
3	35510	63	$< 10K$
4	35620	22	$\geq 10K$
5	35620	22	$< 10K$
6	35620	35	$\geq 10K$
7	35740	45	$< 10K$
8	35740	32	$< 10K$
9	35740	40	$\geq 10K$

Code). The utility assessments would therefore consist in measuring how much from the specific considerations over *Age* and *ZIP Code* we can infer whether the *Salary*'s values belong to subsets: $[0K, 5K[$, $[5K, 10K[$ and $[10K, 15K[$, or are greater or lower than $10K$.

Using SeDR, the key attributes are *ZIP Code* and *Age* and the sensitive attribute is *Salary*. The *semantic partitions* for attribute *Salary* are depicted (1) and (2) and the *semantic partitions* of *ZIP Code* and *Age* are defined by (a), (b) and (c). Hence, for each case (1) or (2) we use the three *semantic partitions* (a), (b) and (c).

Let us now compute our SeDR measurements according to the considered attributes with the considered *semantic partitions*; the results are depicted in Table 6.8 (for the assessment with respect to (1)) and Table 6.9 (for the assessment with respect to (2)).

We can observe that the combination of attributes *ZIP Code* and *Age* in both cases (1) and (2) provide a greater utility (**SeDR = 0.85** for (1) and **SeDR = 0.75** for (2)) than assessments using *Age* only (**SeDR = 0.52** for (1) and **SeDR = 0.42** for (2)) as we are more able to infer the considered subsets of attribute *Salary*. We also observe that most of the values provide the maximum utility (**SeDR = 1**) and can therefore be used to directly grant the subsidies. For example, from the combination of *35510* and *35* we can directly know that the respondent earns between $0K$ and $5K$ (for the *semantic partition* (1)) or less than $10K$ (for the *semantic partition* (2)) and then grant the subsidies.

TABLE 6.8: Utility assessment within the Microaggregated Table 6.6 w.r.t. (1).

X	Y	$DR_X(Y)$
Salary*	22	0.66
Salary*	32	1
Salary*	35	0.85
Salary*	40	1
Salary*	45	1
Salary*	63	1
Salary*	Age	0.52
Salary*	35510/22	1
Salary*	35620/22	0.85
Salary*	35740/32	1
Salary*	35510/35	1
Salary*	35620/35	1
Salary*	35740/40	1
Salary*	35740/45	1
Salary*	35510/63	1
Salary*	ZIP Code/Age	0.85
Salary*	≥ 35	0.57
Salary*	< 35	0.46
Salary*	Age*	0.04

6.5 Formal Definition of the A Posteriori Utility Within a Microdata

From Section 6.4.3 we observe that the *a posteriori* approach requires two steps: (1) a posteriori utility need formulation by defining attributes of interest and *semantic partitions*, (2) computation of the *a posteriori* utility degree by using the SeDR.

6.5.1 A Posteriori Utility Need Formulation

Let us provide our formal definition of the utility need with respect to a microdata set.

Let us first formally define a microdata set. A microdata set Z referring to r respondents with s attributes is a $r \times s$ matrix where Z_{ij} is the value of attribute j ($1 \leq j \leq s$) for respondent i ($1 \leq i \leq r$). Attributes are considered as random variables possibly discrete or continuous.

Let us now define the utility need. As explained in Section 6.4, the utility need should consider:

1. **A set of attributes of interest**
2. **Specific partitions of values for each of the attributes of interest (referring to *semantic partitions*)**

We propose this definition for characterizing a *semantic partition*. Let X be an attribute within Z; and Ω_x its set of outcomes. Let $\mathcal{P}(\Omega_x)$ be the set of partitions of

TABLE 6.9: Utility assessment within the Microaggregated Table 6.7 w.r.t (2).

X	Y	$DR_X(Y)$
Salary**	22	0.66
Salary**	32	1
Salary**	35	0.75
Salary**	40	1
Salary**	45	1
Salary**	63	1
Salary**	Age	0.42
Salary**	35510/22	1
Salary**	35620/22	0.75
Salary**	35740/32	1
Salary**	35510/35	1
Salary**	35620/35	1
Salary**	35740/40	1
Salary**	35740/45	1
Salary**	35510/63	1
Salary**	ZIP Code/Age	0.75
Salary**	≥ 35	0.60
Salary**	< 35	0.41
Salary**	Age*	0.01

Ω_x and $p \in \mathcal{P}(\Omega_x)$, a specific *partition of interest*. We define the following function for characterizing the *semantic partition*:

$$S_p : X \rightarrow X' \tag{6.1}$$

In this definition X' refers to attribute X where some values have been grouped to form new subsets of values with respect to p. For example in Table 6.7 new subsets of attribute *Salary*'s values are formed ("$< 10K$" and "$\geq 10K$" with $p = \{\{4K, 8K\}, \{13K\}\}$) and utility assessment is performed with respect to this new partition.

From this definition we propose the following equation for characterizing the *a posteriori* utility within a microdata Z. Let $X_1, X_2, ..., X_k$ ($k \leq s$) be a set of attributes of interest; and $p_1, p_2, ..., p_k$ their corresponding partitions of interest, we characterize the *a posteriori* utility need as:

$$(S_{p_1}(X_1), S_{p_2}(X_2), ..., S_{p_k}(X_k)) = (X'_1, X'_2, ..., X'_k) \tag{6.2}$$

This equation characterizes the formulation of an *a posteriori* utility need which is depicted by the attributes of interest $(X_1, X_2, ..., X_k)$ and the partitions of interest with respect to each attribute of interest $(p_1, p_2, ..., p_k)$.

Note that the choice of the partitions of interest (semantic partitions) is made according to the data processor's need which is very subjective; our goal here is not to propose a specific way to identify these partitions of interest, but rather to propose a way for characterizing the need.

TABLE 6.10: Attributes of the Adult data set used in the experiment

	Attribute	Type	#Values
1	Age	key attribute	72
2	Education	key attribute	16
3	Native Country	key attribute	41
4	Race	key attribute	5
5	Salary-class	Sensitive	2

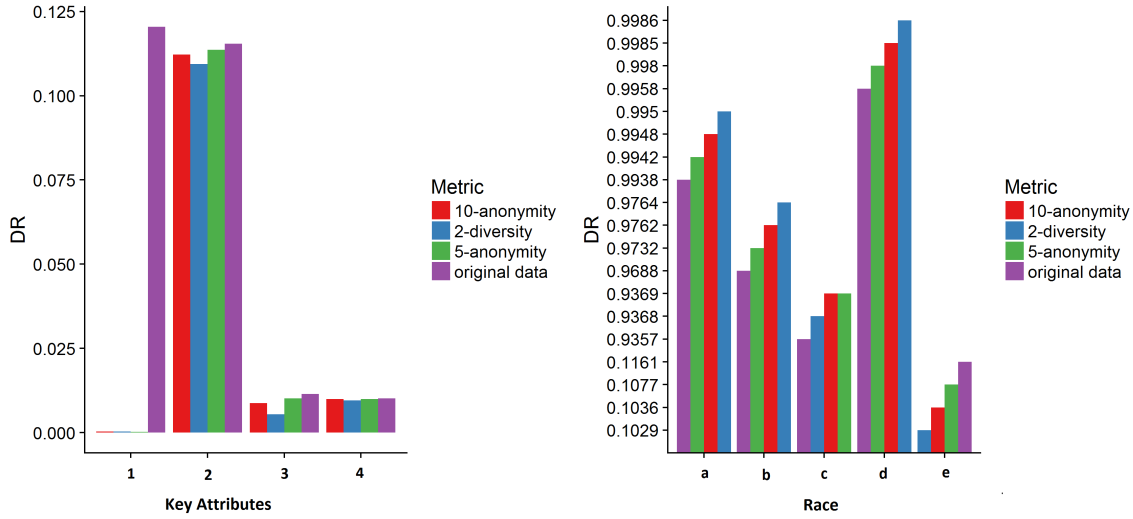

FIGURE 6.1: Utility assessment over sanitized data sets and comparison with original data

6.5.2 A Posteriori Utility Need Computation Using the SeDR

As the SeDR computes the capability of one or more key attributes to refine the semantic partitions of the values of a given sensitive attribute, we define the following function for computing the *a posteriori* utility need based on SeDR:

$$SeDR_{X_i} : (X'_1, X'_2, ..., X'_{i-1}, X'_{i+1}, ..., X'_k) \rightarrow [0, 1] \tag{6.3}$$

with $1 \leq i \leq k$. In this definition, X_i refers to the sensitive attribute which the set of outcomes is our anonymity set. Therefore, this function evaluates the capability of the other attributes (indexed by the set $\{1, 2, ..., k\} \setminus \{i\}$) to refine the set of outcomes of X_i.

For example, in Section 6.4.3, the 3 attributes of interest are "*ZIP Code*", "*Age*" and "*Salary*"; the key attributes are "*ZIP Code*" and "*Age*" and the sensitive attribute is "*Salary*". The *semantic partitions* for "*ZIP Code*" and "*Age*" are depicted by (a), (b) and (c) and the *semantic partitions* of "*Salary*" are depicted by (1) and (2). Then, the SeDR computes the utility degree which is depicted in Tables 6.8 and 6.9.

6.6 Experiment

The goal of this experiment is to assess the *a posteriori* utility over sanitized data sets. Our experiment is based on the well known Adult data set (more than 30.000

TABLE 6.11: Values of attribute "Race" used in the experiment

	Race	#Values
a	Amer-Indian-Eskimo	286
b	Asian-Pac-Islander	895
c	Black	2817
d	Other	231
e	White	25932

TABLE 6.12: Values of attribute "Salary-class" used in the experiment

Salary-class	#Values
$\leq 50K$	22653
$> 50K$	7508

records) (Martínez, Sánchez, and Valls, 2012) (Fung, Wang, and Yu, 2005) from the UC Irvine Machine Learning Repository which contains data from a US census. The goal of this census was to predict whether income exceeds $ 50K per year, based on different attributes including: Age, Education, Race and Native Country.

We assess the *a posteriori* utility according to this goal. We use the *global recoding* mechanism (cf. Section 6.4.2) for sanitization. We generate 2 instantiations of the k-anonymity and one instantiation of l-diversity (Machanavajjhala et al., 2007) over the Adult data set: **5-anonymity**, **10-anonymity** and **2-diversity**. We consider **4 key attributes (Age, Education, Race and Native Country)** and **1 sensitive attribute (Salary-class)** as depicted in Table 6.10. We use the ARX tool (Polonetsky, Tene, and Jerome, 2014) (version 3.5.1) to compute the anonymization techniques (k-anonymity and l-diversity) and the R tool (Chokkathukalam et al., 2013) (version 3.3.1) to compute our DR metric.

We also perform our assessment over the specific values of attribute Race (Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, White) as depicted in Table 6.11.

All the assessments have been performed based on the predefined *semantic partitions* of sanitized data i.e. the *semantic partitions* generated by the ARX tool during the sanitization process. For example, in Section 6.4.2, a predefined *semantic partition* is depicted by "2*", "≥ 40" and "3*"; which are subsets built for this instantiation. Using the ARX tool, the *semantic partitions* are generated automatically by the sanitization algorithm. For the two instantiations of k-anonymity (5-anonymity and 10-anonymity), only the key attributes have been partitioned and no semantic partition is applied on the sensitive attribute (Salary-class), as k-anonymity only acts on key attributes. For the l-diversity (Machanavajjhala et al., 2007) instantiation (2-diversity), the key attributes and the sensitive attribute (Salary-class) are partitioned in order to provide at least 2 different values of the sensitive attribute *Salary-class* for each category of key attributes.

The results are compared to assessments over the original data and are depicted in Figure 6.1.

Interpretation of results

Figure 6.1 depicts the capability of key attributes in Table 6.10 (the graph on the left hand side) and capability of the specific values of attribute Race in Table 6.11 (the graph on the right hand side) to predict the income of respondents within the sanitized Adult Data sets. The assessment is performed according to the different

instantiations (k-anonymity and l-diversity) and the original data. This enables to compare the *a posteriori* assessment with the assessment over original data.

Concerning the capability of the key attributes (the graph on the left hand side), we observe that attribute "2" (*Education* in Table 6.10) is the one which best enables to predict the income as the assessments provides the higher SeDR for any of the sanitized data sets. We also observe that the 2-diverse instantiation provides the weakest capability for predicting the income. Finally, Attribute "1" (*Age*) provides the weakest capability for the sanitized data sets whereas its capability is very high within the original data set. This is due to the chosen taxonomy for recoding attribute Age's values which is not very granular; indeed, the Age's values are splitted into two values: "< 80" and "≥ 80" in the sanitized data, rather than 72 values in the original data (cf. Table 6.10).

Concerning the capability of the specific values of attribute *Race* (the graph on the right hand side), we observe that the value "*d*" (*Other* in Table 6.11) provides the highest capability for predicting the income of respondents, and the value "*e*" (*White*) provides the weakest capability for prediction. The graph globally reflects the measurement over attribute "*4*" (*Race*) referring to the graph on the left hand side on Figure 6.1.

We can also observe that on the graph on the right hand side, some Race's values (*a*, *b* , *c* and *d*), provide less utility for the original data than for the anonymized data. This can be explained by the definition of utility provided in this paper (cf. Section 6.4), which states that, a utility assessment is performed according to a predefined utility need. Therefore, if the anonymization process is performed with respect to the predefined need, anonymization can improve both utility of data and anonymity of users. For example, let us consider Tables 6.2 and 6.3 which are an original data table and its 3-anonymous instantiation respectively. Consider now a study which aims to provide respondent with treatment with respect to their Zip Code* (Table 6.3) instead of Zip Code (Table 6.2), which is more specific. Therefore, for this use case, the anonymous table will provide more utility as it directly refers to the relevant data, and can be considered as a pre-processed instantiation of the original data. This principle also applies on the experiment and explains why on the graph on the right hand side of Figure 6.1, some anonymized data provide more utility than the original data.

This experiment provides a practical fine grained *a posteriori* utility assessment that can be used by data processors to validate the quality of sanitized data.

6.7 Conclusion

Utility is one of the two main goals of data sanitization as data sanitization aims to provide the best trade-off between privacy and utility. However, unlike privacy that can be characterized through a set of well known attacks, utility is very subjective and depends on the need of the data processor. In practice, utility is assessed on the basis of a sanitized data set and a predefined utility need (*a posteriori* assessment). However, due to the complexity of capturing the versatility of utility needs, the current metrics are concerned with assessing how much sanitized data do reflect original data (*a priori* assessment). This latter approach does not directly respond to the data processor's need and requires further investigations for complete assessment. This chapter proposes an *a posteriori* approach for utility assessment which directly

targets the data processor's need. Our model expresses the utility need through the partitioning of attribute values in order to define partitions of interest (*semantic partitions*), and computes the utility degree based on the Semantic Discrimination Rate, which is an information theoretic metric. The model enables to provide fine grained assessment of utility down to specific attributes' values. To the best of our knowledge this is the first work providing such accurate measurements of the *a posteriori* utility of sanitized data sets, which is a very important concern for companies tackling the newly adopted data protection regulation (General Data Protection Regulation) in Europe.

Part IV

Conclusion

Chapter 7

Conclusion and Perspectives

7.1 Conclusion

Data protection has never been as crucial as today; indeed, the new data protection regulation (GDPR) which will take effect on may 2018, will completely transform the way big data - recognized as the new oil of our era - will be processed. GDPR will drastically change how data are collected, stored, exploited, shared and deleted; with for the offenders, some fines up to 20 millions euro or 4 per cent of their previous year's global turnover (whichever is greater). Some technology groups even suggested GDPR could be one of the most expensive pieces of regulation in the internet sector's history (Ram, 2017), as it also implies great changes in terms of data management. However, while the regulation has already been adopted and will be effective after May 2018, some clarifications still remain to be made including:

- **A clear definition of what a personal data is**; indeed, the current regulation text defines personal data as data which are linked to an identifier. However, identifiers are not defined nor even characterized, although they have been recognized by the community as one of the most important concern (Schwartz and Solove, 2011).
- **Clear recommendations about data protection (anonymization)**; indeed, anonymization is one of the most used protecting mechanism for ensuring data privacy, but there is no clear recommendation about how it should be implemented, for example, how to measure the anonymization level, which level of anonymization should be considered for avoiding or enforcing fines.

On the other hand, beyond the fear of paying huge fines in case of privacy breaches (which are not clearly defined), companies also have concerns about:

- **Data utility**, companies are worried about whether their anonymized data will still be useful for their own services as anonymization destroys data by reducing their identification capability. This might reduce significantly the data utility.

This thesis addresses the three previously identified concerns by providing practical methods which enable to:

1. **Precisely define identifiers** and therefore personal data (Chapter 4).
2. **Provide accurate anonymization measures** for providing recommendations about data anonymization (Chapters 4 and 5).

3. **Provide accurate utility measures** for computing the data utility degree while taking into account the specific needs of data processors (Chapter 6).

Chapter 4 presents our first contribution, the Discrimination Rate metric which is an information theoretic metric that measures the identification capability of attributes by computing how much they can refine a set of subjects (an anonymity set). The measurements are scaled between 0 and 1. The DR therefore enables to characterize identifiers as they have a DR equal to 1 and to introduce new notions like *sketchy identifiers*, *partial identifiers* and *zero-identifiers*. This is a novel approach which is different from the existing ones as these latter are either empirical or have other limitations which include: lack of granularity, difficulty to link the measurements to the identification capability, use case specific measures and limited number of variables; the proposed DR is addressing all these limitations. The DR addresses disclosure risk assessment through the *uniqueness* principle (Domingo-Ferrer, Sánchez, and Hajian, 2015a) and enables accurate assessment and comparison of k-anonymity and l-diversity which are two of the most popular anonymization techniques. Thanks to its accuracy, the DR enables an evaluation down to specific attributes' values and to underline specific inconsistencies of the anonymization implementation. As the DR uses a generic approach for assessment, it can be used by the regulator to define wide scope rules for applying the regulation in different application domains.

Chapter 5 presents our second contribution, the Semantic Discrimination Rate metric (SeDR) which is an improvement of the DR metric. The SeDR adds the capability to refine subsets of subjects instead of single subjects through the concept of *semantic partitions* (specific partitions of the sensitive values). This last property provides a greater flexibility for addressing both *uniqueness* and *record linkage* which are the two approaches for computing disclosure risk (cf. Chapter 3). We then use the SeDR to assess and compare l-diversity and t-closeness and we are able to underline the limitation of t-closeness by showing it is not as protective as claimed by authors and that, depending on the case, t-closeness can be worse than l-diversity. Thanks to its improvements, the SeDR provides to the regulator and to companies a full stack metric that can be used to evaluate disclosure risk through its two main approaches which are *uniqueness* and *record linkage*.

Chapter 6 presents our third contribution which is about *a posteriori* utility assessment of anonymized data based on the SeDR metric. We provide utility assessment based only on a utility need and the anonymized data (*a poteriori* assessment) while the existing metrics provide measures based on original data and anonymized data (*a priori* assessment). This approach is more pragmatic as in practice utility assessment is based on a predefined need and a sanitized data set. We are able to formalize the utility need through the concept of *semantic partitions* (introduced with the SeDR) by considering specific partitions of attributes' values, and from this formalization, we provide our measurements with the SeDR metric. This approach applies to sanitized data obtained from any anonymization mechanism and measurements are computed down to specific values providing a very accurate assessment. This is, to the best of our knowledge, the first practical and accurate metric that can apply to sanitized data from any anonymization mechanism. This method for utility assessment can therefore be used by companies to evaluate their anonymized data quality to address their specific needs.

7.2 Perspectives

Several tracks are still conceivable for data assessment either for disclosure assessment or for utility assessment; among which:

- **Disclosure risk assessment of differential privacy anonymized data**: the DR can be used for assessing disclosure risk over differential privacy (DP) data. Indeed, as DP anonymized data do not necessarily contain the similar values than original data, semantic partitions can be used to capture similarity of values within anonymized and original data sets and assessment would be performed accordingly. The semantic partitions computation would require specific analysis as DP anonymized data may contain new values that may bias the assessment. This would be a specific disclosure risk assessment over data that do not contain similar values.

- **Micro data anonymization based on the DR metric**: as DR enables fine grained assessments both for disclosure risk and utility assessment, we can use it to provide well balanced anonymization. Indeed, the DR could assist the existing anonymization mechanisms to guarantee the best trade-off between privacy and utility. For example, some works have been proposed for ensuring this trade-off by composing different anonymization techniques (Soria-Comas et al., 2014); while the approach is interesting and can actually enhance utility, an accurate metric would enable better parametrization for better results. Note that, this approach directly refers to the *a priori* data assessment.

- **Anonymization of real time data**: The DR acts directly on data within a micro data table, and therefore acts on static data. However, many applications use real time data that are continuously recovered for providing services and need metric for assessing both utility and privacy over time (Wang et al., 2016). For example, location services use continuous data to improve lives of mobile devices' users through spatio temporal analysis. Another research field is continuous authentication based on behavioural biometrics which uses continuous data for authenticating people, based on different parameters (walking, swipe gestures...). The DR can apply on this type of data for providing continuous assessment which would evolve over time.

Glossary

a posteriori assessment An approach of assessment which aims to assess how much sanitized data can be used to answer a predefined utility need. 83

a priori assessment An approach of assessment which aims to assess how much sanitized data are similar to original data. 83

anonymization A process that aims to transform data in order to ensure both utility of data and privacy of respondents. 14

categorical attributes Attribute which take values over a finite set and over which arithmetical operations can not be performed. 16

Confidential/Sensitive attributes Attributes which contain sensitive information on the respondent. Examples are: salary, religion, health. 15

continuous attributes Attributes on which numerical and arithmetical operations can be performed. Examples: Age, Salary. 15

Data Disclosure Violation of privacy which occurs by linking different records belonging to the same respondent. 30

data processor Is the one who processes data for either protecting them or extracting information. 3

Discrimination Rate (DR) A novel approach for assessing both utility and privacy and which computes the identification capability of attributes with values scaled between 0 and 1. For example an identifier has a DR equals to 1. 51

Identifiable Natural Person Someone who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. 5

identifier An attribute that can be used to characterize a single respondent among others. Examples of such attributes are: social security numbers, names, fingerprints. 4

k-anonymity A data set is said to satisfy k-anonymity for an integer $k > 1$ if, for each combination of values of quasi-identifier attributes, at least k records exist in the data set sharing that combination. 30

microdata A table where each row (record) contains individual's information splitted into different columns (attributes). A record refers to a single respondent and an attribute is an information shared by all the respondents within the microdata. 13

Original Data Data on which no sanitization mechanism has been applied. 7

personal data Any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person. 3

pseudonymization The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person. 4

Quasi-Identifiers/Key Attributes Attributes which do not completely characterize a respondent but can be combined with others for complete characterization. Examples of such attributes are: zip code, age, gender... 15

Record Linkage A knowledge an attacker can have for linking different records belonging to the same respondent. 30

respondent Is the one to which the data records correspond. 3

Semantic Discrimination Rate (SeDR) An improvement of the DR which takes into account the similarity between values through semantic partitions. 71

semantic partition A partitioning of values of a given attribute reflecting the processor's need. 71

synthetic Data Data that do not contain original data but which are similar to original data with respect of some relevant statistics. 37

Appendix A

Résumé de la thèse en français (long)

A.1 Introduction

Il est difficile d'estimer tous les avantages que peuvent avoir les données personnelles aussi bien pour les fournisseurs de services, que pour les personnes à qui appartiennent ces données. De nos jours, les données personnelles sont utilisées dans presque tous les secteurs d'activité pour l'amélioration et le développement de nouveaux services, entre autres: l'analyse de la consommation, la réduction des coûts de transactions, l'accroissement de la rentabilité des publicités. Fort de ce constat, le commissaire européen à la protection des consommateurs compare en 2009, les données personnelles au pétrole (Spiekermann et al., 2015), pour illustrer sa capacité à créer de la valeur ajoutée pour les entreprises. Un rapport du Boston Consulting Group (Global, 2012), estime les bénéfices produits par les données personnelles à environ 1 milliard d'euro d'ici 2020.

D'autre part, les données personnelles peuvent représenter un fardeau important pour les entreprises, ceci lié au risque inhérent de violation de vie privée. En effet, les données peuvent révéler plus d'information aux fournisseurs de services que souhaité par le propriétaire des données (Domingo-Ferrer, 2007), ce qui mène à des violations de la vie privée pouvant porter atteinte à l'entreprise concernée sur plusieurs plans (réputation, amendes importantes...). Par exemple, en 2006, l'entreprise American Online's (AOL) a publié 20 million de requêtes web sous forme anonymisée à des fins de recherche; par la suite, les données ont été utilisées pour identifier Thelma Arnold, une veuve de 62 ans vivant à Lilburn aux Etats Unis (Barbaro, Zeller, and Hansell, 2006). Pour empêcher de telles violations de se reproduire, l'Europe a adopté une nouvelle loi dénommée Règlement Général sur la Protection des Données (RGPD) qui vise à protéger la vie privée des citoyens européens et qui prendra effet en Mai 2018. L'une des principales nouveautés de ce règlement est la pénalité: elle s'élève à 20 million d'euro ou 4% du chiffre d'affaire annuelle (le plus élevé étant retenu) pour les entreprises fautives.

Cependant, selon notre analyse, deux points essentiels nécessitent encore d'être clarifiés pour la mise en oeuvre du RGPD, comme spécifié dans la Section A.2.1.

A.2 Le manque de clarté du RGPD dans son implémentation

A.2.1 Le besoin de caractériser les identifiants, qui n'est pas clairement pris en compte par le RGPD

La principale difficulté de la gestion de la vie privée provient des textes de loi, et plus précisément de la difficulté à caractériser l'identifiant comme souligné par (Schwartz and Solove, 2011). En effet, le RDPD comme d'autres textes de loi précédent (Schwartz and Solove, 2011), ne fournit pas une définition claire de ce qu'est un identifiant, ce qui peut remettre en question la fiabilité des mesures de protection préconisées. En effet, dans son Article 32, le RGPD définit les moyens à mettre en oeuvre pour assurer la sécurité des traitements, qui peuvent être séparés en 2 groupes:

- Les méthodes permettant d'assurer et d'évaluer la vie privée: pseudonymisation et chiffrement.
- Les méthodes permettant d'assurer les autres aspects de la sécurité (confidentialité, intégrité, disponibilité, résilience...).

Ainsi, le RGPD propose la pseudonymisation comme un moyen de garantir la protection des données personnelles. Cependant, la *pseudonymisation* est définie dans l'Article 4 comme:

"le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable;"

En d'autres termes, la pseudonymisation consiste à transformer les données personnelles de telle sorte qu'elles ne puissent plus être associées à une personne unique.

Les données à caractère personnel sont définies comme:

"toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée "personne concernée"); est réputée être une "personne physique identifiable" une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale;"

Cependant, le terme identifiant n'est pas défini, cette définition fournit une énumération non exhaustive des informations pouvant être considérées comme identifiants. De plus, si on considère les cas considérés, pouvons nous affirmer sans risque de nous tromper, qu'ils s'agisse bien d'identifiants?

Pour illustrer la complexité de cette terminologie, considérons l'exemple suivant qui souligne l'importance de la prise en compte du contexte pour la définition des identifiants:

Supposons qu'il y'ait une voiture mal garée dans une rue et que nous souhaitions identifier le propriétaire de la voiture. Supposons que nous savons que le propriétaire, M. John est dans un bar dans la même rue. Supposons aussi que dans le même bar, il y'a plusieurs personnes du nom de John. Peut-on considérer dans ce cas que le nom est toujours un identifiant, étant donné qu'il ne fait plus référence à une seule personne. D'autre part, supposons que nous savons aussi que le propriétaire de la voiture porte une chemise blanche et que dans le bar il n'y a qu'une personne vêtue d'une chemise blanche. Ainsi, la chemise blanche devient un identifiant dans ce cas puisqu'il est suffisant pour identifier le propriétaire de la voiture.

Cette exemple montre la complexité de la terminologie des identifiants et l'importance de la prise en compte du contexte; ce qui n'est pas le cas dans le texte de loi actuel.

En effet, une définition claire des identifiants est une nécessité étant donné que *les données à caractère personnel* sont définies en fonction des *identifiants* et que la *pseudonymisation* (l'un des mécanismes de protection recommandé) a pour but de protéger les *données à caractère personnel*. Donc, sans une définition claire de ce qu'est un *identifiant*, les entreprises ne seraient pas en mesure de déterminer un *identifiant* et donc, ne seraient pas en mesure de protéger *les données à caractère personnel*.

A.2.2 Le besoin d'utiliser l'Anonymisation plutôt que la Pseudonymisation (Un point manquant du RGPD)

La *pseudonymisation* protège contre des liens directs entre les données et une personne concernée (cf. Section A.2.1), cependant, il a été prouvé que le reste des informations contenues dans les données pseudonymisées, peut encore être utilisé pour re-identifier un sujet (Hansell, 2006) (Barbaro, Zeller, and Hansell, 2006); la pseudonymisation est donc insuffisante pour protéger la vie privée. Dans la littérature, il existe un mécanisme plus générique permettant la protection des données personnelles qui est l'anonymisation. L'anonymisation consiste à transformer les données de telle sorte que les données ne puissent plus être liées à un sujet particulier, qu'elles ne contiennent plus aucune information concernant un utilisateur en particulier, et qu'elles conservent une certaine utilité pour les services. Cette définition inclut aussi bien l'incapacité de lier une identité à un individu, que l'incapacité d'acquérir de l'information sur un individu en particulier.

L'exemple suivant décrit la procédure d'anonymisation. Considérons les deux tables de données: Table A.1 (la table de données originales) et Table A.2 (la table correspondante de données anonymisées). Ceci représente une implémentation spécifique du modèle d'anonymisation "*k-anonymity*" (Samarati and Sweeney, 1998) (Sweeney, 2002).

Le principe de la "*k-anonymity*" est de transformer certaines valeurs d'attributs (attributs clés, ex: *Age*) dans le but de réduire leur capacité d'identification par rapport à un autre type d'attribut (attribut sensible, ex: *Disease*) en formant des sous ensembles de k enregistrements.

Le problème principal de l'anonymisation est de permettre le meilleur compromis entre l'utilité des données pour les services et la vie privée des sujets concernées (Li and Li, 2009) (Loukides and Shao, 2008a). En effet, alors que le RGPD recommande la confidentialité des données, les fournisseurs de services doivent utiliser les données pour améliorer leurs services, puisque l'anonymisation consiste à réduire

TABLE A.1: Original Data Table.

	Age	Disease
1	22	lung cancer
2	22	lung cancer
3	22	lung cancer
4	45	stomach cancer
5	63	diabetes
6	40	flu
7	35	aids
8	35	aids
9	32	diabetes

TABLE A.2: Anonymized Data Table.

	Age*	Disease
1	2*	lung cancer
2	2*	lung cancer
3	2*	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	3*	aids
8	3*	aids
9	3*	diabetes

les informations inhérentes aux attributs, il devrait y avoir un compromis entre ce qui est recommandé par les autorités et ce qui est nécessaire par les fournisseurs de services. L'anonymisation doit donc être évaluée avec précision afin de fournir le bon équilibre entre la confidentialité des données et l'utilité des données.

L'instanciation d'anonymisation précédente est donc destinée à répondre à la fois aux problèmes de vie privée et d'utilité des données. Les deux questions peuvent être formulées en termes de besoins. Nous donnons ici un exemple de besoins possibles:

- (**Vie privée**): quelle est la difficulté des asser des valeurs (*Age**) transformées aux valeurs brutes correspondantes (*Age*)?
- (**Utilité des données**): a partir des groupes d'ages définis, est-il possible d'identifier la maladie correspondante?

En effet, la violation de la vie privée provient de la capacité à ré-identifier un sujet concerné et donc dans notre exemple, de passer de l'attribut transformé *Age** à l'attribut brut *Age*; et mesurer la vie privée consiste à mesurer cette capacité.

D'autre part, l'utilité se réfère à la capacité d'extraire des informations utiles à partir des données. Dans ce cas, mesurer l'utilité consiste à mesurer comment, à partir de l'âge, nous pouvons deviner la maladie dont souffre un sujet.

Cet exemple montre à quel point l'anonymisation peut être complexe, puisqu'elle renvoie à des questions spécifiques qui dépendent du cas considéré.

Cependant, dans la littérature, bien qu'il existe de nombreuses propositions de mécanismes d'anonymisation ("k-anonymity" entre autres), il existe peu de mesures pratiques permettant de quantifier la vie privée et l'utilité des données. De plus, il n'existe pas d'approche uniforme pour la comparaison de différentes techniques/instanciations d'anonymisation. Cette dernière préoccupation étant importante pour la mise en oeuvre d'une réglementation cohérente.

A.3 Les questions pertinentes traitées dans cette thèse

Nous pouvons donc considérer les problèmes suivants:

1. **Comment pouvons-nous caractériser les identifiants par rapport au contexte?** Les identifiants sont au centre de la régulation des données personnelles car les données personnelles sont définies en fonction des identifiants; cependant, la réglementation actuelle ne considère pas le contexte dans sa définition et un manque de caractérisation appropriée peut conduire à des malentendus.
2. **Quelle approche doit-on prendre en compte pour l'évaluation de la vie privée?** Il existe de nombreuses approches pour la mesure de la vie privée, mais peu sont pratiques. De plus, pour définir des règles de régulation cohérentes, une approche uniforme est nécessaire.
3. **Comment pouvons-nous définir une échelle pour comparer divers mécanismes d'anonymisation?** De nombreux mécanismes existent pour protéger la vie privée, mais il n'y a pas d'échelle uniforme pouvant être utilisée pour les comparer.
4. **Comment pouvons-nous mesurer l'utilité en termes de besoins spécifiques?** Comme l'utilité renvoie à une utilisation spécifique des données et donc à une question spécifique, une évaluation précise consisterait à mesurer des données par rapport à un besoin. Ce qui est difficile à mettre en oeuvre à cause de la subjectivité du besoin.

A.4 Contributions de cette thèse

Dans cette thèse, nous proposons une métrique appelée DR (Discrimination Rate) qui permet:

- **Objectif 1:** Une caractérisation fine des identifiants par rapport au contexte.
- **Objectif 2:** Une mesure du degré d'anonymat en termes de capacités d'attributs, permettant une granularité fine et donc utilisable pour différents domaines d'application.
- **Objectif 3:** Une évaluation et une comparaison précises des techniques d'anonymisation existantes en termes de vie privée.
- **Objectif 4:** Une mesure précise de l'utilité en termes de besoins d'utilité spécifiques exprimés.

Le reste de la thèse est organisé en 3 parties:

- **L'etat de l'art**: qui se décline en deux parties: les techniques d'anonymisation et les métriques d'anonymisation.
- **Notre contribution**: il s'agit essentiellement de la métrique DR, qui se décline en plusieurs versions permettant d'évaluer la vie privée sous ses différents aspects (mesure de la re-identification, évaluation du niveau de connaissance qu'on peut avoir sur un individu en particulier) et l'utilité des données en fonction d'un besoin exprimé. Plus précisément il s'agit de 3 contributions:
  1. **Le Discrimination Rate: une métrique centrée sur les attributs pour mesurer la vie privée (Objectifs 1, 2 et 3)**: le Discrimination Rate (DR) est une nouvelle métrique qui fournit une approche centrée sur les attributs pour la mesure de la vie privée et qui est pratique et suffisamment flexible pour s'adapter à divers domaines d'application. Le DR calcule la capacité d'un attribut (évaluée entre 0 et 1) à raffiner un ensemble de sujets; plus un attribut peut affiner un ensemble de sujets, plus son DR est élevé. Par exemple, un identifiant a un DR égal à 1 car il permet d'isoler chacun des sujets de l'ensemble. Grâce au DR, nous fournissons une première évaluation précise ainsi qu'une comparaison de deux des techniques d'anonymisation les plus utilisées, à savoir le k-anonymat et la l-diversité. Ce travail a été publié dans le Journal Annals of Telecommunications, 2017 (Sondeck, Laurent, and Frey, 2017c).
  2. **Le Semantic Discrimination Rate (Objectif 3)**. le Semantic Discrimination Rate (taux de discrimination sémantique) (SeDR), est une amélioration du DR qui prend en compte des considérations sémantiques. Le SeDR permet plus de flexibilité pour ses mesures d'anonymat et est utilisé pour comparer la l-diversité à la la t-proximité qui sont deux des meilleures techniques d'anonymisation de type k-anonymat. De plus, comme la t-proximité est considérée meilleure que la l-diversité, le SeDR montre que, selon les considérations sémantiques, la proximité t-proximité peut être pire que la l-diversité. Ce travail a été publié dans la conférence Security and Cryptography (SECRYPT) en 2017 (Sondeck, Laurent, and Frey, 2017b).
  3. **Evaluation d'utilité a posteriori de données anonymisées avec la mesure Discrimination Rate (Objectif 4)** Après avoir utilisé notre métrique (SeDR) pour la mesure de l'anonymat, nous montrons comment il peut être utilisé pour fournir une évaluation de l'utilité *a posteriori* précise pour tout type de données anonymisées. L'évaluation a posteriori est l'approche la plus pratique car elle est réalisée uniquement à partir des données anonymisées et d'un besoin prédéfini d'utilité alors que l'évaluation de type *a priori* vise à évaluer dans quelle mesure les données désinfectées reflètent les données originales et donc basée sur des données originales et les données anonymisées (qui ne sont pas accessibles).
- **La conclusion et les perspectives**.

A.5 Etat de l'art

A.5.1 Les techniques d'anonymisation

Il existe un grand nombre de mécanismes pour l'anonymisation des données dans les bases de données statistiques, et différentes façons de considérer une base de

données statistiques. Nous analysons ici un type particulier de bases de données appelé *microdonnées* et évaluons les capacités des techniques d'anonymisation en fonction de leur qualité déterministe ou non à répondre au problème de l'anonymisation qui est: le compromis entre vie privée et utilité des données.

Définitions

Une microdonnée: est un fichier généralement représenté par une table où chaque ligne (enregistrement) contient des informations individuelles divisées en différentes colonnes (attributs). Un enregistrement fait référence à un seul sujet et un attribut est une information partagée par tous les sujets au sein de la microdonnée. Par exemple, Table A.1 est une microdonnée avec 2 attributs (*Age* et *Disease*) et 9 enregistrements.

Les attributs dans une microdonnée peuvent être de trois catégories qui ne sont pas nécessairement disjointes:

- **Identifiants**: attributs qui peuvent être utilisés pour caractériser un seul sujet parmi d'autres. Des exemples de tels attributs sont: les numéros de sécurité sociale, les noms, les empreintes digitales.
- **Quasi-identifiants/Attributs clés**: attributs qui ne caractérisent pas complètement un sujet mais peuvent être combinés avec d'autres pour une caractérisation complète. Des exemples de ces attributs sont: code postal, âge, sexe.
- **Attributs confidentiels/sensibles**: attributs qui contiennent des informations sensibles sur le sujet. Les exemples sont: le salaire, la religion, la santé.

Comme précisé dans l'introduction, la définition actuelle de l'identifiant ne prend pas en compte le contexte et nous proposons une définition plus précise comme contribution.

Les mécanismes d'anonymisation déterministes

Il s'agit de mécanismes qui ne prennent pas en compte de génération aléatoires ni l'ajouts de données synthétiques. Ces mécanismes incluent entre autres: la généralisation et suppression (Hundepool et al., 2005), la microaggregation (Domingo-Ferrer and Torra, 2005), (Torra, 2004), la suppression locale (Hundepool et al., 2008). Ces techniques ont l'avantage de permettre un meilleur contrôle sur le processus d'anonymisation et permettent donc un meilleur calibrage des données en fonction du besoin en utilité.

Les mécanismes d'anonymisation non-déterministes

Il s'agit de mécanismes basés sur la génération aléatoire et sur l'ajout de données synthétiques. Ces mécanismes incluent entre autres: la méthode "Post-RAndomization" (PRAM) (Gouweleeuw, Kooiman, and De Wolf, 1998), (Kooiman, Willenborg, and Gouweleeuw, 1997), la génération de données synthétiques (Dwork, 2008), le "swapping" des données (Dalenius and Reiss, 1982) (Reiss, 1984) (Carlson and Salabasis, 2000). Le principal avantage de ces techniques est de fournir une meilleure résistance aux attaques sur la re-identification. Cependant, à cause de l'aléatoire et de l'ajout de données synthétiques dont elles dépendent, elles permettent un contrôle plus faible sur l'utilité des données.

A.5.2 Les métriques d'anonymisation

L'anonymisation a un double objectif (vie privée et utilité des données), ainsi, pour l'évaluer il existe deux grands types de métriques: les métriques de vie privée et les métriques d'utilité des données. Nous analysons ici les métriques existantes dans leur capacité à évaluer d'une part la vie privée et d'autre part l'utilité des données. Nous proposons ensuite une comparaison des métriques de vie privée en fonction de plusieurs critères que nous trouvons pertinents.

Les métriques de vie privée

Dans la littérature, il existe plusieurs métriques d'évaluation de la vie privée. Elles évaluent le degré de vie privée dans un jeu de données anonymisées en mesurant sa capacité à résister aux attaques connues sur la vie privée. Pour ce faire plusieurs propriétés peuvent être considérées:

1. **Lien avec la ré-identification**: Le lien entre les mesures et la capacité de ré-identification est-il direct ou non?
2. **Empirique ou analytique**: les mesures sont-elles empiriques ou analytiques?
3. **Granularité**: Est-il possible d'effectuer des mesures sur plusieurs attributs, en fonction des valeurs d'attributs, d'une combinaison de valeurs d'attributs ...?
4. **Généralité**. La métrique peut-elle être utilisée avec différents mécanismes d'anonymisation? La métrique prend-elle en compte différents types d'attributs? La métrique peut-elle être utilisée pour lier des enregistrements dans des microdonnées qui ne contiennent pas des valeurs d'attributs identiques ou similaires?
5. **Applicabilité et évolutivité**. La métrique est-elle applicable sur de grands ensembles de données?

Nous proposons une comparaison des métriques en fonction de ces critères et montrons qu'aucune métrique existante ne répond à tous ces critères.

Les métriques d'utilité des données

La principale difficulté pour l'évaluation de l'utilité des données est la subjectivité du besoin d'utilité. En effet, le besoin dépend du cas d'utilisation et varie en fonction de l'interprétation du problème. Dans la littérature il existe 2 principales approches pour évaluer l'utilité des données: les métriques pour des besoins spécifiques et le métriques pour des besoins génériques.

Les métriques pour les besoins spécifiques évaluent la capacité des données à répondre à un besoin prédéfinit et utilisent principalement des techniques provenant du domaine de l'intelligence artificielle (Torra, 2017a) (classification, regression, clustering).

Les métriques d'utilité génériques essaient de maximiser la quantité d'information restante dans un jeu de données anonymisées afin de maximiser l'utilisation des données pour différents usages non identifiés à l'avance. Cette méthode utilise des techniques de statistique incluant: l'erreur quadratique moyenne (MSE), l'erreur quadratique absolu (MAE), la variation moyenne (Domingo-Ferrer, Sánchez, and Hajian, 2015a).

Notez que ces métriques utilisent principalement une approche dites *a priori* pour l'évaluation des données, ce qui signifie que les mesures sont effectuées afin de fournir des données anonymes qui reflètent dans une certaine mesure les données d'origine. Une approche *a posteriori* consisterait à faire des mesures sur des données anonymisées par rapport à un besoin d'utilité donné. Cependant, cette dernière approche est plus complexe car elle nécessite un cadre qui capterait le besoin d'utilité, qui est subjectif. Nous proposons comme contribution un cadre pour effectuer une évaluation dites *a posteriori* (en fonction des données anonymisées et d'un besoin d'utilité exprimé).

A.6 Contribution

Notre contribution s'articule autour de la métrique Discrimination Rate (taux de discrimination) pour répondre aux objectifs présentés en introduction (Secction A.3). Le DR se décline 3 version le Simple DR, le Combine DR pour la contribution 1 (Objectifs 1, 2 et 3) et le Semantic DR pour la contribution 2 (Objectif 3). Le SeDR est complété ensuite pour remplir l'objectif 4.

A.6.1 Le Discrimination Rate: une métrique centrée sur les attributs pour mesurer la vie privée (Objectifs 1, 2 et 3)

Cette première contribution vise à répondre aux objectifs 1 2 et 3 (en partie). Ainsi, nous proposons: (1) une définition précise de la notion d'identifiant, (2) une approche générique pour l'évaluation de la vie privée et (3) une évaluation et comparaison précises du k-anonymat et de la l-diversité qui compte parmi les méthodes les plus utilisées pour anonymiser les données.

Le Discrimination Rate

Le but de notre métrique est de calculer la capacité d'identification d'un attribut dans un ensemble de sujets donné, plus précisément de calculer la quantité d'informations d'identification contenue dans un attribut par rapport à cet ensemble de sujets. A des fins de généralité, nous considérons les attributs comme des variables aléatoires et l'ensemble des sujets est défini comme l'ensemble des résultats (valeurs et occurrences) d'une autre variable aléatoire. Pour clarifier notre idée, considérons 2 variables aléatoires X et Y; Y l'attribut dont nous souhaitons mesurer la capacité d'identification et X, les attributs dont l'ensemble des valeurs est notre ensemble d'anonymat. Comme souligné dans (Shin et al., 2012a), l'une des principales préoccupations concernant les mesures d'anonymat basées sur l'entropie est la variable aléatoire considérée. Dans notre cas, nous voulons calculer la quantité d'information portée par une variable aléatoire par rapport à la capacité de raffinement de l'ensemble des résultats d'une autre variable aléatoire. Pour cela, nous considérons $H(X)$ la quantité d'information (incertitude) portée par X comme notre état initial. Nous calculons ensuite l'entropie de X conditionnée sur Y ($H(X|Y)$) car nous souhaitons mesurer l'effet de Y sur X. Cette quantité représente l'incertitude restante à l'intérieur de X, après que Y soit divulgué. Afin de calculer la quantité d'information portée par Y par rapport à X, nous devons soustraire cette quantité de $H(X)$ et ainsi nous obtenons $H(X) - H(X|Y)$ qui est la quantité effective d'information d'identification portée par l'attribut Y par rapport à l'ensemble des

TABLE A.3: Example of data table

Subjects	ZIP Code	Age	Salary	Disease
subject 1	35000	22	4K	cancer
subject 2	35000	35	5K	diabetes
subject 3	35000	63	3K	malaria
subject 4	35000	22	13K	cancer
subject 5	35000	22	8K	cancer
subject 6	35000	35	15K	malaria
subject 7	35000	45	9K	malaria
subject 8	35000	35	7K	diabetes
subject 9	35000	40	11K	diabetes

sujets. Finalement, nous divisons cette quantité par $H(X)$ pour normaliser le résultat.

Nous proposons donc la définition formelle suivante:

Definition 17 ***(Le Simple Discrimination Rate)***
Soit ***X*** *et* ***Y*** *deux variables aléatoires. Le* ***simple Discrimination Rate*** *de* ***Y*** *par rapport à* ***X*** *est la capacité de* ***Y*** *à raffiner l'ensemble des sorties de* ***X*** *et est calculé comme suit:*

$$DR_X(Y) = \frac{H(X) - H(X|Y)}{H(X)} = 1 - \frac{H(X|Y)}{H(X)} \tag{A.1}$$

Dans ce qui suit, **X** est appelé **Attribut sensible** et **Y** est le **Attribut clé**.

Exemple

Considérons l'exemple suivant qui illustre un calcul de DR.

Considérons la table de données A.3 et évaluons la capacité de chacun des attributs à identifier un sujet dans la table.

Supposons que nous souhaitions évaluer la capacité de Age a raffiner l'ensemble de sujets dans la table. L'attribut sensible serait donc **X** = "Subjects" et l'attribut clé **Y** = "Age". Les calculs s'effectuent donc comme suit:

$$\begin{aligned} SDR_X(Y) &= 1 - \frac{H(X|Y)}{H(X)} \\ &= 1 - \frac{-1/3\log_2(1/3) - 1/3\log_2(1/3)}{-\sum_{s=1}^{9} 1/9\log_2(1/9)} \\ &= 1 - \frac{1/3\log_2(3) + 1/3\log_2(3)}{\log_2(9)} \\ &= 0.66 \end{aligned}$$

Pour $H(X|Y)$, la distribution est calculée selon la définition de l'entropie conditionnelle: l'attribut *Age* peut prendre 5 valeurs 22, 35, 40, 45, 63. Cela permet de réduire l'ensemble principal à des sous-ensembles de 3, 3, 1, 1 et 1 sujet(s) respectivement, correspondant à 1/3, 1/3, 1/9, 1/9 et 1/9 de l'ensemble respectivement. Les entropies conditionnelles sont respectivement: $H(X|Y = 22) = -log_2(1/3)$, $H(X|Y = 35) = -log_2(1/3)$, $H(X|Y = 40) = 0$, $H(X|Y = 45) = 0$ et $H(X|Y = 63) = 0$. $H(X|Y)$ est donc la somme de $-1/3\log_2(1/3)$ et $-1/3\log_2(1/3)$.

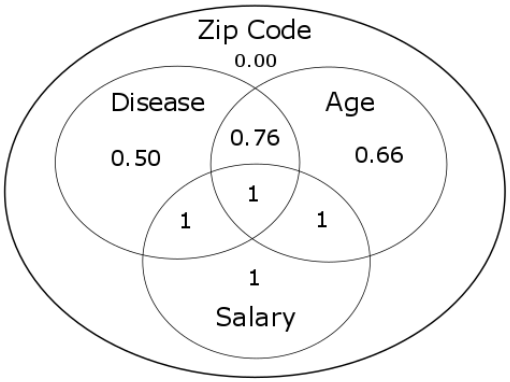

FIGURE A.1: Le Discrimination Rate dans la Table A.3

Nous proposons par la suite le Combined DR qui prend en entrée plusieurs attributs clés comme suit:

Definition 18 ***(Le Combined Discrimination Rate)***
Soit $\boldsymbol{X}$ *et* $Y_1, ..., Y_n$ *des variables aléatoires. Le* ***Combined Discrimination Rate*** *de* $Y_1, ..., Y_n$ *par rapport à* $\boldsymbol{X}$ *est la capacité de* $Y_1, ..., Y_n$ *à raffiner l'ensemble des sorties de* $\boldsymbol{X}$ *et est calculé comme suit:*

$$DR_X(Y_1, ..., Y_n) = 1 - \frac{H(X|Y_1, ..., Y_n)}{H(X)}. \tag{A.2}$$

En reprenant l'exemple précédent, nous évaluons la capacité d'une combinaison d'attributs pour identifier un individu dans la table. Le résultat est illustré par la figure A.1.

Le DR permet de satisfaire les objectifs de la thèse mentionnés à la Section A.4, comme précisé ci-dessous.
Objectif 1

A partir du CDR nous proposons les définitions suivantes pour les identifiants:

Definition 19 ***Identifiant***
Soit X *un attribut sensible et* $Y_1, Y_2, ...Y_n$ *un ensemble d'attributs clés et* $n \in \mathbb{N} \setminus \{0\}$. $(Y_1, Y_2, ...Y_n)$ *est un* ***Identifiant*** *relativement à* X *si, et seulement si:*
$DR_X(Y_1, Y_2, ..., Y_n) = 1$.

Definition 20 ***Sketchy-Identifiant***
Soit X *un attribut sensible et* $Y_1, Y_2, ...Y_n$ *un ensemble d'attributs clés,* $n \in \mathbb{N} \setminus \{0\}$. $(Y_1, Y_2, ...Y_n)$ *est un* ***Sketchy-Identifiant*** *relativement à* X *si, et seulement si:*
$DR_X(Y_1, Y_2, ..., Y_n) \in]0, 1[$.

Definition 21 ***Zero-Identifiant***
Soit X *un attribut sensible et* $Y_1, Y_2, ...Y_n$ *un ensemble d'attributs clés,* $n \in \mathbb{N} \setminus \{0\}$. $(Y_1, Y_2, ...Y_n)$ *est un* ***Zero-Identifiant*** *relativement à* X *si et seulement si:*
$DR_X(Y_1, Y_2, ..., Y_n) = 0$.

Definition 22 ***Identifiant Partiel***
Soit X un attribut sensible et $Y_1, Y_2, ...Y_n$ un ensemble d'attributs clés, $n \in \mathbb{N} \setminus \{0\}$ et $\mathcal{Y}_1, \mathcal{Y}_2, ..., \mathcal{Y}_n$ les ensembles de valeurs possibles de $Y_1, Y_2, ...Y_n$. $(Y_1, Y_2, .., Y_n)$ est un ***Identifiant Partiel*** *par rapport à X, si et seulement si $(Y_1, Y_2, .., Y_n)$ est un* ***Sketchy-identifiant*** *relativement à* ***X*** *et si $\exists (y_1, y_2, ..., y_n) \in (\mathcal{Y}_1, \mathcal{Y}_2, ..., \mathcal{Y}_n)$ /*
$DR_X(y_1, y_2, ..., y_n) = 1$.

Objectif 2

Le DR étant basé sur une définition générique de l'anonymat prenant en compte des variables partagées par tout problème de vie privée: un ensemble de sujets, un attribut sensible, des attributs clés; il fournit une approche générique pour l'évaluation de la vie privée.

Objectif 3

Le DR permet l'évaluation et la comparaison des techniques d'anonymisation (aussi bien différentes instanciations d'un même mécanisme que différents mécanismes). Nous introduisons une approche nouvelle d'évaluation qui consiste à évaluer à quel point un jeu de données est résistant à une attaque donnée.

Nous comparons ainsi différentes instanciations de k-anonymat et une instanciation de k-anonymat avec plusieurs instanciations de l-diversité.

Notre évaluation permet de souligner certaines incohérences avec ce qui est admis théoriquement. Nous montrons par exemple que pour certaines valeurs, la l-diversité peut être pire que le k-anonymat.

Cependant, le DR peut encore être amélioré pour prendre en compte la proximité de certaines valeurs sensibles pour exprimer leur similarité sémantique. En effet, les valeurs sensibles peuvent être regroupées en sous-ensembles pour exprimer le sens, et il peut être intéressant de distinguer ces sous-ensembles particuliers de valeurs plutôt que des valeurs uniques; par exemple pour obtenir l'information qu'une personne souffre d'un cancer, quel que soit le type de cancer. La contribution 2 (Section A.6.2) décrit le taux de discrimination sémantique (SEDR) comme une amélioration du DR; il fournit une évaluation des risques de vie privée plus générique pour améliorer le couplage des enregistrements.

A.6.2 Le Semantic Discrimination Rate (Objectif 3)

Le Semantic Discrimination Rate (SeDR) est une amélioration du DR pour permettre des mesures de vie privée qui tiennent compte d'un besoin donné. Grâce au SeDR nous remettons en question la t-proximité et montrons qu'elle n'est pas meilleure que la l-diversité pourtant reconnue dans la litérature comme étant meilleure. En effet, selon le cas la t-proximité peut être pire que la l-diversité.

Le SeDR considère des partitions spécifiques de valeurs sensibles et effectue des mesures en fonction de ces partitions spécifiques. Par exemple dans la table A.4, nous pouvons considérer la partition suivante de l'attribut "Disease": *{{diabetes, flu,*

TABLE A.4: Original Data Table (Salary/Disease).

	ZIP Code	Age	Salary	Disease
1	35567	22	4K	colon cancer
2	35502	22	5K	stomach cancer
3	35560	22	6K	lung cancer
4	35817	45	7K	stomach cancer
5	35810	63	12K	diabetes
6	35812	40	9K	aids
7	35502	35	8K	aids
8	35568	35	10K	flu
9	35505	32	11K	lung cancer

TABLE A.5: Semantic DR in Table 5.3.

X	Y	$DR_X(Y)$
Disease	22	1
Disease	32	1
Disease	35	1
Disease	40	1
Disease	45	1
Disease	63	1
Disease	Age	1

aids}, {colon cancer, lung cancer, stomach cancer}}. Ainsi, nous pouvons évaluer la capacité de *Age* à raffiner cette partition spécifique au lieu de raffiner chacune des maladies prise séparément.

En appliquant le SeDR à la table A.4 en ayant pour attribut clé l'attribut "Age" et pour attribut sensible l'attribut "Disease" partitionné comme précédemment, on obtient les résultats dans la table A.5.

Nous pouvons constater que suivant cette partition spécifique de "Disease", le SeDR est capable d'extraire le maximum d'information (SeDR = 1) car chacune des valeurs de "Age" correspond parfaitement à chacun des sous ensembles de la partition.

Objectif 3

Cette nouvelle propriété permet d'effectuer des mesures selon le sens accordé à certaines valeurs de l'attribut sensible. Nous sommes donc en mesure, d'évaluer la vie privée de manière plus précise et de prendre en compte un besoin spécifique qui permet de souligner les insuffisances de la t-proximité.

A.6.3 Evaluation d'utilité a posteriori de données anonymisées avec la mesure Discrimination Rate (Objectif 4)

Les métriques d'utilité existantes effectuent pour la plupart des mesures dites à *priori*, qui consistent à mesurer à quelle point les données anonymisées reflètent les données d'origine, elles s'appuient donc sur des données d'origine et des données anonymisées.

En effet, à cause de la subjectivité de la notion d'utilité, il est difficile de trouver un modèle permettant d'exprimer le besoin d'utilité dans sa diversité, encore plus

TABLE A.6: Original data table (Global Recoding)

	ZIP Code	Age	Disease
1	35510	22	cancer
2	35602	35	diabetes
3	35712	63	malaria
4	35510	22	cancer
5	35510	22	cancer
6	35602	35	malaria
7	35715	45	malaria
8	35602	32	diabetes
9	35703	40	diabetes

difficile, de trouver une métrique suffisamment flexible pour s'adapter à tous les besoins d'utilité. Dans cette contribution nous nous attelons à répondre a ces deux préoccupations. Nous proposons une nouvelle approche basée sur la métrique SeDR pour évaluer l'utilité a postériori s'appuyant uniquement sur un besoin d'utilité et un jeu de données anonymisée. Nous sommes en mesure de: (1) évaluer l'utilité tout en tenant compte de la polyvalence des besoins d'utilité; (2) mesurer l'utilité avec précision (en fonction de sous-ensembles de valeurs spécifiques); (3) mesurer l'utilité de n'importe quel type de données anonymisées.

Pour ce faire nous proposons un framework permettant d'exprimer une besoin d'utilité et, à partir de ce besoin exprimé, nous évaluons le degré d'utilité des données avec le SeDR. L'évaluation se fait donc en 2 étapes:

1. Expression d'un besoin dans notre modèle
2. Evaluation du besoin d'utilité à partir du SeDR

Exemple d'évaluation d'un besoin d'utilité a posteriori

Nous proposons un modèle permettant de capturer un besoin d'utilité a posteriori dans une microdonnée (cf. Section A.5.1). Pour ce faire nous considérons 2 critères:

- Un ensemble d'attributs d'intérêt
- Pour chaque attribut d'intérêt, une partition d'intérêt

Ces deux critères nous permettent d'exprimer notre besoin d'utilité qui est par la suite évalué avec le SeDR.

Considérons l'exemple suivant pour qui illustre l'évaluation de l'utilité à posteriori où la table A.7 est un 3-anonymat de la table A.6.

Considérons maintenant une étude qui vise à fournir un traitement aux sujets en fonction de leur âge et qui serait basée sur les données anonymisées de la table A.7. Ainsi, le *besoin de l'utilité* serait de savoir à partir de l'attribut *Age** l'attribut *Disease* correspondant. Par conséquent, les *attributs d'intérêt* sont *Age** et *Disease* et nous pouvons considérer comme partition d'intérêt la partition *prédéfinie* (représentée par les valeurs 2*, ≥ 40 et 3* tels qu'ils sont générés par le mécanisme d'anonymisation) pour évaluer le besoin d'utilité. Notez que nous pouvons aussi définir nos propres partitions sémantiques pour le calcul (voir Section A.6.2).

Considérant la partition prédéfinie nous pouvons évaluer le besoin de d'utilité. Le besoin d'utilité dans ce cas fait référence au degré de corrélation entre la *partition*

TABLE A.7: 3-anonymity Table (Disease) of Table A.6.

	ZIP Code	Age*	Disease
1	355**	2*	cancer
2	355**	2*	cancer
3	355**	2*	cancer
4	356**	≥ 40	malaria
5	356**	≥ 40	diabetes
6	356**	≥ 40	malaria
7	357**	3*	malaria
8	357**	3*	diabetes
9	357**	3*	diabetes

TABLE A.8: Utility assessment within Table A.7.

X	Y	$DR_X(Y)$
Disease	$2*$	1
Disease	≥ 40	0.8
Disease	$3*$	0.8
Disease	Age	0.6

d'intérêt de *Age** et chaque valeur de l'attribut *Disease*. Nous utilisons le SeDR pour calculer la capacité des valeurs de *Age** à affiner les valeurs de *Disease* pour calculer ce degré de corrélation. Le résultat est représenté dans la table A.8.

Nous observons que la capacité globale (SeDR) à répondre à l'utilité, c'est-à-dire de prescrire un traitement selon l'âge est **0.6**. Nous pouvons également observer que la valeur "2*" fournit l'utilité la plus élevée (SeDR = **1**) comme pour les sujets qui ont une vingtaine d'années, nous pouvons prescrire sans ambiguïté le traitement du cancer.

Ceci illustre notre méthode d'évaluation que nous formalisons par la suite;

Cette méthode s'applique a toutes les mécanismes d'anonymisation car il suffit d'avoir la microdonnée, de choisir les attributs d'intérêt et les partitions d'intérêt.

A.7 Conclusion et perspectives

La protection des données n'a jamais été aussi cruciale qu'aujourd'hui. En effet, le nouveau règlement sur la protection des données (RGPD), qui entrera en vigueur en mai 2018, transformera complètement la manière dont les données de masse - reconnues comme la nouveau pétrole de notre époque - seront traitées. Le RGPD modifiera radicalement la façon dont les données sont collectées, stockées, exploitées, partagées et supprimées; avec pour les contrevenants, des amendes allant jusqu'à 20 millions d'euros, ou 4% du chiffre d'affaires global de l'année précédente (le plus élevé des deux étant retenu). Certains groupes de travail considèrent le RGPD comme l'un des règlements les plus coûteux de l'histoire de la réglementation numérique (Ram, 2017), car il implique également de grands changements en termes de gestion des données.

Cependant, alors que le règlement a déjà été adopté et entrera en vigueur en mai 2018, certains points restent à éclaircir, notamment:

- **Une définition claire de ce qu'est une donnée personnelle**; En effet, le texte actuel de la réglementation définit les données personnelles comme des données liées à un identifiant. Cependant, les identifiants ne sont ni définis ni même caractérisés, bien qu'ils aient été reconnus par la communauté comme l'un des sujets les plus importants (Schwartz and Solove, 2011).
- **Des recommandations claires sur la protection des données (anonymisation)**; En effet, l'anonymisation est l'une des méthodes de protection des données personnelles les plus utilisées, mais il n'y a pas de recommandation claire quant à sa mise en œuvre, par exemple comment mesurer le niveau d'anonymisation, quel niveau d'anonymisation faut-il envisager pour éviter des amendes?

D'autre part, au-delà de la crainte de payer d'énormes amendes en cas d'atteintes à la vie privée (qui ne sont pas clairement définies), les entreprises s'inquiètent également pour:

- **L'utilité de données**, les entreprises s'inquiètent de savoir si leurs données anonymisées seront toujours utiles pour leurs propres services, car l'anonymisation détruit les données en réduisant leur capacité d'identification. Cela pourrait réduire considérablement l'utilité des données.

Cette thèse aborde les trois préoccupations précédemment identifiées en fournissant des méthodes pratiques qui permettent de:

1. **Définir précisément les identifiants** et donc les données personnelles (Section A.6.1).
2. **Fournir des mesures d'anonymisation précises** permettant des recommandations claires quant à l'anonymisation des données (Sections A.6.1 et A.6.2).
3. **Fournir des mesures d'utilité précises** pour calculer le degré d'utilité des données tout en prenant en compte les besoins spécifiques des processeurs de données (Section A.6.3).

Bibliography

Abadi, Martín et al. (2016). "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, pp. 308–318.

Abril, Daniel, Guillermo Navarro-Arribas, and Vicenç Torra (2010). "Towards semantic microaggregation of categorical data for confidential documents". In: *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, pp. 266–276.

Agrawal, Rakesh and Ramakrishnan Srikant (2000). "Privacy-preserving data mining". In: *ACM Sigmod Record*. Vol. 29. 2. ACM, pp. 439–450.

Bapna, Sanjay and Aryya Gangopadhyay (2006). "A Wavelet-Based Approach to Preserve Privacy for Classification Mining". In: *Decision Sciences* 37.4, pp. 623–642.

Barbaro, Michael, Tom Zeller, and Saul Hansell (2006). "A face is exposed for AOL searcher no. 4417749". In: *New York Times* 9.2008, 8For.

Batet, Montserrat et al. (2013). "Utility preserving query log anonymization via semantic microaggregation". In: *Information Sciences* 242, pp. 49–63.

Bayardo, Roberto J and Rakesh Agrawal (2005). "Data privacy through optimal k-anonymization". In: *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*. IEEE, pp. 217–228.

Bernardo, JM et al. (2003). "Assessing the risk of disclosure of confidential categorical data". In: *Bayesian statistics 7: Proceedings of the seventh Valencia international meeting*. Oxford University Press, USA, p. 125.

Bindschaedler, Vincent, Reza Shokri, and Carl A Gunter (2017). "Plausible deniability for privacy-preserving data synthesis". In: *Proceedings of the VLDB Endowment* 10.5, pp. 481–492.

Brand, Ruth (2002). "Microdata protection through noise addition". In: *Inference control in statistical databases* 2316, pp. 97–116.

Brickell, Justin and Vitaly Shmatikov (2008). "The cost of privacy: destruction of data-mining utility in anonymized data publishing". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 70–78.

Carlson, Michael and Mickael Salabasis (2000). *A Data-swapping Technique Using Ranks: A Method for Disclosure Control*. Univ., Department of Statistics.

Chen, Rui et al. (2011). "Publishing set-valued data via differential privacy". In: *Proceedings of the VLDB Endowment* 4.11, pp. 1087–1098.

Chin, Francis Y and Gultekin Ozsoyoglu (1982). "Auditing and inference control in statistical databases". In: *IEEE Transactions on Software Engineering* 6, pp. 574–582.

Chokkathukalam, Achuthanunni et al. (2013). "mzMatch–ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data". In: *Bioinformatics* 29.2, pp. 281–283.

Ciriani, Valentina et al. (2007). "Microdata protection". In: *Secure data management in decentralized systems*. Springer, pp. 291–321.

Clifton, Chris and Tamir Tassa (2013). "On syntactic anonymity and differential privacy". In: *Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference on*. IEEE, pp. 88–93.

Dalenius, Tore (1974). "The invasion of privacy problem and statistics productionan overview". In: *Statistik Tidskrift* 12, pp. 213–225.

Dalenius, Tore and Steven P Reiss (1982). "Data-swapping: A technique for disclosure control". In: *Journal of statistical planning and inference* 6.1, pp. 73–85.

Dandekar, Ramesh A and Lawrence H Cox (2002). "Synthetic tabular data-an alternative to complementary cell suppression". In: *Manuscript, Energy Information Administration, US Department of Energy*.

Defays, D and Ph Nanopoulos (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.

Denning, DE (1982). "Inference controls". In: *Cryptography and data security*, pp. 331–392.

Denning, Dorothy E, Peter J Denning, and Mayer D Schwartz (1979). "The tracker: A threat to statistical database security". In: *ACM Transactions on Database Systems (TODS)* 4.1, pp. 76–96.

Diaz, Claudia, Carmela Troncoso, and George Danezis (2007). "Does additional information always reduce anonymity?" In: *Proceedings of the 2007 ACM workshop on Privacy in electronic society*. ACM, pp. 72–75.

Domingo-Ferrer, Josep (2006). "Microaggregation: achieving k-anonymity with quasi-optimal data quality". In: *European Conference on Quality in Survey Statistics*.

— (2007). "A three-dimensional conceptual framework for database privacy". In: *Workshop on Secure Data Management*. Springer, pp. 193–202.

— (2008a). "A survey of inference control methods for privacy-preserving data mining". In: *Privacy-preserving data mining*. Springer, pp. 53–80.

— (2008b). "A survey of inference control methods for privacy-preserving data mining". In: *Privacy-preserving data mining*. Springer, pp. 53–80.

Domingo-Ferrer, Josep, Josep M Mateo-Sanz, and Vicenç Torra (2001). "Comparing SDC methods for microdata on the basis of information loss and disclosure risk". In: *Pre-proceedings of ETK-NTTS*. Vol. 2, pp. 807–826.

Domingo-Ferrer, Josep and Josep Maria Mateo-Sanz (2002). "Practical data-oriented microaggregation for statistical disclosure control". In: *IEEE Transactions on Knowledge and data Engineering* 14.1, pp. 189–201.

Domingo-Ferrer, Josep, David Sánchez, and Sara Hajian (2015a). "Database Privacy". In: *Privacy in a Digital, Networked World*. Springer, pp. 9–35.

— (2015b). "Database Privacy". In: *Privacy in a Digital, Networked World*. Springer, pp. 9–35.

Domingo-Ferrer, Josep, Francesc Sebé, and Agusti Solanas (2008). "A polynomial-time approximation to optimal multivariate microaggregation". In: *Computers & Mathematics with Applications* 55.4, pp. 714–732.

Domingo-Ferrer, Josep and Vicenc Torra (2001a). "A quantitative comparison of disclosure control methods for microdata". In: *Confidentiality, disclosure and data access: theory and practical applications for statistical agencies*, pp. 111–134.

Domingo-Ferrer, Josep and Vicenc Torra (2001b). "Disclosure control methods and information loss for microdata". In: *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies*, pp. 91–110.

Domingo-Ferrer, Josep and Vicenç Torra (2002). "Distance-based and probabilistic record linkage for re-identification of records with categorical variables". In: *Butlletí de lACIA, Associació Catalana dIntelligència Artificial*, pp. 243–250.

— (2003). "Disclosure risk assessment in statistical microdata protection via advanced record linkage". In: *Statistics and Computing* 13.4, pp. 343–354.
— (2005). "Ordinal, continuous and heterogeneous k-anonymity through microaggregation". In: *Data Mining and Knowledge Discovery* 11.2, pp. 195–212.
Domingo-Ferrer, Josep and Vicenç Torra (2008). "A critique of k-anonymity and some of its enhancements". In: *Availability, Reliability and Security, 2008. ARES 08. Third International Conference on*. IEEE, pp. 990–993.
Dwork, Cynthia (2008). "Differential privacy: A survey of results". In: *International Conference on Theory and Applications of Models of Computation*. Springer, pp. 1–19.
— (2011). "Differential privacy". In: *Encyclopedia of Cryptography and Security*. Springer, pp. 338–340.
Dwork, Cynthia et al. (2006). "Calibrating noise to sensitivity in private data analysis". In: *Theory of Cryptography Conference*. Springer, pp. 265–284.
Edwards, Anthony WF and L Luka Cavalli-Sforza (1965). "A method for cluster analysis". In: *Biometrics*, pp. 362–375.
Elamir, Elsayed AH and Chris Skinner (2006). "Record level measures of disclosure risk for survey microdata". In: *Journal of Official Statistics* 22.3, p. 525.
Elliot, Mark J, Anna M Manning, and Rupert W Ford (2002). "A computational algorithm for handling the special uniques problem". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05, pp. 493–509.
Elliot, MJ, Chris J Skinner, and A Dale (1998). "Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk". In: *Research in Official Statistics* 1, pp. 53–67.
Erola, Arnau et al. (2010). "Semantic microaggregation for the anonymization of query logs". In: *International Conference on Privacy in Statistical Databases*. Springer, pp. 127–137.
EU (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. URL: http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016R0679.
Feldman, Dan et al. (2009). "Private coresets". In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, pp. 361–370.
Fellegi, Ivan P and Alan B Sunter (1969). "A theory for record linkage". In: *Journal of the American Statistical Association* 64.328, pp. 1183–1210.
Fischetti, Matteo and Juan José Salazar (2000). "Complementary cell suppression for statistical disclosure control in tabular data with linear constraints". In:
Fung, Benjamin CM, Ke Wang, and S Yu Philip (2007). "Anonymizing classification data for privacy preservation". In: *IEEE transactions on knowledge and data engineering* 19.5.
Fung, Benjamin CM, Ke Wang, and Philip S Yu (2005). "Top-down specialization for information and privacy preservation". In: *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*. IEEE, pp. 205–216.
Ganta, Srivatsava Ranjit, Shiva Prasad Kasiviswanathan, and Adam Smith (2008). "Composition attacks and auxiliary information in data privacy". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 265–273.
Global, Liberty (2012). *The value of our digital identity. boston consulting group*.
Goldstone, Robert L (1994). "Similarity, interactive activation, and mapping." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20.1, p. 3.

Gomatam, Shanti, Alan F Karr, and Ashish P Sanil (2005). "Data swapping as a decision problem". In: *Journal of Official Statistics* 21.4, p. 635.

Gordon, AD and JT Henderson (1977). "An algorithm for Euclidean sum of squares classification". In: *Biometrics*, pp. 355–362.

Gouweleeuw, J M, Peter Kooiman, and PP De Wolf (1998). "Post randomisation for statistical disclosure control: Theory and implementation". In: *Journal of official Statistics* 14.4, p. 463.

Hansell, Saul (2006). "AOL removes search data on vast group of web users". In: *New York Times* 8, p. C4.

Hansen, Pierre, Brigitte Jaumard, and Nenad Mladenovic (1998). "Minimum sum of squares clustering in a low dimensional space". In: *Journal of Classification* 15.1, pp. 37–55.

Hansen, Stephen Lee and Sumitra Mukherjee (2003). "A polynomial algorithm for optimal univariate microaggregation". In: *IEEE Transactions on Knowledge and Data Engineering* 15.4, pp. 1043–1044.

Hardt, Moritz, Katrina Ligett, and Frank McSherry (2012). "A simple and practical algorithm for differentially private data release". In: *Advances in Neural Information Processing Systems*, pp. 2339–2347.

Hundepool, A et al. (2005). "μ-ARGUS version 4.0 Software and Users Manual". In: *Statistics Netherlands, Voorburg NL*.

Hundepool, Anco et al. (2008). "Josep Domingo and Vicenc Torra (Numerical micro aggregation and rank swapping) Ruth Brand and Sarah Giessing (Sullivan Masking)". In:

Hundepool, Anco et al. (2012). *Statistical disclosure control*. John Wiley & Sons.

Jaro, Matthew A (1989). "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida". In: *Journal of the American Statistical Association* 84.406, pp. 414–420.

Karr, Alan F et al. (2006). "A framework for evaluating the utility of data altered to protect confidentiality". In: *The American Statistician* 60.3, pp. 224–232.

Kim, Jay J (1986). "A method for limiting disclosure in microdata based on random noise and transformation". In: *Proceedings of the section on survey research methods*. American Statistical Association, pp. 303–308.

Kolmogorov, A (1956). "On the Shannon theory of information transmission in the case of continuous signals". In: *IRE Transactions on Information Theory* 4.2, pp. 102–108.

Kooiman, Peter, Leon Cornelis Roelof Johannes Willenborg, and Josephine Margaretha Gouweleeuw (1997). *PRAM: A method for disclosure limitation of microdata*. CBS.

Laszlo, Michael and Sumitra Mukherjee (2005). "Minimum spanning tree partitioning algorithm for microaggregation". In: *IEEE Transactions on Knowledge and Data Engineering* 17.7, pp. 902–911.

Lee, Jaewoo and Chris Clifton (2011). "How much is enough? choosing ε for differential privacy". In: *International Conference on Information Security*. Springer, pp. 325–340.

LeFevre, Kristen, David J DeWitt, and Raghu Ramakrishnan (2005). "Incognito: Efficient full-domain k-anonymity". In: *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, pp. 49–60.

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t-closeness: Privacy beyond k-anonymity and l-diversity". In: *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, pp. 106–115.

Li, Ninghui, Wahbeh H Qardaji, and Dong Su (2011). "Provably private data anonymization: Or, k-anonymity meets differential privacy". In: *Arxiv preprint*.

Li, Tiancheng and Ninghui Li (2009). "On the tradeoff between privacy and utility in data publishing". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 517–526.

Loukides, Grigorios and Jianhua Shao (2008a). "Data utility and privacy protection trade-off in k-anonymisation". In: *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*. ACM, pp. 36–45.

— (2008b). "Data utility and privacy protection trade-off in k-anonymisation". In: *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*. ACM, pp. 36–45.

Machanavajjhala, Ashwin et al. (2007). "l-diversity: Privacy beyond k-anonymity". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1, p. 3.

Makhdoumi, Ali and Nadia Fawaz (2013). "Privacy-utility tradeoff under statistical uncertainty". In: *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*. IEEE, pp. 1627–1634.

Martínez, Sergio, David Sánchez, and Aida Valls (2012). "Semantic adaptive microaggregation of categorical microdata". In: *Computers & Security* 31.5, pp. 653–672.

Muralidhar, Krishnamurty and Rathindra Sarathy (2005). "An enhanced data perturbation approach for small data sets". In: *Decision Sciences* 36.3, pp. 513–529.

Murdoch, Steven J (2014). "Quantifying and measuring anonymity". In: *Data Privacy Management and Autonomous Spontaneous Security*. Springer, pp. 3–13.

Neumann, Dean A and Victor T Norton (1986). "Clustering and isolation in the consensus problem for partitions". In: *Journal of classification* 3.2, pp. 281–297.

Ni, Sang, Mengbo Xie, and Quan Qian (2017). "Clustering Based K-anonymity Algorithm for Privacy Preservation." In: *IJ Network Security* 19.6, pp. 1062–1071.

Nin, Jordi and Vicenç Torra (2009). "Analysis of the univariate microaggregation disclosure risk". In: *New generation computing* 27.3, pp. 197–214.

Pagliuca, D and G Seri (1999). "Some results of individual ranking method on the system of enterprise accounts annual survey". In: *Esprit SDC Project, Deliverable MI-3 D* 2, p. 1999.

Pfitzmann, Andreas and Marit Hansen (2010). *A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management*.

Polonetsky, Jules, Omer Tene, and Joseph Jerome (2014). "Benefit-Risk Analysis for Big Data Projects". In: *Future of Privacy Forum*.

Prasser, Fabian et al. (2014). "Arx-a comprehensive tool for anonymizing biomedical data". In: *AMIA Annual Symposium Proceedings*. Vol. 2014. American Medical Informatics Association, p. 984.

Raghunathan, Trivellore E, Jerome P Reiter, and Donald B Rubin (2003). "Multiple imputation for statistical disclosure limitation". In: *Journal of official statistics* 19.1, p. 1.

Ram, Aliya (2017). "Tech sector struggles to prepare for new EU data protection laws". In: *Finacial Times*. URL: https://www.ft.com/content/5365c1fa-8369-11e7-94e2-c5b903247afd.

Rebollo-Monedero, David, Jordi Forne, and Josep Domingo-Ferrer (2010). "From t-closeness-like privacy to postrandomization via information theory". In: *Knowledge and Data Engineering, IEEE Transactions on* 22.11, pp. 1623–1636.

Regner, Tobias and Gerhard Riener (2017). "Privacy Is Precious: On the Attempt to Lift Anonymity on the Internet to Increase Revenue". In: *Journal of Economics & Management Strategy* 26.2, pp. 318–336.

Reiss, Steven P (1984). "Practical data-swapping: The first steps". In: *ACM Transactions on Database Systems (TODS)* 9.1, pp. 20–37.

Reiter, Jerome P (2005). "Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168.1, pp. 185–205.

Rodriguez-Carrion, Alicia et al. (2015). "Entropy-based privacy against profiling of user mobility". In: *Entropy* 17.6, pp. 3913–3946.

Rubin, Donald B (1993). "Discussion statistical disclosure limitation". In: *Journal of official Statistics* 9.2, p. 461.

Sakuma, Jun and Tatsuya Osame (2017). "Recommendation with k-anonymized Ratings". In: *arXiv preprint arXiv:1707.03334*.

Samarati, Pierangela (2001). "Protecting respondents identities in microdata release". In: *Knowledge and Data Engineering, IEEE Transactions on* 13.6, pp. 1010–1027.

Samarati, Pierangela and Latanya Sweeney (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Tech. rep. Technical report, SRI International.

Sánchez, David et al. (2012). "Ontology-based semantic similarity: A new feature-based approach". In: *Expert Systems with Applications* 39.9, pp. 7718–7728.

Sankar, Lalitha, S Raj Rajagopalan, and H Vincent Poor (2013). "Utility-privacy tradeoffs in databases: An information-theoretic approach". In: *IEEE Transactions on Information Forensics and Security* 8.6, pp. 838–852.

Sarathy, Rathindra and Krishnamurty Muralidhar (2011). "Evaluating Laplace noise addition to satisfy differential privacy for numeric data." In: *Trans. Data Privacy* 4.1, pp. 1–17.

Schlörer, Jan et al. (1975). "Identification and retrieval of personal records from a statistical data bank." In: *Methods Archive* 14, pp. 7–13.

Schwartz, Paul M and Daniel J Solove (2011). "The PII problem: Privacy and a new concept of personally identifiable information". In: *NYUL rev.* 86, p. 1814.

Shin, Kang G et al. (2012a). "Privacy protection for users of location-based services". In: *IEEE Wireless Communications* 19.1.

— (2012b). "Privacy protection for users of location-based services". In: *Wireless Communications, IEEE* 19.1, pp. 30–39.

Singh, A, F Yu, and G Dunteman (2003). "MASSC: A new data mask for limiting statistical information loss and disclosure". In: *Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality*, pp. 373–394.

Singh, Amardeep, Divya Bansal, and Sanjeev Sofat (2014). "Privacy Preserving Techniques in Social Networks Data Publishing-A Review". In: *International Journal of Computer Applications* 87.15.

Skinner, Chris, Catherine Marsh, and Colin Wymer (1994). "Disclosure control for census microdata". In: *Journal of Official Statistics* 10.1, p. 31.

Sondeck, Louis Philippe, Maryline Laurent, and Vincent Frey (2017a). "Discrimination rate: an attribute-centric metric to measure privacy". In: *Annals of Telecommunications*, pp. 1–12.

— (2017b). "The Semantic Discrimination Rate Metric for Privacy Measurements which Questions the Benefit of t-closeness over l-diversity". In: *Proceedings of the 14th International Joint Conference on e-Business and Telecommunications - Volume 6: SECRYPT, (ICETE 2017)*. INSTICC. SciTePress, pp. 285–294. ISBN: 978-989-758-259-2. DOI: 10.5220/0006418002850294.

Sondeck, LP., M. Laurent, and V. Frey (2017c). "Discrimination rate: an attribute-centric metric to measure privacy". In: *Annals of Telecommunications journal DOI: 10.1007/s12243-017-0581-8.*

Soria-Comas, Jordi et al. (2014). "Enhancing data utility in differential privacy via microaggregation-based k-anonymity". In: *The VLDB Journal* 23.5, pp. 771–794.

Spiekermann, Sarah et al. (2015). "The challenges of personal data markets and privacy". In: *Electronic Markets* 25.2, pp. 161–167.

Sweeney, Latanya (2002). "Achieving k-anonymity privacy protection using generalization and suppression". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05, pp. 571–588.

Templ, Matthias, Bernhard Meindl, and Alexander Kowarik (2013). "Introduction to statistical disclosure control (sdc)". In: *Project: Relative to the testing of SDC algorithms and provision of practical SDC, data analysis OG*.

Tendick, Patrick (1991). "Optimal noise addition for preserving confidentiality in multivariate data". In: *Journal of Statistical Planning and Inference* 27.3, pp. 341–353.

Tendick, Patrick and Norman Matloff (1994). "A modified random perturbation method for database security". In: *ACM Transactions on Database Systems (TODS)* 19.1, pp. 47–63.

Torra, Vicenç (2004). "Microaggregation for categorical variables: a median based approach". In: *Privacy in statistical databases*. Springer, pp. 518–518.

— (2017a). "Information Loss: Evaluation and Measures". In: *Data Privacy: Foundations, New Developments and the Big Data Challenge*. Springer, pp. 239–253.

— (2017b). "Machine and Statistical Learning". In: Springer International Publishing, pp. 23–54.

Torra, Vicenç and Josep Domingo-Ferrer (2003). "Record linkage methods for multidatabase data mining". In: *Information fusion in data mining*. Springer, pp. 101–132.

Torra, Vicenç and Sadaaki Miyamoto (2004). "Evaluating fuzzy clustering algorithms for microdata protection". In: *Privacy in Statistical Databases*. Vol. 3050. Springer, pp. 175–186.

Tóth, Gergely, Zoltán Hornák, and Ferenc Vajda (2004). "Measuring anonymity revisited". In: *Proceedings of the Ninth Nordic Workshop on Secure IT Systems*. Espoo, Finland, pp. 85–90.

Wang, Qian et al. (2016). "Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy". In: *IEEE Transactions on Dependable and Secure Computing*.

Wernke, Marius et al. (2014). "A classification of location privacy attacks and approaches". In: *Personal and Ubiquitous Computing* 18.1, pp. 163–175.

Willenborg, Leon and Ton De Waal (2012a). *Elements of statistical disclosure control*. Vol. 155. Springer Science & Business Media.

— (2012b). *Elements of statistical disclosure control*. Vol. 155. Springer Science & Business Media.

Winkler, William E (2004). "Re-identification methods for masked microdata". In: *International Workshop on Privacy in Statistical Databases*. Springer, pp. 216–230.

Wolf, Peter-Paul de (2006). "Risk, utility and PRAM". In: *Privacy in Statistical Databases*. Springer, pp. 189–204.

Wong, Raymond Chi-Wing et al. (2006). "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 754–759.

Xiao, Xiaokui and Yufei Tao (2006). "Anatomy: Simple and effective privacy preservation". In: *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, pp. 139–150.

Xu, Jian et al. (2006). "Utility-based anonymization using local recoding". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 785–790.

Xu, Lei et al. (2015). "Privacy or utility in data collection? A contract theoretic approach". In: *IEEE Journal of Selected Topics in Signal Processing* 9.7, pp. 1256–1269.

Xue, Minhui et al. (2017). "Characterizing user behaviors in location-based find-and-flirt services: Anonymity and demographics". In: *Peer-to-Peer Networking and Applications* 10.2, pp. 357–367.