# Chapter 1

# Introduction

## 1.1 Motivations

Statistical Disclosure Control (SDC) is the discipline concerned with the anonymization of the statistical data containing confidential information about individual entities such as persons or enterprises. Normally, data anonymization is achieved by modifying data values. The aim of SDC is to prevent third parties working from this data to recognize individuals and disclosing confidential information about them. Here, we understand *third parties* as the data users outside the statistical agencies (*e.g.* policy makers, academic researchers and general public).

Typically, data published by statistical agencies can be classified as tabular data and microdata files. Tabular data contains aggregated values and their utility is limited. In contrast, microdata files (*i.e.* records which contain information about individuals) have much more utility due to their flexibility to allow the user to perform a wide range of data analysis (*i.e.* regressions). For this reason, third parties have increased their demand for statistical data according to this latter form. This issue motivates statistical agencies to increase the release of microdata files.

In both scenarios, statistical agencies have to be careful when releasing statistical data since they have an important responsibility towards the respondents. Moreover, international and local law seek to ensure that confidential data is managed in a correct (and private) manner. They have to make (almost) impossible for third parties to acquire sensitive information about respondents from the released microdata file.

A closely related research line where privacy is involved is Privacy Preserving Data Mining (PPDM). PPDM tackles the problem of developing data mining techniques where the privacy of the individuals is preserved. In a very similar way to SDC, PPDM modifies individual data records in such a way that the results of a mining process are (almost) the same as those obtained when using the real data.

In both cases (SDC and PPDM) the privacy of the individuals through data protection methods should be ensured. These methods modify the original microdata file or data set[1], adding some noise in the original data. Of course, the aim of such methods is to preserve the statistical utility of the protected data as much as possible. This is equivalent to modify the information as little as possible. However, protected data have to be altered enough to obfuscate the identity of the respondents.

Protection methods solve in some way the problem of the privacy of the respondents. Nevertheless, an important and challenging problem arises: the evaluation of such methods. This evaluation has two clear components. On the one hand, the loss of statistical utility of the protected data (*information loss*) and on the other hand, the risk that third parties discover the identity of certain respondents (*disclosure risk*).

Information loss measures can be general or specific. General information loss measures roughly reflect the amount of information loss for a reasonable range of data uses. On the other hand, specific information loss measures evaluate the loss of statistical utility for a particular data analysis. Normally, the first kind of measures are used to compare protection methods and the second ones are used to evaluate in an accurate way the real effect of a protection method for a concrete statistical analysis.

Disclosure risk, the main topic of this thesis, evaluates the privacy of the respondents against possible malicious uses that third parties (sometimes called intruders) could do with the information released. Disclosure risk measures evaluate the number of respondents whose identity is revealed. Normally, these measures are computed in several scenarios where the intruder has partial knowledge of the original data. In order to compute the disclosure risk, general methods for re-identification are used. These methods find relationships (*i.e.* links) between the protected data and the partial knowledge which the intruder is assumed to have.

In the real world, the disclosure risk is bounded by the best re-identification method that an intruder is able to conceive. Finding this method is a challenging task as the intruder can exploit any weakness of the protection method or any extra information about the original data. Therefore, the computation of the real disclosure risk is a very hard issue since lots of considerations must be taken into account. This thesis is focused on this matter. The aim of this work is to provide a set of techniques for statistical agencies and data providers in general to determine the disclosure risk in the most accurate way.

## 1.2 Contributions

The research done in this thesis contributes in three different aspects.

Firstly, it contributes to the area of disclosure risk evaluation. We introduce several re-identification methods to compute the disclosure risk of different data

[1]Microdata file is the term used in SDC to refer to the raw data, and data set is usually the term used in PPDM to refer to the same concept. In this thesis we will use both terms.

protection methods. The new re-identification methods show that up to now the real disclosure risk of such protection methods was underestimated. These methods demonstrate that an intruder can increase the amount of correctly re-identified respondents by considering the protection method applied in the anonymization process. Therefore, the disclosure risk of these methods rises accordingly. We also define a different disclosure risk scenario where the intruder has no access to the original data. However, under some assumptions, we prove that it is still possible for the intruder to re-identify some of the respondents of such protected data set.

The second contribution is included in the area of data protection methods. We introduce several protection methods which solve the drawbacks presented in the disclosure risk evaluation. These new methods improve the privacy of the respondents. The methods showed in this thesis avoid that an intruder may exploit the knowledge of the protection method used. We also define a new measure to evaluate, in an empirical way, the anonymity level achieved using a specific configuration of a protection method and assuming that the intruder has access to the original values of a subset of the protected attributes.

Finally, we present a suite of techniques for time series anonymization and re-identification. The idea underlaying this approach is that data accumulation through consecutive statistical surveys enables to perform temporal analysis over such data (*e.g.* forecasting). However, this temporal information can be also used by the intruder to increase the disclosure risk of this new accumulated survey. Under this scenario, we also define new information loss measures which consider temporal analysis that third parties can perform in the accumulated data set.

## 1.3 Structure of the Document

This document is organized in three parts with five chapters: preliminaries and related work (Chapter 2), our contributions (Chapters 3 to 6) and, finally, conclusions and future directions (Chapter 7).

- **Chapter 2.** We explain some preliminaries needed later on. These preliminaries are divided in six sections:
  - **Aggregation functions.** We begin the preliminaries explaining some basic concepts about aggregation functions. Such description includes the definition of the OWA (Ordered Weighted Averaging) operator and some fuzzy integrals, in particular, the Choquet, Sugeno and twofold integrals.
  - **Time series.** We introduce some notions about time series as, for instance several time series distances and forecasting models.
  - **Re-identification methods.** We give a brief introduction of classical re-identification methods and explain in more detail record linkage (RL) methods. RL methods are specific cases of the re-identification methods.

  - **Microdata protection methods.** We show the general problem of data privacy, the re-identification scenario and we give two classifications of protection methods. We also explain in detail two specific data protection methods: rank swapping and microaggregation.
  - **Information loss and disclosure risk.** We present some information loss and disclosure risk measures and a framework for evaluating a data protection method.
  - **Data sets description.** We give an exhaustive description of the data sets used in the experiments performed in this thesis.

- **Chapter 3.** We explain some contributions about specific microaggregation disclosure risk measures. We also present two new variants of the generic microaggregation algorithm.

- **Chapter 4.** Three ad-hoc record linkage methods are presented. These methods consider the protection method applied on the original data, and due to this, they achieve a larger number of re-identifications than generic record linkage methods.

- **Chapter 5.** We study an alternative scenario for record linkage methods where attributes in the original and the protected data set are not the same.

- **Chapter 6.** We present some results about time series protection and re-identification. We also present some information loss measures for the evaluation of time series protection methods.

- **Chapter 7.** This thesis concludes with some conclusions and a description of future work.