Privacy Protection Methods for Documents and Risk Evaluation for Microdata

by

Daniel Abril Castellano

Advisor:
Vicenç Torra Reventós

Tutor:
Jordi Herrera Joancomartí

A dissertation presented in partial fulfillment for
the degree of PhD in Computer Science

Universitat Autònoma
de Barcelona

January 21, 2015
Universitat Autònoma de Barcelona
Departament d'Enginyeria de la Informació i de les Comunicacions

Abstract

The capability to collect and store digital information by statistical agencies, governments or individuals has created huge opportunities to analyze and build knowledge-based models. With the rise of Internet many services and companies have exploited these opportunities collecting huge amounts of data, which most of the cases are considered confidential. This causes the need to develop methods that allow the dissemination of confidential data for data mining purposes while preserving individuals' private information. Thus, personal data could be collected, transferred or sold to third parties ensuring the individuals' confidentiality, but still being statistically useful.

Internet is full of unstructured textual data like posts or documents with a large content of information that can be extracted and analyzed. Documents are especially difficult to protect due to their lack of structure. In this thesis we distinguish two different models to protect documents. On the one hand, we consider the protection of a collection of documents, so this set of documents can be analyzed by means of text mining and machine learning techniques. On the other hand, we consider the protection of single documents by means of the documents' sanitization. This is the process of detecting and removing the parts of the text considered sensitive. When dealing with governmental classified information, sanitization attempts to reduce the sensitiveness of the document, possibly yielding a non-classified document. In this way, governments show they uphold the freedom of information while the national security is not jeopardised.

This thesis presents a set of different methods and experiments for the protection of unstructured textual data protection and besides, it introduces an advanced method to evaluate the security of microdata protection methods. The main contributions are:

- The development of a new semi-automatic method to assist documents' declassification.
- The definition of two specific metrics for evaluating the information loss and disclosure risk of sanitized documents.
- The development of two cluster-based approaches based on the k-anonymity principle to anonymize vector space models. One exploits the sparsity and the other exploits the possible semantic relations of those vectors.

- The study of advanced methods to evaluate the disclosure risk of microdata protection methods. In particular, we have developed a general supervised metric learning approach for distance-based record linkage. Moreover, we have reviewed the suitability of a set of parameterized distance functions that can be used together with the supervised approach.

Acknowlegements

Aquest treball no hagués sigut possible sense l'ajuda de moltes persones, per tant voldria dedicar unes línies per expressar el meu sincer i profund agraïment a totes aquelles persones que directament o indirectament han col.laborat en la realització d'aquest treball.

Primer de tot m'agradaria donar a les gràcies al meu director de tesis, en Vicenç Torra, amb el qual sense la seva ajuda això no hagués estat possible. Gràcies per la confiança, la paciència i la dedicació contínua al llarg d'aquests quatre anys. En aquesta mateixa línia també m'agradaria agrair molt especialment a en Guillermo Navarro. Gràcies pel teu suport i orientació tant en l'àmbit de la recerca com en el personal. També m'agradaria agrair a en David Nettleton, en Javier Murillo i en Yasuo Narukawa per les seves respectives col.laboracions en aquesta tesi.

En segon lloc haig d'agrair al CSIC i al IIIA per brindar-me l'oportunitat de gaudir d'una beca pre-doctoral (JAE) amb la qual he pogut dedicar-me exclusivament a la realització d'aquest treball.

M'agradaria fer una menció especial a totes aquelles persones que m'han pogut acompanyar al llarg d'aquests últims anys. Gràcies als meus companys de departament Jordi, Sergi i Marc per tots els moments compartits. També voldria agrair l'amabilitat i simpatia rebudes per part de tots i cada un dels membres del IIIA i de la universitat Milano-Bicocca.

Finalment, m'agradaria fer un reconeixement molt especial per la comprensió, paciència i ànims rebuts per part de la meva família, amics i sobretot de la meva parella, la Beatriz.

A tots vosaltres, moltes gràcies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The capability to collect and store digital information by corporations, governments or individuals has created huge opportunities to analyze and build knowledge-based models. Both easy capability of collecting data and cheap and powerful computers are the key points for the knowledge discovery process. That is, we are able to analyze an enormous set of data and get its hidden and useful information that can provide us with knowledge. The clearest example of collecting and mining data is the Internet, where there is a booming of services such as social networks, electronic commerce, forums and many others. Most of the collected data is considered personal information, initially used to build personal profiles and then, used to analyze, determine and even predict individuals' behaviours, interests and habits. This fact has made companies and organizations realize about the powerfulness of data analytics, so they have started to collect and mine data for their own benefit.

Inevitably, these information extraction processes have created a huge debate concerning to the individuals' private information since many data collectors are likely to share or sold the collected information to third parties. Besides these sharing or leaking problems, individuals are not completely aware of the potential abuses the companies can practice with their personal data. Therefore, new mechanisms have been introduced to make people aware of the misuse and transferences of their personal data for a variety of purposes. Several areas like data mining, cryptography and information hiding have been developed in order to provide data mining tools in a privacy preserving way.

In this dissertation, we focus on the development of new methods for data protection. With these methods, the practitioners will be able to release or share their collected data so that they can be analyzed controlling the disclosure risk of individual's information. Moreover, we provide a set of mechanisms to evaluate the released protected data. These mechanisms are responsible of assessing the prevailing statistical utility and also the risk of individual's privacy disclosure.

1.1 Motivation

The most common way to share confidential information between several parts is using cryptographic techniques. The data is encrypted and then released and just those who know the correct decoding procedure can recover the original data. Thus, the data has no analytical utility for users without clearance. However, by using these methods the owner of the data has to blindly trust in the statistical agencies and people with which he has shared the data. This is difficult to control due to the amount of agencies that could be involved in the sharing process. Therefore, unfortunately it is foreseeable to expect unauthorized copies. In fact, one of the largest releases of classified data online occurred on the 28th of November 2010, when WikiLeaks [Wikileaks, 2010], a non-profit organization, published more than $250,000$ United States diplomatic documents. From this large set of published documents there were over $115,000$ labeled as *confidential* or *secret* and the remaining ones were not confidential, i.e, considered safe, by the official security criteria. According to the United States government the documents are classified in 4 levels: *Top secret*, *Secret*, *Classified* and *not confidential.* These categories are assigned by evaluating the presence of information in a document whose unauthorized disclosure could reasonably be expected to cause identifiable or describable damage to the national security [E.O. 13526, 2009]. This type of information includes military plans, weapons systems, operations, intelligence activities, cryptology, foreign relations, storage of nuclear materials, and weapons of mass destruction. On the other hand, some of this information is often directly related to national and international events which affect millions of people in the world. Thus, people in a democracy may wish to know the decision making processes of their elected representatives, ensuring a transparent and open government. Therefore, releasing such amount of confidential data caused a great debate between those who uphold the freedom of information and those who defend the right to withhold information.

All the US Embassy cables were published on the Internet fully unedited, in a “raw state. That means that they included all kinds of confidential information such as emails, telephone numbers, names of individuals and certain topics. Nonetheless, their absence may not have significantly impaired the informative value of the documents with respect to what are now considered the most important revelations of the Cables. So, it is fundamental to provide privacy to data against disclosure of confidential information. The importance of this problem has attracted the attention of some international agencies. For example, the DARPA, the Defense Advanced Research Projects Agency of the United States Department of Defense, solicited for new technologies to support declassification of confidential documents [DARPA, 2010]. The maturity of these technologies would permit partial or complete declassification of documents. In this way, documents could be transferred to third parties without any confidentiality problem, or with the only information really required by the third party aiming to make the possibility of sensitive information leakage minimal. These technologies will also help the capability of departments to identify still sensitive information and to make declassified information available to the public.

Many government agencies and companies are collecting massive amounts of confidential data such as medical records, income credit rating, search queries or even several types of test scores in order to perform different kind of studies. These databases are analyzed by their owners or more commonly by third parties, creating a conflict between individual's right to privacy and the society's need to know. A well known example of this conflict is the AOL case. On the 4th of August 2006, AOL in an attempt to help the information retrieval research community publicly released several million search queries made by AOL users. Twenty million search queries for over 650, 000 users over a period of three months were released after performing a very simplistic anonymization, giving an anonymous identifier to each user. However, as proven later search queries contain personally identifiable information. Although AOL retired the dataset three days after the release, the data was mirrored and distributed on the Internet. Whereupon, through analysis of text and linking attributes from the queries to public data, the user *4417749* was re-identified [Barbaro and Zeller, 2006].

Privacy Preserving Data Mining (PPDM) [Agrawal and Srikant, 2000] and Statistical Disclosure Control (SDC) [Willenborg and Waal, 2001] research on methods and tools for ensuring the privacy of databases. Unlike encryption methods, which provide a maximum protection but no utility for unauthorized users, these two disciplines research for new protection methods entailing some degree of data modification. That is, they seek to protect data in such a way that it can be publicly released or shared with third parties such as statistical agencies, so this data can be analyzed and mined without giving away private information that can be linked to specific individuals or entities. Figure 1.1 depicts this process. This is an important application in several areas, such as official statistics, health statistics, e-commerce, etc.

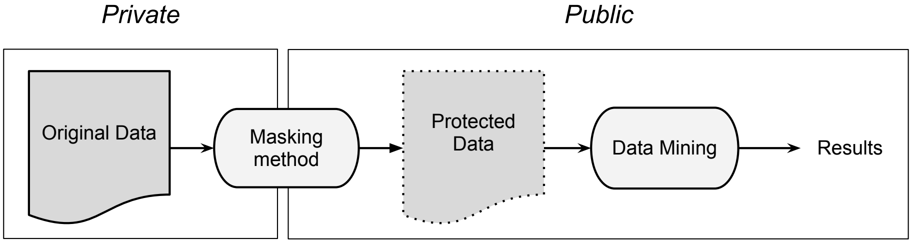


Figure 1.1: PPDM and SDC process.

Both disciplines are very similar although they differ in their origin. On the one hand, Statistical Disclosure Control (also known as Inference Control in Databases) has its origin in the National Statistical Offices and the need of publishing the data obtained from census and questionnaires for researchers or policy makers. Typically, these agencies deal with statistical individual data, also called microdata, due to its flexibility to perform a wide range of data analysis. On the other hand, Privacy Preserving Data Mining has its origin in the data mining community, and methods are more oriented to companies that need to share the data either with other companies or with researchers.

One line of research in both areas focus on the development of data protection methods that ensure the privacy of the data individuals. These methods achieve protection by applying modifications or transformations to the original data. In this case, the challenge is to achieve protection with minimum loss of the accuracy sought by data individuals, while ensuring a low risk. Hence, the evaluation of such methods is performed by two concepts: the *information loss* (or the opposite concept the *data utility*) and the *disclosure risk*.

Information loss measures evaluate the statistical utility of the protected data. These measures can be divided in two types: general or specific measures. General information loss measures roughly reflect the amount of information loss for a reasonable range of data uses. Whereas, specific information loss measures evaluate the amount of statistical information loss for a specific data analysis. Normally, the first kind of measures are used to compare protection methods or evaluate the protection in a general way and the second ones are used to evaluate in a more accurate way, i.e., to study the real effect of protection method for a particular statistical analysis.

Disclosure risk measures evaluate the capacity of an intruder to obtain some information about the original data set from the protected one. Some of these measures evaluate the number of respondents (or data individuals) whose identity is revealed. In order to compute the disclosure risk of a protected data set, general methods for re-identification can be used. Mainly, these methods rely on record linkage approaches, in which they try to find relationships between original and protected records that belong to the same individual. In the real world, the disclosure risk is bounded by the best re-identification method an intruder is able to conceive. Because of that, this approach is a challenging so the intruder can exploit any weakness of the protection method and can exploit any extra information about the original data.

The goal for all protection methods is to find a trade-off between these two concepts, since they are inversely proportional. That is, when the information loss decreases, the disclosure risk increases and vice versa. The task of finding the optimal trade-off between these two concepts is difficult and challenging. This has made that many researchers put their efforts into the development of accurate disclosure risk and information loss measures. In this thesis we focus on all these challenges, from developing new and advanced information loss and disclosure risk measures to the implementation of novel protection methods that satisfy a good trade-off between disclosure risk and data utility.

1.2 Contributions

In this dissertation we contribute with three different topics of privacy preserving data mining and statistical disclosure: *(i)* the development of new protection methods, *(ii)* the introduction of new measures to evaluate the loss of information produced by a protection method and, finally, *(iii)* the development of new measures to evaluate the disclosure risk of a protected data set.

Our first contribution is in the area of declassification technologies. We propose a novel semi-automatic technique for confidential documents sanitization. This technique aims to assist document's declassification by making this process faster and less tedious. Sanitization techniques are required for the declassification process. That is, they involve identification and removal of confidential individual information of the information that third parties are not authorized to know. Our approach automatically identifies and anonymizes confidential information that is directly related to individuals; names, phones or e-mails are some examples of user confidential information. Besides, it is able to identify those words, sentences or paragraphs that contain information that should not be shared. These sensitive text parts should be reviewed and deleted by an expert. Additionally, we evaluate the effect of document sanitization by proposing a couple of measures in order to measure the information loss and the risk of disclosure of sanitized (or declassified) documents. In this way, declassification practitioners are able to automatically sanitize and evaluate their sanitized documents.

The second contribution is in the area of data protection methods. We introduce a protection method for unstructured textual data collections. By following traditional information retrieval steps a set of documents can be represented as a matrix data set. Each row corresponds to an individual document of the collections and each column expresses the relevance of words within a document. This matrix-like representations resembles to microdata. Thus, new methods for unstructured textual data collections can be developed according to standard methods in the PPDM and SDC disciplines.

We propose two protection methods relying on different document characteristics. On the one hand, we present a protection method for sparse and high-dimensional datasets such as document-term matrix representations. Typically, most documents (rows) contain only a small subset of the total number of words and hence they are very sparse, a sparsity about 90% is common. Current protection methods do not take into account this data distribution. After the protection we get an anonymous document-term matrix which is a protected model generated from the original one. Recall, this model representations are used in many machine learning and information retrieval techniques like clustering, classification, text indexing, etc. On the other hand, we propose another protection method for textual datasets. This has the singularity to consider the semantic relations of document words. It considers the same representation than in the previous approach, but now it relies its effectiveness on a given external database. This is a lexical database which contains semantic word relations. Hence, it allows us to perform different operations between words that take into account the semantics of those words. For instance, we are able to compute words similarities or find semantic-based word generalizations. Like the previous method, at the end of the protection process, we get a protected document-term matrix, so it is also easy to analyze with several mining algorithms.

Finally, our last contribution is in the area of record linkage as a disclosure risk evaluation method. We present a novel technique for distance-based record linkage. The aim of this technique is to assess the disclosure risk of a protected dataset and improve the results obtained by the current distance-based record linkage methods. In addition, it also can be used by an intruder to re-identify individuals and get some confidential information. This method consists of a supervised learning method for distance-based record linkage. First, we describe it as a general problem and then, we introduce a set of parameterized distance functions that can be used together with the supervised learning approach. We formalize the problem as an optimization problem. Its goal is to find the function parameters that maximize the number of re-identifications. The key point of the method is the parameterized function used. Because of that, we study the different characteristics and behaviours of each proposed parameterized function. Additionally, we evaluate their re-identification accuracies and consuming times when they are used in the supervised learning method. Apart from evaluating the disclosure risk, this method provides the function parameters. They provide useful information about which are the attributes with weak or strong protection levels. That is, they identify those attributes that are likely to generate a data security breach. Therefore, data practitioners should consider the analysis of these parameters and so, avoid publishing data protected with an inappropriate method.

1.3 Structure of the Document

This thesis has been structured in three main parts:

The first part (Chapter 2) consists of an introduction of all topics related to this thesis.

Chapter 2 introduces the state-of-the-art as well as some background knowledge needed to understand the following chapters. These preliminaries are divided into five different sections.

- **Aggregation Operators and Fuzzy Measures.** We begin giving some basic definitions and properties about aggregation operators. We also define fuzzy measure and we review some of its interesting properties and transformations. Finally, we introduce the Choquet integral aggregation operator, which permits integrating a function with respect to a fuzzy measure.

- **Record Linkage.** We review the state-of-the-art of record linkage. We also describe the general process and two different types of record linkage techniques: distance and probabilistic based techniques. Finally, we describe the possible applications and scenarios where record linkage techniques can be applied in the context of data privacy.

- **Microdata Protection and Evaluation Methods.** We describe some protection methods like microaggregation, rank swapping and additive noise. Finally, we introduce a set of mechanisms to evaluate microdata protection methods. These mechanisms consist of evaluating the information loss and the disclosure risk of the released protected data.

- **Metric Learning.** We introduce the problem and review some state-of-the-art techniques of metric learning. Moreover, we introduce some basic concepts for solving optimization problems.

- **Unstructured Text Data Protection.** We review some related work on unstructured textual data protection such as sanitization and privacy preserving text mining techniques. In addition, we describe some basic techniques for document pre-processing and representation. Finally, we describe WordNet [WordNet, 2010], a lexical database with which we are able to find semantic relations between words.

The second part (Chapters 3, 4 and 5) focuses on our contributions.

Chapter 3 describes some contributions about specific measures to evaluate the information loss and the disclosure risk of sanitized documents. In addition to these evaluation mechanisms, we present a semi-automatic sanitization method for document declassification following the protocols stated by US Administration.

In **Chapter 4** we address the problem of how to release a set of confidential documents. To that end we propose a couple of methods that provide some anonymized metadata of these documents that can be released and used for analysis and mining purposes. Relying on document-term matrix document representation, we present two protection methods: the spherical microaggregation and the semantic microaggregation. We also present some specific and general techniques for their evaluation. These measures are defined in terms of the information that has been lost in the protection process.

In **Chapter 5** we explain some contributions for the disclosure risk assessment of protected datasets. We present a general supervised metric learning approach for distance-based record linkage. This is a Mixed Integer Linear Problem (MILP). Moreover, we review a set of parameterized distance functions that can be used together with the supervised approach. They are the weighted arithmetic mean, a symmetric bilinear function and the Choquet integral. All of them will be studied and compared emphasizing the importance of their parameters. We present two different ways of solving for the optimization problem: using a commercial solver, [IBM ILOG CPLEX, 2010a], and also using a heuristic method, which consists of a gradient descent algorithm.

The last part of the thesis, **Chapter 6**, summarizes our conclusions and provides some directions for future work.

Chapter 2

Preliminaries

In this chapter we introduce the state-of-the-art as well as some basic background knowledge needed to understand the following chapters. First, in Section 2.1 we explain some basics about aggregation functions and its integration with fuzzy measures. Next, in Section 2.2 we review the origins of record linkage and some actual applications as well as the two main existing approaches: distance-based and probabilistic-based record linkage, of which the former will be further studied and extended in Chapter 5. Then, a brief description of microdata and its protection methods are given in Section 2.4. We pay special attention to microaggregation, which will be modified in the following chapters. In Section 2.5 we introduce a general evaluation for protected microdata file; this evaluation relies on a measure to quantify the information loss in the protection process and a disclosure risk measure. Afterwards, we review the state-of-the-art related to different metric learning approaches. Finally, in Section 2.7 we review some research lines related to the unstructured text protection methods and additionally we introduce some basics about algebraic models for representing collections of documents.

2.1 Aggregation Operators and Fuzzy Measures

This section defines some basic terms in the field of information fusion and integration. We review the definition of fuzzy measure and some of its most interesting properties and the Möbius transform. We will also review belief functions a type of fuzzy measure that is relevant for the definition of the distance. The section finishes with the definition of the Choquet integral and an example of its application.

According to the information fusion field, the term aggregation operator is described as those operators (also called means operators) corresponding to a particular mathematical function used for information fusion. Generally, these mathematical functions combine n values in a given domain D and return a value in the same domain.

Definition 2.1. *Let $X = \{x_1, \cdots, x_n\}$ be a set of information sources, and let $f(x_i)$ be the function that models the value c_i supplied by the i-th information source x_i. Then an aggregation operator is a mathematical function of the form, $\mathbb{C} : D^n \to D$, which usually requires satisfying the following properties,*

(i) $\mathbb{C}(c, \cdots, c) = c$ *(idempotency)*

(ii) $\mathbb{C}(c_1, \cdots, c_n) \geq \mathbb{C}(c'_1, \cdots, c'_n) \quad \forall\ c_i \geq c'_i, \quad i = \{1, \cdots, n\}$ *(monotonicity)*

As long as properties (i) and (ii) are hold, the aggregation operators also hold,

(iii) $\min_i c_i \leq \mathbb{C}(c_1, \cdots, c_n) \leq \max_i c_i$ (internality)

Two of the most extended aggregation operators are the arithmetic mean (AM) and the weighted mean (WM). See their corresponding functions below,

- $AM(c_1, \cdots, c_n) = \frac{1}{n} \sum_{i=1}^{n} c_i$
- $WM_p(c_1, \cdots, c_n) = \frac{1}{n} \sum_{i=1}^{n} p_i \cdot c_i, \ \forall\ p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$

As the weighted mean does, most aggregation operators fuse a set of input values taking into account some information about the sources. That is, operators are parametric and thus that additional background knowledge is considered in the fusion process. These parameters play an important role in the applications using aggregation operators. They can express the reliability of the information sources. Usually, parametric aggregation operators are expressed by $\mathbb{C}_p$, where p represent the parameters of the operator.

One of the most well known aggregation operators is the Ordered Weighted Averaging operator, also known as OWA operator. This was introduced by Yager in [Yager, 1988] to provide a mean for aggregating scores associated with the satisfaction of multiple criteria.

Definition 2.2. *An Ordered Weighted Averaging (OWA) operator of dimension n is a mapping $OWA : D^n \to D$ that has associated a weighting vector $p = (p_1, \cdots, p_n)$ such that,*

$$OWA_p(c_1, \cdots, c_n) = \sum_{i=1}^{n} p_i \cdot c_{\sigma(i)}$$

where σ defines a permutation of $1, ..., n$ such that $c_{\sigma(i)} \geq c_{\sigma(i+1)}$ for all $i \geq 1$. The weights are all non negative ($p_i \geq 0$) and their sum is equal to one ($\sum_{i=1}^{n} p_i = 1$). Remark that the obtained aggregated value is always between the maximum and the minimum of the input values.

Additionally to the previously defined aggregation operator properties, (i), (ii) and (iii) from Definition 2.1, the OWA operator has another interesting property,

(iv) It is a symmetric operator for all permutation π on $\{1, \cdots, n\}$:

$$OWA_p(c_1, \cdots, c_n) = OWA_p(c_{\pi(1)}, \cdots, c_{\pi(n)})$$

As noted in the literature [Yager, 1993], the particularity of this aggregator is that provides a parameterized family of aggregation operators, which include a notable set of mean operators depending on the particular weights chosen. Some operators' examples are the maximum, the minimum, the k-order statistics, the median and the arithmetic mean.

Among all the existing types of aggregation parameters, *fuzzy measures* are a rich and an important family. They are used in conjunction with several fuzzy integrals, such as Sugeno integral [Sugeno, 1974] or Choquet integral [Choquet, 1953] (explained below, see Definition 2.7), for aggregation purposes.

Assuming that the set over which the fuzzy measure is defined is finite, as this is the usual case with aggregation operators, we define fuzzy measures (also known as non-additive measures or capacities) as:

Definition 2.3. *A non-additive (fuzzy) measure μ on a finite set X is a set function $\mu : \wp(X) \to [0,1]$ satisfying the following axioms:*

(i) $\mu(\emptyset) = 0$, $\mu(X) = 1$ (boundary conditions)

(ii) $A \subseteq B$ implies $\mu(A) \leq \mu(B)$ (monotonicity)

The upper boundary requirement, $\mu(X) = 1$, is an arbitrary convention and any other value can be used. However, this is a convenient condition for aggregation purposes, and it is, in fact, a condition analogous to the one for the weighted means to have weights that add to 1.

In addition, from this definition we can observe that non-additive measures are a general case of probability distributions, since they replace the axiom of additivity, satisfied by probability measures, by a more general one, monotonicity. Therefore, probability distributions correspond to a specific type of fuzzy measures, those measures that satisfy $\mu(A \cup B) = \mu(A) + \mu(B)$, which are called additive fuzzy measures.

The interest of using non-additive measures is that they permit us to represent interactions between the elements. For example, we might have $\mu(A \cup B) < \mu(A) + \mu(B)$ (negative interaction between A and B), and $\mu(A \cup B) > \mu(A) + \mu(B)$ (positive interaction between A and B).

Another interesting property of non-additive measures is *submodularity.* A fuzzy measure μ is submodular if the following condition is satisfied for all $A, B \subseteq X$:

$$\mu(A) + \mu(B) \geq \mu(A \cup B) + \mu(A \cap B) \tag{2.1}$$

Any fuzzy measure satisfying the submodularity property will be a *subadditive* fuzzy measure. That is,

$$\mu(A) + \mu(B) \geq \mu(A \cup B)$$

Fuzzy measures can be rewritten alternatively through the Möbius transform. Some particularities of these representations are that it can be used to give an indication of which subsets of X interact with one another and in addition, as we will present below, concepts like k-additivity measures arise naturally.

The Möbius transform of a fuzzy measure μ on a finite set X is a function $m : \wp(X) \rightarrow \mathbb{R}$ that satisfies the following conditions:

(i) $m(\emptyset) = 0$

(ii) $\sum_{A \subseteq X} m(A) = 1$

(iii) if $A \subset B$, then $\sum_{C \subseteq A} m(C) \leq \sum_{C \subseteq B} m(C)$

Definition 2.4. *The Möbius transform m of a fuzzy measure μ is defined as*

$$m_\mu(A) := \sum_{B \subseteq A} (-1)^{|A|-|B|} \mu(B)$$

for all $A \subset X$.

Note that the function m is not restricted to the $[0, 1]$ interval and consequently the Möbius representation of a measure, $m_\mu(A)$, can take negative values for all $A \subseteq X$ with $|A| > 1$.

Then, it is possible to reconstruct the original fuzzy measure μ if the Möbius transform m is given for each $a \subseteq X$.

Definition 2.5. *The Möbius inverse transform is the Zeta transform and it is expressed by:*

$$\mu(A) = \sum_{B \subseteq A} m(B)$$

for all $A \subset X$.

Note that when a measure is additive, its Möbius transform on the singletons corresponds to a probability distribution, and it is zero for non-singletons, i.e., $m(A) = 0$ for all $A \subseteq X$ such that $|A| > 1$.

Taking into account the Möbius transform, it is possible to define a family of fuzzy measures on the basis of the largest set A with non-null $m(A)$. This family of fuzzy measures is called *k-order additive fuzzy measures*, where k is the cardinality of such largest set A.

Definition 2.6. *Let μ be a fuzzy measure and let m be its Möbius transform, then, μ is a k-order additive fuzzy measure if $m(A) = 0$ for any $A \subseteq X$ such that $|A| > k$, and there exists at least one $A \subseteq X$ with $|A| = k$ such that $m(A) \neq 0$.*

Therefore, with its corresponding Möbius transformation, any fuzzy measure can be represented as a k-order additive fuzzy measure with an appropriate k value. This family of measures can be seen as a generalization of additive ones, so if we set $k = 1$, then the measure will be additive. In fact, understanding the Möbius transform as a function that makes explicit the interactions between the information sources, k-order additive fuzzy measure stands for measures where the interactions can only be expressed up to dimension k. For instance, when k is set to 2, only binary interactions are allowed.

We finish this section reviewing another tool for aggregation, the Choquet integral, which permits to integrate a function with respect to a fuzzy measure. The Choquet integral generalizes additive operators as the OWA or the weighted mean.

The Choquet integral is defined as the integral of a function f with respect to a fuzzy measure μ. Both the function and the fuzzy measure are based on the set of information sources $X = \{x_1, \cdots, x_n\}$. The function $f : X \to \mathbb{R}^+$ corresponds to the value that the sources supply and the fuzzy measure assigns importance to subsets of X.

Definition 2.7. *Let μ be a fuzzy measure on X; then, the* Choquet integral *of a function $f : X \to \mathbb{R}^+$ with respect to the fuzzy measure μ is defined by*

$$(C)\int f d\mu = \sum_{i=1}^{N} [f(x_{s(i)}) - f(x_{s(i-1)})]\mu(A_{s(i)}), \tag{2.2}$$

where $f(x_{s(i)})$ indicates that the indices have been permuted so that $0 \leq f(x_{s(1)}) \leq \cdots \leq f(x_{s(N)}) \leq 1$, and where $f(x_{s(0)}) = 0$ and $A_{s(i)} = \{x_{s(i)}, \ldots, x_{s(N)}\}$.

For the sake of simplicity, given a reference set $X = \{x_1, \ldots, x_n\}$ and a fuzzy measure μ on this set, we will use in this paper the notation $CI_\mu(c_1, \ldots, c_n)$ to denote the Choquet integral of the function $f(x_i) = c_i$ with respect to μ.

As a final remark, we show how the weighted mean can also be seen as an aggregated value computed as the integral of a function with respect to a measure. That is, a weighted mean with a weighting vector $p = (p_1, \cdots, p_n)$ can be interpreted as the integral with respect to an additive measure defined on the singletons by $\mu(\{x_i\}) = p_i$ and $\mu(A) = \sum_{x \in A} \mu(x)$.

Definition 2.8. *Let μ be an additive fuzzy measure, then, the integral of a function $f : X \to \mathbb{R}^+$ (with $c_i = f(x_i)$) with respect to μ is*

$$WM_p(c_1, \ldots, c_n) = (C)\int f d\mu = \int f d\mu = \sum_{x \in X} f(x)\mu(\{x\}).$$

where, $p = (\mu(\{x_1\}), \cdots, \mu(\{x_n\}))$.

2.2 Record Linkage

Record linkage is the process of finding quickly and accurately two or more records distributed in different databases (or data sources in general) that correspond to the same entity or individual. The entities under consideration most commonly refer to people, such as patients, customer, tax payers, etc., but they can also refer to companies, governments, publications or even consumer products.

Record Linkage was initially introduced in the public health area when files of individual patients were brought together using name, date-of-birth and other information. It was originally used by Dunn [Dunn, 1946] to describe the idea of assembling a *book of life* for every individual in the world. Dunn defined this book as : "*each person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name of the process of assembling the pages of this Book into a volume*" and he realized that having such information for all individuals will allow governments to improve their national statistics and also the identification of those individuals. In the following years, advances have yielded computer systems that incorporate sophisticated ideas from computer science, statistics, and operations research.

The ideas of modern record linkage were originated by the Canadian geneticist Howard Newcombe et al. [Newcombe et al., 1959, Newcombe and Kennedy, 1962] relying on the full implications of extending the principle to the arrangement of personal files into family histories. Newcombe was the first that undertook a software version that is used in many epidemiological applications and often relies on odds ratios of frequencies that have been computed a priori using large national health files and also the decision rules for delineating matches and non-matches. He also developed the basic ideas of the probabilistic record linkage approach. His approach for deciding whether two records belong to the same person is based on a total computed weight which represents a measure of probability that two records match or not. Later, based on Newcombe's ideas, Ivan Fellegi and Alan Sunter presented in [Fellegi and Sunter, 1969] a mathematical model developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events. This theory demonstrated the optimality of the decision rules used by Newcombe and introduced a variety of ways of estimating crucial matching probabilities (weights) directly from the files being matched. This pioneering work has been the basis for many data matching systems and software products, and even today is still used.

Since the advent of databases, record linkage is one of the existing preprocessing techniques used for data cleaning [Mccallum and Wellner, 2003, Winkler, 2003], and also, it is used to control the quality of the data [Batini and Scannapieco, 2006]. In this way, data sources could be analyzed to deal with dirty data like duplicate records [Elmagarmid et al., 2007], data entry mistakes, transcription errors, lack of standards for recording data fields, etc.

Moreover, it is nowadays a popular technique employed by statistical agencies, research communities, and corporations to integrate different data sets that provide information regarding to the same entities [Defays and Nanopoulos, 1993, Colledge, 1995, Canada, 2010]. For example, a census could be linked to a health dataset in order to extract person-oriented health statistic.

Some of the work was originated in epidemiological and survey applications, but then, this technique was extended to other areas, in which merging different data sources produces a new data with a higher value. A clear example of this database integration are the initiatives launched by governments such as the United States of America or the United Kingdom [data.gov, 2010, data.gov.uk, 2010], to make all their data available as RDF (Resource Description Framework) with the purpose of enabling data to be linked together. Since these two pioneer governments started publishing data an increasing number of other countries have also committed to open up data. Moreover, the Open Knowledge Foundation has ranked a set of 70 countries according to their data openness by looking at ten key areas. Figure 2.1 shows the scores given to the 30 governments with the higher score values in the open data index 2013.

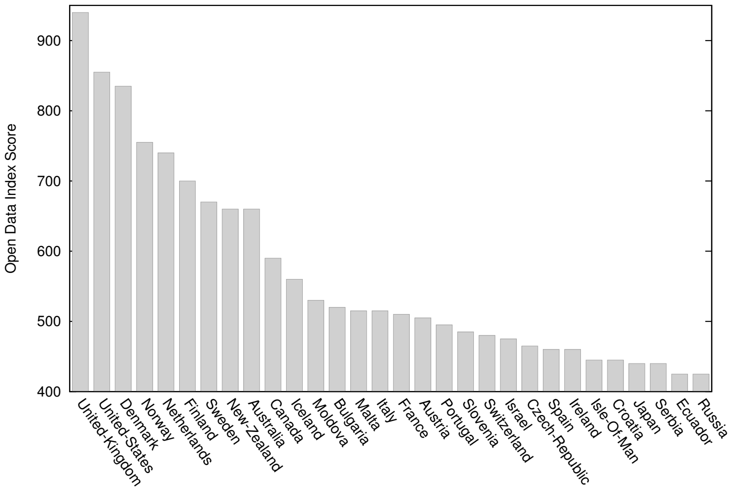


Figure 2.1: Snapshot from October 28th, 2013 of the state of open data release by national governments. For the sake of simplicity, this figure only shows the 30 first countries of the ranking provided by the Open Data Index. See more detailed information on [OKF, 2013].

In the last years, record linkage techniques have also emerged in the data privacy context. Many governments agencies and companies need to collect and analyze data about individuals in order to plan, for example, several kinds of activities or marketing studies. All this information therefore contains confidential information such as income credit rating, types of diseases, or test scores of individuals and it is typically stored in on-line databases, which are analyzed using sophisticated database management systems by the owners or in general, third parties, creating a conflict between individual's right to privacy and the society's need to know and process information. So, it is fundamental to provide security to statistical databases against disclosure of confidential information. Privacy preserving data mining [Agrawal and Srikant, 2000] and Statistical Disclosure Control [Willenborg and Waal, 2001] research on methods and tools for ensuring the privacy of this data. One of the applications of record linkage in this area is the evaluation of disclosure risk of protected data [Torra et al., 2006, Winkler, 2004]. By identifying links of records that belong to the same individual between protected and original databases we can evaluate the re-identification risk of a protected database. [Domingo-Ferrer and Torra, 2001b] define a score function to assess masking methods. This score function relies on a combination of disclosure risk and data utility evaluation techniques. Hence, using analytical measures (either generic or data-use-specific) the score function quantifies the risk of re-identification as well as the information loss of the masked data. The authors also create a ranking taking into account the score (a trade-off between disclosure risk and information loss) of each masking method.

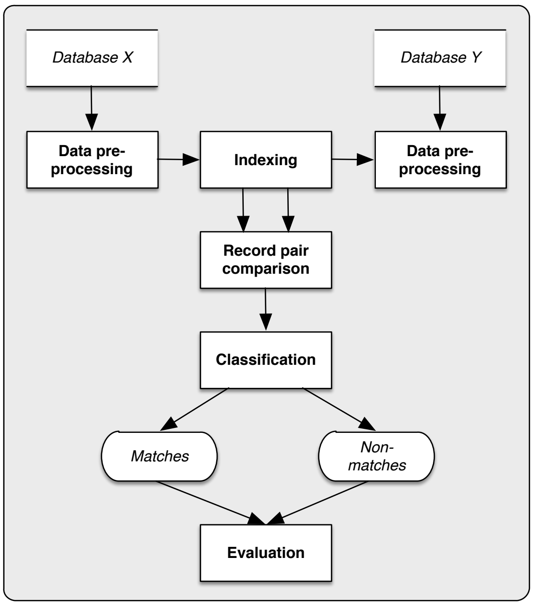


Figure 2.2: Record Linkage process.

We consider that the record linkage process is formed by a set of tasks, as it is shown in Figure 2.2. The process starts with a data sources *pre-processing* task. It is important to make sure that all attributes in both input databases have been appropriately cleaned and standardized. This is a crucial step to achieve successful matchings [Herzog et al., 2007].

The cleaned and standardized database table (or files) are now ready to be matched. In practice, each record from one database needs to be compared with all records in the other database to obtain all the similarities between two records from different databases, thus, the comparison between large databases becomes difficult and sometimes unfeasible. To reduce the possibly very large number of pairs of records that need to be compared, *indexing* techniques are commonly applied [Christen, 2012]. These techniques filter out those pair of records that are very unlikely to match.

Once the candidate record pairs were generated the *pairwise comparison* is computed to determine all pair matches, i.e., all pairs that belong to the same individual. There are several strategies to compare records. The most common ones are based on computing distances and conditional probabilities. Both strategies are explained respectively in Section 2.2.1 and 2.2.2. Then, all compared record pairs are *classified* to be either a *match* or a *non-match*.

Finally, it is necessary to analyze the quality of the list of matching pairs. That is, how many of the classified matches correspond to records that belong to the same individual. This last step requires human intervention or a data *ground-truth* to evaluate the obtained results. Accuracy measures such as precision and recall are used to assess matching quality.

2.2.1 Distance Based Record Linkage (DBRL)

Distance based record linkage (DBRL) consists of linking records by means of computing the distances between all database X and database Y records. Then, the pair of records at the minimum distance are considered a correct link, whereas the remaining pairs are considered not linked pairs. The main point in distance based record linkage is the definition of a distance function to match correctly as many records as possible.

Different distances can be found in the literature, each obtaining different results. In this section we start reviewing two of the most frequently used distances on record linkage, the Euclidean and the Mahalanobis distances.

We adopt the definition of distance function and metric from [Deza and Deza, 2009]. Where a distance function is defined in a less restrictive way than a metric.

Definition 2.9. *Let X be a set. A function $d : X \times X \rightarrow \mathbb{R}$ is called a* **distance** *(or* **dissimilarity***) on X if, for all $a, b, \in X$, there holds:*

(i) $d(a, b) \geq 0$ *(non-negativity)*

(ii) $d(a, a) = 0$ *(reflexivity)*

(iii) $d(a, b) = d(b, a)$ *(symmetry)*

Definition 2.10. *Let X be a set. A function $d : X \times X \rightarrow \mathbb{R}$ is called a* **metric** *on X if, for all $a, b, c \in X$, there holds:*

(i) $d(a,b) \geq 0$ *(non-negativity)*

(ii) $d(a,b) = 0$ *iff* $a = b$ *(identity of indiscernibles)*

(iii) $d(a,b) = d(b,a)$ *(symmetry)*

(iv) $d(a,b) \leq d(a,c) + d(c,b)$ *(triangle inequality)*

Finally, we consider the term *pseudo-distance* to refer other functions that satisfy other small sets of combinations of these properties. Note that other works may consider the terms *metric* and *distance function* as the same concept described in Definition 2.10. Then, those works are using terms such as *pseudo-metric* or *pre-metric* in order to denote Definition 2.9.

Now that we have reviewed the properties required by a metric and a distance function, we are going to survey some metrics used in record linkage. To do so we will use $V_1^X, \ldots, V_n^X$ and $V_1^Y, \ldots, V_n^Y$ to denote the set of variables of file X and Y, respectively. Using this notation, we express the values of each variable of a record a in X as $a = (V_1^X(a), \ldots, V_n^X(a))$ and of a record b in Y as $b = (V_1^Y(b), \ldots, V_n^Y(b))$. $\overline{V_i^X}$ corresponds to the mean of the values of variable V_i^X.

[Pagliuca and Seri, 1999] were the first to use an Euclidean distance based record linkage (Definition 2.11) in the context of data privacy.

Definition 2.11. *Given two datasets X and Y, the squared of the Euclidean distance between two records $a \in X$ and $b \in Y$ for variable-standardized data is defined by:*

$$d^2ED(a,b) = \sum_{i=1}^{n} \left(\frac{V_i^X(a) - \overline{V_i^X}}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V_i^Y}}{\sigma(V_i^Y)} \right)^2$$

where $\sigma(V_i^X)$ and $\overline{V_i^X}$ are the standard deviation and the mean of all the values of variable V_i in the dataset X, respectively.

It is well known that in the Euclidean distance all the variables contribute equally to the computation of the distance. Because of that all points with the same Euclidean distance to the origin define a sphere. However, there are other metrics were this property does not hold. For example, the Mahalanobis distance [Mahalanobis, 1936] allows us to calculate distances taking into account a different variable contribution by means of weighting these variables. These weights are obtained from the correlations between data variables. Because of this rescaling, points at the same Mahalanobis distance define an ellipse around the mean of the set of variables. Torra et al. considered the Mahalanobis distance, also in the data privacy context for disclosure risk assessment [Torra et al., 2006].

Definition 2.12. *Given two datasets X and Y, the square of the Mahalanobis distance between two records $a \in X$ and $b \in Y$ is defined by:*

$$d^2MD_\Sigma(a,b) = (a-b)^T \Sigma^{-1} (a-b)$$

where $(a-b)^T$ is the transposed of $(a-b)$ and Σ is the covariance matrix, computed by $[Var(V^X)+Var(V^Y)-2Cov(V^X,V^Y)]$, where $Var(V^X)$ is the variance of variables V^X, $Var(V^Y)$ is the variance of variables V^Y and $Cov(V^X,V^Y)$ is the covariance between variables V^X and V^Y.

Where, all covariance matrices satisfy the following two properties:

- $\Sigma = \Sigma^T$ (symmetry)
- $\Sigma \succeq 0$ (positive semi-definiteness, see Definition 2.13)

Definition 2.13. *A symmetric matrix Σ is called positive semi-definite if the following property holds: $\vec{x}^T\Sigma\vec{x} \geq 0, \quad \forall\vec{x}$. This is denoted as $\Sigma \succeq 0$.*

Definition 2.14. *If the condition in Definition 2.13 holds with a strict inequality, then Σ is called positive definite, $\Sigma \succ 0$.*

Therefore, knowing that any covariance matrix is at least a positive semi-definite matrix and the inverse of a positive semi-definite matrix is always positive semi-definite, it is straightforward to verify that the Mahalanobis distance, parameterized by a matrix which is not positive definite is not a metric. It does not satisfy the identity of indiscernibles metric property. Conversely, when Σ is positive definite it is a metric.

Notice that Euclidean distance is a special case of the Mahalanobis distance when Σ is the identity matrix, $\Sigma = I$.

2.2.2 Probabilistic Record Linkage (PRL)

The probabilistic record linkage (PRL) algorithm uses a linear sum assignment model to choose which pairs of the original and protected records must be matched. In order to compute this model, the EM (Expectation - Maximization) algorithm [Hartley, 1958, Dempster et al., 1977, McLachlan and Krishnan, 1997] is normally used.

For each pair of records (a,b) where $a \in X$ and $b \in Y$, it is defined a coincidence vector, $\gamma(a,b) = (\gamma_1(a,b)\ldots\gamma_n(a,b))$, where

- $\gamma_i(a,b) = 1$, if $V_i(a) = V_i(b)$,
- $\gamma_i(a,b) = 0$, if $V_i(a) \neq V_i(b)$,

According to some criterion defined over these coincidence vectors, pairs are classified as linked pairs or non-linked pairs. This method was introduced in [Jaro, 1989].

From a computation point of view, probabilistic record linkage is much more complex compared to distance based record linkage. In [Domingo-Ferrer and Torra, 2002] the authors conclude that both methods provide very similar results for categorical data. Moreover, their results also show that both methods are complementary and the best results were obtained when both record linkage methods are combined. In contrast, [Torra and Domingo-Ferrer, 2003] compares both methods and conclude that probabilistic based method is more appropriate for categorical, while distance based seems more appropriate for numerical data.

2.3 Record Linkage in Data Privacy

In the context of record linkage for data privacy, we can distinguish two interesting scenarios in which it is possible to apply record linkage techniques. Although, formally speaking one is a specific case of the other, they are used for different purposes. The fist scenario concerns to how an intruder with some prior knowledge about a set of individuals can extract new valuable information about them or other individuals in the protected file, which unlike its original version it is freely available. The second scenario is focused on the analysis and estimation of the risk of disclosure of a protected data file. That is, it is estimated the possible disclosure of sensitive information of a protected file assuming an attack by an intruder with previous knowledge. Both application scenarios are described below.

- In this first scenario it is considered a possible attacker with some prior knowledge of the original data, a set of original individual records with some common attributes with the protected data file. The attacker is able to find data patterns that will help to link his/her original information to the protected one. Applying record linkage techniques according to a set of constraints given by the attacker will give him/her this statistical information. The sets of prior constraints are generated by indicating which are the links between his/her prior knowledge and the public protected data file. Figure 2.3 in Section 2.4 illustrates this scenario.

- In this second scenario the attacker knows who is in the original database, and has information of all the attributes in the database, and also he is able to link all of them information with the released protected data file. As you may notice this is a very special situation which can only happen if the attacker is the owner of the original data file, since this is confidential and non-public. Therefore, the data owner is the only one that is able to introduce the whole set of constraints to the problem. That is, by knowing all data possible correct links we can evaluate which is the total disclosure risk by applying record linkage techniques. In other words, it is possible to evaluate the *worst case*, where an attacker is able to link all records from both data files in order to obtain sensitive information. This scenario is called the worst case scenario.

2.4 Microdata Protection Methods

Microdata in the context of statistical analysis is information at the level of individual respondents from census or surveys. This data is composed of individual records containing information collected on persons or other kind of entities. The responses of each individual are recorded in separate attributes. For instance, a national census might collect attributes such as age, address, salary, etc., and those attributes are recorded separately for every respondent.

More formally, we define a microdata set X as a matrix with N rows (*records*) and V columns (*attributes*), where each row refers to a single individual. The set of attributes in a dataset can be classified in three different categories, depending on their capability to identify unique individuals, as follows:

- *Identifiers*: attributes that can be used to identify the individual unambiguously. A typical example of identifier is the passport number.

- *Quasi-identifiers*: attributes that are not able to identify a single individual when they are used alone. However, when combining several quasi-identifier attributes, they can unequivocally identify an individual.

- *Others*: otherwise, attributes are classified in this general category.

Another possible attribute classification is distinguishing by its content. That is,

- *Confidential*: attributes which contain sensitive information about the individual. An example of confidential attribute might be any information related to individual's health.

- *Non-confidential*: attributes which do not contain sensitive information, or in other words, attributes which contain public (or easy to access) information. An example of non-confidential attribute would be the zip code.

Among the quasi-identifier attributes, we distinguish between *confidential* (X_c) and *non-confidential* (X_{nc}), depending on the kind of sensitive information that they contain.

Among all possible types of attributes, in this thesis we just have considered three different types: identifiers(id), non-confidential quasi-identifiers (nc) and confidential quasi-identifiers (c).

Table 2.1 shows an example of microdata file. In this example, the attribute Id is an identifier, while the rest of the attributes are quasi-identifiers. From the set of quasi-identifiers we can distinguish between the non-confidential such as *Age*, *Gender*, *Zip code* and *salary*, while the confidential is the *Cancer* attribute, which identifies which of the individuals has cancer.

Following the attribute classification, a dataset X is defined as $X = X_{id}||X_{nc}||X_c$, where X_{id} are the identifiers, X_{nc} and X_c are the non-confidential and confidential quasi-identifiers, respectively, and where $||$ is the operator that combines two datasets by columns.

Before releasing the data, a protection method ρ is applied, leading to a protected dataset Y. Indeed, we will assume the following typical scenario: (i) identifier attributes in X are either removed or encrypted; (ii) confidential quasi-identifier attributes X_c are not modified, and so we have $Y_c = X_c$; (iii)

Id	Age	Gender	Zip Code	Salary	Cancer
1231	23	Female	08191	25,000	No
6273	18	Female	08221	10,000	Yes
7564	58	Male	08191	12,000	No
1188	46	Male	08221	30,000	Yes
0909	18	Female	08046	10,000	No
8761	23	Male	08225	14,000	No

Table 2.1: Example of a microdata file.

the protection method itself is applied to non-confidential quasi-identifier attributes, in order to preserve the privacy of the individuals whose confidential data is being released. Therefore, we have $Y_{nc} = \rho(X_{nc})$. This scenario allows the release of the protected data, $Y = Y_{nc}||Y_c$, to third parties and so they can have precise information on confidential data without revealing to whom the confidential data belongs to. This scenario, which was first used in [Domingo-Ferrer and Torra, 2001b, Sweeney, 2002] to compare several protection methods, has also been adopted in other works like [Winkler, 2004].

Figure 2.3 shows the scenario where an intruder with some extra knowledge, Z, including some individuals from a released protected dataset Y, tries to re-identify them to extract their confidential information. By applying record linkage techniques between the protected attributes, Y_{nc} and the same attributes obtained possibly from other non-protected data sources, $Z = (Z_{id}, X_{id})||(Z_{nc}, X_{nc})$, the intruder might be able to establish some correct links between both datasets. Therefore, for those correct matchings the intruder is able to link the confidential information, X_c, with a piece of its information, $X_{id}|X_{nc}$. This is what protection methods try to prevent.

At present, different protection (also called masking or anonymization) methods have been developed. Protection procedures can be classified [Torra, 2010] into data-driven (or general purpose), computation-driven (or specific purpose), and result-driven. This classification focuses on the intended or known use of the data.

Data-driven or general purpose: In this case, it is not known the intended use of the data, and protection should take into account that the user might apply to the data a large range of tools. For instance, some users might apply regression, other classification or association rules. Perturbative methods are appropriate for this purpose.

Computation-driven or specific purpose: In this case, it is known the type of analysis to be performed on the data. For example, this would be the case if we know that the researcher will apply association rule mining to the data. In this case, protection can be done so that the results on the protected data are the same as (or as similar as possible to) on the original data. In this case, the best approach is that the data owner and the data analyzer agree on a cryptographic protocol [Yao, 1982] so that the analysis

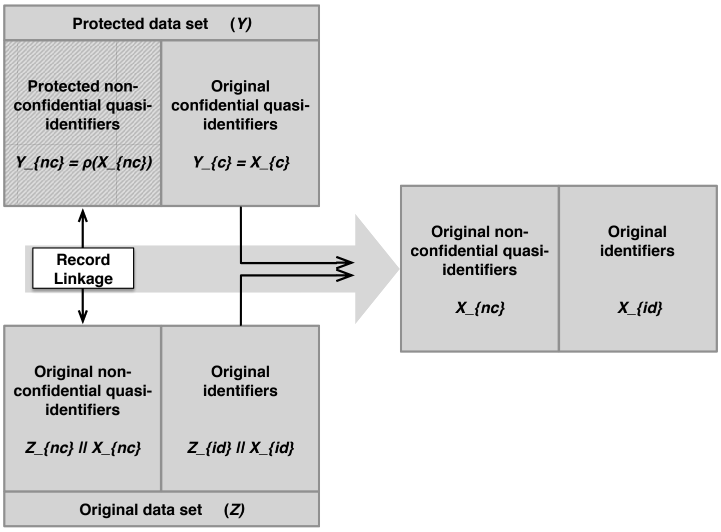


Figure 2.3: Re-identification scenario.

can be done with no information loss. The case of distributed data with a specific goal falls also in this class.

Result-driven: In this case, the concern about privacy is about the results of applying a particular data mining method to some particular data. Protection methods have been designed so that, for instance, the resulting association rules from a data set do not disclose sensitive information for a particular individual.

Note that for a particular application, data-driven and result-driven are not exclusive aspects to be taken into account.

Data-driven protection methods are methods that given a data set build another one based on the former adding some imprecision or noise. Most of these methods can be classified into the following three categories: Perturbative methods, Non-perturbative methods and Synthetic data generators.

In perturbative masking methods data is distorted by adding noise, so combinations of quasi-identifiers which unambiguously identify an individual disappear, while new combinations appear in the perturbed data. This data perturbation preserves individuals confidential information. Since it is an irreversible operation it makes difficult for an intruder to obtain the original values. A perturbative method should ensure a significant degree of preservation of the original statistical information in the protected data. Non-perturbative methods do partial suppressions or reductions of detail (generalizations) on the original data attributes values which unambiguously can identify an individual. Synthetic data generator methods build a statistical model from the original data and afterwards they generate a new random data set constrained by the computed model.

In the following sections we briefly review the most relevant perturbative masking methods for this thesis: microaggregation, rank swapping and additive noise methods for continuous data attributes. Recall, continuous attributes are those numerical attributes where arithmetic operators can be performed on them.

2.4.1 Microaggregation

Microaggregation is a well known anonymization technique that provides privacy by means of partitioning the data into small clusters and then replacing the original data by the representatives of the corresponding clusters. Figure 2.4 depicts microaggregation general behaviour.

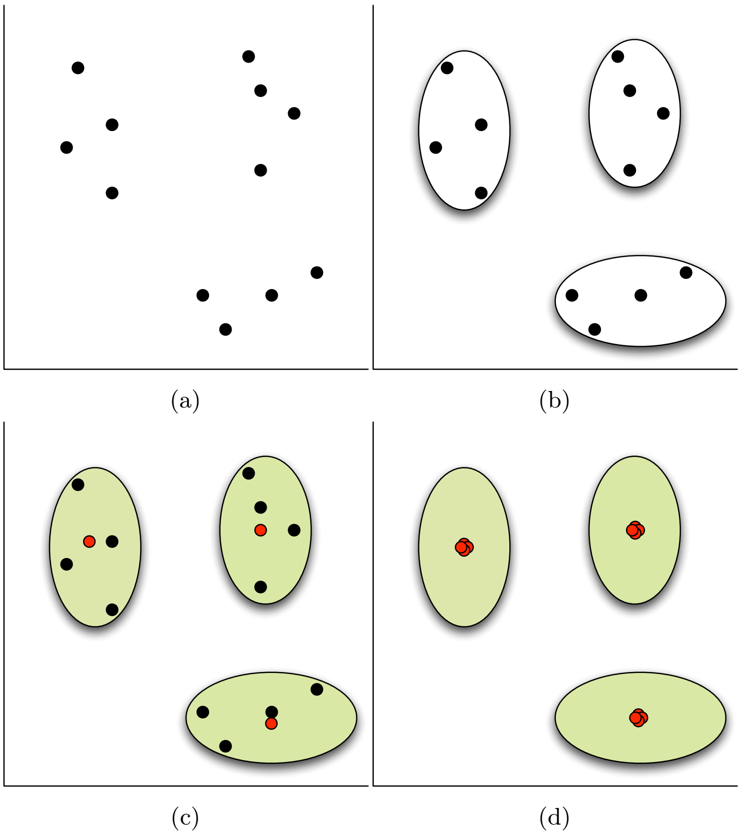


Figure 2.4: Microaggregation.

Microaggregation was originally [Defays and Nanopoulos, 1993] defined for numerical attributes, but later extended to other domains, e.g., to categorical data in [Torra, 2004] (see also [Domingo-Ferrer and Torra, 2005]), and in constrained domains in [Torra, 2008].

From the operational point of view, microaggregation is defined in terms of partition and aggregation:

- **Partition**: Records are partitioned into several clusters, each of them consisting of at least k records. Note that to minimize information loss, clusters should be as homogeneous as possible.

- **Aggregation**: For each of the clusters a representative (the centroid) is computed, and then original records are replaced by the representative of the cluster to which they belong to.

From a formal point of view, microaggregation can be defined as an optimization problem with some constraints, so it should minimize the information loss resulting from this replacement process, in other words, the difference between all the elements of each cluster (x_j) and its corresponding centroid (v_i) should be as minimum as possible. We give a formalization below using u_{ij} to describe the partition of the records in the sensitive data set X with n records. That is, $u_{ij} = 1$ if record j is assigned to the ith cluster and $u_{ij} = 0$ otherwise. Let v_i be the representative (centroid) of the ith cluster, then a general formulation of microaggregation with g clusters and a given k is as follows:

$$Minimize \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij} (d(x_j, v_i))^2 \tag{2.3}$$

$$Subject\ to:$$

$$\sum_{i=1}^{g} u_{ij} = 1, \quad \forall j = 1, \ldots, n \tag{2.4}$$

$$2k \geq \sum_{j=1}^{n} u_{ij} \geq k, \quad \forall i = 1, \cdots, g \tag{2.5}$$

$$u_{ij} \in \{0, 1\}. \tag{2.6}$$

For numerical data it is usual to require that $d(x, v)$ is the Euclidean distance. In the general case, when attributes $\mathbf{V} = (V_1, \ldots, V_s)$ are considered, x and v are vectors and d becomes $d^2(x, v) = \sum_{V_i \in \mathbf{V}} (x_i - v_i)^2$. In addition, it is also common to require for numerical data that v_i is defined as the arithmetic mean of the records in the cluster. That is,

$$v_i = \frac{\sum_{j=1}^{n} u_{ij} x_i}{\sum_{j=1}^{n} u_{ij}}.$$

In the rest of this section we will explain three different algorithms that have been proposed for microaggregation and which we will use them later on, in the experimental analysis of Chapter 5.

- *Individual Ranking* (MicIR): Each attribute is grouped independently to the other attributes.
- *Multivariate Ranking*: All attributes (or subsets of attributes) are grouped together. As the solution of this problem is NP-Hard [Oganian and Domingo-Ferrer, 2001] when we consider more than one variable at a time (multivariate microaggregation), several heuristic methods can be found in the literature.
    - *Projection microaggregation*: The multivariate data first are ranked by projecting them onto a single axis and then are aggregated into groups of at least k. In Chapter 5, we have considered these two projection variants:
        - *Z-scores projection (MicZ).*
        - *Principal component projection (MicPCP).*
    - *Heuristic microaggregation*: We have considered the numerical implementation of the *MDAV* algorithm algorithm [Domingo-Ferrer and Mateo-Sanz, 2002]:
        - *Maximum Distance to Average Vector (MDAV)*: this is a heuristic algorithm for clustering record in a data set X with n records so that each cluster is constrained to have at least k records. This algorithm can be used for univariate and multivariate microaggregation. It is described in more detail in Algorithm 1.

Table 2.2 shows an example of applying microaggregation using the MDAV algorithm with $k = 2$. This algorithm is performed on the numerical attributes *Age* and *Salary* from the microdata example, see Table 2.1.

The rationale behind microaggregation is that privacy is achieved ensuring that all clusters have at least a predefined number of elements, say k. Subsequently, all records in a specific cluster are replaced by the corresponding computed representative (centroid) record. Therefore, if we consider a multivariate microaggregation using all possible data attributes at the same time, then the resultant masked data there will have at least k indistinguishable records. This fact leads us to review k-anonymity [Samarati, 2001, Sweeney, 2002].

Definition 2.15. *A masked dataset Y is said to satisfy k-anonymity if for every individual record $b_i \in Y$, there are at least $k-1$ individual records equals to b_i.*

Multivariate microaggregation is a method that permits to achieve k-anonymity when it is applied to all attributes at the same time. Note that when subsets of attributes are microaggregated independently (as is the case of individual ranking) k-anonymity is not ensured.

2.4.2 Rank Swapping

Rank swapping is another of the most popular perturbative methods. Although the idea of swapping data values for disclosure control comes

```
Data: $X$: original data set, $k$: integer
Result: $X'$: protected data set
begin
    while ($|X| \geq 3 * k$) do
        Compute average record $\bar{x}$ of all records in $X$;
        Consider the most distant record $x_r$ to the average record $\bar{x}$;
        Form a cluster around $x_r$. The cluster contains $x_r$ together with
        the $k - 1$ closest records to $x_r$;
        Remove these records from data set $X$;
        Find the most distant record $x_s$ from record $x_r$;
        Form a cluster around $x_s$. The cluster contains $x_s$ together with
        the $k - 1$ closest records to $x_s$;
        Remove these records from data set $X$;
    end
    if ($|X| >= 2 * k$) then
        Compute the average record $\bar{x}$ of all records in $X$;
        Consider the most distant record $x_r$ to the average record $\bar{x}$;
        Form a cluster around $x_r$. The cluster contains $x_r$ together with
        the $k - 1$ closest records to $x_r$;
        Remove these records from data set $X$;
    end
    Form a cluster with the remaining records;
end
```

Algorithm 1: Maximum Distance to Average Vector algorithm.

from [Dalenius and Reiss, 1982], it was not until 1996, when [Moore, 1996] proposed this technique. This was first designed for ordinal attributes, but it can also be used for numerical attributes.

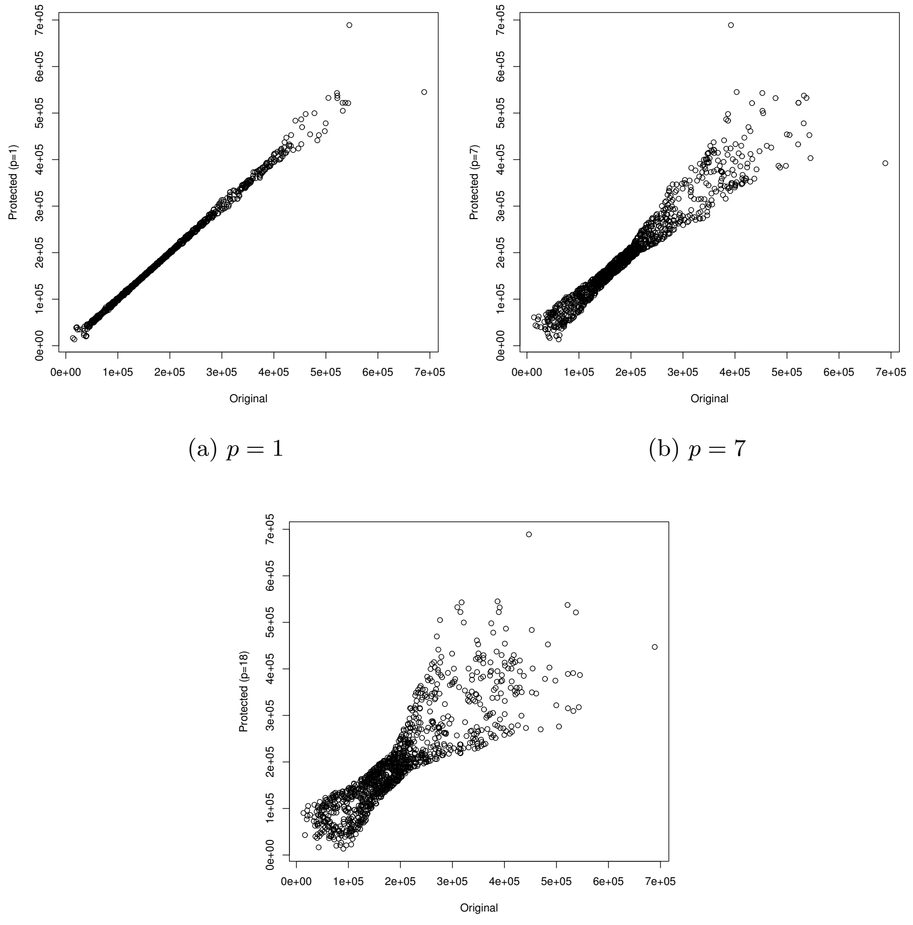
Rank swapping is a univariate masking method that leaves means and variances unchanged but may seriously affect the correlation structure of the data. We can define its process with respect to a parameter p and one attribute column, V_j^X of the original data X as follows:

- All the records values of a variable V_j^X are ranked in ascending order. That is, $V_j^X(a_i) \leq V_j^X(a_l)$ for all $1 \leq i < l \leq n$.
- Each value $V_j^X(a_i)$ is randomly and uniformly swapped with another value $V_j^X(a_l)$ chosen from a restricted range $i < l \leq i+p$. Hence, the rank of two swapped values cannot differ by more than p percent of the total number of records.

The parameter p is used to control the swap range. That is, when p increases, the difference between $V_j^X(a_i)$ and $V_j^X(a_l)$ may increase accordingly. This fact makes re-identification more difficult, but it also produces a higher loss of information.

Original Microdata		Protected Microdata	
Age	Salary	Age	Salary
23	25,000	34.5	27,500
18	10,000	18.0	10,000
58	12,000	40.5	13,000
46	30,000	34.5	27,500
18	10,000	18.0	10,000
23	14,000	40.5	13,000

Table 2.2: Microaggregation example using MDAV with $k = 2$.



(a) $p = 1$

(b) $p = 7$

(c) $p = 18$

Figure 2.5: Comparing protected and non-protected values for variable V_1 using Rank Swapping for different values of p.

Original Microdata		Protected Microdata	
Age	Salary	Age	Salary
23	25,000	18	30,000
18	10,000	23	12,000
58	12,000	18	25,000
46	30,000	58	10,000
18	10,000	23	12,000
23	14,000	18	10,000

Table 2.3: Rank swapping with a percentage of 40% probability of swapping.

To show how the protection methods affect or distort the original data, we provide three plots in Figure 2.5 which compare the protected and original values of variable V_1 using the Rank Swapping protection method. The original value is shown (axis x) versus the protected value (axis y) for different values of the parameter p.

Finally, Table 2.3 shows an example of applying rank swapping with a percentage of 40% probability of swapping. This algorithm is performed on the numerical attributes *Age* and *Salary* from the microdata example, see Table 2.1.

2.4.3 Additive Noise

Microdata protection by adding noise is being discussed since several years ago. During this period several algorithms implementing different characteristics have been developed. Simpler algorithms consist of adding white (i.e., *Gausian*) to data. More sophisticated methods rely on adding random noise with the same correlation structure as the original unmasked data, others use different transformations of the data and complex error-matrices to improve the results. Some works give an overview over different noise addition algorithms and discus their properties in terms of protected data utility and the protection level [Brand, 2002, Domingo-Ferrer et al., 2004].

Additive noise consists of adding Gausian noise completely at random to each variable according to a normal distribution $N(0, p\sigma)$, where σ is the standard deviation of the original data, and p is the parameter of the algorithm indicating the amount of noise in percentage added to the unmasked data.

Table 2.4 shows an example of applying additive noise with correlated noise. This algorithm is performed on the numerical attributes *Age* and *Salary* from the microdata example, see Table 2.1.

2.5 Protected Microdata Assessment

In the context of privacy preserving data mining, the goal of every protection method is to minimize the disclosure risk (DR) as well as the amount of information loss (IL). Due to the relevance of these two concepts many researchers have

Original Microdata		Protected Microdata	
Age	Salary	Age	Salary
23	25,000	26	26,461
18	10,000	17	9,785
58	12,000	55	12,995
46	30,000	48	30,831
18	10,000	17	9,038
23	14,000	22	14,494

Table 2.4: Additive noise with correlated noise (rounding decimal values).

put their efforts into the development of accurate measures. These two concepts are inversely proportional. That is, when the information loss decreases, the disclosure risk increases and vice versa. Therefore, the task of finding the optimal trade-off between these two concepts is difficult and challenging.

Having such measures any protection practitioner could analyse and evaluate the protection level and the statistical utility of their protected data. As data uses are very diverse and it may be hard to identify all data uses at the release moment, so it is desirable that these measures are developed in a generic way. Then protection practitioners could decide if the obtained measure values are appropriate for their data and its purposes before its releasing.

In the following sections we present the most common information loss and disclosure risk measures. Finally, it is a generic score to evaluate protected data sets.

2.5.1 Information Loss

Information loss is described as the difference between the analytical structures of the original and the protected datasets. Then, small information loss values mean that both analytical structures are very similar. The fact of preserving the original dataset structure in the protected dataset is to ensure that it will be analytically valid and statistically interesting.

It is desirable that this measure is as generic as possible, because protection practitioners will not know the intended use of the released data. Following this statement some different approaches were proposed. [Domingo-Ferrer and Torra, 2001a] calculated the average difference between some statistics computed on both original and protected microdata. This set of measures is reviewed below.

Assume a microdata file of N individuals and n attributes. Let X be the matrix representing the original microdata set and Y be the matrix representing the masked data. Recall, matrix records correspond to the individuals and the columns to the attributes. Components of matrices are represented by the corresponding lower case. For example, $x_{i,j}$ is the j-th attribute of the i-th individual. The following information loss measures have been described,

- Mean variation between the original matrix X and the protected one Y.

$$IL_1 = \frac{\sum_{j=1}^{n} \sum_{i=1}^{N} \frac{|x_{ij} - y_{ij}|}{|x_{ij}|}}{Nn}$$

- Mean absolute between the average of attributes.

$$IL_2 = \frac{\sum_{j=1}^{n} \frac{|\overline{x_j} - \overline{y_j}|}{|\overline{x_j}|}}{n}$$

  where $\overline{x_j}$ is the average of all values of the j-th attribute, i.e., $\overline{x_j} = \frac{1}{N} \sum_{i_1}^{N} x_{ij}$.

- Mean variation between the covariance matrices, VX and VY, of the original matrix X and the protected one Y, respectively.

$$IL_3 = \frac{\sum_{j=1}^{n} \sum_{1 \leq i \leq j} \frac{|vx_{ij} - vy_{ij}|}{|vx_{ij}|}}{\frac{(n+1)n}{2}}$$

- Mean variation between the attribute variances of the original matrix X and the protected one Y. These values are the diagonal of both covariance matrices, VX and VY.

$$IL_4 = \frac{\sum_{j=1}^{n} \frac{|vx_{jj} - vy_{jj}|}{|vx_{jj}|}}{n}$$

- Mean variation between the correlation matrices, RX and RY, of the original matrix X and the protected one Y, respectively.

$$IL_5 = \frac{\sum_{j=1}^{n} \sum_{1 \leq i \leq j} \frac{|rx_{ij} - ry_{ij}|}{|rx_{ij}|}}{\frac{(n-1)n}{2}}$$

In [Domingo-Ferrer and Torra, 2001b] prosed the overall information loss measure by averaging all these means variations and then multiplying the result by 100. That is,

$$IL = 100 \left(\frac{IL_1 + IL_2 + IL_3 + IL_4 + IL_5}{5} \right) \tag{2.7}$$

A probabilistic variation of these measures were presented in [Mateo-Sanz et al., 2005] to ensure that information loss value was always within the interval $[0, 1]$.

Furthermore, there are information loss measures for specific masking methods. A clear example is microaggregation, which is commonly evaluated using a set of intra and inter cluster sum of squares errors [Domingo-Ferrer and Mateo-Sanz, 2002, Jero, 2012]. In clustering, this is a well known criteria to measure the clusters' quality.

The microaggregation algorithm aims to find centroids that minimize the inertia, or intra-cluster sum of squared errors, see its previously defined objective functions in Equation (2.3). Therefore, assuming that microaggregation uses the Euclidean distance, it is straightforward to measure the information loss in terms of the Sum of Squares Errors (SSE) to evaluate the homogeneity of the with-in clusters, the Sum of Absolute Square Errors (SSA) to evaluate the between-groups cluster and the Total Sum of Squares (SST), that are defined as follows:

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} ((x_{ij} - \bar{v}_i)^T (x_{ij} - \bar{v}_i)) \tag{2.8}$$

where g is the number of groups and n_i the number of records in the ith group, so $n_i \geq k$ and $n = \sum_{i=1}^{g} n_i$. Note, $(x_{ij} - v_i)^T$ is the transposed of $(x_{ij} - v_i)$. In the same way x_{ij} is the jth record in the ith group and $\bar{v}_i$ denotes the average data record over the ith group (cluster representative or centroid). The lower SSE, the higher the within-group homogeneity.

$$SSA = \sum_{i=1}^{g} n_i ((\bar{v}_i - \bar{v})^T (\bar{v}_i - \bar{v})) \tag{2.9}$$

where $\bar{v}$ is the average record over the whole set of n vectors. The higher SSA, the lower the between-groups homogeneity.

$$SST = SSE + SSA = \sum_{i=1}^{g} \sum_{j=1}^{n_i} ((x_{ij} - \bar{v})^T (x_{ij} - \bar{v})) \tag{2.10}$$

The last measure is the normalized information loss, defined as the relation between the sum of squares of with-in group and the total sum of squares. That is,

$$IL = \frac{SSE}{SST} \tag{2.11}$$

The optimal p-partition is defined by the one that minimizes the SSE measure (i.e., maximizes the within-group homogeneity) and maximizes the SSA measure (i.e, minimizes the between-group homogeneity).

2.5.2 Disclosure Risk

Disclosure risk is a way to evaluate the protection degree of a masked microdata set. Then, high disclosure risk values will mean that the masking method performed on the original data has introduced a weak protection layer. Hence, an attack performed on the masked data could provide original sensitive information to that intruder.

There are different approaches for measuring disclosure risk. There are some analytical and some empirical measures. Some of them are described below.

The *Interval Disclosure* (ID) [Domingo-Ferrer and Torra, 2001b] is an example of an analytical approach. This is the average percentage that an attacker is able to identify an original value within an interval defined around the corresponding masked value. More details of this measure are given below.

For a given masked record r the computation of rank intervals is as follows: For each attribute it is defined a rank interval around the value the attribute takes on the given record. The center value of each rank interval should correspond to the value of the attribute in the record r, and these rank intervals should differ less than p percent of the total number of records. If so, the proportion of original values that fall into the interval centered around their corresponding masked value is a measure of disclosure risk.

A way to evaluate empirically the disclosure risk is to use record linkage between the protected file and the original data file. In particular we can define a measure of the risk as a percentage of correctly linked records using record linkage approaches. For this purpose we can use the methods explained in Section 2.2: Distance Based Record Linkage (DBRL) and Probabilistic Record Linkage (PRL).

Using record linkage for disclosure risk assessment follows Figure 2.2. In this case, the intruder knows who is in the original microdata set, and all the information related with all these individual's attributes. The information of the intruder can be seen as a subset of the original file. In the worst case, the intruder knows all data in the original file. The original and protected files correspond respectively to files X and Y in Figure 2.2. In addition, the intruder also knows all the information related to the masking process. That is, method and parameters. Therefore, in this scenario we use record linkage to find correct links between the whole original dataset X and its protection version Y. When all these conditions hold the re-identification scenario is known as the *worst case* scenario.

A final disclosure risk measure is computed by combining the results of all presented disclosure risk approaches (ID, DBRL and PRL). This is formalized as follows,

$$DR = \frac{\left(\frac{DBRL+PRL}{2}\right) + ID}{2} \tag{2.12}$$

2.5.3 Generic Score

A combination of information loss and disclosure risk measures was introduced in [Domingo-Ferrer and Torra, 2001b] to evaluate different protected data sets generated from a single data set. This is,

$$Score(X,Y) = \frac{IL(X,Y) + DR(X,Y)}{2} \tag{2.13}$$

The authors used this score to evaluate different protection methods and present a ranking of such methods.

An alternative way to illustrate the trade-off is the R-U confidentiality map [Duncan et al., 2001], which provides a graphical representation of disclosure risk (R) and information loss (U). In these maps, the outcomes of alternative

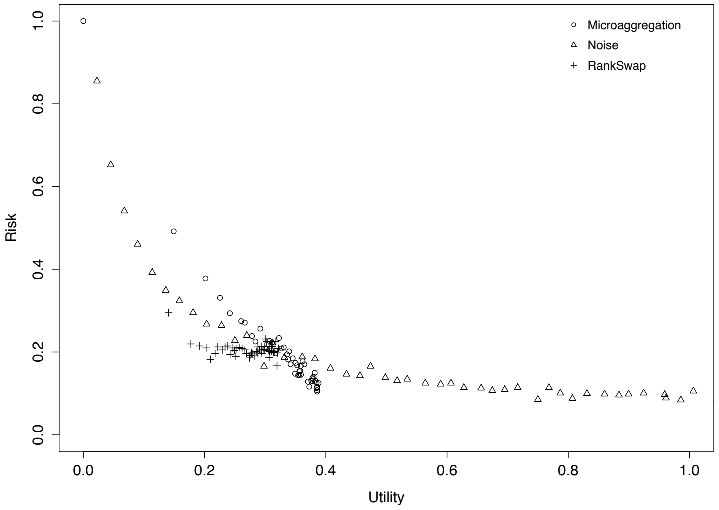


Figure 2.6: R-U map for microaggregation, rank swapping and additive noise.

protection methods are visualized in a 2-dimensional plot, where one axis represents the disclosure risk and the other axis the information loss. Note also that the data sets close to the $(0, 0)$ point are those that provide less information loss and disclosure risk, and thus, the best ones.

Figure 2.6 is an example of an R-U map for different parametrizations of microaggregation (Section 2.4.1), rank swapping (Section 2.4.2) and additive noise (Section 2.4.3) methods. These maps are easy to visualize the balance between information loss and disclosure risk

2.6 Metric Learning

During the last years, some researchers have noticed that the poor results obtained by a set of different machine learning algorithms were due to the use of simple distances, such as the standard Euclidean distance. The main weakness of this distance is that it treats all data features equally, independently of their relations, and thus, it fails to exploit the structure information which is embedded in the analyzed data. Because of this fact, the scientific community researched and proposed new methods to automate and learn task-specific distance functions in a supervised manner. In this section we briefly review some state-of-the-art approaches that have been proposed for distance metric learning and how these methods can improve the performance of different machine learning algorithms such as classification, clustering and regression.

Broadly speaking, metric learning is the task of learning a pairwise real-valued metric functions of the problem of interest, using the information brought by the training examples. Most methods learn the parameters of a metric in a weakly-supervised way from pair or triplet-based constraints of the following form:

- *Must-link / cannot-link* constraints (also called *positive / negative pairs*):
  - $\mathcal{S} = (a, b)$: a and b should be similar,
  - $\mathcal{D} = (a, b)$: a and b should be dissimilar.
- Relative constraints (sometimes called training triplets):
  - $\mathcal{R} = (a, b, c)$: a should be more similar to b than to c.

The metric learning problem is typically formulated as an optimization problem, which given a set of constraints, $(\mathcal{S}, \mathcal{D}, \mathcal{R})$, and an objective function that usually incurs a penalty when the constraints are violated, finds the metric parameters such that it best agrees with the training constraints. Depending on the task the sets of constraints will be more or less complex. For example, one of the most studied metric learning problems is the one that just considers pair-wise metrics, so the set of given constraints consists of must-link or cannot-link constraints, $(\mathcal{S}, \mathcal{D})$. However, there are more complex problems that also consider other kinds of side information like relative comparison such as $(\mathcal{S}, \mathcal{D}, \mathcal{R})$.

The origins of metric learning can be traced earlier in [Hastie and Tibshirani, 1996, Friedman, 1994]. This literature is very wide so, we review some of the most important contributions related with our research. [Xing et al., 2003] parameterize the Euclidean distance using a symmetric positive semi-definite matrix $\Sigma \succeq 0$ to ensure the non-negativity of the metric. That is,

$$d_\Sigma(a, b) = \sqrt{(a - b)'\Sigma(a - b)}$$

Note that setting $\Sigma = I$, the identity matrix, results the Euclidean distance, and if we restrict Σ to be diagonal, this corresponds to learning a metric in which the different axes are given different weights, i.e, we are learning the parameters of the weighted mean. More generally, Σ parameterizes a family of Mahalanobis distances over $\mathbb{R}^n$. Learning this distance metric is equivalent to finding a rescaling of the data that replaces each point a with $\Sigma^{1/2}a$ and applying the standard Euclidean distance to that rescaled data.

Current methods for distance metric learning can be roughly divided into two categories: unsupervised and supervised metric learning.

Most of the unsupervised distance metric learning approaches attempt to find low-dimensions embedding from high-dimensional input spaces. Some well known linear examples are the Principal component analysis (PCA) [Fukunaga, 1990] and Euclidean Multidimensional scaling [Cox and Cox, 1994], while ISOMAP [Tenenbaum et al., 2000] is also a well known non-linear dimensionality reduction approach, it aims to preserve the geodesic distance for all data points.

Supervised distance metric learning can be subdivided into two categories as follows:

- *Fully-supervised*: the metric learning algorithm has access to a set of labeled training instances, where each training example is a tuple, composed of an instance and a label or class. In practice, the label information is often used to generate specific sets of pair/triplet constraints $\mathcal{S}, \mathcal{D}, \mathcal{R}$ for instance based on a notion of neighborhood. Some earlier works that optimize the metric with class labels for classification tasks are [Hastie and Tibshirani, 1996] and [Goldberger et al., 2004]. Recently, in [Weinberger et al., 2006] the authors proposed a new classification algorithm, the Large Margin k-nearest neighbour (LMNN), in which a Mahalanobis distance is learnt. This metric is trained with the goal that the k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin. However, [Sun and Chen, 2011] show that LMNN cannot satisfactorily represent the local metrics which are respectively optimal in different regions of the input space and they propose a local distance metric learning method, a hierarchical distance metric learning for LMNN, which first groups data points in a hierarchical structure and then learns the distance metric for each hierarchy.

- *Weakly-supervised*: the metric learning algorithm has no access to the individual labels of the training instances. The algorithm is only provided with side information in the form of sets of constrains $\mathcal{S}, \mathcal{D}, \mathcal{R}$. This label information at the pair/triplet level is meaningful for application where labeled data is costly to obtain or it is not available as happens in clustering tasks, and so it is easy to get this such side information. In this context many approaches exploiting pairwise-link constraints have focused on using the generalizations of the Mahalanobis metric. As was mentioned above one of the first works was introduced by [Xing et al., 2003]. They presented an algorithm that maximizes the sum of distances between dissimilar points, while keeping closer the set of distances between similar points. This was formalized as follows,

$$Maximize_{\Sigma} \sum_{(x_i,x_j)\in\mathcal{D}} d_{\Sigma}(x_i,x_j) \tag{2.14}$$

$$Subject\ to:$$

$$\sum_{(x_i,x_j)\in\mathcal{S}} d^2_{\Sigma}(x_i,x_j) \geq 1 \tag{2.15}$$

$$\Sigma \succeq 0 \tag{2.16}$$

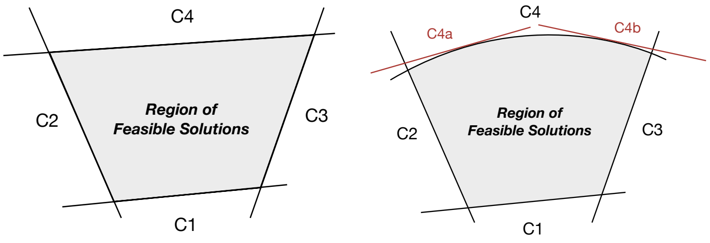
where $\mathcal{D}$ and $\mathcal{S}$ is the sets of dissimilar and similar pairs, respectively.

The algorithm proposed to solve the problem consists of an iterative gradient ascend step to optimize Equation (2.14), followed by an iterative projection step to ensure that Constraints (2.15) and (2.16) hold. Specially, the projection to ensure the latter constraint, Equation (2.16), is performed setting negative eigenvalues to 0. However, despite its simplicity, the method is not a scalable problem, because it has to perform many eigenvalue decompositions. [Schultz and Joachims, 2004] proposed a method for learning distance metrics from relative comparisons such as a is closer to a than a is to c. This relies on a less general Mahalanobis distance learning in which for a given matrix a, only a diagonal matrix W is learnt such that $\Sigma = A'WA$. More recently, [Halkidi et al., 2008] proposed a framework for learning the weighted Euclidean subspace based on pairwise constrains and cluster validity, where the best clustering can be achieved. Other types of metrics were contemplated, e.g., [Beliakov et al., 2011] considered the problem of metric learning in semi-supervised clustering defining the Choquet integral with respect to a fuzzy measure as an aggregation distance. The authors investigate necessary and sufficient conditions for the discrete Choquet integral to define a metric.

More detailed information can be found in [Bellet et al., 2013], a metric learning survey. Besides, introducing the topic and detailing some of the most important metric learning algorithm, they also identify and describe five key properties of metric learning algorithms which were used to provide a taxonomy of the different methods in the literature.

One of the most important challenges associated with supervised metric learning approaches, specially in Mahalanobis-based distances is the satisfaction of the positive semi-definiteness. In the literature there are different approximations, from several matrix simplifications to modern *semi-definite programming* methods within the operations research field. Some Σ simplifications force it to be diagonal and so Σ is positive semi-definite if and only if all diagonal entries are non-negative. This simplification reduces the number of parameters drastically and makes the optimization problem a linear program. In [Higham, 2002] Higham proposed an algorithm to find the nearest correlation matrix, symmetric positive semi-definite matrix with unit diagonal, to a given symmetric matrix by means of a projection from the symmetric matrices onto the correlation matrices, with respect to a weighted Frobenius form. Semi-definite programming (SDP), is a kind of convex programming which evolved from linear programming. While, a linear programming problem is defined as the problem of maximizing or minimizing a linear function subject to a set of linear constraints, semidefinite programming is defined as the problem of maximizing or minimizing a linear function subject to a set of linear constraints and a “semi-definite” constraint, a special form of non-linear constraints. Therefore, the semi-definite constraint is what differentiates SDPs from LPs. In addition, whereas in LP the solutions feasible region is a convex polyhedron, i.e., the intersection of all defined linear constraints, in SDP the non-linear constraint produces a non-flat face in the solutions feasible region. Figure 2.7 shows an example of LP and SDP feasible regions. Interestingly, this non-linear constraint can be interpreted as an infinite number of linear constraints in the solutions feasible region.

Two well known techniques are used to solve these problems. Whereas Simplex algorithm is commonly used to solve linear problems, interior point algo-



(a) Linear problem.

(b) Semi-definite problem.

Figure 2.7: Example of LP and SDP feasible regions.

rithms use a different way to solve linear and nonlinear problems. We briefly describe them below.

In [Dantzig, 1951] it was introduced *Simplex*, an algorithm which remains widely used today by many optimization solvers. Broadly speaking, the simplex algorithm proceeds by going from one vertex to another of the polyhedron defined by the problem constraints. At each step, it goes to a vertex that is better with respect to the objective function. The algorithm will either determine that the constraints are unsatisfiable, determine that the objective function is unbounded, or reach a vertex from which it cannot make progress, which necessarily optimizes the objective function.

The Interior point method are radically different from the simplex methods. They start identifying a feasible trial solution. At each iteration, they move from the current trial solution to a better trial solution within the feasible region. They continue the process until reaches a trial solution that is optimal. The biggest difference between Simplex and interior point methods lies in the nature of these trial solutions. While Simplex movement is along edges on the boundary of the feasible region, interior point methods move along points in the feasible region. [Karmarkar, 1984] is an example of algorithm to solve linear programming problems, but they can be extended to solve nonlinear problems. See [Luenberger and Ye, 2008] for further details.

2.7 Unstructured Text Data Protection

Data privacy has become more important in recent years because of the increasing ability to store individuals' data. This ease of collecting individuals data is because the Cloud (Internet services) is used as a pervasive communication system. Not only generic information and public information is exchanged, but also private and confidential information is published. This information is normally intended for a reduced audience with appropriate clearance. For instance, one may consider internal collaborative reports in a corporate Intranet, personal electronic health records, law-suits, research project proposals, etc. This is compounded by the increasing sophistication of data mining algorithms to leverage this information.

In this section we review the main topics related work to the protection of unstructured textual data. For the sake of simplicity we have divided the problem of unstructured textual data protection into two areas.

On the one hand, and probably the best known, there is the text document *sanitization* (also known as *redaction*) field. This consists of a set of techniques that attempt to reduce the document's classification level, possibly yielding an unclassified document, by allowing the selective disclosure of information in a document while keeping other parts of the document secret. On the other hand, there is text anonymization while permitting text mining field. The primary task of text mining, similarly to data mining, is the development of models that are analyzed to discover patterns or relationships between variables. Then, the goal of text anonymization is to generate anonymized text models from the original data in order they can be released and analyzed by third parties.

In the following sections we review the state-of-the-art as well as some related background knowledge.

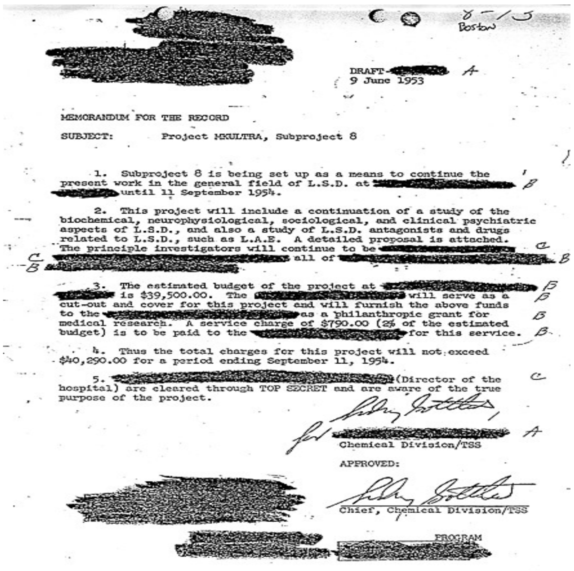
2.7.1 Data Sanitization

There are several reasons for sanitizing a document. Each government has its own protocols and laws explaining how to declassify documents by removing their sensitive information before releasing them. In hospitals, medical records have also to be sanitized to cloak patients identity information or diagnoses of deadly diseases, etc. Companies have the need to sanitize their documents, for instance, to prevent inadvertent disclosure of proprietary information while sharing data with outsourced operations.

Example 1. *Figure 2.8 is an example of a US government document that has been manually sanitized prior to release. In recent years there have been many efforts to automate or help people to perform the anonymization process saving time and getting more accurate results.*

Traditionally, documents were sanitized manually, which meant a slow, tedious and inefficient process and if we also consider the amount of digital textual information made available daily, one can realize of the need of automatic text sanitization methods. The US Department of Energy's OpenNet initiative [DOE/OSTI, 2014] which requires of sanitizing millions of documents yearly or the use of Internet information services, are a couple of examples of this amount of shared information. Additionally, it is foreseeable to expect unauthorized copies or the release of classified information. In fact, the international non-profit organization, WikiLeaks [Wikileaks, 2010], has been publishing thousands of classified information (about military and diplomatic issues) of many countries of the world.

The importance of this problem has attracted the attention of some international agencies requesting for new technologies to support declassification of confidential documents. E.g., the *DARPA*, the Defense Advanced Research Projects Agency of the United States Department of De-



[redacted]

8-15
Boston

DRAFT-[redacted] A
9 June 1953

MEMORANDUM FOR THE RECORD

SUBJECT: Project MKULTRA, Subproject 8

1. Subproject 8 is being set up as a means to continue the present work in the general field of L.S.D. at [redacted] until 11 September 1954. B

2. This project will include a continuation of a study of the biochemical, neurophysiological, sociological, and clinical psychiatric aspects of L.S.D., and also a study of L.S.D. antagonists and drugs related to L.S.D., such as L.A.E. A detailed proposal is attached. The principle investigators will continue to be [redacted] all of [redacted]. C B

3. The estimated budget of the project at [redacted] is $39,500.00. The [redacted] will serve as a cut-out and cover for this project and will furnish the above funds to the [redacted] as a philanthropic grant for medical research. A service charge of $790.00 (2% of the estimated budget) is to be paid to the [redacted] for this service. B

4. Thus the total charges for this project will not exceed $40,290.00 for a period ending September 11, 1954.

5. [redacted] (Director of the hospital) are cleared through TOP SECRET and are aware of the true purpose of the project. C

for [redacted] A
Chemical Division/TSS

APPROVED:

Chief, Chemical Division/TSS

[redacted]

PROGRAM

Figure 2.8: Sanitization example (source: Wikipedia).

fense [DARPA, 2010] or the *CHIR*, the Consortium for Healthcare Informatics Research [Meystre et al., 2010].

Contrary to microdata anonymization, raw text does not necessarily contain explicitly identified sensitive attributes, e.g., identifiers and quasi-identifiers. Therefore, document sanitization consists of two tasks: (i) The detection of sensitive data within the text; and then, (ii) the hiding or deletion of such detected information with the aim of minimizing the disclosure risk, while causing the least distortion to the document content. We briefly comment some state-of-the-art sanitization works below.

[Chakaravarthy et al., 2008] present *ERASE* (Efficient RedAction for Securing Entities) system for the automatic sanitization of unstructured text documents. The system prevents disclosure of protected entities by removing certain terms from the document, which are selected in such a way that no protected entity can be inferred as being mentioned in the document by matching the remaining terms with the entity database. Each entity in the database is associated with a set of terms related to the entity; this set is termed the context of the entity.

[Cumby and Ghani, 2011] present a privacy framework for protecting sensitive information in text data, while preserving known utility information. The authors consider the detection of a sensitive concept as a multi-class classification problem, inspired in feature selection techniques, and present several algorithms that allow varying levels of sanitization. They define a set D of documents, where each $d \in D$ can be associated with a sensitive category $s \in S$, and with a finite subset of non-sensitive utility categories $U_d \subset U$. They define a privacy level similar to k-anonymity [Sweeney, 2002], called k-confusability, in terms of the document classes.

[Hong et al., 2011] present a heuristic data sanitization approach based on term frequency and inverse document frequency (commonly used in the text mining field to evaluate how significant a word in a corpus is to a document). [Samelin et al., 2012] present an RSS (redactable signature scheme) for ordered linear documents which allows for the separate redaction of content and structure. In [Chow et al., 2011] is presented a patent for a document sanitization method, which determines the privacy risk for a term by determining a confidence measure $c_s(t_1)$ for a term t_1 in the modified version of the document relative to sensitive topics s. In the context of the sanitization of textual health data, [Neamatullah et al., 2008] presents an automated de-identification system for free-text medical records, such as nursing notes, discharge summaries, X-ray reports, and so on.

In [Anandan et al., 2012], the authors focused on the protection of already detected entities, trying to preserve the utility of sanitized documents by means using an ontology to replace sensitive nouns with other words semantically more general. The authors use the WordNet [WordNet, 2010] ontology to generalize the entities (see Section 2.7.4 to know more about WordNet). They also introduce a measure, t-plausibility, to evaluate the quality of sanitized documents from a privacy protection point of view. That is, a protected document holds t-plausibility principle if at least t base documents can be generalized to such a sanitized document where a base text refers to one that has not been sanitized in any way.

2.7.2 Privacy Preserving Text Mining

In this section we review some related work concerning to the second line of research, text anonymization preserving text mining. This research line shares the same goals that the two established microdata anonymization disciplines, Statistical Disclosure Control (SDC) [Willenborg and Waal, 2001] and Privacy Preserving Data Mining (PPDM) [Agrawal and Srikant, 2000]. Building general models from unstructured textual data sets, developing anonymization techniques that preserve textual relations and properties and developing disclosure risk and data utility are the main research tasks.

Most of the research were geared towards the problem of anonymizing search engine query logs after the highly publicized AOL case in 2006 [Barbaro and Zeller, 2006]. For instance, [Navarro-Arribas et al., 2012] argued that removing some queries from the log does not preserve and acceptable privacy degree and they present a technique for query-anonymization, they ensure the k-anonymity in query logs by aggregating them. In the same direction, [Erola et al., 2010] introduced a variation of the microaggregation method, which enforced user k-anonymity by taking into account the semantic similarity between user queries relying on a hierarchical ontology, such as the Open Directory Project (ODP). The cluster representatives are made by selecting specific queries form each user in the group, that is, queries that are semantically close and/or that are in deeper ODP levels.

Other related works have been focused on semantic microaggregation variations by means of introducing external knowledge databases, such as WordNet [Miller, 1995]. The authors in [Martínez et al., 2012a, Martínez et al., 2012b] extend the microaggregation algorithm to support non-numerical (categorical) attributes defining a distance and an aggregation operator. They introduce a weighted semantic distance and an aggregation operator that integrates the distribution and the semantics of the data.

In [Liu and Wang, 2013] the authors focus on preserving privacy in publishing vocabularies, that is, very sparse bag-valued data extracted from a web query log. They extend the k-anonymity principle to ensure that every vocabulary for a given granularity is indistinguishable from at least $k-1$ other vocabularies. They call this principle $v$ocabulary k-anonymity. They propose a semantic similarity based on clustering for retaining enough data utility, relying on the minimum path distance over a semantic network, such as WordNet, between a pair of terms. Unlike previous authors, they substituted the terms with semantic similar terms, because they stated that for sparse data the generalization operation suffers from a high loss of information.

Part of this dissertation is focused on the protection of vector spaces (i.e., texts are mapped onto document-term matrices), which are supported by lots of traditional information retrieval and data mining analysis algorithms. These text mappings lead to very sparse and high-dimensional data matrices and, although the application of anonymization to vector spaces is recent, other researchers have also been focused towards the anonymization of high-dimensional spaces. [Ghinita et al., 2008] proposed an anonymization technique which combines the advantages of both generalization and permutation whose main idea is to first group closer transactions and then associate each group to a set of diversified sensitive values. [Lasko and Vinterbo, 2010] introduced the term *spectral anonymization* to refer to an algorithm that uses a spectral basis for anonymization instead of an original data. They also presented two spectral anonymization examples, one based on data swapping and the other based on Recursive Histogram Sanitization, a microaggregation method.

2.7.3 Document Representation Models

A bag-of-words is a simplified document representation that considers a document as an unordered collection of words. Take into account that unlike a set of words, a bag-of-words allow word duplicates. Neither the original word order nor the word syntax and grammar are considered. However, this fact makes the representation a very simple model, which can be easily extended by means of extracting different term-weights for each of the words in a document. The objective of such weighting representation schemes is to enhance discrimination between various document vectors and to enhance retrieval effectiveness [Salton and Buckley, 1988].

Each document can be represented as a term-weight vector and therefore the whole collection of unstructured texts can be represented as a document-term matrix, where the rows are the documents and the columns are the weights corresponding to the meaningfulness of each term in each document. This representation is the basis of the Vector Space Model (VSM) and it is commonly used in the information retrieval and computer vision areas due to the ease to compare vectors. The VSM is a simple model based on linear algebra, which allows the computation of continuous distances between documents, ranking them in order of relevance to a given query, partial matching, etc.

In the next section we present some basic document preprocessing methods in order to represent each document as a bag of words. From the set of bag of words and a weighting scheme we are able to represent all documents into a term-document matrix, i.e, vector space model.

Document processing

Figure 2.9 shows the document preprocessing step. This is composed of a set of tasks and techniques. Among all of those tasks we just have considered the most extended measures used in the Information retrieval and text analysis field.

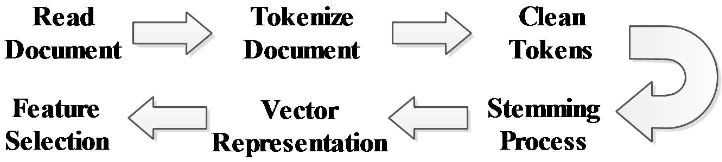


Figure 2.9: Document preprocessing techniques.

In order to represent each document as a bag of words, each documents should be read and tokenized into individual words. However, not all the tokens included in a document are useful when we want to perform text classification or other information retrieval techniques. Thus, a cleaning process is needed. This process eliminates all those tokens which are considered not useful such as numbers, punctuation symbols and some words. An example of useless words in text mining are language specific functional words which carry no information. Some of them are pronouns, prepositions or conjunctions. In text analysis these words are called *stop words*. At the end of this cleaning process a token normalization is performed. This consists in reducing all remaining words to upper or lower case.

Once the bag of words is cleaned, we can use word reduction techniques. One of the most extended techniques is the stemming, such as the Porter algorithm [Porter, 1980], an English stemming algorithm. This algorithm considers all words with the same stem as the same word, producing a reduction in the size of the feature set. The purpose of this method is to remove various word's suffixes to find the stem/root word. Thus, we are able to reduce the number of words just considering words with the same matching stem. Saving memory space and time are two of the most clear advantages. Figure 2.10 shows an example of stemming. Note that these feature reduction measures are not compulsory, so data owners should decide its application depending on their data sets and goals.

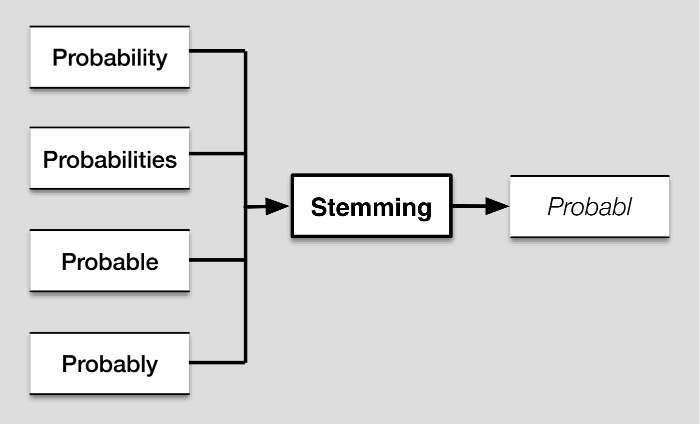


Figure 2.10: Example of stemming process.

The document vectorization is done by means of selecting a value/weight to represent the meaningfulness of each word within a document or collection of documents. There are different ways to measure the meaningfulness of words, such as a binary representation, the frequency of each word, the information gain, etc. In this dissertation we have used the term frequency-inverse document frequency, known as *tf-idf* [Manning et al., 2008a]. The *tf-idf* increases proportionally to the number of times a term appears in a document, but it is countered by the frequency that term appears in the text collection, also called corpus. Equation (2.17) shows the relation between the term frequency of a term in a single document and the frequency of the term in the corpus. That is,

$$\textit{tf-idf}(t, d, D) = tf(t, d) \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{2.17}$$

where t is a term of the corpus, d is a document and D is the set of documents. $tf(t, d)$ denotes the raw frequency of a term in a document.

Generally, in the VSM models it is beneficial to abstract out the magnitude of the term weights because it takes out the influence of the document length: only the relative weights across documents are important and not how big the document is. For that reason each document can be normalized to have unit norm, same direction but with length 1. That is, the division of each of the document vector, $\vec{d_i}$, by its L^2-norm:

$$\vec{v'} = \frac{\vec{d_i}}{\|\vec{d_i}\|} = \frac{\vec{d_i}}{\sqrt{\sum_j \vec{d_{ij}}^2}} \tag{2.18}$$

The last step is the feature selection. This is a useful step when we are dealing with a big corpus and a reduction of features can improve the model. Therefore, the features, i.e. all the vocabulary terms, are sorted in terms of the given weights and the most meaningful ones are selected while the others are discarded.

Finally, the set of all document vectors can be seen as a document-term matrix, where the rows represent each document and the columns are the corresponding term weights.

2.7.4 WordNet Database

WordNet [WordNet, 2010, Miller, 1995] is a general lexical database of the English language. It structures nouns, verbs, adjectives, and adverbs, into sets of cognitive synonyms called *synsets* which express concrete word concepts. Each synset is accompanied with its specific definition and a set of case examples illustrating its use. In addition, these synsets are interlinked by several conceptual-semantic and lexical relations. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules, and the relations between different types have some differences. The majority of the WordNets relations connect words of the same type. Thus, there are four subnets, one for each word type, with few relations

The biggest structure correspond to the nouns, in which there are the following synsets relations:

- Is-a relation: This is the most common relation. It states a super-subordinate relation between two synsets.
  - Hypernyms: a synset s_1 is a hypernym of s_2 if every s_1 is more generic than s_2.
  - Hyponyms: a synset s_1 is a hyponym of s_2 if s_2 is more specific than s_1.
- Part-whole relation: This relation states for words that are part of a larger whole.
  - Meronym: a synset s_1 is a meronym of s_2 if s_1 is a part of s_2.
  - Holonym: a synset s_1 is a holonym of s_2 if s_2 is a part of s_1.

Example 2. *Words will often have more than one sense. For instance, given the word* car*, these are the possible synsets returned by the WordNet database:*

1. *'Car_n.01': a motor vehicle with four wheels; usually propelled by an internal combustion engine.*
2. *'Car_n.02': a wheeled vehicle adapted to the rails of railroad.*
3. *'Car_n.03': the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant.*

4. 'Car_n.04': where passengers ride up and down

Figure 2.11 shows a small portion of the concept hierarchy from the first synset of the word car 'car_n.01'.

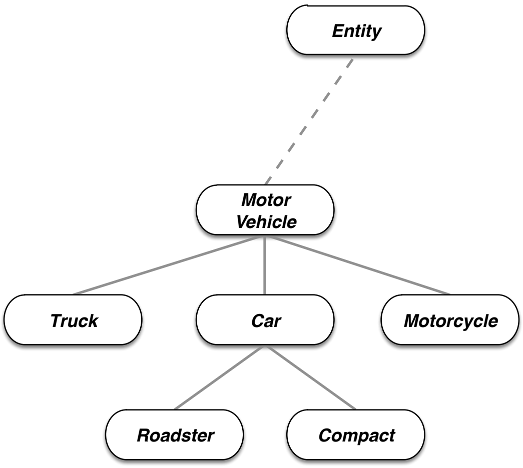


Figure 2.11: Fragment of WordNet concepts hierarchy.

Chapter 3

An Information Retrieval Approach to Document Sanitization

Motivated by the WikiLeaks scandal in which $115,000$ United States diplomatic documents considered as confidential or secrets were released, and also by the tedious, slow and inefficient processes to manually sanitize documents in this chapter we present a novel technique to assist document sanitization as well as a couple of measures to evaluate the quality of sanitized documents. These measures rely on the assessment of two known concepts in data privacy; the risk of sensitive information disclosure and the amount of lost (or removed) information.

In Figure 2.8 it was showed an example of a U.S. sanitized document. It is clear to see that the information was removed following some criteria and not randomly or perturbing the data with some kind noise. Thus, sanitization techniques need a previous definition of this information that should be considered sensitive. This information is usually imposed by the owner of the document such as companies or governmental organizations.

To evaluate our proposal we have selected a set of U.S. Confidential documents which were released by WikiLeaks [Wikileaks, 2010]. Therefore, to perform sanitization we have followed the criteria imposed by the U.S. laws. According to the United States government the documents are classified at four levels: Top secret, Secret, Classified and not confidential. These categories are assigned by evaluating the presence of information in a document whose unauthorized disclosure could reasonably be expected to cause identifiable or describable damage to the national security [E.O. 13526, 2009]. This type of information includes military plans, weapons systems, operations, intelligence activities, cryptology, foreign relations, storage of nuclear materials, and weapons of mass destruction. Note, some of this information is often directly related to national and international events which affect millions of people in the world, who in a democracy may wish to know the decision making processes of their elected representatives, ensuring a transparent and open government.

This chapter is organized as follows. Section 3.1 introduces a two-step semi-automatic method to assist sanitization of confidential unstructured textual data. Afterwards, in Section 3.2 we propose a mechanism to evaluate the information loss and the disclosure risk of a sanitized document. This mechanism relies on traditional information retrieval metrics which evaluates both the information loss and the risk of disclosure of a sanitized dataset, by means of query comparisons. Then, Section 3.3 presents the method followed to extract a set of WikiLeaks U.S. confidential documents as well as an empirical evaluation of this extracted set of selected documents. This evaluation is performed in terms of the proposed information loss and disclosure risk metrics. Finally, in Section 3.4 we summarize the work done and present some conclusions.

3.1 Sanitization Method

In this section, a supervised sanitization method based on entity recognition and pattern-matching techniques is presented. The purpose of this method is to identify and delete all those entities and sensitive information within classified documents that could disclose some sensitive information previously defined. Figure 3.1 depicts the two-step sanitization method. The main goal of the first step is to identify and anonymize all those terms considered as clear identifiers of certain individuals or places. The second step is focused on the identification of some established topics, that is, the identification of parts of the text that contain concepts considered as *risky*. These text parts will be later manually reviewed and eliminated. Both steps are explained in detail in Sections 3.1.1 and 3.1.2, respectively.

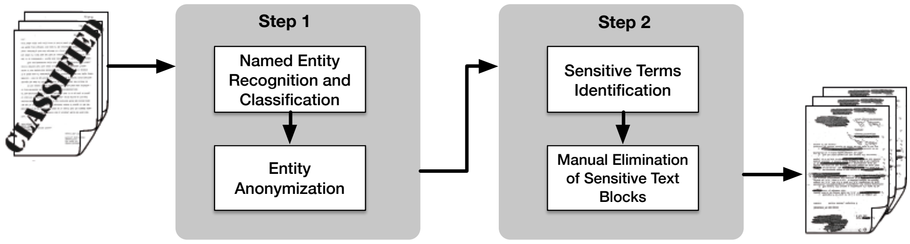

CLASSIFIED
Step 1
Named Entity Recognition and Classification
Entity Anonymization
Step 2
Sensitive Terms Identification
Manual Elimination of Sensitive Text Blocks

Figure 3.1: Scheme for document sanitization.

3.1.1 Step 1: Anonymization of Names and Personal Information of Individuals

To perform this first task we have used *Pingar* [http://www.pingar.com, 2014], a very powerful set of tools for text analytics. Its *Metadata Extractor* tool uses different techniques of natural language processing in order to detect, recognize and classify different entities such as organizations or companies, people, locations, addresses, e-mails, account numbers, dates, phone numbers and many other custom created entities. In addition, it is able to match similar terms, misspellings of words or equivalent spelling in different variations of English.

The entity anonymization is the second task of this step and it is also performed by using Pingar tools. This anonymization process is carried out by replacing the identified sensitive information by its category, provided by Pingar, plus an identification number. This identification number allow us to read the text distinguishing when an entity is mentioned, but without knowing the name of the entity, just the category. That is, $\{Pers_1, Pers_2, \cdots\}, \{Loc_1, Loc_2, \cdots\}, \{Date_1, Date_2, \cdots\}$ and so on. See Example 3.

Example 3. *Given the following piece of text,*
"Prof. Smith asked Imagine Inc. to start the project in New York, where the NY University could provide a laboratory near the Washington Square Park. Prof. Smith wants to study the birds of the Washington Square Park."

Then, the following sanitized text is obtained:
"$Pers_1$ asked Org_1 to start the project in Loc_1, where the Org_2 could provide a laboratory near the Loc_2. $Pers_1$ wants to study the birds of the Loc_2."

The names of countries (Iran, United States, Russia, Italy, etc.) and places (London, Abu Dhabi, Guantanamo, etc.) are unchanged in this process.

3.1.2 Step 2: Elimination of Text Blocks of *Risky Text*

This second step is divided in two sub-tasks; the identification of *risky* text blocks, which are those which contain the *risky* concepts, and the manual elimination of them. Unlike the first step, which hides/removes clear identifiers, such as personal information or locations, the goal of this second step, which is independent from the first step, is to detect and remove parts of the texts which contain risk terms. Due to the elimination of blocks of risk text, the main document information loss is incurred in this step.

The risk concepts are represented by 30 keywords extracted from Section 1.4 of *Executive Order 13526* [E.O. 13526, 2009]. This section includes eight points (a) to (h) defining the topics that the US government considers of risk in terms of national security. In Table 3.1 there is the list of the first 30 initial risk terms. As a list of 30 concepts are not enough to figure out if a text makes reference to any of the stated points and so we have used the WordNet ontology database [WordNet, 2010, Miller, 1995] to extend it. Thus, for each of these initial concepts we have extracted from WordNet database a set of new words related to its sense, such as synonyms and hyponyms (see Section 2.7.4). That is, words written different but with the same meaning and words whose semantic field is included within that given words. For example, the word *weapon* would give the following set of words {*knife, sling, bow, arrow, rock, stick, missile,*

ID_q	Keywords (risk queries)	ID_{Rd}	Categories, $a-h$, see [E.O. 13526, 2009]
rq_1	{military, plan, weapon, systems}	r_1	(a)
rq_2	{intelligence, covert, action, sources}	r_2	(b)
rq_3	{cryptology, cryptogram, encrypt}	r_3	(c)
rq_4	{sources, confidential, foreign, relations, activity}	r_4	(d)
rq_5	{science, scientific, technology, economy, national, security}	r_5	(e)
rq_6	{safeguard, nuclear, material, facility}	r_6	(f)
rq_7	{protection, service, national, security}	r_7	(g)
rq_8	{develop, production, use, weapon, mass, destruction}	r_8	(h)
rq_9	All terms → $\{rq_1, \cdots, rq_8\}$	r_9	$(a)-(h)$

Table 3.1: Queries used to test risk of disclosure extracted from point (a)-(h) of [E.O. 13526, 2009]. ID_q is the identification name for a specific set of keywords, and ID_{Rd} is the disclosure risk document set identification name.

cannon, gun, bomb, gas, nuclear, biological, $\cdots\}$. After extracting all synonyms and hyponyms from the given initial set we have obtained a list with a total of 655 risk terms (original + synonyms + hyponyms). We note that in this extraction process the word sense disambiguation was performed manually.

Afterwards, we process the documents and by using patter-matching techniques we identify all the text words that are in the list of risky words. Syntactic variations were also identified such as misspellings (*e-mail/e-amil*), singular or plurals (*e-mail/e-mails*), separators (*e-mail/email*). Additionally, we applied a stemming process (using the Porter Stemming algorithm version 3 [Porter, 1980]) to the keyword list and the words in the documents in order to match as many possible variants as possible of the root term. For each identified word the relative distance from the start of the file is given. We cluster these distances for each file and use the information to signal documents with text areas that have a high density of risk keywords, which would be candidates to be eliminated from the file.

Finally, we manually revise the labeled files, using the clustered distance information for support, and then deleting those paragraphs (or text blocks) identified as having the highest density of *risky terms.*

3.2 Information Loss and Risk Evaluation

In this section we present the information loss and risk metrics as well as a vectorial model search engine. We note that the same metrics are used to measure information loss and disclosure risk. However, these metrics are applied using different sets of queries (utility and risk queries), which give different evaluations and correspond to different interpretations. The utility queries consist of terms about the general topic of each document set and the risk queries consist of terms that define sensitive concepts.

3.2.1 Search Engine

We have implemented our own search engine with the following main characteristics: an inverted index to store the relation between terms and documents and a hash-table to efficiently store the terms (vocabulary). Moreover, we also have applied text pre-processing techniques such as elimination of stop-words and stemming (see Section 2.7.3). We have also implemented a Vectorial Model formula to calculate the relevance of a set of terms (query) with respect to the corpus of documents. This is based on the calculation of term frequency, inverted document frequency, root of the sum of weights for the terms in each document. Refer to [Baeza-Yates and Ribeiro-Neto, 2011] for a complete description of the Vectorial model and the formula used.

We observe that the queries are by default OR. That is, if we formulate the query "$term_1$ $term_2$ $term_3$", as search engines do by default, an OR is made of the terms and the documents are returned which contain at least one of the three given terms, complying with "$term_1$ OR $term_2$ OR $term_3$".

3.2.2 Information Loss and Risk of Disclosure Metrics

As a starting point, we have used a set of well-known information retrieval metrics, which are listed in Table 3.2 and briefly described below. The formulas are defined in terms of the following sets of documents:

- *true_relevant_documents* is the unchanged, non-sanitized, document set retrieved by the corresponding query by the Vectorial search engine.
- *retrieved_documents* is the set returned by the search engine in reply to a given query that is above the relevance threshold.
- *relevant_documents*, are the documents above the relevance threshold which are members of the *true_relevant_documents* set.
- *true_relevant_docs_returned* are the documents in *true_relevant_documents* that are returned by the search engine in any position (above or below the threshold).
- *false_relevant_docs* are the documents not members of *true_relevant_documents* but which are returned above the relevance threshold.

The degree of relevance of a document with respect to a query is calculated as a quantified value by the Vectorial model search engine. The assignment of this relevance thresholds is explained in Section 3.3.2.

Precision	$P = \frac{\lvert\{relevant\_docs\} \cap \{retrieved\_docs\}\rvert}{\lvert\{retrieved\_docs\}\rvert}$
Recall	$R = \frac{\lvert\{relevant\_docs\} \cap \{retrieved\_docs\}\rvert}{\lvert\{true\_relevant\_docs\}\rvert}$
F-measure	$F = 2 \cdot \frac{P \cdot R}{P + R}$
Coverage	$C = \frac{\lvert\{true\_relevant\_docs\_returned\}\rvert}{\lvert\{true\_relevant\_docs\}\rvert}$
Novelty	$N = \frac{\lvert\{false\_relevant\_docs\}\rvert}{\lvert\{total\_relevant\_docs\}\rvert + \lvert\{false\_relevant\_docs\}\rvert}$

Table 3.2: Information retrieval metrics.

- The Precision is considered as the percentage of retrieved documents above the relevance threshold that are relevant to the informational query.

- The Recall, on the other hand, is considered as the percentage of retrieved documents above the relevance threshold that are defined as truly relevant.

- The F-measure (or balanced F-score) combines precision and recall and mathematically represents the harmonic mean of the two values.

- The Coverage is the proportion of relevant documents retrieved out of the total true relevant documents, documents known previously as being the correct document set for a given search.

- The Novelty is the proportion of documents retrieved and considered relevant which previously were not relevant for that query. That is, it measures the new information introduced for a given query. We interpret novelty as undesirable with respect to the quality of the results, because we assume that we have correctly identified the set of all true relevant documents.

As well as the four metrics listed in Table 3.2, we also consider four other measures:

- The average relevance of the documents whose relevance is above the relevance threshold.

- The total number of documents returned by the query whose relevance is greater than zero.

- The number of random documents which are members of the set of relevant documents for a given query.

- NMI (Normalized Mutual Information), we use the NMI metric [Manning et al., 2008b] for counting documents assignments to query document sets before and after sanitization. That is, we compare the results of the document assignments to query sets by identifying the documents in each query document set before sanitization, and the documents which are in the same corresponding query document set after sanitization.

$$NMI = \frac{I(\Omega; Q)}{[H(\Omega) + H(Q)]/2}$$

where $I(\Omega; Q)$ is the mutual information of documents before and after sanitization (Ω) and a query document set (Q). $H(\Omega)$ and $H(Q)$ are the corresponding entropies.

Quantification of information loss and risk. In order to obtain a single resulting value, we have studied all the parameters presented and defined a formula in terms of the factors which showed the highest correlation between the original and sanitized document metrics: F = F-measure, C = coverage, N = novelty, TR = total number of documents returned, PR = percentage of random documents in the relevant document set, and the NMI value. Hence, the information loss IL is calculated as:

$$IL = DR = \frac{(2F) + C - N + TR - PR - 2NMI}{8} \tag{3.1}$$

We observe that of the six terms in the formula, F and NMI are given a relative weight of 25%, and the other four terms are given a relative weight of 12.5%. The weighting was assigned by evaluating the relative correlations of the values before and after document sanitization for each factor. As the F-measure and the Normalized Mutual Information were the factors that showed the highest correlation between the original and sanitized document, we gave them a higher weight according to their correlation value with respect to the other values.

Not that for the risk of disclosure, RD, we use the same Equation (3.1) and terms, however the interpretation is different: for IL a negative result represents a reduction in information, and for RD a negative result represents a reduction in risk.

3.3 Experimental Analysis

In this section we describe the set of U.S confidential documents used and how we have obtained them. Then, we explain which is the value relevance threshold and how was computed. Finally, we present the results for information loss and risk of disclosure, comparing query results between the original and the sanitized dataset by means of the presented metrics.

3.3.1 Document Extraction

In order to test the proposed sanitization and evaluation techniques we have extracted a set of confidential documents from the online Wikileaks Cable repository [Wikileaks, 2010]. Since in this online repository there are lots

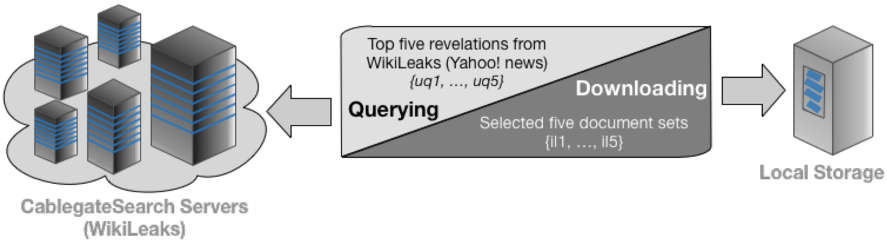


Figure 3.2: Scheme for document extraction and querying.

of documents related to different subjects, we selected the first five topics from the top ten revelations published by Yahoo! News [Yahoo! News, 2010]. From these five topics we derived five queries, one per each topic. They are showed in Table 3.3. Then, we searched using these queries as keywords on www.cablegatesearch.net [Wikileaks, 2010] to find the corresponding cables, thus obtaining a set of documents for each query. Figure 3.2 shows a schematic representation of this process.

We observe in Table 3.3 that there is a sixth document set, i_6. This set of documents was randomly chosen from [Wikileaks, 2010] for benchmarking purposes. These extracted five queries will also be used to test the information loss (utility) of the sanitized documents, see Section 3.3.3.

With respect to the risk queries, we used the same 30 seed terms extracted from the eight risk points defined in Section 1.4 of the US Executive Order 13526 [E.O. 13526, 2009], shown in Table 3.1. Hence, for each stated point we designated a risk query; $\{rq_1, \cdots, rq_8\}$. We identify eight sets of documents corresponding to each queries; $\{r_1, \ldots, r_8\}$. These risk queries were used in our sanitization processing to detect risk text blocks, and now are also employed to define eight different queries which are used to evaluate the risk. Note that we also defined a ninth query, rq_9, composed of all the terms from queries rq_1 to rq_8, whose corresponding document set is r_9.

3.3.2 Computing the Relevance Inflexion Point

In this section we define the relevant threshold value for informational document sets as well as for risk document sets.

Relevance threshold value for informational document sets. In order to apply the same criteria to all the search results, after studying the distributions in general of the relevance of the different queries, we chose a relevance of 0.0422 as the threshold. That is, we identify an inflexion point between the relevant documents (relevance greater or equal to 0.0422) and non-relevant documents (relevance less than 0.0422).

Example 4. *Table 3.4 shows an example for the search results of a given query uq_{5-1}, for which the first seven ranked documents (highlighted in grey) are above the relevance threshold. Figure 3.3 also shows the search results.*

ID_q	Keywords (utility queries)	TC, CH	ID_{Id}	Top five news item revelations (Yahoo!)[17]
uq_1	{saudi, qatar, jordan, UAE, concern, iran, nuclear, program}	35,10	u_1	”Middle Eastern nations are more concerned about Iran’s nuclear program than they’ve publicly admitted”
uq_2	{china, korea, reunify, business, united, states}	3,3	u_2	”U.S. ambassador to Seoul said that the right business deals might get China to acquiesce to a reunified Korea, if the newly unified power were allied with the United States”
uq_3	{guantanamo, incentives, countries, detainees}	12,10	u_3	”The Obama administration offered incentives to try to get other countries to take Guantanamo detainees, as part of its plan to progressively close down the prison”
uq_4	{diplomats, information, foreign, counterparts}	6,6	u_4	”Secretary of State Hillary Clinton ordered diplomats to assemble information on their foreign counterparts”
uq_{5-1}	{putin, berlusconi, relations}	97,10	u_5	”Russian Premier Vladimir Putin and Italian Premier Silvio Berlusconi have more intimate relations than was previously known”
uq_{5-2}	{russia, italy, relations}			
–	–	10,10	u_6	–

Table 3.3: Queries and documents used to test Information Loss. Remark, il_6 represents a set of randomly chosen documents to be used as a benchmark, ID_q is the identification name for a specific set of keywords, TC is the number of total cables, CH is the number of cables chosen, and ID_u is the informational document set identification name.

Rank	**Doc id**	**Relevance**
1	$u_{5.6}$	0.26248
2	$u_{5.1}$	0.21050
3	$u_{5.2}$	0.10709
4	$u_{5.3}$	0.09852
5	$u_{5.4}$	0.08784
6	$u_{3.7}$	0.07626
7	$u_{5.8}$	0.05202
8	$u_{5.10}$	0.02243
...	...	...
44	$u_{5.9}$	0.000034

Table 3.4: Example search results for the query uq_{5-1}, "putin berlusconi relations".

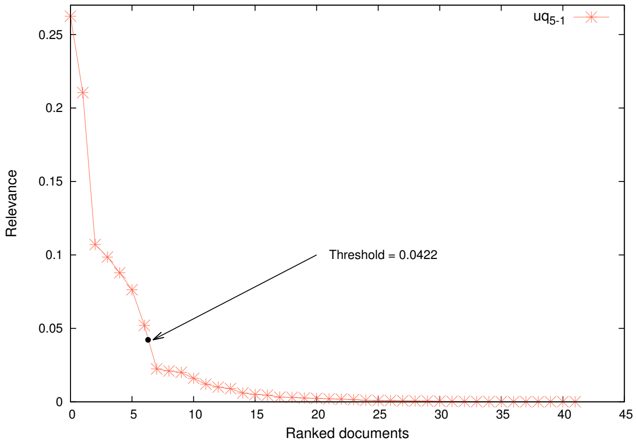


Figure 3.3: Example distribution of relevance (x-axis) of ranked documents (y-axis) corresponding to the query of Table 3.4.

	uq_1	uq_2	uq_3	uq_4	uq_{5-1}	uq_{5-2}
Step 1	0.00	0.00	0.00	0.00	100.00	0.00
Step 2	11.00	0.00	14.00	**50.00**	**100.00**	0.00

Table 3.5: Information Loss: percentage (%) differences of NMI metric for original and sanitized document corpuses (steps 1+2)

For this example, and with reference to the definitions given in Table 3.2, the information loss metrics are calculated as follows: (i) precision $= 6/7 = 0.8571$. *That is, there were* 6 *known relevant documents from a total of* 7 *above the relevance threshold; (ii) recall* $= 6/10 = 0.6$. *That is, six of the* 10 *known relevant documents were returned above the relevance threshold; (iii)* $F - measure = 2((0.85710.6)/(0.8571 + 0.6)) = 0.7058$*, where the precision is* 0.8571 *and the recall is* 0.6*; (iv) coverage* $= 10/10 = 1.0$*, because all* 10 *known relevant documents were returned among the* 44 *results of the search engine; (v) novelty* $= 1/(10 + 1) = 0.0909$*, where there are* 10 *known documents relevant to the query (Table 3.3) and in the list of relevant documents (relevance* ≥ 0.0422*), one of the documents (*$u_{3.7}$*, ranked sixth) is not in the set of* 10 *known documents.*

Relevance threshold value for risk document sets. After studying the distributions of the relevance for each risk document set returned by the search engine, we assigned the relevance threshold of 0.010 for all the results sets, with the exception of the result sets r_9, r_1 and r_2 which were assigned a threshold of 0.020. The metric calculations then followed the same process as for the informational document sets.

3.3.3 Information Loss

In Table 3.5 we show the NMI metric applied to the original and sanitized document query sets. We see only a small reduction in correspondence for the majority of query document sets, except for uq_4 and uq_{5-1}, however, the latter is due to the loss of the named query terms in the documents ('Putin' and 'Berlusconi' were masked as named entities in Step 1 of the sanitization process).

In the case of uq_4, a value of 50% for Step 2 means that 50% of the relevant documents from the original document set returned by the search engine, are to be found in the relevant documents from the sanitized document set returned by the search engine.

Table 3.6 shows the percentage change for each metric value and informational document set, of the original documents and the sanitized documents processed by Steps 1 and 2. We observe that the indicators used in the information loss, Equation (3.1) are high-lighted in grey. The information loss calculated using Equation (3.1) is shown in the rightmost column ($\Delta(IL)$) from the below part of Table 3.6, giving an average value of 26.1%, included query uq_{5-1} and a value of 16.1% excluding query uq_{5-1}.

	$\Delta(P)$	$\Delta(R)$	$\Delta(F)$	$\Delta(C)$	$\Delta(N)$
uq_1	-1.56	-12.50	-0.08	0.00	0.00
uq_2	-40.00	0.00	-0.25	0.00	40.00
uq_3	0.00	-14.29	-0.09	0.00	0.00
uq_4	-62.50	-75.00	-0.70	0.00	33.33
uq_{5-1}	-100.00	-100.00	-1.00	-100.00	-100.00
uq_{5-2}	-11.11	0.00	-0.05	0.00	38.46

	$\Delta(AR)$	$\Delta(TR)$	$\Delta(PR)$	$\Delta(IL)$
uq_1	-38.15	-15.38	0.00	-6.625
uq_2	-0.38	-4.76	20.00	-14.37
uq_3	3.77	-12.50	0.00	-7.375
uq_4	9.80	-10.81	25.00	-38.62
uq_{5-1}	-100.00	-4.55	0.00	-75.62
uq_{5-2}	-5.03	0.00	0.00	-13.75

Table 3.6: Information Loss: percentage (%) differences (Δ) of statistics for original and sanitized document corpuses (steps 1+2). Where, P=precision, R=recall, F=F measure, C=coverage, N=novelty, AR=Average relevance for documents above threshold, TR= total docs. returned, PR=percentage of random docs in relevant doc set, IL=percentage information loss calculated using Equation (3.1)

With reference to query uq_{5-1}, recall that the names of two persons, "berlusconi" and "putin", were substituted by their respective categories, "$Pers_1$" and "$Pers_2$". As they were essential for the successful retrieval by this query of the corresponding documents, this resulted in a total loss of retrieval. In Table 3.6 we also observe that the $\Delta(F)$ measure (which is a ratio of precision and recall) has reduced for uq_2 and uq_4, and the novelty ($\Delta(N)$) and percentage of random documents ($\Delta(PR)$) have increased. Novelty is considered a negative aspect, given that we interpret it as the entry of irrelevant documents into the set of relevant documents (above the threshold). We observe that the information loss is highest for query uq_4 (-38.62) and lowest for queries uq_1 and uq_3. If we look again at the terms which correspond to these queries (Table 3.3), those of queries uq_1 and uq_3 are more specific whereas those of query uq_4 are more general. Another observation is the correlation of the different metrics with the information loss (IL). Again, excluding query uq_{5-1}, we see that $\Delta(PR)$, $\Delta(N)$ and $\Delta(F)$ correlate with the maximum values of $\Delta(IL)$ (-14.37 and -38.62), whereas $\Delta(C)$ is invariant; $\Delta(TR)$ appears to be query dependant and has little correlation with the other metrics.

To summarize, Step 1 (*anonymization of names and personal information of individuals*) has little or no effect on the success of the informational queries, except those which contain specific names of people. However, this is an important required process because it is necessary to preserve the confidentiality of the individuals who appear in these documents. On the other hand, Step 2 (*elim-*

rq_1	rq_2	rq_3	rq_4	rq_5	uq_6	rq_7	rq_8	rq_9
60.00	**67.00**	–	36.00	**25.00**	56.00	63.00	**70.00**	58.00

Table 3.7: Risk of Disclosure: percentage (%) differences of statistics for original and sanitized document corpuses (steps 1+2).

ination of risk text) inevitably had a higher impact, given that blocks of text are eliminated from the documents. From the results of Table 3.6, we see that the information loss is query dependent, the $\Delta(F)$ and $\Delta(TR)$ indicators being the most invariant. By manual inspection of the documents, we can conclude in general that a worse value is due to the loss of key textual information relevant to the query. Also, queries with more general terms incur a higher information loss.

3.3.4 Disclosure Risk

We recall that the NMI metric measures the degree of correspondence between different groups. In Table 3.7 this metric is applied to the original and sanitized document query sets. A significant reduction can be seen in the correspondence, which contrasts with the results for the same metric applied to the information loss of query document sets. Table 3.8 shows the percentage change for each of the metrics we described in Section 3.2.2, for each of the nine risk queries, for the original documents and the sanitized documents of processing step 2. The risk calculated using Equation (3.1) is shown in the rightmost column ($\Delta(RD)$) from the below part of Table 3.8, and it decreases a percentage average value of −47.26%. In general, we see a significantly greater percentage change in comparison to the information loss results of Table 3.6.

We observe that the greatest risk reduction is for queries rq_2, rq_7, rq_8, rq_1 and rq_6, with values of −50.75, −49.00, −48.87, −47.37 and −45.37, respectively. On the other hand, the risk reduction was least for queries rq_4 and rq_5, with values of −19.5 and −28.87, respectively. If we look at the terms which correspond to these risk queries (Table 3.1), we see that words from queries rq_4 and rq_5 are more general and neutral, whereas that words of the risk queries which corresponded to a greater risk reduction had more specific terms. Note that query rq_3 did not retrieve any documents, although we included it in the results as it corresponds to point (c) of [E.O. 13526, 2009].

By observing the relative ranking of the documents returned by the queries, we saw that some documents with risk terms actually went up the ranking. After inspecting the corresponding documents, we found that this was due to the presence of terms such as *nuclear*, but in a peaceful (energy) context, and *war* with reference to conflicts such as the Balkans, which had no relation to U.S. national security. However, we re-checked our editing of the documents corresponding to query rq_5, given the increased presence of these documents in the highest ranked positions. We confirmed that the sanitization was consistent with the other document groups.

	$\Delta(P)$	$\Delta(R)$	$\Delta(F)$	$\Delta(C)$	$\Delta(N)$
rq_1	−66.67	−60.00	−0.64	−16.67	40.00
rq_2	−66.67	−66.67	−0.67	−33.33	40.00
rq_3	0.00	0.00	0.00	0.00	0.00
rq_4	−18.18	−35.71	−0.28	−7.14	15.38
rq_5	−57.14	−25.00	−0.45	−12.50	50.00
rq_6	−60.00	−55.56	−0.58	−22.22	40.00
rq_7	−71.43	−50.00	−0.64	−12.50	55.56
rq_8	−50.00	−70.00	−0.63	−50.00	23.08
rq_9	−54.55	−58.33	−0.57	0.00	35.29

	$\Delta(AR)$	$\Delta(TR)$	$\Delta(PR)$	$\Delta(RD)$
rq_1	−26.94	−44.44	30.0	−47.37
rq_2	27.07	−48.39	16.7	−50.75
rq_3	0.00	0.00	0	−
rq_4	17.80	−4.17	1.96	−19.5
rq_5	11.74	−18.60	8.90	−28.87
rq_6	8.07	−55.26	17.8	−45.37
rq_7	−0.49	−33.33	35.7	−49.00
rq_8	−39.31	−29.41	23.3	−48.87
rq_9	−14.29	−10.20	9.9	−35.62

Table 3.8: Risk of Disclosure: percentage (%) differences (Δ) of statistics for original and sanitized document corpuses (steps 1+2). Where, P=precision, R=recall, F=F measure, C=coverage, N=novelty, AR=Average relevance for documents above threshold, TR= total docs. returned, PR=percentage of random docs in relevant doc set, RD=percentage risk disclosure calculated using Equation (3.1).

3.4 Conclusions

In this chapter we have used information retrieval metrics to evaluate information loss and disclosure risk for a set of sanitized documents. In order to evaluate these two values, we implemented a vectorial model search engine and also defined a formula to evaluate the information loss and disclosure risk by means of querying both document sets. Additionally, we developed a semi-supervised method to assist the sanitization of confidential unstructured textual documents. Finally, we tested our sanitization method on a set of real U.S. confidential documents and then, we evaluated the data utility and the risk of the sanitized documents. The results show a relatively low overall information loss (16% excluding query uq_{5-1}) for the utility queries (uq_1 to uq_5), whereas an average risk reduction of 47% was found for the risk queries (rq_1 to rq_9).

Chapter 4

Vector Space Model Anonymization

In this chapter we address the problem of how to release a set of confidential documents. To that end we propose a couple of methods that from a given set of confidential and thus private documents provide some anonymized metadata which can be released and used for analysis and text mining purposes.

To conduct this problem we have relied on a well known data representation of a set of documents, the Vector Space Model (VSM) [Salton and Buckley, 1988], which is widely used in information retrieval and text mining. This algebraic model for representing documents as vectors of numeric weights associated with words was previously described in Section 2.7.3. Our proposals can be summarized as providing an anonymous VSM. So that, the anonymous vector model can be publicly released, while the original documents are kept secret. As many information retrieval and text mining tasks rely on this text representation model, third parties will be able to analyze these released anonymous VSM and perform several techniques such as classification, clustering, etc.

Both presented methods are inspired by microaggregation, a popular protection method from statistical disclosure control which was introduced in Section 2.4. As we showed, microaggregation ensures a certain privacy degree by satisfying the k-anonymity principle. Therefore, the protected data is completely k-anonymous in the sense that there are k vectors completely indistinguishable. Contrary to other works on k-anonymity, there is no need for additional protection mechanisms regarding unprotected sensitive attributes such as l-diversity [Machanavajjhala et al., 2007]. This ensures that re-identification algorithms [Narayanan and Shmatikov, 2008] will have a re-identification probability bounded by the number of indistinguishable elements.

Our first approach is motivated by the high-dimensional and sparsity of vector space document text representations, which usually have a few thousand dimensions and a high percentage of sparsity, about 90%. We propose the *spherical microaggregation*, a microaggregation variation to deal with these sparse and high-dimensional data sets. This variation consists of an adaptation of the partition and aggregation functions of microaggregation in order to improve the data quality of the output protected data. That is, the anonymization method has to protect an original data set ensuring the entities' confidentiality and achieve it with the lowest loss of information, so that this anonymous data could reflect as much as possible the original data distribution. Therefore, given that microaggregation is a distance-based clustering approach we propose to use the cosine dissimilarity, instead of the Euclidean distance, in order to exploit the sparsity of the data. Moreover, we also improve the way clusters' representatives should be computed. A cluster representative is the vector which is the closest in terms of cosine distance (in average) to all data vectors belonging to its respective partition. A mathematical proof is provided to support our proposition. Finally, the evaluation of the method is conducted by performing a large set of experiments on two different sparse and high-dimensional data sets.

The second protection proposal is focused on the semantic mining of the words contained in documents. Radically different as the previous approach the vector space models used for this approach are small, usually by applying harder feature reduction techniques to select the most representative document words. We introduce the *semantic microaggregation*, a microaggregation variation exploiting the semantics of the words. This approach relies on a generic lexical database such as WordNet, which provides tools to define distances and relations between words. It is important to remark that although we have used WordNet, depending on the domain we are working on, other domain specific databases can be used. Unified Medical Language System (UMLS) [Bodenreider, 2004] is an example of a biomedical database which provides semantic relations between terms in the health vocabulary.

It is important to clarify what do we consider as private information with respect to a set of documents. In this work we have focused on the concrete protection of the document owner, creator, or the entity to which the document is explicitly related. We try to prevent the ability of an attacker to correctly link a given document (or document representation) to a concrete entity (individual, organization, . . .). In Section 4.1 we discuss several possible scenarios that present this particularity or threat, some examples are a set of health patient records, research project proposals, individual posts to a private internet forum, etc. The rest of this chapter is organized as follows. In Section 4.2 we provide the anonymous VSM definition. Section 4.3 presents the spherical microaggregation proposal and formalization. Then, in Section 4.4 we provide an evaluation of this first microaggregation variation. These evaluations are performed by means of different traditional information retrieval techniques. Afterwards, in Section 4.5 we present the second microaggregation variation, the semantic microaggregation. Section 4.6 evaluates this second proposal in terms of intra and inter cluster sum of square errors. Finally, Section 4.7 concludes the chapter.

4.1 Scenarios

To better shape our proposals we present here three motivating scenarios. In short, the presented anonymization techniques are suitable for scenarios involving a set of confidential documents in which each document is directly or indirectly related to one or a set of different entities. A direct relation is when the document contains sensitive information of the specific person or institution that must be anonymized, while an indirect relation is when the document does not contain explicit information about the entity to be anonymized, but also there is an implicit relation between the entity and the document that can be inferred through some other document properties. We describe three cases where our proposals have a direct application.

Private textual datasets for generic research. A clear application scenario is within the research community, in the information retrieval and text mining fields. Several organizations present their research at scientific journals or conferences. As usual their research proposals are supported by a set of experiments, however, unlike university researchers, their experiments are performed on organization's data. Thus, when other researchers want to reproduce those experiments it becomes impossible, since the datasets are private and cannot be shared due to they could contain confidential information. Examples are a set of patient health records, user posts to a private Internet forum, a set of user profiles from a social network, or even a set of user queries made to a search engine (recall the infamous AOL search data leak [Barbaro and Zeller, 2006]).

A possible solution is to publish an anonymized data that represents the original dataset and can be used to reproduce to some extent the research made on the original dataset. This is straightforward in text mining research where the VSM is frequently used to represent a set of documents, but other similar data structures can be envisioned with the same purpose for more specific tasks.

Private profiling data for advertising. Personalized online advertisement is another possible area where anonymization should be considered. Lots of web services are offering their services for free in exchange of introducing advertisements on their services. Google, Twitter or Facebook are some examples of companies, which collect and store thousands of users' confidential information in order to analyze and offer targeted advertisements [Rutkin, 2013, Simonite, 2013]. E-mails, user's posts or even personal documents are some clear examples. In some cases these data could be transferred to specialized companies, which analyze all data in order to define advertisement strategies in a user base.

These user data might be considered confidential, and might not be directly transferable to other parties. Therefore, the solution is to anonymize the data before its transference. The idea behind this approach is that the advertisement company will not be able to distinguish a unique user from a set of k of them. Hence, the advertisements selected for a single user are actually extracted from a mix of several user's profiles.

Anonymized metadata from public tender. As a last example, we consider a government agency managing applications to public research project funding. Such applications should be kept private, but at the same time it can be interesting to be able to give some information about the applications and more precisely of the projects presented by the applicants. This becomes specially difficult if we assume that the projects are written in a free-form text. This information is interesting not only to the community applying for funding but also to the administration and politicians. They may be interested in information such as: "this geographic area applies for projects about this topic", or "this methodology is proposed by a given percentage of researchers from these given topics". While this information can be valuable it normally does not reveal specific and private information.

4.2 Anonymous VSM

The purpose of both methods is to provide an anonymous or privacy-preserving version for a given VSM, so it satisfies the k-anonymity property with respect to document owners or entities to which the document is related.

In this sense we define a k-anonymous VSM as follows.

Definition 4.1. *A VSM is said to satisfy k-anonymity if for every vector $\vec{v}$ in VSM, there are at least $k-1$ vectors equals to $\vec{v}$.*

Microaggregation ensures this property by means of building groups of k similar vectors. For each group, microaggregation generates a representative vector with similar statistical properties to that group. Then, each original vector is replaced by its corresponding representative. Thus, each group consists of k indistinguishable vectors. Although, such representatives do not represent the true data vectors, they are useful for most modeling purposes, since they reflect the original information of the vectors.

For instance, imagine a VSM representing a set of patient records, where each row corresponds to a unique patient. Its k-anonymous VSM version will consist of sets of k equal patients. Therefore, the probability of linking a given record to a given patient is bounded by $1/k$. Note that each of these anonymous records can be seen as a representative of a group of patients from the original data.

4.3 Spherical Microaggregation

In this section we present which are the changes that should be applied to the classic microaggregation algorithm (Section 2.4) in order to obtain better results when dealing with sparse and high-dimensional data, such as vector spaces.

It was shown in Section 2.4 that microaggregation works as a distance-based clustering algorithm. However, it was originally defined for low-dimensional and dense data. Our goal is to extend its functionality to vector space models of large document collections, or in other words, very sparse and high-dimensional models.

It is well known how in distance-based clustering algorithms the use of different distances lead to different data partitions and so, depending on the data type used one distance can obtain better results than others. [Strehl et al., 2000] and [Dhillon and Modha, 2001] demonstrate that Euclidean distance is a weak discriminant when data is multidimensional and sparse. Therefore, in order to adapt the microaggregation algorithm a new distance function has to be considered, and as a consequence, we are forced to modify the aggregation function so that it fits better with the considered distance. That is, the resultant aggregated vector, or centroid, should minimize the sum of distances between all the cluster members and itself.

To do so, we have focused on the spherical k-means clustering algorithm, presented by [Dhillon and Modha, 2001], due to its similarities with the presented problem. Their objective was to adapt the k-means clustering algorithm results for vector spaces by means of using the cosine similarity and an aggregation function that computes the best representative for a given cluster, according to the cosine similarity.

In the next sections are explained the necessary modifications to adapt the microaggregation to vector space models. Section 4.3.1 introduces the cosine similarity and in Section 4.3.2 we present the aggregation function and its relevant property.

4.3.1 Distance function

We assume that all document vectors have been normalized using the L^2-norm, as described in Equation (2.18). This means that vectors are represented on a high dimensional unit sphere. Furthermore, for most of the weighting schemes all the document vectors are non-negative, and hence all document vectors can be represented in the first quadrant of the unit sphere. For those vectors, the dot product is a natural measure of similarity. This measure is known as cosine similarity and it is widely used in text mining and information retrieval due to its easy interpretation and simple computation for sparse vectors. The similarity between two given document vectors, $\vec{d_1}$ and $\vec{d_2}$, is given by:

$$s_{cos}(d_1, d_2) = \cos(\theta(\vec{d_1}, \vec{d_2})) = \frac{\langle \vec{d_1}, \vec{d_2} \rangle}{|\vec{d_1}||\vec{d_2}|} \tag{4.1}$$

where $\langle \vec{d_1}, \vec{d_2} \rangle$ and $\theta(\vec{d_1}, \vec{d_2})$ are the inner product and the angle between these two vectors, respectively.

The resulting similarity ranges from -1, meaning exactly opposite vectors, to 1, meaning exactly the same. However, as we have assumed non-negative vectors the similarity ranges from 0 to 1. The maximum similarity, 1, is achieved when there is a complete match between both vectors, and the minimum, 0, when both vectors are orthogonal, that is the angle between both vectors is 90°.

It is easy to transform Equation (4.1) into a distance d_{cos} (Equation (4.3)).

$$d_{cos}(\vec{d_1}, \vec{d_2}) = 1 - \frac{\langle \vec{d_1}, \vec{d_2} \rangle}{|\vec{d_1}||\vec{d_2}|} \tag{4.2}$$

Following the assumption that all document vectors have magnitude 1, vector norms can be removed from the previous equation. That is,

$$d_{cos}(\vec{d_1}, \vec{d_2}) = 1 - \langle \vec{d_1}, \vec{d_2} \rangle \tag{4.3}$$

4.3.2 Aggregation function

The aggregation step is defined by a function that given a partition of document vectors, π_j, returns the corresponding representative vector, c_j, which in average is closest, in terms of cosine similarity, to all document vectors belonging to that partition.

Given a set of non negative document vectors, $\vec{d_1}, \ldots, \vec{d_N}$, which have unit norm and a set of p disjoint clusters on the vectors, $\pi_1, \ldots, \pi_p$, Dhillon and Modha [Dhillon and Modha, 2001] proposed to compute the centroid of each π_j first computing the vector $\vec{m_j}$ as follows,

$$\vec{m_j} = \frac{1}{N_j} \sum_{d_i \in \pi_j} \vec{d_i}, \tag{4.4}$$

where N_j is the number of document vectors in the cluster π_j. And, then, as the resulting vector does not have a unit norm they define the centroid of π_j normalizing the vector by its L^2 norm. That is:

$$\vec{c_j} = \frac{\vec{m_j}}{\|\vec{m_j}\|} \tag{4.5}$$

Proposition 4.1. *Let S be the set of vectors in the unit sphere and let π_j be a cluster containing a set of documents vectors $\vec{d_1}$, …, $\vec{d_N} \in S$. The average vector of S defined by*

$$\mathbb{C}(\vec{d_1}, \ldots, \vec{d_N}) = argmin_{\vec{c_j} \in S} \{ \sum_{\vec{d_i} \in \pi_j} d_{cos}(\vec{c_j}, \vec{d_i}) \} \tag{4.6}$$

can be computed using the following expression

$$\mathbb{C}(\vec{d_1}, \ldots, \vec{d_N}) = \frac{\sum_{\vec{d_i} \in \pi_j} \vec{d}_{ir}}{\sqrt{\sum_{r=1}^{R} (\sum_{\vec{d_i} \in \pi_j} \vec{d}_{ir})^2)}}. \tag{4.7}$$

where R is the number of dimensions of the vectors.

Proof. Starting from the assumption that all the elements have a unit norm, including the centroid, we can express the cosine distance as the half of the squared euclidean distance.

$$\begin{aligned}\|\vec{d_i} - \vec{d_j}\|^2 &= (\vec{d_i} - \vec{d_j})^T(\vec{d_i} - \vec{d_j}) \\ &= \|\vec{d_i}\|^2 + \|\vec{d_j}\|^2 - 2\vec{d_i}^T\vec{d_j} \\ &= 2(1 - cos(\theta(\vec{d_i}, \vec{d_j})))\end{aligned}$$

so,

$$d_{cos}(d_i, d_j) = \frac{1}{2}\|\vec{d_i} - \vec{d_j}\|^2 \tag{4.8}$$

We can express the minimization problem, stated in Equation (4.6) and its corresponding constraint, $\|\vec{c_j}\| = 1$, by means of Lagrange multipliers,

$$L = \frac{1}{2}\sum_{\vec{d_i}\in\pi_j} ||\vec{c_j} - \vec{d_i}||^2 - \lambda(\vec{c_j}^T\vec{c_j} - 1).$$

We can rewrite this expression in terms of the components:

$$L = \frac{1}{2}\sum_{\vec{d_i}\in\pi_j}\sum_{r=1}^{R}(\vec{c_j}_r - \vec{d_i}_r)^2 - \lambda(\sum_{r=1}^{R}\vec{c_j}_r^2 - 1)$$

In order to obtain an expression for $\vec{c_j}$, we obtain an expression for each dimension of it, $\vec{c_j}_r$, using $\frac{\partial L}{\partial \vec{c_j}_r} = 0$. That is,

$$\frac{\partial L}{\partial \vec{c_j}_r} = \sum_{\vec{d_i}\in\pi_j}\frac{1}{2}2(\vec{c_j}_r - \vec{d_i}_r)(1) - \lambda 2\vec{c_j}_r = 0, \quad \forall\ r = 1 \ldots R$$

From this expression, we get

$$-\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r + \vec{c_j}_r|\pi_j| - 2\lambda\vec{c_j}_r = 0, \quad \forall\ r = 1 \ldots R,$$

Therefore, each dimension of $\vec{c_j}$ is expressed as follows,

$$\vec{c_j}_r = \frac{\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r}{|\pi_j| - 2\lambda}, \quad \forall\ r = 1 \ldots R. \tag{4.9}$$

Now, we consider $\frac{\partial L}{\partial \lambda} = 0$. That is,

$$\frac{\partial L}{\partial \lambda} = -\left(\sum_{r=1}^{R}\vec{c_j}_r^2 - 1\right) = 0.$$

Hence, using the expression in Equation (4.9), we get

$$\sum_{r=1}^{R}\left(\frac{\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r}{|\pi_j|-2\lambda}\right)^2=1.$$

So,

$$\sum_{r=1}^{R}(\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r)^2=(|\pi_j|-2\lambda)^2.$$

So, in order to obtain λ, we have to solve the following second degree equation:

$$4\lambda^2-4\lambda|\pi_j|+|\pi_j|^2-\sum_{r=1}^{R}\left(\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r\right)^2=0$$

This results into the following expression for λ:

$$\lambda=\frac{|\pi_j|\pm\sqrt{\sum_{r=1}^{R}(\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r)^2}}{2}.$$

If we replace now this expression for λ in Expression 4.9 we get for all $r=1\ldots R$,

$$\begin{aligned}\vec{c_j}_r &= \frac{\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r}{|\pi_j|-2\left(\frac{|\pi_j|\pm\sqrt{\sum_{r=1}^{R}(\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r)^2}}{2}\right)}\\ &= \frac{\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r}{\sqrt{\sum_{r=1}^{R}(\sum_{\vec{d_i}\in\pi_j}\vec{d_i}_r)^2)}}.\end{aligned}$$

As the last equation is the aggregation operator that returns the vector which is closest in terms of the cosine distance (in average) to all document vectors belonging to the cluster π_j, we have proven the proposition. Note that Expression 4.7 corresponds to the process described by Dhillon and Modha. □

4.4 Spherical Experimental Results

In this section we evaluate the anonymous data generated by the presented microaggregation variation. This assessment relies on a set of evaluation experiments to estimate how much information has been lost in the anonymization process and so, to determine the utility of this anonymized data. These evaluations are performed by means of a comparison between the results obtained by the original data and its respective anonymizations when different information retrieval techniques are applied.

We have divided this section in three different subsections. The first two, 4.4.1 and 4.4.2, describe the data sets used in the experiments before and after its anonymization and in Section 4.4.3 we quantify how much information has been lost after the anonymization process by analyzing the results obtained by a set of machine learning techniques.

4.4.1 Original Data

We have selected two well known datasets, the Reuters-21578 and the Movie Reviews Corpus. The former, Reuters, is currently the most widely used test collection for text categorization research. It consists of 10K documents classified in 52 different categories. To make it simple, we have avoided documents which are related to more than one category, and also, we have reduced the number of document features by means of removing the terms occurring in just one document. This reduction is based on the assumption that rare words do not affect category prediction. Hence, after these simplifications the corpus has 7346 documents classified in 7 different categories.

The latter, Movie Reviews, is a collection of documents used for sentiment analysis, so all documents are labeled with respect to their sentiment polarity (positive or negative). It consists on 2000 movie reviews, classified on 1000 positive reviews and 1000 negative reviews. As in the Reuters corpus, we have applied a feature selection process relying on the deletion of the words appearing just in one document.

In both cases, the set of document pre-processing techniques were applied, see Figure 2.9 in Section 2.7.3 for more details. Besides, we have considered an additional process within the cleaning task. It consists in removing all the words which are not in the WordNet ontology. This process adds and additional protection level, so proper names and very specific terms of a particular field are removed. Note, in case that the analysis to be carried on requires such words this process can be avoided. In any case, the presented anonymization algorithm ensures k-anonymity on the resulting VSM. The normalized term frequency-inverse document frequency was the weighting scheme used to represent the relevance of words within documents and corpus, see Equations (2.17) and (2.18).

In Section 4.4.3 we used two different supervised classification methods to estimate the quality of the protected models with respect to the original ones. Therefore, data should be split in two sets, training and test data. Hence, both data sets are divided with a proportion 70-30. The 70% is the training data and the remaining 30% is the test data. In the case of the Reuters corpus this partition is given by the data owners, otherwise, in the case of the Movie Reviews corpus this partition is random. To simplify the data sets name, we will refer to Reuters and Movie Reviews as R and MR, respectively. Besides, when we refer to one of its partitions, we will add its corresponding subscript, tr and ts for the training and test case respectively (e.g., R_{tr}, Reuters training set).

A summary of all vector space models used for the spherical microaggregation experiments is shown in Table 4.1. In this table, besides the number of documents and words of the datasets used, we show an indicator of the term-

Corpus	Split	N	d	$Avg(d_{nz})$	K	*Balance*
Reuters	*All*	7346	5473	29.4	7	0.0131
	Train	5255	5343	30.5	7	0.0152
	Test	2091	4152	26.7	7	0.0093
Movie Review	*All*	2000	12431	186.5	2	1
	Train	1420	12188	185	2	0.96
	Test	580	10479	190.3	2	0.91

Table 4.1: Summary of all vector spaces used. (*For each dataset, N is the number of documents, d is the number of words after removing stop-words, $Avg(d_{nz})$ is an average of the number of words per document, K is the total number of classes and Balance is the ratio of the number of documents belonging to the smallest class to the number of documents belonging to the largest class.*)

document matrix sparsity level by means of an average of the non-zero weight terms ($Avg(d_{nz})$) and also the ratio of the number of documents in the smallest class to the number of documents in the largest class (*Balance*). So a value close to 1 indicates a very balanced data set and a value close to 0 indicates completely the opposite.

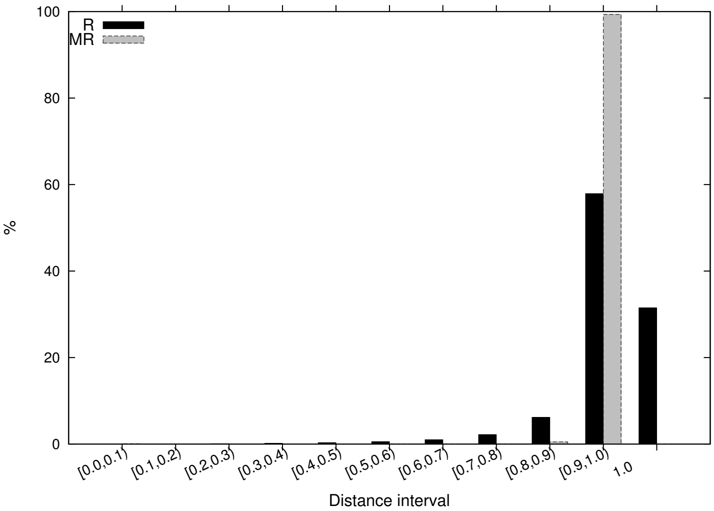
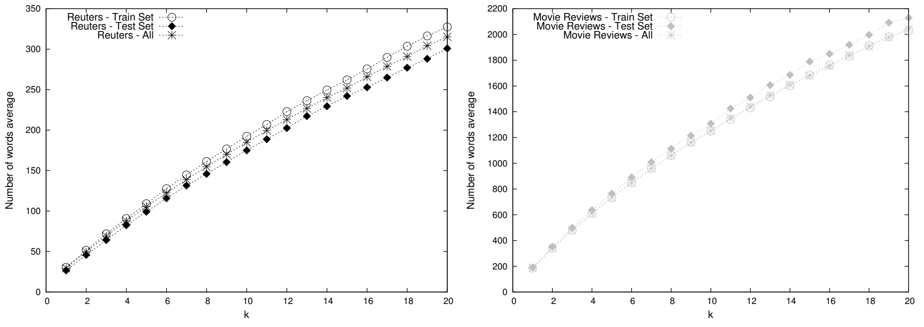


Figure 4.1: Distance distribution by intervals.

In the histogram of Figure 4.1 we show two distance distributions by intervals of 0.1. Black bars corresponds to the Reuters data set (R) and the grey ones represent the distance distribution of the Movie Reviews data set (MR). At first glance, it is clearly appreciable that almost all distances computed are located in the intervals with higher distance values. Specifically, the 99.3% and 89.4% of the distances are located in the two last intervals for the Movie Reviews and Reuters data set, respectively. This means that most document vectors of both data sets are far from each other in terms of cosine distance. Moreover, if we focus on the last interval, with distance equal to one, we will realize that while Movie Reviews has a percentage of almost zero, Reuters has much more distances located in that interval and so, it has more document vectors with non-overlapped terms. In detail, Reuters has the 57.9% of the distances in the $[0.9, 1)$ interval and the 31.5% in the interval with distances equal to one. In general *MR* vectors are much more scattered than the *R* vectors and as we will see in the following sections, this fact has an important impact in the quality of the protected data.

4.4.2 Anonymized Data

In order to anonymize the data we have implemented the microaggregation heuristic method MDAV, Algorithm 1, with the corresponding distance and aggregation functions proposed in Sections 4.3.1 and 4.3.2. We have used the implemented method to anonymize both datasets and their partitions independently, the training and test sets. The fact of protecting independently the training and test sets is because in Section 4.4.3 we have considered a set of three different scenarios relaying on supervised learning techniques to evaluate the quality of the protected data.



(a) Reuters datasets

(b) Movie Reviews datasets

Figure 4.2: Average of the number of words with a non-zero weight.

In addition, we have considered different values for the k microaggregation parameter, which ranges from 2 to 20, and so for each original dataset we have 19 protected variations. Thereby, we will study the variation effects of this parameter in order to decide which should be the best anonymization value.

All the protected datasets have the same number of documents. However, as the k parameter is increased the anonymity clusters are growing and therefore the protected vectors are increasing the number of words due to the aggregation operation. The sparsity of the vector space or term-document matrix is decreased, or inversely, as it is shown in Figure 4.2, the average of the number

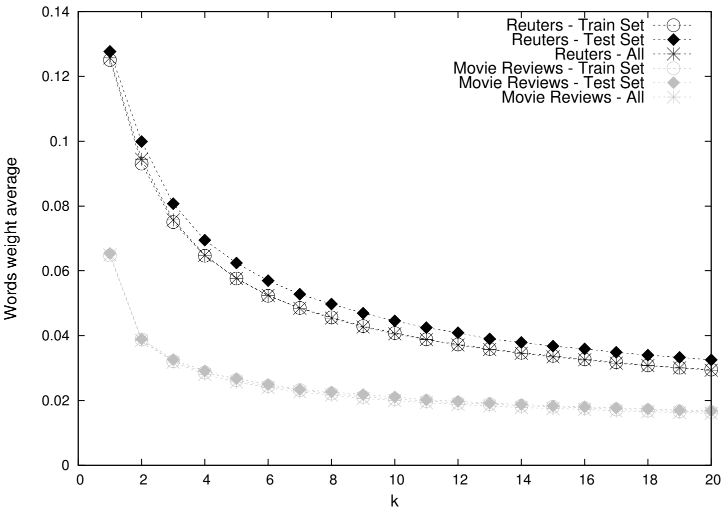
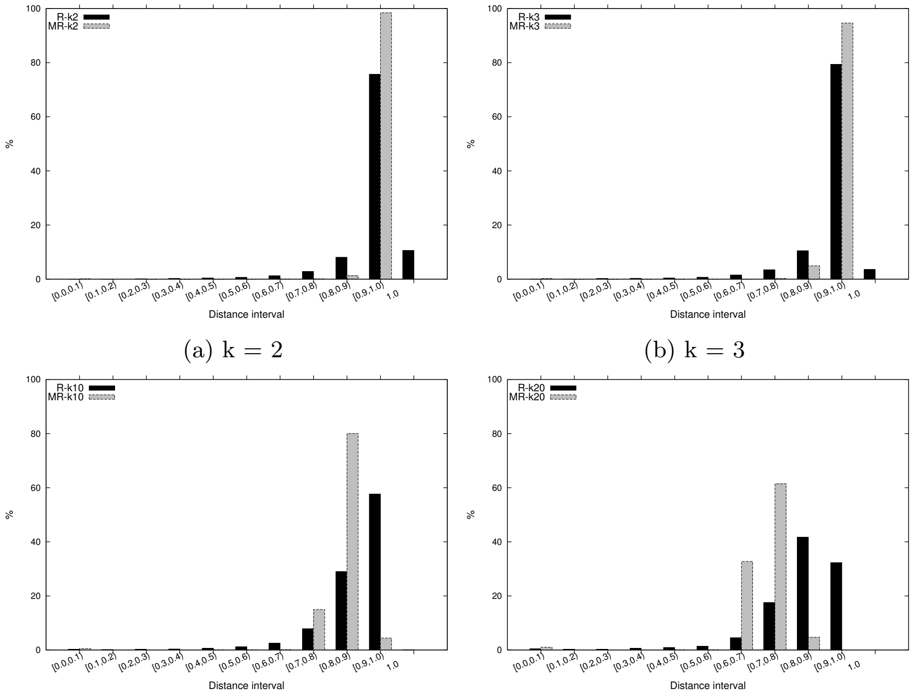


Figure 4.3: Global word weight average.

of words reaches a maximum value when the protection degree is equal to 20. It increases about a 5% in the Reuters case and about 16% in the Movie Reviews case. Furthermore, increasing the number of words per document also implies a decrease in all vector weight values. This behavior is shown in Figure 4.3. Note that in all plots the data point with $k = 1$ is referring to the original data.

Figure 4.4 consists of a set of histograms for three different protection degrees, $k = \{2, 3, 10, 20\}$, showing the distance distributions by intervals. If we compare the original distribution, Figure 4.1, with the smallest protection degrees, Figures 4.4a and 4.4b, we realize that both data sets show different behaviors. In the Reuters case it is clearly appreciable the reduction of higher distances. In particular, the interval with the highest distance has decreased from the 31.5%, originally, to the 3.6% with $k = 3$, and most of those distances are spread in lower intervals, such as in the $[0.9, 1)$ interval. However, in the Movie Reviews case, this effect is not happening despite the nature of the protection algorithm. This fact gives us an idea of how far are data vectors between them, since the data representatives are as far as the real data vectors. In Figures 4.4c and 4.4d, in which protection degrees are much higher than before, the nature of the algorithm is clearer. As before, Reuters dataset shows a clear distance redistribution where higher distance values are decreasing in favor to the next interval with lower distance values. Nevertheless, as in Movie Reviews almost all distances were originally very high, the protected representatives are not as accurate as in the Reuters corpus and this produces extreme changes on distributions of distances.



(a) k = 2

(b) k = 3

(c) k = 10

(d) k = 20

Figure 4.4: Distance distribution by intervals.

4.4.3 Evaluation

As the presented anonymization algorithm is considered a general purposed protection technique (i.e., it is not known the intended use of the protected data), we have selected a set of different but basic information retrieval techniques, so different aspects of the anonymized data can be evaluated. This evaluation is conducted by comparing the results obtained by the original data with the results obtained by the protected data. Firstly, we use a basic clustering evaluation, relying on intra and inter cluster similarity; secondly, we provide an evaluation by comparison of search query result sets; and finally, we consider an evaluation by comparing result sets when different classification and clustering techniques are performed.

Intra/Inter-Cluster Sum of Errors

The first evaluation relies on the set of measures presented in Section 2.5.1 to evaluate the information loss produced by microaggregation. As these measures were directly related to the distance and the aggregation functions used by classical microaggregation (Euclidean distance and arithmetic mean), we should also re-define them to fit with the presented spherical microaggregation. Hence, we evaluate the information loss in terms of the intra/inter-cluster sum of errors. That is,

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (d_{cos}(\vec{d_{ij}}, \bar{\vec{d_i}})) \tag{4.10}$$

where g is the number of groups and n_i the number of vectors in the ith group, so $n_i \geq k$ and $n = \sum_{i=1}^{g} n_i$. In the same way $\vec{d_{ij}}$ is the jth record in the ith group and $\bar{\vec{d_i}}$ denotes the average data vector over the ith group.

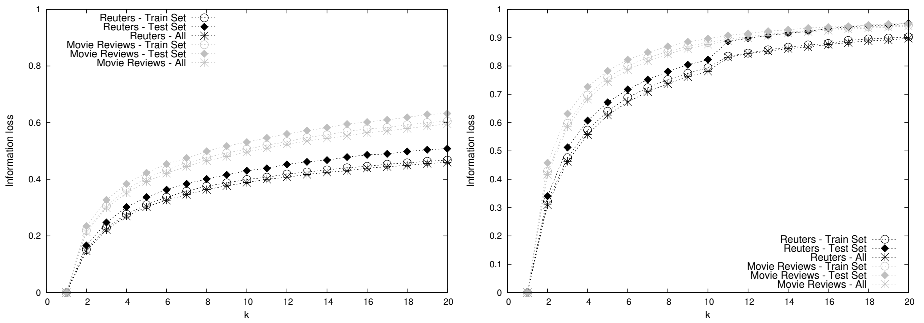
$$SSA = \sum_{i=1}^{g} n_i (d_{cos}(\bar{\vec{d_i}}, \bar{\vec{d}})) \tag{4.11}$$

where $\bar{\vec{d}}$ is the average vector over the whole set of n vectors.

Finally, the normalized information loss is defined as follows,

$$IL = \frac{SSE}{SSE + SSA} \tag{4.12}$$

Figure 4.5 shows the relation between the normalized information loss and the protection degree k for the proposed microaggregation, Spherical microaggregation (Figure 4.5a) and the original microaggregation method (Figure 4.5b). Each method was evaluated with its specific information loss measure, Equations (4.12) for the Spherical microaggregation and Equation (2.11) for the original microaggregation. As expected, in both methods data utility is decreasing as the protection degree is increasing. Moreover, we can see that, in general, vector spaces with more vectors (bigger datasets) have lost less information. Both



(a) Spherical Microaggregation

(b) Classical Microaggregation

Figure 4.5: Information Loss.

the original and the Spherical microaggregation have been used to protect the same normalized datasets. As the data was normalized the Euclidean and the cosine distances are equivalent (see Equation (4.8)) and hence, both methods built the same data partitions. However, unlike the original microaggregation, which generates the cluster representatives using the mean vector, the Spherical microaggregation generates normalized cluster representatives and so, a normalized protected dataset. This difference is relevant when we are dealing with sparse data and so, we are not interested in preserving the magnitude of the vectors. If we compare the information loss produced by both methods, we see a significant improvement achieved by the Spherical microaggregation. As we can appreciate in Figures 4.5a and 4.5b the information loss produced by the Spherical microaggregation is always much lower than the one produced by the original microaggregation. If we focus on the minimum and maximum information loss values we see a significant difference. On the one hand, the Spherical microaggregation obtained a minimum ratio of information loss of 0.17, while the original microaggregation obtained a ratio of 0.31. On the other hand, the Spherical microaggregation produced a maximum ratio of information loss of 0.65, while the maximum ratio of information loss produced by the original microaggregation is about 0.95.

Querying Vector Spaces

A common use of the VSM is to use it as an index to be queried. In order to measure the loss of information introduced by our method we can compare the results of querying the original vector space with the results obtained in the protected versions.

To that end we have built a simple and generic index from each vector space model, making use of an inverted index from the vector space and a cosine score. An inverted index consists of a list of all the unique words that appear in any document, and for each word, a list of the documents in which it appears. The query is represented as a vector $\vec{q}$ from the vector space, and the search engine returns

q0	the n most frequent terms in the collection.
q1	the n least frequent terms in the collection.
q2	n random terms.
q3	n random terms.
q4	the n terms with higher average weight per document.
q5	the n terms with higher weight.

Table 4.2: Test queries.

a ranked list of documents, where the rank is determined by the cosine distance between the query vector and the document vectors. See [Manning et al., 2008a] for more details. Notice that searching the same query in a set of inverted index built from different vector spaces will return a result in almost the same time.

We have used 6 different queries described in Table 4.2. These are divided in three different types. The first type, $q0$ and $q1$, are queries related to the number of times a term appears in the corpus. The second type, $q2$ and $q3$, are queries with terms extracted randomly from the corpus. And finally, the third type of queries are related to the weight given to each word in the feature selection step, in this case the weight is the tf-idf of each term.

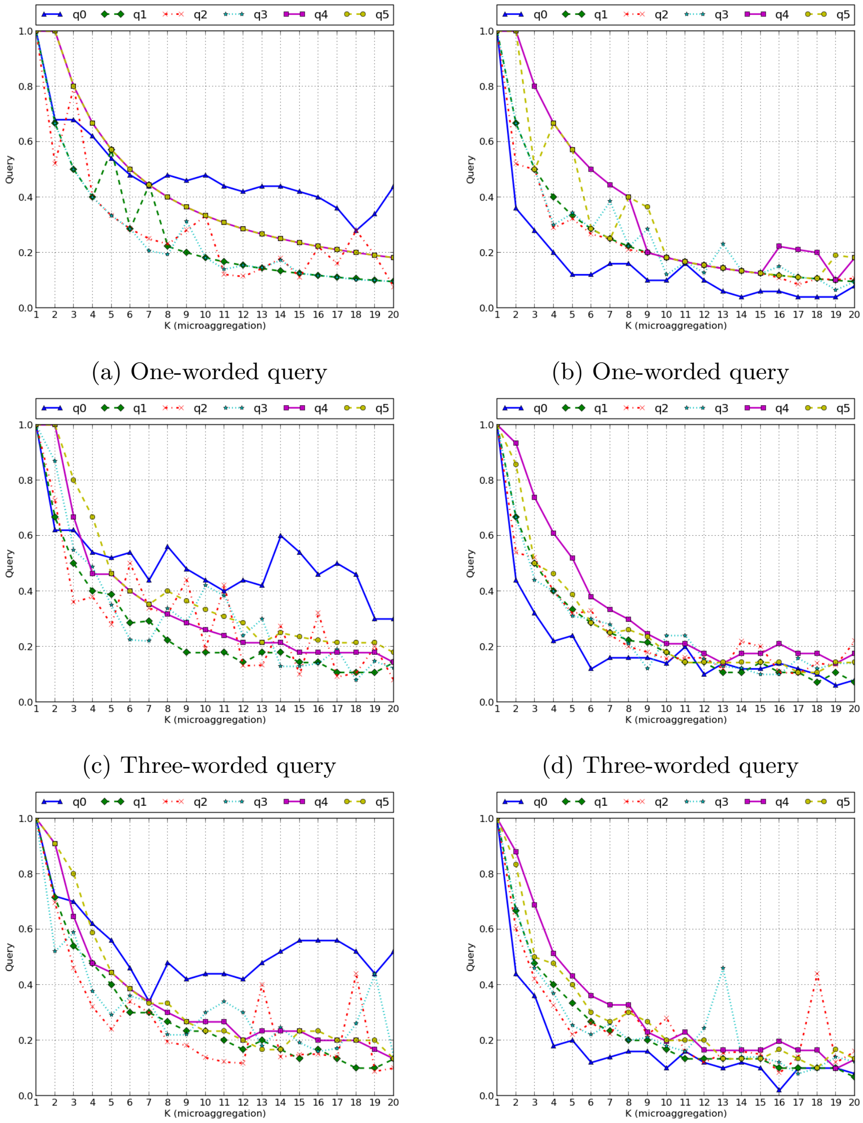
The queries are fed to the search engine which returns a list of the r documents with the highest rank. And we compare the resulting list R from the original vector space with the protected one R' by means of the harmonic mean between precision and recall. That is, the *F-measure* (also known as $F1$-*score*) which was described in Table 3.2.

Figure 4.6 shows how the F_1-score ratio has a decreasing tendency while the protection degree, k, gets higher. The left hand side of this figure corresponds to the results when the Reuters corpus is queried with queries of sizes 1, 3 and 5 and the right hand side corresponds to the results when the Movie Reviews corpus is queried.

Classification

This last part is focused on the comparison of results given by different classification and clustering algorithms, two widely extended techniques used within the information retrieval and text mining field. As in the previous section, the data utility evaluation is done by a comparison between the results obtained with the original data and its protected version when a classification or a clustering algorithm is performed. Two different metrics were used. For the classification algorithms we have used the Jaccard index [Tan et al., 2005], which is defined as the size of the intersection divided by the size of the union of two label sets. While for the clustering algorithms we have used the Adjusted Rand Index [Hubert and Arabie, 1985]. The Rand Index considers whether pairs of elements are in the same or different clusters in two clustering partition (in our case, the cluster partitions obtained by the original data and the protected data).

To evaluate the utility of the protected data, and taking into account



(a) One-worded query

(b) One-worded query

(c) Three-worded query

(d) Three-worded query

(e) Five-worded query

(f) Five-worded query

Figure 4.6: F_1-Score ration of equal documents returned querying the original and the protected data sets for different query lengths. The left column corresponds to the results of the Reuters and the right column tho the Movie Reviews corpus.

that both data sets are labeled, we have considered two well known classifier algorithms: the K-Nearest Neighbors [Duda et al., 2012] and the Naive Bayes [Manning et al., 2008a]. Besides, we have also considered the well known K-Means clustering algorithm, more precisely its spherical version [Dhillon and Modha, 2001].

In order to perform the evaluation by the **classification** algorithms two data partitions are necessary, training and test. As it was explained in Section 4.4.1, two partitions were extracted from the two initial corpus. So that for each corpus there are 3 vector spaces, the whole vector space and its two corresponding partitions, the training and the test sets. Then, all these data sets were independently anonymized. Note, that in a reverse process, where the partitions are done after data protection, some errors could be introduced due to a separation of a k-anonymous cluster. These data partitions also allow us to define a set of scenarios to estimate how much information has been lost in the anonymization process.

Scenario 1. The algorithm is trained with an original training data partition and then the model is tested with a protected test data partition. An example of this scenario would be when the data owners build their model from their original data. Then they can use this model to classify data provided by a third party and thus anonymized. Results in Figure 4.7.

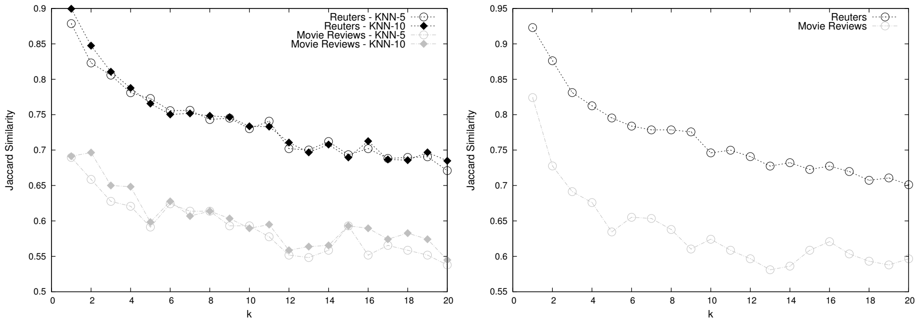
Scenario 2. The training and test partitions are extracted from the same protected data set. We assumed that the released protected data is totally or partially labeled. Results in Figure 4.8.

Scenario 3. The algorithm is trained with a protected training data partition and then the model is tested with an original test data partition. This is probably the most strange scenario, however, we have considered it because it provides a good evaluation of the protected data. That is, we can evaluate how good is the model obtained from the protected data when compared with the one obtained from the original data. Results in Figure 4.9.

Figures 4.7, 4.8 and 4.9 show the results of the defined scenarios. In the left hand side of all these figures we show the *KNN* results using 5 and 10 as the number of selected neighbors to decide the class of the tested document vector. In the right hand side there are the plots corresponding to the *Bayes* classification results.

In the case of the **clustering** algorithm, the spherical k-means, we have just evaluated *Scenario 2*, where the data owner releases the whole vector space after being protected. Thus, the evaluation is expressed in terms of how different are the anonymized clusters with respect to the original ones. The spherical k-means experimental setup has been performed as follows.

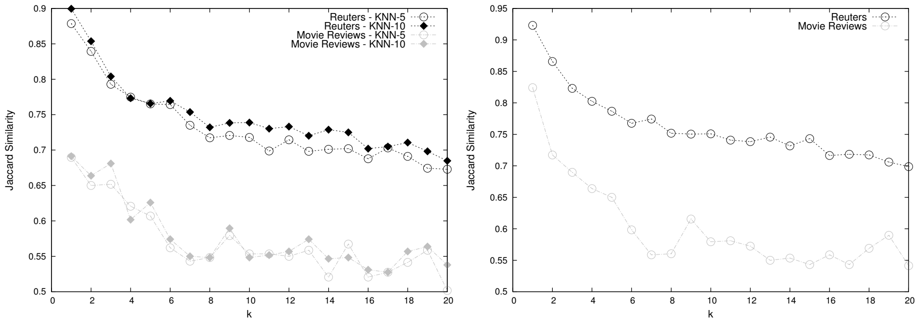
We have executed the spherical k-means clustering for both original vector spaces, Reuters and Movie Reviews, and also for all their 19 anonymized variations. For the k-means we have considered the number of clusters parameter



(a) K-Nearest Neighbor

(b) Naive Bayes

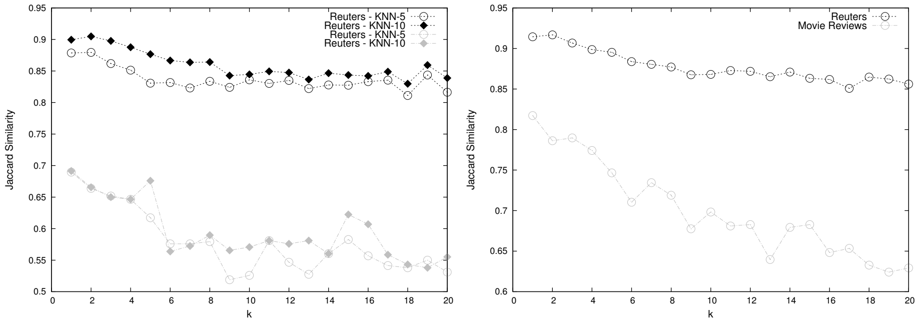
Figure 4.7: First scenario - KNN - BAYES



(a) K-Nearest Neighbor

(b) Naive Bayes

Figure 4.8: Second scenario - KNN - BAYES



(a) K-Nearest Neighbor

(b) Naive Bayes

Figure 4.9: Third scenario - KNN - BAYES

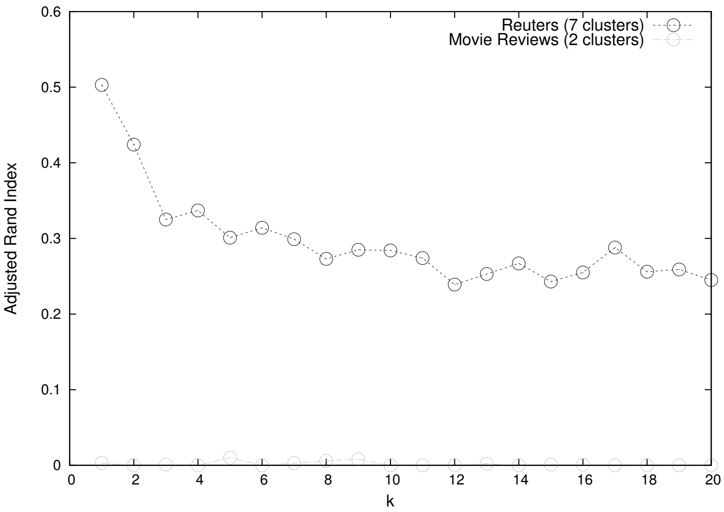


Figure 4.10: Comparison between the spherical k-means and the expected labels.

as 5, 10, 15, 20 and 25 for all the executions. Besides, in order to avoid that the algorithm converges to a local minimum we have repeated each execution 10 times and we have got the best cluster partition. In addition, since both corpus are labeled, we have also considered the number of different labels as the number of clusters considered by the algorithm, which are 7 and 2 for the Reuters and Movie Reviews, respectively. However, as it is shown in Figure 4.10 in both instances the cluster partition and the labels partition do not match, especially in the Movie Reviews example.

Figure 4.11 shows how similar are the cluster partitions, in terms of the adjusted rand index, between the original vector spaces and their respective 19th protection variations. Although the Reuters results are not so bad the Movie Reviews results are in some way surprising. However, this big difference between the original clustering and the protected one is due to the distance between all Movie Reviews data vectors. As we have shown, in Figure 4.1 almost all distances between all data pairs were in the $[0.9, 1)$ interval. This means that all documents are dissimilar in terms of cosine distance, and therefore the centroids built from each group are too general. Hence, the data distribution can limit the performance of a distance-based anonymization method. Note that this is an extreme case, since we are considering 12000 features per vector. It is known that extremely high dimensional data has problems with k-anonymity based methods [Aggarwal, 2005], which usually rely on distance functions.

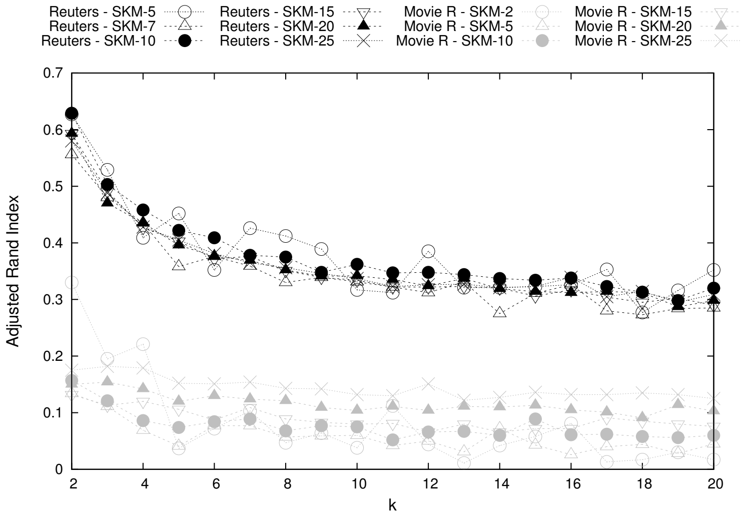


Figure 4.11: Comparison between original and protected clustering partitions.

4.5 Semantic-based Microaggregation

In this section we present which are the changes that should be applied to the classic microaggregation algorithm (Section 2.4) in order to improve their functionalities in a semantic level.

The main disadvantage of VSM is that it cannot exploit semantic words relations because it is based on weighting schemes such as term frequency only capture the number of occurrences in the document. However, in some scenarios these semantic relations could be a useful information to improve text analysis techniques. A simple example showing this disadvantage is described below.

Example 5. *Given two example sentences,*

- *'I like rock'*
- *'I love music'*

The meaning of both sentences are clearly close, even we are able to say that the second one has a more general meaning than the first. However, by using Euclidean or Cosine distance with both vectors showed in the binary weighting based vector space text representation (Table 4.3), we are not able to reach the previous conclusions; one sentence is a generalization of the other.

In order to provide words semantic relations to microaggregation we have implemented an adaptation of microaggregation relying on the WordNet lexical database. It provides a net of interlinked word cognitive synonyms (synsets) by means of its conceptual semantic and lexical relations. In Section 2.7.4 we described WordNet and its semantic word relations.

I	like	love	music	rock
1	1	0	0	1
1	0	1	1	0

Table 4.3: VSM representation based on a binary scheme.

Therefore, by exploiting the WordNet structure it is possible to define a dissimilarity measure between pairs of words, this is defined in Section 4.5.1. Then, relying on this dissimilarity we define a document vector distance in Section 4.5.2 and finally, Section 4.5.3 describes the aggregation function also relying on WordNet structures.

4.5.1 Term dissimilarity

We rely on the Wu & Palmer measure [Wu and Palmer, 1994], which provides a similarity function between two concepts in an ontology. This similarity principle is based on the edge counting and in the WordNet context, given two synsets, s_1 and s_2, it is defined as follows,

$$sim_{wup}(s_1, s_2) = \frac{2\ depth(lcs(s_1, s_2))}{depth(s_1) + depth(s_2)} \tag{4.13}$$

where $lcs(s_1, s_2)$ denotes the least common subsumer (i.e., the most specific ancestor node) between two synsets s_1 and s_2 in a WordNet taxonomy. $depth(s)$ is the length of the path from s to the root of the taxonomy. Given that multiple inheritance is allowed in WordNet taxonomies, there might be more than one candidate for $lcs(s_1, s_2)$, in this case the deepest one in the taxonomy is chosen. Note that this similarity ranges from 1 (equal synsets) to 0 (actually never reaches 0, which is only assigned to non-comparable synsets). An example relying on the WordNet structure of Figure 2.11 is shown below,

Example 6. *Given two synsets, s_1 = 'compact' and s_2 = 'truck' we compute its similarity as*

$$sim_{wup}(s_1, s_2) = \frac{2\ depth(lcs(s_1, s_2))}{depth(s_1) + depth(s_2)} = \frac{2n_3}{n_1 + n_2} = 0.88$$

where n_1, n_2 and n_3 are the depths of 'compact', 'truck' and 'Motor vehicle' respectively. Note that 'Motor vehicle' is the most specific ancestor of synsets 'compact' and 'truck'. Figure 4.12 shows the necessary fragment of WordNet.

We can easily convert sim_{wup} into a dissimilarity function. That is,

$$dst_{wup}(s_1, s_2) = 1 - sim_{wup}(s_1, s_2) \tag{4.14}$$

To carry out a term (or word) dissimilarity we should search which are the conceptual meanings of the couple of words we are going to compare. However, homonym words, i.e., words that share the same spelling and pronunciation but

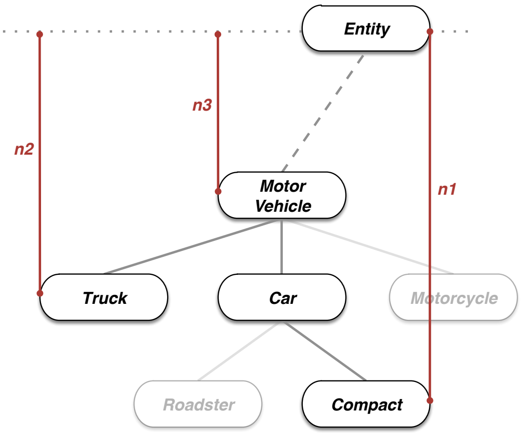


Figure 4.12: Wu & Palmer example.

have different meanings, are a frequent issue. This fact makes that in WordNet we found several conceptual meanings (synsets) for almost each word. Thus, as we have not considered a word sense disambiguation task in the pre-processing step, we opted for considering all term synsets. Then, the dissimilarity between two terms is defined as the minimum dissimilarity between both sets of synsets, each set corresponds to a term.

If we denote the set of synsets of the term t as $syns(t)$, we define the term dissimilarity dst_t between two terms, t_1 and t_2, as follows:

$$dst_t(t_1, t_2) = \min\{dst_{wup}(s_i, s_j) \mid (s_i, s_j) \in syns(t_1) \times syns(t_2)\} \tag{4.15}$$

Example 7. *Given two word such as 'butterfly' and 'computer', we look for all their noun synsets. Their definitions according to WordNet are the followings:*

- *⟨computer.n.01⟩: a machine for performing calculations automatically.*
- *⟨calculator.n.01⟩: an expert at calculation (or at operating calculating machines).*
- *⟨butterfly.n.01⟩: diurnal insect typically having a slender body with knobbed antennae and broad colorful wings.*
- *⟨butterfly.n.02⟩: a swimming stroke in which the arms are thrown forward together out of the water while the feet kick up and down.*

Note that synsets are denoted with the word name followed by the letter 'n', denoting it is a noun and finally an identification number.

Table 4.4 shows all possible synsets dissimilarities and the minimum one is highlighted in bold. Therefore, the term dissimilarity between 'butterfly' and *'computer' is* 0.4286*, which is the dissimilarity between synsets* ⟨*butterfly.n.01*⟩ *and* ⟨*calculator.n.01*⟩*.*

	⟨computer.n.01⟩	⟨calculator.n.01⟩
⟨butterfly.n.01⟩	0.6190	**0.4286**
⟨butterfly.n.02⟩	0.9048	0.8888

Table 4.4: Example of Wu-Palmer dissimilarity between all synsets of 'computer' and 'butterfly'.

Despite of having different meanings, the selected synsets which provide the minimum dissimilarity have a common concept, which is that both are organisms.

4.5.2 Semantic dissimilarity

Once we have defined the semantic dissimilarity between two terms, we can use it to define a semantic measure between two document vectors, which is the first essential point to perform the partition task of the microaggregation process.

Document vector representation are large vectors of words, hence in order to compute the dissimilarity between two document word-vectors, we have selected those words with a non-null weight and then, we have sorted all its elements in a descending order. Then, we obtain a vector of tuples (weight, term). In that way, the jth document of the collection is expressed as a term document vector $\vec{d_j}$. Formally defined as follows,

$$d_j = ((t_{\sigma(1),j}, \omega_{\sigma(1),j}), (t_{\sigma(2),j}, \omega_{\sigma(2),j}), \ldots, (t_{\sigma(n),j}, \omega_{\sigma(n),j})) \tag{4.16}$$

where n is the number of words of the document and σ is a permutation such that $\omega_{\sigma(i),j} \geq \omega_{\sigma(i+1),j}$ for all $i = 1, .., n-1$.

The dissimilarity between two terms was described as the minimum dissimilarity between both sets of synsets corresponding one for each term. Therefore, when we are dealing with big sets of words lots of computations are required; first extracting all synsets for each word and then computing the pair-wise dissimilarity between all sets of synsets. One solution would be to pre-compute and save a database with all dissimilarities between all synsets, but currently WordNet has $117,000$ synsets which means $117,000^2$ dissimilarity calculations. Thus, we opted to consider another more naive strategy. We compute dissimilarities between two documents taking advantage of the weighting-order of defined term vectors. Weighting schemes such as the term frequency allows us to define which is the relevance of each word within a document. Thus, we measure the term dissimilarities between the first term in one of the vectors and all terms of the other vector. Then, those pair of terms which achieve a minimum dissimilarity are not considered in the following measure computation for the terms of the other vectors. That is, when a relationship is established between two terms, they cannot be used again in other relationships. This process is repeated until there are no pair of terms to measure. Finally, the mean of these minimum dissimilarities is returned. Formally, this is defined as:

$$dstd_{\nu}(\vec{v_1}, \vec{v_2}) = \frac{1}{len(\vec{v_1})} \sum_{t_i \in \vec{v_1}} \alpha(t_i, t_j) dst_t(t_i, t_j) \tag{4.17}$$

where t_j is the $\arg\min_{t_j \in \nu_2} dst_t(t_i, t_j)$, the $\alpha(t_i, t_j)$ is the frequency mean of t_i and t_j and $len(d_1)$ is the length of the first vector.

Unfortunately, this function is not commutative. In order to make it commutative, we define the distance as the minimum of the two possible parameterizations. That is,

$$dst_{\nu}(\vec{v_1}, \vec{v_2}) = \min\{dstd_{\nu}(\vec{v_1}, \vec{v_2}), dstd_{\nu}(\vec{v_2}, \vec{v_1})\} \tag{4.18}$$

Example 8. *We consider two simple term vectors $\vec{v_1}$ and $\vec{v_2}$ with four terms each one with their respectively frequencies:*

- $\vec{v_1} = (('butterfly', 0.4), ('performance', 0.2), ('pen', 0.1), ('dog', 0.3))$
- $\vec{v_2} = (('computer', 0.1), ('cat', 0.2), ('approach', 0.2), ('beetle', 0.5))$

Table 4.5 shows all the dissimilarities between the terms of the different vectors and emphasizes the minimum ones. Note that the dashes in this table denote relationships with terms that cannot be considered. That is, when a term of the second vector is used, it cannot be used again.

	'beetle'	'cat'	'approach'	'computer'
'butterfly'	**0.130**	0.454	0.238	0.428
'dog'	-	**0.143**	0.375	0.22
'performance'	-	-	**0.25**	0.60
'pen'	-	-	-	**0.333**

Table 4.5: Dissimilarities between terms of two vectors.

Finally, we calculate the mean of the products between these minimum dissimilarities and the frequency mean of both terms (i.e., t_i in document 1 and the most similar *t_j in document 2).*

$$dstd_{\nu}(\vec{v_1}, \vec{v_2}) = \tfrac{1}{4}(0.059 + 0.050 + 0.036 + 0.0333) = 0.044$$

4.5.3 Semantic Aggregation

The second operation of the microaggregation is the aggregation, which computes a new vector, that represents the cluster representative or centroid. In this case, we need to form this centroid taking into account the semantic meaning of the different elements of the vectors.

The semantic aggregation process for two different term vectors is defined as the aggregation function $\mathbb{C}_\nu$:

$$\begin{aligned} \mathbb{C}_\nu(\vec{v_1}, \vec{v_2}) = \bigcup_{t_i \in \vec{v_1}} \{ & lch(t_i, \arg\min_{t_j \in \vec{v_2}} dst_t(t_i, t_j)), \\ & \alpha(t_i, \arg\min_{t_j \in \vec{v_2}} dst_t(t_i, t_j))\} \end{aligned} \tag{4.19}$$

where $lch(t_i, t_j)$ (the lowest common hypernym) denotes the lowest term in the WordNet hierarchy, which both terms, t_i and t_j, have in common, and $\alpha(t_i, t_j)$ is the mean frequency of both terms. This ensures the preservation of the frequencies in the microaggregated data. Note that the term dissimilarity is used again to find a semantically closer relation between terms of both vectors, in order to generalize the meaning of each pair in one term using the function lch.

As in Equation (4.17) ($dstd_\nu$), the aggregation operation is not commutative. Again, to make it commutative, we compute the minimum of the two parameterizations.

The definition of $\mathbb{C}_\nu$ is only for two term vectors, although it can be generalized easily. To do aggregations for more than two term vectors, we iterate the process aggregating the new computed aggregated vector with the remaining vectors. This process is bootstrapped with the aggregation of the first two term vectors.

Example 9. *Let us consider the given term vectors in Example 8, $\vec{v_1}$ and $\vec{v_2}$. The aggregation of both vectors $\mathbb{C}(\vec{v_1}, \vec{v_2})$ is showed in Table 4.6. That is, the lowest term in common in the WordNet hierarchy is computed by those terms with minimum distance. Additionally, the mean frequency of those terms is showed.*

	’beetle’, 0.5	’cat’, 0.2	’approach’, 0.2	’computer’, 0.1
’butterfly’, 0.4	’insect’, 0.45	-	-	
’dog’, 0.3	-	’carnivore’, 0.25	-	
’performance’, 0.2	-	-	’action’, 0.2	
’pen’, 0.2	-	-	-	’instrumentality’, 0.1

Table 4.6: Aggregation between two term vectors.

4.5.4 Illustrative example

In order to understand better the semantic microaggregation process explained above, we give a toy example, using an original small dataset integrated by four documents as input of the process.

Table 4.7 (top) shows firstly the original file, integrated by four documents with three terms and their respective term frequency for each of them. The table also shows the protected output file obtained after the microaggregation process with a k value of 2. As you can see the output file has four documents as the original file, but it only has two different records. The centroid of the first document represents the set of documents that are related to computers parts, and the second one join the two original documents that are related to different animals. Therefore, we can say that with the protected file we can deduce the general topics of the documents, but we cannot know the specific topics of the original dataset.

Original Data File
((’keyboard’, 0.3), (’laptop’, 0.4), (’software’, 0.3))
((’horse’, 0.7), (’dog’, 0.2), (’cat’, 0.1))
((’hardware’, 0.3), (’screen’, 0.3), (’computer’, 0.4))
((’lion’, 0.5), (’monkey’, 0.3), (’tiger’, 0.2))

Protected Data File
((’abstraction’, 0.3), (’computer’, 0.4), (’instrumentality’, 0.3))
((’big_cat’, 0.3), (’carnivore’, 0.2), (’placental’, 0.5))
((’abstraction’, 0.3), (’computer’, 0.4), (’instrumentality’, 0.3))
((’big_cat’, 0.3), (’carnivore’, 0.2), (’placental’, 0.5))

Table 4.7: Example of semantic microaggregation with $k = 2$.

4.6 Semantic Experimental Results

In this section we evaluate the anonymous data generated by the purposed semantic variation of microaggregation masking method. This assessment relies on a modification of the microaggregation-oriented information loss measure.

This section was divided in two parts; Section 4.6.1 describes the data that have been used to test our proposal and in Section 4.6.2 we present the results obtained when the masked data is evaluated in accordance with an intra/inter cluster comparison.

4.6.1 Original Data

In order to evaluate the semantic microaggregation described we have selected 50 papers published during three years (2007, 2008 and 2009) in the *Modeling Decisions for Artificial Intelligence* (MDAI) conference. From this first dataset we have created two different datasets. On the one hand, we have considered the 50 most relevant terms and on the other hand, we have built another dataset considering the 100 most relevant terms. We have called them *f50x50* and *f100x50*, respectively.

To do so, we have followed the pre-processing step described in Section 2.7.3. Reading, tokenizing and cleaning stop-words tasks were applied, the stemming process were not applied because WordNet is not able to recognize the word stems in its hierarchy. We have also considered a new task within the cleaning process. This task consists of the elimination of words which are not included in the WordNet database. This results in some minor loss of information, leading the loss of some common names such as from the bibliography of each paper, or some very specific and technical terms. Note that this process of removing names and very specific words adds and additional protection level.

Afterwards, we have computed the normalized term frequency-inverse document frequency, tf-idf, Equations (2.17) and (2.18), in order to generate the VSM with the tf-idf weighting scheme. Finally, we have applied a feature reduction, by selecting on the one hand the 50 words with the highest weights, and on the other hand the 100 words with the highest weights.

Regarding to the additional token cleaning task, we have compared both datasets, *f50x50* and *f100x50*, with and without considering this removing task. The averaged similarity computed by the Jaccard similarity[1] between both datasets is 0.769557 for *f50x50* and 0.771153 for *f100x50*. Recall, that this experiment is just an illustrative experiment that could be improved by considering domain-specific ontologies.

In order to anonymize the data we have implemented the microaggregation heuristic method MDAV, Algorithm 1, with the corresponding changes in the partition and aggregation parts proposed in Sections 4.5.2 and 4.5.3. Then, both data files have been protected with different values of the parameter k in the range from 2 to 10. We have not computed values of k greater than 10 due to the limited size or the test dataset, and to the fact that as we will see, with $k = 10$ we already have a high degree of information loss.

4.6.2 Evaluation

We have evaluated the information loss produced by the semantic microaggregation by means of the specific information loss measure which was described in Section 2.5.1. Nonetheless, as we did in Section 4.4.3, we have to modify the original set of measures to correctly evaluate the microaggregation variation proposal. Thus, the intra and inter cluster evaluations are now defined in terms of $dstd_\nu$ and $\mathbb{C}_\nu$ as follows,

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (dstd_\nu(\vec{d_{ij}}, \bar{\vec{d_i}}))^2 \tag{4.20}$$

where g is the number of groups and n_i the number of individuals in the ith group. Naturally, ($n_i \geq k$ and $n = \sum_{i=1}^{g} n_i$). In the same way $\vec{d_{ij}}$ is the jth record in the ith group and $\bar{\vec{d_i}}$ denotes the average data vector over the ith group.

$$SSA = \sum_{i=1}^{g} n_i (dstd_\nu(\bar{\vec{d_i}}, \bar{\vec{d}}))^2 \tag{4.21}$$

where $\bar{\vec{d}}$ is the average vector over the whole set of n individuals.

[1] The Jaccard similarity coefficient measures the similarity between two sets A and B as $\frac{|A \cap B|}{|A \cup B|}$

Then, the normalized information loss is defined in term of the previous SSE and SSA functions as,

$$IL = \frac{SSE}{SSE + SSA} \tag{4.22}$$

k	Data set	SSE	SSA	IL
2	f50x50	4.938	30.929	13.766
	f100x50	4.936	37.119	11.736
3	f50x50	11.407	21.390	34.780
	f100x50	12.049	29.733	28.838
4	f50x50	15.693	21.556	42.131
	f100x50	16.647	22.759	42.245
5	f50x50	20.404	11.890	63.181
	f100x50	21.070	19.157	52.377
6	f50x50	23.072	17.372	57.046
	f100x50	24.516	18.336	57.212
7	f50x50	25.109	11.332	68.903
	f100x50	26.712	18.981	58.560
8	f50x50	27.034	8.986	75.053
	f100x50	27.662	16.101	63.209
9	f50x50	28.529	10.085	73.883
	f100x50	30.107	11.657	72.088
10	f50x50	31.670	5.680	84.793
	f100x50	31.455	10.857	74.341

Table 4.8: Semantic microaggregation evaluation.

Table 4.8 shows the evaluation values defining how optimal is the k-partition for each one of these protected files. As expected, the SSE values increases as k increases. It means that within-group homogeneity decreases when the number of documents per cluster increases. On the contrary, SSA values decrease when k decrease. This is reasonable because when k grows, there are less centroids and homogeneity between clusters decreases. Finally, we focus on the information loss. As expected, when k increases, the information loss also increases. Moreover, we can appreciate that the dataset with 50 terms, *f50x50*, results into a higher information loss than the dataset with 100 terms. You can see it clearly in Figure 4.13.

After the analysis, we can say that the best parameter is the k with values between 3 and 5, because they are the ones with a lower information loss value. At this point, we do not consider 2 as an acceptable value for k, because in this case the protection level is too weak for ensuring data confidentiality.

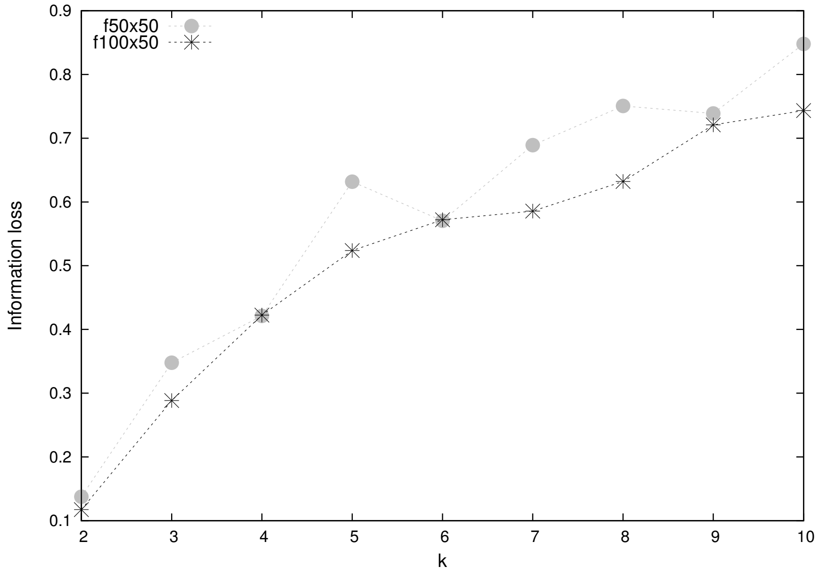


Figure 4.13: Information loss (IL) vs. privacy level k.

4.7 Conclusions

In this chapter we have introduced two protection methods to anonymize VSM, an algebraic representation commonly used in information retrieval to represent textual data as a term-document matrix. These anonymizations are two variations of a well known clustering-based masking method, microaggregation. While one is focused on improving the protection in sparse and high-dimensional datasets, the second approach is focused on exploiting the existing semantic relations between words.

The first approach was motivated by the sparsity and high-dimensional nature of VSMs. Microaggregation is a clustering-based masking method based on the Euclidean distance and many authors demonstrate that this distance is a weak discriminant when data has the VSMs properties. Therefore, we proposed spherical microaggregation, a variation of microaggregation based on the cosine distance. In order to be consistent to that distance we also had to adapt the aggregation functions of the algorithm, so this function has to compute the cluster representative that minimizes the sum of distances between all cluster members and itself. Finally, we have implemented and evaluated a variation of the heuristic microaggregation algorithm, MDAV, with the proposed modifications.

Our results show how in most of the presented evaluation tests the data is highly useful, specially in the classification tests and the sum of errors. The results obtained by the traditional evaluation, sum of errors, were satisfactory despite the sparsity of the data, we have obtained less than a 50% of information loss for protection degrees lowers than 10 for the Movie Reviews data sets, and

20 for the Reuters data sets. In addition, in the classification tests the prediction results of protected data are quite similar to the ones obtained by the original data. Specifically, the maximum error obtained by the highest protection degree ($k = 20$) in the Reuters data is about a 20% and about a 15% in the Movie Reviews. Otherwise, the results obtained by the clustering algorithm are much more discouraging. The partitions obtained with the protected data are very dissimilar to the original, specially, in the Movie Reviews data sets due to its distance distribution.

The second approach is another variation of microaggregation that exploits the semantic relations between words. One of the drawbacks of VSM is that it cannot exploit the semantic word relations because it is based on weighting schemes. On that account, we proposed the semantic microaggregation, a variation of this popular masking method that allows dealing with these word semantic relations. The key point of this method is the ability to integrate word hierarchies. That is, by means of using ontologies like WordNet, in which words are linked by its conceptual meanings, we are able to define word similarities as well as word generalizations. Thus, we have implemented the MDAV algorithm according to a defined dissimilarity and aggregation functions. Our current approach can be seen as a generic base, which can be better adapted if the information retrieval task to be applied to the data is known beforehand.

Chapter 5

Supervised Learning for Record Linkage

In this chapter a new supervised metric learning approach for distance-based record linkage is introduced. The main goal of this approach is to achieve better estimates of the disclosure risk of sensitive information of protected data files. This supervised learning approach finds out the set of metric parameters according to the corresponding linkage constraints such that we will obtain a higher number of correct re-identification (matches) in the linkage process.

Our contribution is two fold. On the one hand, it improves the accuracy of standard disclosure risk evaluations, which are based on distance-based record linkage methods. On the other hand, by learning those distance parameters it is possible to extract insightful information about the relevance of attributes and sets of attributes in the linkage process. So, variables spotted as very relevant for the re-identification are the ones that should be considered riskier due to its information leakage. Therefore, once a set of variables are detected and classified as risky the data owner may take preventive actions to reinforce the anonymization in those data parts.

The performance of this approach depends critically on a given metric. The choice of a distance metric over an input space always has been a key issue in many machine learning algorithms. Hence, the use of convenient and intuitive distances, such as the commonly used Euclidean distance is not always the best choice, because it fails to capture the idiosyncrasies of the data of interest considering each feature equally important and independent from the others. Because of all these problems and motivated by the distance metric learning research field, described in Section 2.6, we propose four parameterized distance-based aggregation operators of different types and complexities to be used in the supervised metric learning problem. By automatically learning the metric parameters from the specified constraints we expect to obtain a higher number of correct re-identifications between an original and a masked data file.

The parameterized metric functions studied in this chapter are the following:

a weighted average operator, an ordered weighted average (OWA) operator, a symmetric bilinear form and a fuzzy integral. Note that they are ordered here according the number of parameters considered, from the fewer to the larger set.

Supervised learning for record linkage is a powerful technique to be used in both scenarios described in Section 2.3. Note hat these parameters are the ones that maximize the number of re-identifications. With this information an attacker can tune his algorithm to get a higher number of correct matches, and the data owner can determine which are the weaker and stronger attributes in terms of risk, and then apply other stronger protections to them. Note that depending on the parametric function used, this information is more or less accurate and so different scenarios lead to different solutions. In this chapter we evaluate the advantages and disadvantages of the proposed functions in different scenarios.

This chapter is structured as follows. In Section 5.1 we present the general optimization problem in general terms as well as its formalization for any aggregation function. All the distance-based aggregation functions are presented in Section 5.2 as well as their implementation in the general optimization problem. In addition, we study their suitability for defining a distance or metric function. Section 5.3 introduces two different ways to solve the optimization problem, one by using global optimization techniques and the other based on a local optimization method. In Section 5.4 a set of experimental tests is presented to show the performance of our general proposal for each of the introduced aggregation functions. Finally, in Section 5.5 we present some conclusions.

5.1 General Supervised Learning Approach for Record Linkage

In this section we present and formalize the supervised metric learning problem for distance-based record linkage. This formalization is presented as a generalization of the record linkage problem regardless of the parameterized function used. Defining the problem in a general form allows us to create multiple variations of the problem depending on the parameterized distance function used.

The problem is modeled as a Mixed Integer Linear mathematical optimization (MILP). More formally, the stated problem is expressed with a linear objective function and it is subject to a set of linear equality and inequality constraints. The difference between MILP and Linear Programming (LP) lies in the type of the variables considered. LP just considers real-valued variables whereas, MILP involves problems in which only some variables are constrained to be integers and the other variables are allowed to be non-integers (real). This fact makes MILPs harder problems. That is, LPs can be solved in polynomial time while, MILPs there are NP-complete problems [Schrijver, 1986] and therefore, there is no known polynomial-time algorithm. In Section 5.3 we give more details about solving MILPs problems.

For the sake of simplicity in the formalization of the process, we assume that

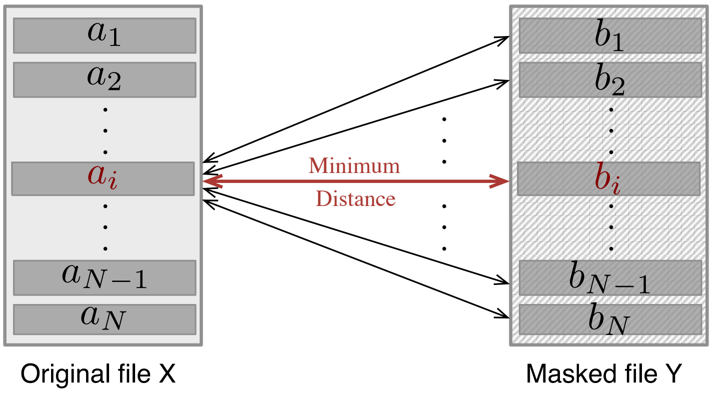


Figure 5.1: Distances between aligned records should be minimum.

each record b_i of Y is the protected version of a_i of X. That is, files are aligned. Then, two records are correctly linked using a parameterized aggregation function, $\mathbb{C}_p$, when the distance between the records a_i and b_i is smaller than the distance between the records a_i and b_j for all other j different from i. So, records belonging to the same entity are considered less distant in terms of the aggregation function. Figure 5.1 shows an illustration of this scenario. Formally, we have that a record a_i is correctly matched when the following equation holds for all $i \neq j$.

$$\mathbb{C}_p(a_i, b_i) < \mathbb{C}_p(a_i, b_j) \tag{5.1}$$

In optimal conditions these inequalities should be true for all records a_i. Nevertheless, we cannot expect this to hold because of the errors in the data caused by the protection method. Then, the learning process is formalized as an optimization problem with an objective function and some constraints.

To formalize the optimization problem and permit that the solution violates the constraint described in Equation (5.1), we have followed the divide and conquer rule, so the problem is relaxed by dividing it in several parts, called blocks. We consider a block as the set of equations concerning record a_i. Therefore, we define a block as the set of all distances between one record of the original data and all the records of the protected data. Then, we assign to each block a control variable K_i and hence, there are as many K_i as the number of rows of the original file. Besides, a constant C is needed for the formalization, it multiplies each K_i to overcome the inconsistencies and satisfy the constraints. Table 5.1 shows a graphical example of the problem division and the information needed for the learning process (which are the links that correspond to the same individuals).

The rationale of this approach is as follows. The variable K_i indicates, for each block, if all the corresponding constraints are accomplished ($K_i = 0$) or

Block	Aggregator function	Label
K_1	$\mathbb{C}_p(a_1, b_1)$	must-link
	$\mathbb{C}_p(a_1, b_i)$	cannot-link
	$\mathbb{C}_p(a_1, b_N)$	cannot-link
$\vdots$	$\vdots$	$\vdots$
K_i	$\mathbb{C}_p(a_i, b_1)$	cannot-link
	$\mathbb{C}_p(a_i, b_i)$	must-link
	$\mathbb{C}_p(a_i, b_N)$	cannot-link
$\vdots$	$\vdots$	$\vdots$
K_N	$\mathbb{C}_p(a_N, b_1)$	cannot-link
	$\mathbb{C}_p(a_N, b_i)$	cannot-link
	$\mathbb{C}_p(a_N, b_N)$	must-link

Table 5.1: Data to be considered in the learning process.

not ($K_i = 1$). That is, if for a record a_i, Equation (5.1) is violated for a certain record b_j, then, it does not matter that other records b_h, where $h \neq j \neq i$, also violate the same equation for the same record a_i. This is so because record a_i will not be re-identified. Then, the goal of the proposed optimization problem is to minimize the number of blocks non compliant with the constraints. This way, the optimization problem is able to find the combination of weights that minimize the number of violations, or in other words, those weights that maximize the number of re-identifications between the original and protected data.

Thus, Equation (5.1) can be rewritten with the control variables, K_i, and a constant C, as follows,

$$\mathbb{C}_p(a_i, b_j) - \mathbb{C}_p(a_i, b_i) + CK_i > 0.$$

for all $i \neq j$.

As K_i is a binary variable, $K_i = \{0, 1\}$, we use the constant C as the factor needed to really overcome the constraint. In fact, the constant C expresses the *minimum distance* we require between the correct link and other incorrect links. The larger it is, the more correct links are distinguished from incorrect links.

Using these constraints the optimization problem for a given parameterized aggregation function $\mathbb{C}_p$ is defined as:

$$\begin{aligned}
&Minimize \sum_{i=1}^{N} K_i && (5.2)\\
&Subject\ to:\\
&\mathbb{C}_p(a_i, b_j) - \mathbb{C}_p(a_i, b_i) + CK_i > 0, \quad \forall i, j = 1, \ldots, N, i \neq j && (5.3)\\
&K_i \in \{0, 1\} && (5.4)
\end{aligned}$$

This is an optimization problem with a linear objective function, Equation (5.2), and linear constraints, Equations (5.3) and (5.4). However, depending on which aggregation function $\mathbb{C}_p$ is going to be used, some additional constraints related to that aggregation function and its parameters should be considered and added to the problem. In addition, we have to pay special attention to which is the polynomial degree of the aggregation operator we want to use and the parameter constraints, because it could lead us to deal with non-linear or quadratic programming problems.

If N is the number of records, and n the number of variables of the two data sets X and Y. Then, the objective function, Equation (5.2), consists of a summation of N control variables, one per each defined distances' block, i.e., K_i for all $i = 1 \dots N$. With respect to the total number of problem constraints; there are $(N(N-1))$ constraints concerning to Equation (5.3) and N constraints defining the control variables, Equation (5.4). Therefore, there are a total of $(N(N-1)) + N$ constraints. Note that depending on the aggregation function $\mathbb{C}_p$ used, there will be more constraints in the problem. We will discuss the number of such constraints in the particular problems below.

5.2 Parameterized Distance-Based Record Linkage

As stated above, we have considered different parametric distances. Each of them have different complexity. That is, different number of parameters. The advantage of using functions with higher complexity levels is because they will learn more information about the data and the given constraints, so they can easily overcome the number re-identifications of simpler non-parameterized distances, such as the Euclidean distance (Definition 2.11). However, higher complexities also means more consumption of computing time and system resources. Accordingly, we should choose a specific distance depending on the situation and the precision on the number of re-identifications we want to achieve. For instance, data owner will be more interested in using an aggregation operator with larger number of parameters, so the analysis of the disclosure risk of the protected database will be more accurate, and the study of weaker and stronger protected attributes to withstand attacks will also be more accurate, providing much more information than distances with smaller number of parameters.

Figure 5.2 illustrates the classification of the different distances we have considered and that we will be explained in the following sections. As you can see the arithmetic mean is a particular case of the weighted mean when all the weights are balanced, which in turn is a specific case of OWA operators and the Mahalanobis distance. In addition, Choquet fuzzy integral is a general case of OWA operators, and the Mahalanobis distance is a symmetric bilinear form. Some more details about these relationships can be found in [Torra and Narukawa, 2012].

Before delving into the particular definitions of parameterized distance functions, we first introduce and formalize a generic definition for all parameterized aggregation functions. This generalization is based on the fact that the Eu-

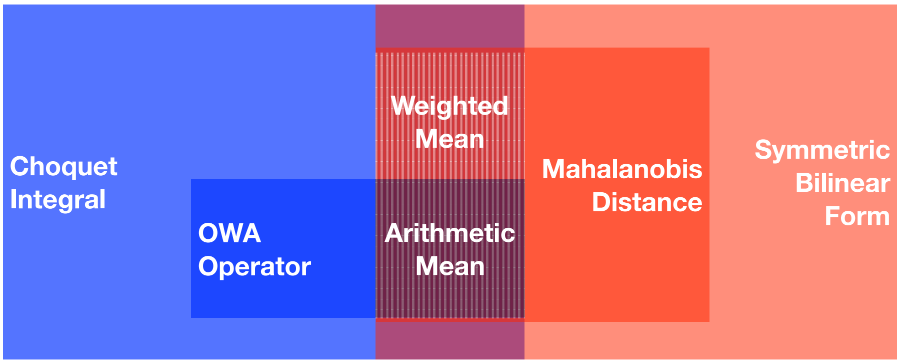


Figure 5.2: Distances classifications.

clidean distance has the same results when it is multiplied it by a *constant* and it will not change the results of any record linkage algorithm.

To do so we will use the same notation as in Figure 5.1, where all b_i records from file Y are the masked version of the records a_i from the original file X. Besides, for the sake of simplicity we consider the square of the Euclidean distance, $d^2ED(a,b)$. Although $d^2ED(a,b)$ is not a metric (it does not satisfy the triangular inequality) it is a distance according to Definition 2.9 (Section 2.2.1).

In a formal way, we redefine $d^2ED(a,b)$ as follows:

$$d^2(a,b) = \sum_{i=1}^{n} \frac{1}{n} \left(diff_i(a,b)\right)^2$$

where we define the difference between two variables from two records taking into account the normalization of data as follows,

$$diff_i(a,b) = \frac{V_i^X(a) - \overline{V_i^X}}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V_i^Y}}{\sigma(V_i^Y)} \tag{5.5}$$

In addition, we will refer to each squared term of this distance as

$$d_i^2(a,b) = \left(diff_i(a,b)\right)^2 \tag{5.6}$$

Using these expressions we can define the square of the Euclidean distance as follows,

Definition 5.1. *Given two datasets X and Y the square of the Euclidean distance for variable-standardized data is defined by:*

$$d^2AM(a,b) = AM(d_1^2(a,b), \ldots, d_n^2(a,b)),$$

where AM is the arithmetic mean $AM(c_1, \ldots, c_n) = \sum_i c_i/n$.

In general, any aggregation operator $\mathbb{C}$ might be used in the place of arithmetic mean, although it is important to note that not all the aggregation operators will satisfy all the metric properties. However, for all the proposed parameterized functions in the following sections we will give some small tricks and modifications in order that these functions satisfy as much metric properties as possible.

We can consider the following generic function to evaluate the distance between record a and b.

$$d^2\mathbb{C}(a,b) = \mathbb{C}(d_1^2(a,b),\ldots,d_n^2(a,b)) \tag{5.7}$$

for a given aggregation operator $\mathbb{C}$.

In the sections that follow we will discuss the case of using Equation (5.7) in the supervised metric learning problem for record linkage when different aggregation operators $\mathbb{C}$. In particular, we consider the weighted mean and the OWA operator (Section5.2.1), the symmetric bilinear form (Section 5.2.2) and the Choquet integral (Sectionr̃efsec:ci). In all of these sections are discussed their advantages and disadvantages as well as their metric's definition. That is, an explanation of which additional constraints should be added to each problem to consider the aggregation operator a distance metric.

5.2.1 Weighted Mean and OWA Operator

It is straightforward to consider weighted versions of Definition 5.1 using Equation (5.7). The first example is the weighted mean, which is defined as follows,

Definition 5.2. *Let $p = (p_1,\ldots,p_n)$ be a weighting vector (i.e., $p_i \geq 0$ and $\sum_i p_i = 1$). Then, the square of the weighted mean is defined as:*

$$d^2WM_p(a,b) = WM_p(d_1^2(a,b),\ldots,d_n^2(a,b)),$$

where $WM_p = (c_1,\ldots,c_n) = \sum_i p_i \cdot c_i$ and $d_i^2(a,b)$ is the squared difference between the two ith attributes of a and b, see Equation (5.6).

The interest of this definition is that we do not need to assume that all attributes are equally important in the re-identification process, since there is a weight for each attribute expressing its relevance in the re-identification process.

This definition does not satisfy all the properties of a metric given in Definition 2.10. First, the identity of indiscernibles property is not satisfied when null weights are considered. Second, the triangle inequality is not satisfied because of the square.

Another straightforward generalization of the Euclidean distance is based on the Ordered Weighted Averaging (OWA) operator. It is defined as follows.

Definition 5.3. *Let $p = (p_1,\ldots,p_n)$ be a weighting vector (i.e., $p_i \geq 0$ and $\sum_i p_i = 1$). Then, the square of the ordered weighted average operator is defined as:*

$$d^2OWA_p(a,b) = OWA_p(d_1^2(a,b),\ldots,d_n^2(a,b)),$$

where $OWA_p = (c_1, \ldots, c_n) = \sum_i p_i \cdot c_{\sigma(i)}$*, and* σ *defines a permutation of* $1, \ldots, n$ *such that* $c_{\sigma(i)} \geq c_{\sigma(i+1)}$ *for all* $i > 1$.

The interest of this operator is that while the weighted mean permits to assign relevance to attributes, the OWA operator (because of the ordering σ) permits assigning relevance to either larger or smaller values. In this way, we might give importance to extreme values or central values. Note that as it happens in the weighted mean, if null weights or the squared of this function are considered this function cannot be considered a metric. As before, and for the same reasons the identity of indiscernibles and the triangle inequality properties do not always hold.

Let us now consider the application of these definitions into the general optimization problem. Since, both operators have the same structure we express the minimization problem in this general way:

$$Minimize \sum_{i=1}^{N} K_i \tag{5.8}$$

$$Subject\ to:$$

$$d^2\mathbb{C}_p(a_i, b_j) - d^2\mathbb{C}_p(a_i, b_i) + CK_i > 0, \quad \forall i, j = 1, \ldots, N, i \neq j \tag{5.9}$$

$$K_i \in \{0, 1\} \tag{5.10}$$

$$\sum_{i=1}^{n} p_i = 1 \tag{5.11}$$

$$p_i \geq 0 \tag{5.12}$$

where $d^2\mathbb{C}_p(a, b)$ is a general notation for the weighted mean and the OWA defined operators. Hence, depending on which operator its is going to be used $d^2\mathbb{C}_p(a, b) = d^2WM_p(a, b)$ or $d^2\mathbb{C}_p(a, b) = d^2OWA_p(a, b)$. Note that in this problem formalization we have taken into account null weights. However, in the experimental Section 5.4.3 we also consider the case with non-null weights. To do so, Equation (5.12) is replaced by a more restrictive inequality: $p_i > 0$.

Remark. For any of the two proposed minimization problems we have to learn the same number of parameters, $p = \{p_1, \cdots, p_n\}$. In both cases, the problem has the same number of constraints. In addition to the $N(N-1)+N$ constraints of the general problem, we have to add 1 more constraint, Equation (5.11) and n more constraints due to the consideration of positive weights, Equation (5.12). Therefore, the minimization problem has $N(N-1) + N + 1 + n$ constraints.

5.2.2 Symmetric Bilinear Form

In this section we introduce a new and more complex parameterized operator, a symmetric bilinear form, and besides, we present the corresponding adaptation to the supervised metric learning for record linkage.

As Xing et al. shown in [Xing et al., 2003] it is possible to parameterize the Euclidean distance using a symmetric positive semi-definite matrix, $\Sigma \succeq 0$, see Section 2.6. The interest of using such an operator is because by learning a weighting matrix we obtain much more information about the data linkage, than just using a vector of n weights, as it does the weighted mean or the OWA operator.

Definition 5.4. *Given a vector space V over a field F, a bilinear form is a function $B : V \times V \rightarrow F$ which satisfies the following axioms for all $w, v, u \in V$:*

1. $B(v + u, w) = B(v, w) + B(u, w)$
2. $B(w, v + u) = B(w, v) + B(w, u)$
3. $B(\alpha v, w) = B(v, \alpha w) = \alpha B(v, w)$
4. $B(v, w) = B(w, v)$

Given a square matrix Σ, we define a bilinear form for all $v, w \in V$ as $B(v, w) = v'\Sigma w$. Note that the matrix Σ of a symmetric bilinear form must be itself symmetric. The symmetric bilinear functions can be considered a generalization of the Mahalanobis distance.

Then, we can use this symmetric bilinear form on the light of previous definitions as:

Definition 5.5. *Let Σ be an $n \times n$ symmetric weighting matrix. Then, the square of a symmetric bilinear form is defined as:*

$$d^2SB(a, b) = SB_\Sigma(diff_1(a, b), ..., diff_n(a, b))$$

where $SB_\Sigma(c_1, ..., c_n) = (c_1, ..., c_n)'\Sigma(c_1, ..., c_n)$.

Learning the symmetric weighting matrix Σ allows us to find which are the attributes and tuples of attributes that are more relevant in the re-identification process. That is, the diagonal expresses the relevance of each single attribute, while the upper or lower values of the weighting matrix correspond to the weights that evaluate all the interactions between each pair of attributes in the re-identification process.

If the matrix Σ satisfies the symmetry and the positive definiteness property all the metric properties of Definition 2.10, are satisfied. On the contrary, if this matrix restriction is weaker, the matrix is positive semi-definite, the identity of indiscernibles is not fulfilled. Thus, there will be situations where $d(a, b) = 0$ for all $a \neq b$, and then, the defined operator cannot be considered a metric anymore, it is a distance. A clear example is when Σ is completely null. Finally, if Σ is neither positive definite neither positive semi-definite, i.e, negative definite, then the defined operator is a pseudo-distance.

As we do not want negative distance values, the only requirement on Σ we have considered is that it should be at least a positive semi-definite matrix.

After the operator definition, we present the variation of the general optimization problem considering the symmetric bilinear function formalized in Definition 5.5. Then, this minimization problem is expressed as:

$$Minimize \sum_{i=1}^{N} K_i \tag{5.13}$$

$$Subject\ to:$$

$$d^2 SB_\Sigma(a_i, b_j) - d^2 SB_\Sigma(a_i, b_i) + CK_i > 0, \quad \forall i, j = 1, \ldots, N, i \neq j \tag{5.14}$$

$$K_i \in \{0, 1\} \tag{5.15}$$

where, as before, N is the number of records, and n the number of attributes of the input files.

The optimization problem will find the symmetric matrix Σ that ensures the maximum number of re-identification between the original and the protected data. Thus, we have increased the number of parameters learned by the problem from n, in the case of the weighted mean or OWA operators, to $n(n+1)/2$. This corresponds to the diagonal and the upper (or lower) triangle of the matrix Σ.

Note that in the formalized optimization problem there are no additional constraints referring to required matrix property: the positive semi-definiteness ($\Sigma \succeq 0$). A basic technique to force a matrix being positive semi-definite is ensuring its symmetry and that it has all the eigenvalues non-negative [Johnson, 1970]. Nevertheless, this approach is not feasible with linear constraints, and using non-linear constrains will highly increase the problem complexity. Modern techniques suggest modeling the problem using semi-definite programming (SDP). With SDP we can keep the linear objective function and the general linear constraints, but we should add a non-linear constraint representing the semi-definiteness of the matrix, with which our problem will be harder to solve. To avoid the non-linear constraints we have considered two approximations.

The first approximation consists in adding the following additional linear constraint to the previous optimization problem:

$$d^2 SB_\Sigma(a_i, b_j) \geq 0, \quad \forall i, j = 1, \ldots, N \tag{5.16}$$

Equation (5.16) forces the distance to be semi-positive for all pairs of records (a_i, b_j) in the input set. Although, the Σ positive semi-definite is not ensured, this approximation ensures that non-negativity will be satisfied for the input dataset.

The second approximation is to solve the problem estated above and do a post-processing of the resulting matrix Σ. We apply the Higham's algorithm [Higham, 2002] to the matrix Σ. This algorithm computes the nearest semi-definite matrix from a non-positive definite.

Although both proposed approximations have to determine the same number of parameters, $n(n+1)/2$, they differ in the number of constraints. The first approach consists of a linear objective function plus $N(N-1) + N + N^2$ constraints, the general plus all constraints related to Equation (5.16). While, the second approach just consider the same number of constraints as the general optimization problem: $N(N-1) + N$.

5.2.3 Choquet Integral

The last function proposed for the supervised metric learning problem is the Choquet integral, Definition 2.7 (Section 2.1). From a definitional point of view, its main difference with respect to the other functions is that it aggregates two data records with respect to a fuzzy measure. Fuzzy measures permit us to represent, in the computation of a distance, information like redundancy, complementariness, and interactions among data attributes.

Unlike previous aggregators, such as the weighted mean or the OWA operator, which treats all data attributes independently, Choquet integral as well as does the previously presented symmetric bilinear function is able to reveal better the structure information embedded in data. Fuzzy measures allows us to represent much more information concerning to the re-identification process than previous methods. It takes into account information singletons and all combinations of data attributes, i.e., for all $A \subseteq V$, where V is the set of all data attributes. This attribute information could also be understood as which are the weaker attributes or combination attributes, in terms of risk.

Definition 5.6. *Let μ be an unconstrained fuzzy measure on the set of attributes V (i.e. $\mu(\emptyset) = 0$, $\mu(V) = 1$, and $\mu(A) \leq \mu(B)$ when $A \subseteq B$ for $A \subseteq V$, and $B \subseteq V$). Then, the square of the Choquet integral distance is defined as:*

$$d^2CI_\mu(a,b) = CI_\mu(d_1^2(a,b), \ldots, d_n^2(a,b)),$$

where $CI_\mu(c_1, \ldots, c_n) = \sum_{i=1}^{n}(c_{s(i)} - c_{s(i-1)})\mu(A_{s(i)})$, given that $c_{s(i)}$ indicates a permutation of the indices so that $0 \leq c_{s(1)} \leq \ldots \leq c_{s(i-1)}$, $c_{s(0)} = 0$, and $A_{s(i)} = \{c_{s(i)}, \ldots, c_{s(n)}\}$.

Note that $CI_\mu(c_1, \ldots, c_n)$ corresponds to the Choquet integral of the function $f(x_i) = c_i$ which in this case corresponds to the Choquet integral of $f(x_i) = (a_i - b_i)^2$ with respect to the fuzzy measure μ. Definition 5.6 was first introduced by [Pham and Yan, 1999] in the context of color image segmentation with $n = 3$ corresponding to the three RGB colours.

In the case of the Choquet integral based distance d^2CI_μ the minimization problem is defined as follows:

$$Minimize \sum_{i=1}^{N} K_i \tag{5.17}$$

$$Subject\ to:$$

$$d^2CI_\mu(a_i,b_j) - d^2CI_\mu(a_i,b_i) + CK_i > 0, \quad \forall i,j = 1, \ldots, N, i \neq j \tag{5.18}$$

$$K_i \in \{0,1\} \tag{5.19}$$

$$\mu(\emptyset) = 0 \tag{5.20}$$

$$\mu(V) = 1 \tag{5.21}$$

$$\mu(A) \leq \mu(B) \text{ when } A \subseteq B \tag{5.22}$$

where, Equations (5.17), (5.18) and (5.19) are the formalization of the record linkage when the Choquet integral with respect to a fuzzy measure are used and the next constraints make reference to the fuzzy measures conditions. That is, Equations (5.20) and (5.21) are fuzzy measure boundary conditions and Equation (5.22) corresponds to all monotonicity conditions.

The Choquet integral of unconstrained fuzzy measures is not, in general, a metric because it does not satisfy the triangle inequality and the identity of indiscernibles metric's properties. It is shown in [Narukawa, 2007] that the Choquet integral with respect to a submodular measure results in a convex function and it can be used to define a metric. That is, it is possible to use the Choquet integral as a metric by adding a new constraint to the previous optimization problem. To ensure the satisfaction of the submodularity condition by the fuzzy measure, Equation (2.1).

To formulate the problem we use the Möbius transform of the fuzzy measure instead of the measure itself. Möbius representation is a useful representation of non-additive measures, which can be used to give an indication of which subsets of attributes, $A \subseteq V$, interact with one another. So, we have rewritten the previously stated optimization problem in terms of the Möbius transformation, following [Torra and Narukawa, 2007] (Chapter 8). Recall, that the Möbius transform, m_μ of a fuzzy measure μ is not restricted to the interval $[0, 1]$.

Now, we introduce some useful notations for a fuzzy measure identification problem. We denote the set of variables by $V = \{v_1, \ldots, v_n\}$ (recall that $|V| = n$). Then, instead of using $\mu(A)$ to denote the measures of subsets A of V, we will consider μ_r, with $r \in \{0, \ldots, 2^n - 1\}$. To do so, we need a mapping between μ_r and all subsets $A \subseteq V$. This is achieved by representing an integer into base 2, also known as the dyadic representation (or the binary representation) of the integer r. See more details in Example 10. Then, μ_r denotes the measure of the following set

$$A = \{v_l \in V \mid \delta_l^r = 1 \text{ for } l = 1, \ldots, n\}$$

where $\delta_n^r \delta_{n-1}^r \ldots \delta_1^r$ is the dyadic system representation of an integer r. That is,

$$r = 2^{n-1}\delta_n^r + 2^{n-2}\delta_{n-1}^r + \cdots + 2\delta_2^r + \delta_1^r, \quad \delta_l^r \in \{0, 1\}$$

Example 10. *Let $V = \{v_1, v_2, v_3, v_4, v_5\}$, then $\mu(\{v_1, v_3, v_5\})$ is represented by μ_{21}, because the dyadic representation of 21 is 10101. In particular,*

$$\delta_5^{21}\delta_4^{21}\delta_3^{21}\delta_2^{21}\delta_1^{21} = 10101$$

That is, $\delta_5^{21} = 1, \delta_4^{21} = 0, \delta_3^{21} = 1, \delta_2^{21} = 0, \delta_1^{21} = 1$ and therefore,

$$r = 2^4\delta_5^{21} + 2^3\delta_4^{21} + 2^2\delta_3^{21} + 2\delta_2^{21} + \delta_1^{21} = 21$$

By the notation, a fuzzy measure can be represented by the vector $(\mu_0, \mu_1, \cdots, \mu_{2^n-1})$. That is,

$$
\begin{aligned}
\mu_0 &= \mu(\{\emptyset\}),\\
\mu_1 &= \mu(\{v_1\}),\\
\mu_2 &= \mu(\{v_2\}),\\
\mu_3 &= \mu(\{v_1, v_2\}),\\
\mu_4 &= \mu(\{v_3\}),\\
\mu_5 &= \mu(\{v_1, v_3\}),\\
\mu_6 &= \mu(\{v_2, v_3\}),\\
\mu_7 &= \mu(\{v_1, v_2, v_3\}),\\
&\vdots\\
\mu_{2^{n-1}} &= \mu(V).
\end{aligned}
$$

As μ_0 is always 0 there is no need considered it as a parameter to be learned and so the fuzzy measure vector has the following form $(\mu_1, \cdots, \mu_{2^n-1})$, denoted by μ^+. In a similar way, we can consider the vector $m^+ = (\mu_1, \cdots, \mu_{2^n-1})$ that corresponds to the Möbius transform of the fuzzy measure, μ.

Then, the Choquet integral defined in Definition 5.6 can be rewritten in terms of the Möbius transform of the fuzzy measure as:

$$
\begin{aligned}
CI_\mu(c_1, \dots, c_n) &= \sum_{i=1}^{n} (c_{s(i)} - c_{s(i-1)}) (\sum_{A \subset A_{s(i)}} m(A))\\
&= \sum_{i=1}^{n} \sum_{A \subset A_{s(i)}} ((c_{s(i)} - c_{s(i-1)}) m(A))
\end{aligned}
$$

where $c_{s(i)}$ indicates a permutation of the indices so that $0 \leq c_{s(1)} \leq \ldots \leq c_{s(i-1)}$, $c_{s(0)} = 0$, and $A_{s(i)} = \{c_{s(i)}, \dots, c_{s(n)}\}$.

Recall that the data vector c is the vector of distances between attributes of two records a and b, when Equation (5.6) is used. That is, $c = (c_1, \cdots, c_n) = (d_1^2(a,b), \dots, d_n^2(a,b))$. Then, we can define the vector $\mathbf{d}^+(a,b) = (d_1^+(a,b), \dots, d_{2^n-1}^+(a,b))$, where each element corresponds to:

$$
d_r^+(a,b) = \sum_{i=1}^{n} \left(d_{s(i)}^2(a,b) - d_{s(i-1)}^2(a,b) \right) \cdot \tau_{i,r}
$$

for $r = 1, \dots, 2^n - 1$, where

$$
\tau_{i,r} = \begin{cases} 1 & \text{if } \delta(B) = r \text{ for } B \subseteq A_{s(i)}, \\ 0 & \text{otherwise }, \end{cases}
$$

Finally, the Choquet integral with respect to the Möbius representation of a fuzzy measure can be defined in terms of $\mathbf{d}^+$ and $\mathbf{m}^+$ as follows,

$$d^2CI_m(a,b) = CI_m(c_1, \cdots, c_n) = CI_m(d_1^2(a,b), \cdots, d_n^2(a,b)) = \mathbf{d}^+(a,b) \cdot \mathbf{m}^+$$

Now, following the defined notation we can express the record linkage minimization problem as follows,

$$
\begin{aligned}
&Minimize \sum_{i=1}^{N} K_i && (5.23)\\
&Subject\ to:\\
&\qquad (\mathbf{d}^+(a_i,b_j) - \mathbf{d}^+(a_i,b_i)) \cdot \mathbf{m}^+ + CK_i > 0\\
&\qquad\quad \forall i,j = 1,\ldots,N, i \neq j && (5.24)\\
&\qquad K_i \in \{0,1\} && (5.25)\\
&\qquad \sum_{k=1}^{2^n-1} m_k = 1 && (5.26)\\
&\qquad \sum_{B' \subset B} m_k(B') - \sum_{A' \subset A} m_k(A') \geq 0 \text{ for all } A \subset B. && (5.27)
\end{aligned}
$$

Note that the condition $m(\emptyset) = 0$ is not needed, as m_0 is not included in the model and the previous constraints. Equations (5.20) and (5.20) are replaced by the appropriate constraints on m by Equations (5.26) and (5.27).

Following the notation used in the whole chapter, in which N is the number of records, and n the number of data attributes, we can define the total number of constraints of this last optimization problem. As in previous problem definitions there are $N(N-1)+N$ constraints for Equations (5.24) and (5.25). In this case, we have one more constraint concerning to the restriction established by Equation (5.26) and $\sum_{k=2}^{n}\binom{n}{k}k$ more constraints added by Equation (5.27). That is, a total number of $N(N-1)+N+1+\sum_{k=2}^{n}\binom{n}{k}k$ constraints.

As a final remark, notice that this implementation does not hold the metric properties and thus, it cannot be considered a distance metric unless we force the fuzzy measure to satisfy the submodularity condition. Therefore, in case we would like to work with a metric we should add a new inequality constraint to the previous optimization problem. This corresponds to Equation (2.1) (Section 2.1) expressed in terms of the Möbius representation.

$$\sum_{B' \subset B} m_k(B') + \sum_{A' \subset A} m_k(A') - \sum_{C' \subset A \cup B} m_k(C') - \sum_{D' \subset A \cap B} m_k(D') \geq 0,$$

for all $A, B \subseteq V$.

This optimization problem has the same structure as the ones that were presented in the previous sections. There is a mixed integer optimization problem with linear constraints and a linear objective function and therefore, as it is showed in the subsequent section all of presented problems can be solved using the same method.

5.3 Resolution

In this section we describe two different ways to solve the presented optimization problems. Firstly, in Section 5.3.1 we describe the implementation details to solve the mixed integer linear problem by means of a well known commercial solver, IBM ILOG CPLEX [IBM ILOG CPLEX, 2010a]. Since Choquet integral optimization problems are high time consuming and computationally expensive to find the optimal solution (see Section 5.4.3 for more details), we implemented a local optimization algorithm. Section 5.3.2 presents the first-order optimization algorithm. This relies on an adaptation of the gradient descent algorithm proposed by Grabisch in [Grabisch, 1995].

In any of these cases implementations the inputs and outputs are the same. Both take as input an original and its masked data files, which should be aligned and they have common variables. Both outputs are the weights, or the coefficients, of the fuzzy measures that maximize the number of re-identification between both files.

5.3.1 Global Optimization: CPLEX

Given an original file and its masked version, we pre-process and build the problem structure by means of a series of R functions, then following this formalized structure the problem is expressed into MPS (Mathematical Programming System) file format. MPS is a file format to represent and store Linear Programming (LP) and Mixed Integer Programming (MIP) problems. This file format was one of the standard ASCII medium among most of the LP solvers, although, it is an old file format it is accepted by almost all commercial LP and MIP solvers.

Then, each file is processed with an optimization solver. We perform our experiments with one of the most used commercial solvers, the IBM ILOG CPLEX tool [IBM ILOG CPLEX, 2010a]. This software allows us to solve linear programming, mixed integer programming, quadratic programming, and quadratically constrained programming problems. Although the mathematical model for MILP is the LP model with additional constraints indicating that some variables must have integer values, MILP is not that easy to solve than a LP model. That is, by constraining those variables into integer values we are reducing the feasible solution region and at first sight seems to make the problem easier, but it is not the case.

Roughly speaking, when the CPLEX MILP optimizer is invoked it applies several pre-processes to reduce the size of the problem in order to strengthen the initial linear relaxation and to decrease the overall size of the mixed integer program. Then, it transforms the mixed integer linear problem into a linear problem by dropping all variable integer restrictions. This process is known as Linear programming relaxation. The resulting LP problem is known as the root LP relaxation and can be solved efficiently by algorithms like Simplex. The resultant LP optimal solution provides a lower bound on the MILP problem. In case the obtained solution satisfies the dropped variable integer restrictions, then this is the optimal solution for the MILP as well. On the contrary, the Branch and Cut algorithm [Mitchell, 2002] is used to find the optimal solution. This algorithm manages a search tree consisting of nodes. Every node represents a LP subproblem to be solved and then, checked for integer restrictions and perhaps further analyzed. CPLEX processes all nodes in the tree until either there are no more nodes or some limit has been reached. More information about how CPLEX solves MILP problems is provided in [IBM ILOG CPLEX, 2010b].

5.3.2 Local Optimization: Gradient Descent Algorithm

Inspired by Heuristic Least Mean Squares (HLMS), a gradient descent algorithm, introduced by Grabisch in [Grabisch, 1995], we introduce a new record linkage process relying on it. HLMS takes as input a training dataset P like the following:

$$P = \begin{pmatrix} V_1^X(a_1) & \dots & V_i^X(a_1) & \dots & V_n^X(a_1) & T_1 \\ \vdots & & \ddots & & \vdots & \\ V_1^X(a_z) & \dots & V_i^X(a_z) & \dots & V_n^X(a_z) & T_z \\ \vdots & & \ddots & & \vdots & \\ V_1^X(a_N) & \dots & V_i^X(a_N) & \dots & V_n^X(a_N) & T_N \end{pmatrix}$$

where $V_i^X(a_j)$ is the value of sample j for attribute i from the data set X, and T^j its target value. The algorithm finds the fuzzy measure μ that minimizes the difference $\mathcal{C}_\mu(\{V_1^X(a_j), V_2^X(a_j), \cdots, V_n^X(a_j)\}) - T_j \ \forall j$. The error made in the approximation can be calculated as:

$$E(\mu) = \sum_{j=1}^{N} (\mathcal{C}_\mu(\{V_1^X(a_j), V_2^X(a_j), \cdots, V_n^X(a_j)\}) - T_j)^2$$

The formula represents simply the squared difference between the target T_j and the Choquet integral of sample j using μ, summed over all training examples. The direction of the steepest descent along the error surface can be found by computing the derivative of E with respect to each component of the vector μ.

$$\bigtriangledown E(\mu) \equiv [\frac{\delta E}{\delta \mu_{(1)}}, \frac{\delta E}{\delta \mu_{(2)}}, \dots, \frac{\delta E}{\delta \mu_{(n)}}]$$

Since the gradient specifies the direction of the steepest increase of E, the training rule for gradient descent is:

$$\mu_{(i)} \leftarrow \mu_{(i)} - \lambda \bigtriangledown E(\mu_{(i)})$$

Here λ is a positive constant called the learning rate, which determines the step size in the gradient descent search. The negative sign is present because we want to move the attributes of the aggregation operator in the direction that decreases E.

As the record linkage problem cannot be addressed directly with the HLMS since the target value is unknown, we had to adapt it following the same strategy presented in the previous section, dividing the problem in blocks. Therefore, given two files of N records, an original file $X = a_1, \cdots, a_N$ and its masked version $Y = b_1, \cdots, b_N$, each block k is defined as the following matrix:

$$D_k = \begin{pmatrix} d_1^2(a_k, b_1) & \dots & d_i^2(a_k, b_1) & \dots & d_n^2(a_k, b_1) \\ \vdots & & \ddots & & \vdots \\ d_1^2(a_k, b_z) & \dots & d_i^2(a_k, b_z) & \dots & d_n^2(a_k, b_z) \\ \vdots & & \ddots & & \vdots \\ d_1^2(a_k, b_N) & \dots & d_i^2(a_k, b_N) & \dots & d_n^2(a_k, b_N) \end{pmatrix}$$

where $d_i^2(a, b)$ is the squared difference between the ith dimension from two records a and b. See Equation (5.6) for more details.

As for each record in the original file X, we created a matrix representing a block. There will be N different blocks.

The aim of the heuristic algorithm is to find the fuzzy measure μ that makes for a block k the value of

$$\mathcal{C}_\mu(\{d_1^2(a_k, b_z), \dots, d_i^2(a_k, b_z), \dots, d_n^2(a_k, b_z)\}) \quad \forall z \in 1 \cdots N \tag{5.28}$$

to be minimum when $k = z$.

The approach used for each block k is the following:
The fuzzy measure is initialized to the equilibrium state ($\mu_i = \frac{|i|}{n}$). The Choquet integral of each row in D_k is calculated. If the minimum of the Choquet integral is for row k, then proceed with the next block. If the minimum of the Choquet integral is not for row k, calculate the gradient direction that makes the value of the Choquet minimum increases and the gradient of the Choquet integral for row k decreases.

The algorithm for this approach is shown in Algorithm (2).

The algorithm does not guarantee the convergence to a global minimum.

5.4 Experiments

We have applied the supervised learning approaches described in the previous sections. The results obtained are described in the next sections.

In Section 5.4.2 we conduct a disclosure risk evaluation of a large set of protected data files. The common characteristic of these protected data files is that all of them were protected considering a uniform protection: all attributes or groups of attributes are protected with the same protection level. We perform the supervised metric learning for record linkage using a weighted mean to evaluate the disclosure risk of all this data set. Besides, we analyze the time taken to solve each of the record linkage problems. In addition, we compare the percentage of correct matches obtained by the proposed supervised learning approach and the percentage obtained by standard distance-based record linkage.

Section 5.4.3 presents a comparison in terms of accuracy and computing time between all supervised metric learning for record linkage proposals. That

```
Let $X$ be the original database and $X'$ the protected one with $N$ samples
and $n$ attributes each.
  ———— Initialization ————
  for $i \in \mathcal{P}(X)$ do
    $\mu_i = \frac{|i|}{|X|}$
  end for
  ———— For each Block ————
  for $i \in [1..N]$ do
    ——— For each row in $X_i \in X$ ———
    $d_j \leftarrow (X_i - X'_j)^2 \; \forall j \in [1 \ldots N]$
    $s = \{j | \mathcal{C}(d_j) \leq \mathcal{C}(d_i) \; \forall \, j \in [1 \ldots N]\}$
    ———— Update step ————
    for all j $\in$ s do
      Update the fuzzy measure, so that the difference $\mathcal{C}(d_i) - \mathcal{C}(d_j)$ decreases
    end for
    ——— Monotonicity check ———
    Check monotonicity
  end for
  return $\mu$
```

Algorithm 2: Description of the heuristic algorithm for record linkage

is, the weighted mean, the OWA operator, the symmetric bilinear form and the Choquet integral. We also compare their results to the ones obtained by the standard distance-based record linkage and the Mahalanobis distance. In this section non-uniform protections were considered.

As some supervised methods are time consuming, specially the Choquet integral, in Section 5.3.2 we proposed a heuristic approach. Section 5.4.4 compares this heuristic approach to the optimization problem when a Choquet integral is used. Apart from an evaluation considering the worst case scenario, we consider another evaluation scenario, where an attacker with some prior knowledge wants to get more new information by means of linking his/her information with the protected one. By means of analyzing the behaviours of the Choquet integral-based optimization problem, and the heuristic approach we are able to determine if the problem suffers from overfitting.

Finally, in Section 5.4.5 we provide a study of the information provided by the parameters found. Recall, that the optimization problem finds the parameters that lead to a maximum number of correct matches between an original data file and its protected variation. Therefore, these parameters give a lot of information about which are the attributes that are more or less relevant in the re-identification process. We analyze how much information give the weights of all proposed parametric functions.

Note that we focus our work on distance-based record linkage, by comparing our proposals to the standard distance-based record linkage. This allows us to test our proposal directly with very similar approaches, which use the same techniques and strategies. A comparison with probabilistic record linkages is not considered because of the following main reason. Because all our proposals are numerical oriented, and so all experiments are performed on numerical data bases, and [Torra and Domingo-Ferrer, 2003] showed that distance-based record linkage methods are more appropriate than probabilistic-based record linkage for such kind of data. So, with this precedent we can expect probabilistic record linkage performs worse or similarly than standard distance-based record linkage.

5.4.1 Test Data

We have used the Census dataset from the European CASC project [Brand et al., 2002], which contains 1080 records and 13 variables, and has been extensively used in other works [Domingo-Ferrer et al., 2006, Laszlo and Mukherjee, 2005, Domingo-Ferrer and Torra, 2005, Yancey et al., 2002, Domingo-Ferrer et al., 2001].

Attr.	Name	Description
V_1	AFNLWGT	Final weight (2 implied decimal places)
V_2	AGI	Adjusted gross income
V_3	EMCONTRB	Employer contribution for health insurance
V_4	ERNVAL	Business or farm net earnings
V_5	FEDTAX	Federal income tax liability
V_6	FICA	Social security retirement payroll deduction
V_7	INTVAL	Amount of interest income

Table 5.2: Attributes of the Census dataset. All of them are real valued numeric attributes.

Attr.	Mean	Std dev (σ)
V_1	196,039.8	101,251.417
V_2	56,222.76	24,674.843
V_3	3,173.135	1,401.832
V_4	7,544.656	4,905.200
V_5	45,230.84	21,323.470
V_6	2,597.184	1,826.436
V_7	39,712.95	21,224.161

Table 5.3: Mean and standard deviation (σ) for each column attribute.

The dataset was extracted using the Data Extraction System of the U.S. Census Bureau [Census Bureau, 1995]. The records correspond to the *Current Population Survey of the year 1995*, and more precisely to the file-group *March Questionnaire Supplement – Person Data Files*. All the records with zero or missing values for at least one of the 13 attributes were discarded. The final file has 1080 records.

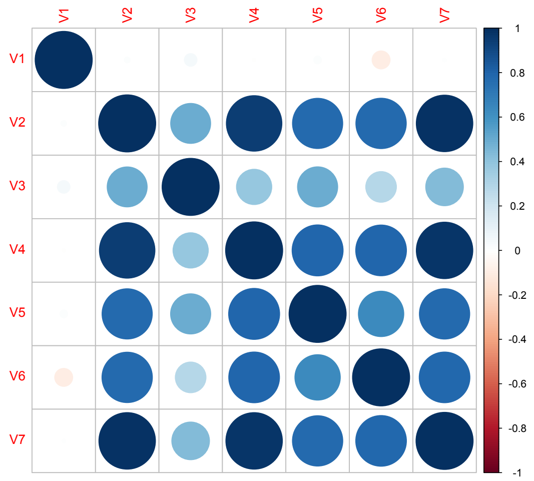


Figure 5.3: Graphical representation of the Census data set correlation matrix.

This dataset has some interesting properties. Records where selected so the number of repeated values was low. Furthermore, the 13 variables were selected so values of each one span a wide range.

For the sake of simplicity and time saving, for our experiments we will select at most 7 out of the 13 data attributes. Specifically, we will test our experiments on the first 7 data attributes. Four of them V_1, V_2, V_3, and V_5 have no repeated values. Regarding this issue we wanted to provide a generic record linkage process, so approximately half of the variables had repeated values. Selecting all 7 variables without repeated values, could provide better results, although the scenario will be less realistic, since repeated values are normally expected in this kind of data.

All attributes used are numerical real valued and are described in Table 5.2. In Table 5.3 we provided some basic statistical information, such as the mean and the standard deviation for each data column. From it we can see how different are the data attributes in terms of their means, and also how spread out are the data points over a large range of values. In addition, in Figure 5.3 we show a graphical representation of the Pearson correlation coefficient, which indicates a degree of linear relationship between all pairs of attributes.

5.4.2 Improvements on Uniformly Protected Files

In this section we present a large test set of evaluations comparing the standard distance-based record linkage (d^2AM), Definition 5.1, to the proposed record linkage weighted mean parameterization (d^2WM), Definition 5.2. To evaluate this comparison we have protected the Census data set with a set of different protection techniques. For each protection method, there were selected and protected the first 7 attributes and 400 randomly selected records from the 1080 records of the Census dataset. Additionally, this process was repeated 10 times, so that we have 10 different protected datasets for each masking method. All masking methods considered are enumerated below,

- **Microaggregation**, (Section 2.4). We have used the Euclidean distance to form the clusters, and the arithmetic mean to compute the centroids. The following variants of microaggregation have been considered:
  - Individual ranking (*MicIR*).
  - Z-scores projection (*MicZ*).
  - Principal components projection (*MicPCP*)
  - Multivariate microaggregation:
    - *Mic2*: microaggregation is applied in groups of two consecutive attributes independently. That is, three groups of two attributes and one attribute alone.
    - *Mic3*: microaggregation is applied in groups of three consecutive attributes independently. That is, two groups of three attributes and a single attribute.
    - *Mic4*: microaggregation is applied in groups of four consecutive attributes independently. That is, one group of four attributes and a group of tree attributes.
    - *Mic5*: microaggregation is applied in groups of five consecutive attributes independently. That is, one group of five attributes and a group of two attributes.
    - *Mic6*: microaggregation is applied in groups of six consecutive attributes independently. That is, one group of six attributes and one single attribute.
    - *MicAll*: All attributes are protected at the same a time. K-anonymity is satisfied.

  For each microaggregation method we have considered values of k from 3 to 20.

- **Rank Swapping**, (Section 2.4.2). We have considered values of p from 1 to 7.

- **Additive Noise**, (Section 2.4.3). We have considered values of $p = \{1, 2, 4, 6, 8, 10, 12, 14, 16\}$.

Summarizing, from these 11 methods we have considered a set of different parameter's configurations; 7 different values for rank swapping, 9 for additive noise and 18 parameter variations for each microaggregation method. These parameter's configurations lead to a total of 34 different protections. In addition, for each of them we have protected 10 different combinations of the Census dataset (400 randomly selected records). Therefore, after all this process there will be $1,780$ protected datasets, i.e., $7 \cdot 10 + 9 \cdot 10 + (18 \cdot 10) \cdot 9$.

Then, the supervised metric learning record linkage with respect to a weighted mean has been performed on all these protected datasets, as described in Sections 5.1 and 5.2.1. Thus, for each dataset our approach finds the optimal combination of weights to achieve the maximum number of re-identifications between the original and the protected data.

These experiments were performed in the Finis Terrae computer from the supercomputing center of Galicia [CESGA, 2010]. Finis Terrae is composed of 142 HP Integrity rx7640 computing nodes with 16 Itanium Montvale cores and 128 GB of memory each, one HP Integrity Superdome node, with 128 Itanium Montvale cores and 1.024 GB of memory, and 1 HP Integrity Superdome node, with 128 Itanium 2 cores and 384 GB of memory. From the Finis Terrae computer we used 16 cores and 32GB of ram memory for a maximum of 300 hours per each execution (limit imposed by CESGA). To solve the optimization problems we used the commercial solver IBM ILOG CLPEX [IBM ILOG CPLEX, 2010a] (version 12.1).

Protection method	%	Protection method	%
MicIR	100%	*Mic5*	12.33%
MicZ	100%	*Mic6*	2.22%
MicPCP	100%	*MicAll*	100%
Mic2	100%	*Rank Swapping*	100%
Mic3	62.78%	*Additive Noise*	100%
Mic4	26.67%		

Table 5.4: Percentage of datasets evaluated for each protection method.

In the end, just $1,247$ from the $1,780$ protected datasets could be fully evaluated, due to the CESGA execution time limitation. Table 5.4 summarizes the percentage of finished protected datasets per protection method.

Table 5.5 shows a general comparison between the results obtained by the standard distance-based record linkage (d^2AM) and the weighted mean (d^2WM) using the weights obtained by the proposed supervised record linkage. Columns $\%d^2AM$ and $\%d^2WM$ provide the averaged percentage of correct links obtained by the arithmetic mean and the weighted mean, respectively. These percentages are the average of all protections' parameterization and its corresponding datasets variations and hence they provide a general overview of those masking methods with a lower risk of disclosure (according to both distance-based re-identification methods). From all evaluated methods, *MicZ* and *MicPCP* are the methods with the higher failure of re-identifications and so it should be considered the best methods in terms of data protection. However, as was shown in the protection methods ranking presented in [Domingo-Ferrer and Torra, 2001b], despite of both methods provide a high protection to data, they also introduce a huge amount of noise to data, making the data difficult to analyze, and so the

Protection method	$\%d^2AM$	$\%d^2WM$	$(\%d^2WM - \%d^2AM)$
MicIR	99.975%	100%	0.025
MicZ	0.2677%	0.447%	0.1793
MicPCP	0.263%	0.446%	0.183
Mic2	96.942%	98.396%	1.454
Mic3	89.332%	92.186%	2.854
Mic4	86.802%	88.865%	2.063
Mic5	76.568%	79.557%	2.989
Mic6	30.875%	32.938%	2.063
MicAll	14.999%	12.557%	−2.442
Rank Swapping	93.637%	95.611%	1.974
Additive Noise	97.428%	98.2%	0.772

Table 5.5: Averaged re-identification percentage per protection method.

Rank Swap

p	d^2AM	$\sigma(d^2AM)$	$\epsilon(d^2AM)$	d^2WM	$\sigma(d^2WM)$	$\epsilon(d^2WM)$
1	99.650	0.300	0.095	100.000	0.000	0.000
2	98.525	0.493	0.156	99.725	0.175	0.055
3	96.975	0.964	0.305	98.850	0.515	0.163
4	94.650	0.673	0.213	97.150	0.450	0.142
5	92.850	1.384	0.438	95.325	1.245	0.394
6	88.000	1.167	0.369	90.825	0.613	0.194
7	84.800	2.006	0.634	87.400	1.441	0.456
avg	*93.636*	*0.998*	*0.316*	*95.611*	*0.634*	*0.201*

Table 5.6: Re-identification percentages comparison between d^2AM and d^2WM for the *Rank Swap* protection method. σ and ϵ correspond to the standard deviation and standard error, respectively. p is the rank swapping parameter indicating the percent difference allowed in ranks (see Section 2.4.2).

trade-off between utility and risk is not well balanced.

In the last column of Table 5.5 we provide the percentage difference between the results obtained by both re-identification methods. It is clear to see that the improvement achieved by the weighted mean is really small. In average, the maximum re-identification improvement is achieved in the *Mic5* files set with a difference of a 3%, i.e. 12 more correct links with respect to the arithmetic mean. Finally, note the negative difference for the *MicAll* protections, in this case the arithmetic mean achieve in average better results than weighted mean. This is due to the random factor trying to match the correct link from a set of records that have the same distances. *MicAll* satisfy k-anonymity and therefore in the protected dataset there are at least k indistinguishable records.

In Tables 5.6, 5.7, and 5.8 we present the results obtained by three protection methods showing the true positive rates (percentage of re-identified records) of the weighted mean with optimal weights (d^2WM) and the standard record link-

Mic2

k	d^2AM	$\sigma(d^2AM)$	$\epsilon(d^2AM)$	d^2WM	$\sigma(d^2WM)$	$\epsilon(d^2WM)$
3	99.975	0.079	0.025	100.000	0.000	0.000
4	99.650	0.269	0.085	99.900	0.211	0.067
5	99.300	0.511	0.162	100.000	0.000	0.000
6	99.275	0.463	0.147	99.700	0.329	0.104
7	99.350	0.412	0.130	99.825	0.265	0.084
8	98.150	0.580	0.183	99.700	0.284	0.090
9	98.425	0.528	0.167	99.525	0.322	0.102
10	98.375	0.377	0.119	99.425	0.206	0.0651
11	97.200	0.632	0.200	98.725	0.381	0.120
12	96.900	0.592	0.187	98.525	0.343	0.108
13	96.775	0.786	0.248	98.375	0.580	0.184
14	96.525	0.924	0.292	98.100	0.615	0.194
15	95.875	0.637	0.202	97.975	0.448	0.142
16	95.850	0.709	0.224	98.150	0.648	0.205
17	94.500	1.041	0.329	96.750	0.920	0.291
18	93.475	0.901	0.285	96.175	0.727	0.230
19	92.925	1.444	0.457	95.325	1.155	0.365
20	92.425	1.068	0.338	94.950	0.632	0.200
avg	*96.942*	*0.664*	*0.210*	*98.396*	*0.474*	*0.142*

Table 5.7: Re-identification percentages comparison between d^2AM and d^2WM for the *Mic2* protection method. σ and ϵ correspond to the standard deviation and standard error, respectively. Here, k is the minim cluster size for the microaggregation (see Section 2.4.1).

Additive Noise

p	d^2AM	$\sigma(d^2AM)$	$\epsilon(d^2AM)$	d^2WM	$\sigma(d^2WM)$	$\epsilon(d^2WM)$
1	100.000	0.000	0.000	100.00	0.000	0.000
2	100.000	0.000	0.000	100.00	0.000	0.000
4	100.000	0.000	0.000	100.00	0.000	0.000
6	99.875	0.177	0.056	100.00	0.000	0.000
8	99.450	0.369	0.117	99.9	0.129	0.041
10	98.050	0.632	0.200	99.1	0.428	0.135
12	95.600	0.637	0.201	97.05	0.387	0.122
14	93.850	0.899	0.284	95.45	0.550	0.174
16	90.025	1.102	0.349	92.3	0.654	0.207
avg	*97.428*	*0.424*	*0.134*	*98.200*	*0.239*	*0.076*

Table 5.8: Re-identification percentages comparison between d^2AM and d^2WM for the *Additive Noise* protection method. σ and ϵ correspond to the standard deviation and standard error, respectively. Here p is the parameter of the additive noise (see Section 2.4.3).

age (d^2AM). Additionally, for each protection method and parameter's configuration we give the average of the 10 protected dataset variations re-identification percentage, the standard deviation (σ) and the standard error (ϵ), computed as $\epsilon = \sigma/\sqrt(10)$, where σ is the standard deviation of the 10 dataset variation. We show the results for the protection methods rank swapping (*Rank Swap*) in Table 5.6, microaggregation (*Mic2*) in Table 5.7, and additive noise (*Additive Noise*) in Table 5.8 as an example. Moreover, the average value of the re-identification rate, the standard deviation and the standard error is also given for each protection method as *avg*.

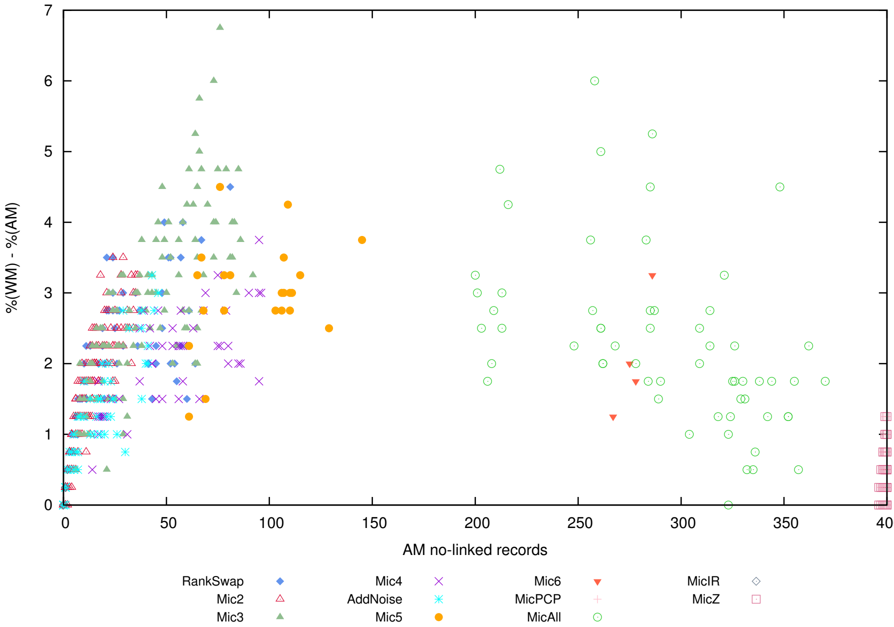


Figure 5.4: Improvement for all cases, shown as the difference between the standard record linkage, d^2AM, and the proposed supervised learning record linkage approach using dWM^2.

In Figure 5.4 we showed the percentage of improvement obtained by the supervised learning for record linkage using a weighted mean for all protected files. The x axe shows the difference on the re-identification percentage between the proposed supervised record linkage (using dWM^2) and the standard one (d^2AM) and the y axe shows the number of non-identified links determined by the standard method. In general for lower or higher d^2AM re-identification percentages the improvement achieved by d^2WM is very low, while for medium percentages, the re-identification percentage increases. Figure 5.5 shows a left-hand crop of Figure 5.4, showing for simpler cases, which standard distance-based record linkage is able to re-identify almost all records, the weighted mean approach also achieves the same good results. However, when the problem complexity increases the standard approach starts loosing its effectiveness, and the weighted mean starts achieving slightly improvements. The maximum improvement is however relatively small (about 6%).

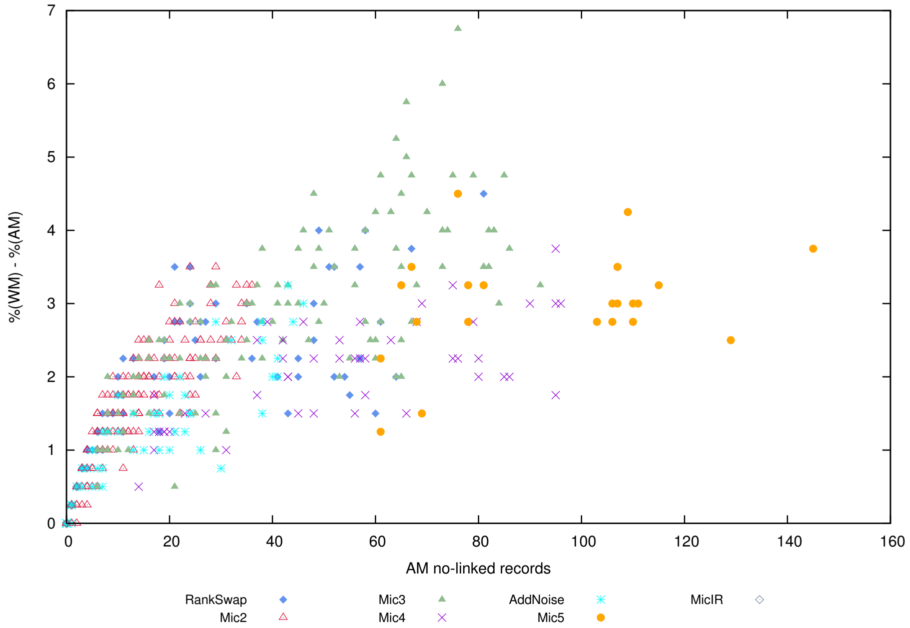


Figure 5.5: Improvement for *Mic2*, *Mic3*, *Mic4*, *Mic5*, *MicIR*, *Rank Swap* and *Additive Noise*, shown as the difference between the standard record linkage, D^2AM, and the proposed supervised learning record linkage approach using dWM^2. This is a corp of Figure 5.4.

Our experiments also show an interesting behavior regarding the computation cost used to find the optimal combination of parameters for the weighted mean, d^2WM. Figure 5.6 shows the time needed by the solver to find the solution for all the protected datasets, note that time is given in a logarithmic scale. We can observe that the computation cost depends on the percentage of re-identifications (as determined by the objective function in the x axis). With low and high percentages of re-identifications the cost is very low, even negligible as the percentages reach 0% or 100%. At the same time with medium number of the re-identifications there is a high computational cost, which reaches more than one week for some cases. Recall, that the maximum time allowed by CESGA is 300 hours (i.e., $\approx 1,080,000$ seconds).

Combining these results with the method accuracies, described before, we have that a significant improvement of record linkage occurs precisely for the data files where the computational cost is high (1 week or more of computational cost).

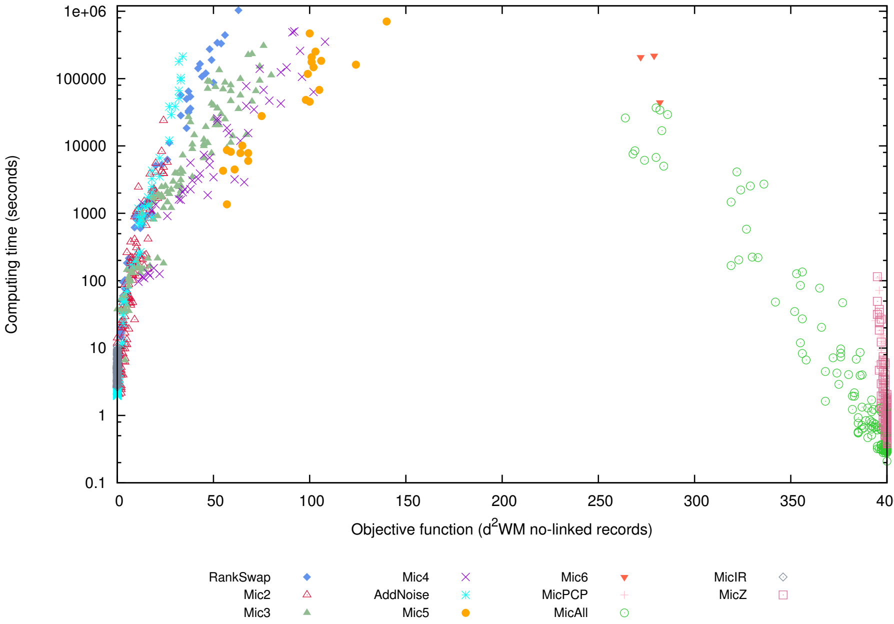


Figure 5.6: Computing time for all cases, in terms of the number of non correct matches records (optimization problem objective function).

As it can be appreciated, the proposed weighted mean approach achieves a very slight improvement with respect to the standard distance record linkage. This leads us to conclude that it is relatively meaningful to use equal weights for estimating the disclosure risk in the scenarios discussed here, especially if we take into account the computation cost (see Figure 5.6). However, in the following sections we show other situations where learning weights leads us to a better understanding of the protection, its strengths and weaknesses, and so to an increase on the number of correct re-identifications.

5.4.3 Improvements on Non-Uniformly Protected Files

As we have seen in the general case, determining optimal weights for the distance-based record linkage does not provide a substantial improvement in the re-identification percentage. There are some cases, where the fact that some attributes are more weighted can have an important impact in the re-identification. This is the case of non-uniformly protected files. That is, files where some attributes have a higher protection degree than others. This means that some attributes have less perturbation and thus are more useful for re-identification than others.

To illustrate this issue, we have evaluated all our proposals with a set of protected datasets applying different protection degrees to attributes or sets of attributes independently. We have protected the 400 randomly selected records from the Census dataset considering different values for the parameter k of the microaggregation method. We have considered files with the following combination of protection parameters:

- *M4-33*: 4 attributes microaggregated in groups of 2 with $k = 3$.
- *M4-28*: 4 attributes, first 2 attributes with $k = 2$, and last 2 with $k = 8$.
- *M4-82*: 4 attributes, first 2 attributes with $k = 8$, and last 2 with $k = 2$.
- *M5-38*: 5 attributes, first 3 attributes with $k = 3$, and last 2 with $k = 8$.
- *M6-385*: 6 attributes, first 2 attributes with $k = 3$, next 2 attributes with $k = 8$, and last 2 with $k = 5$.
- *M6-853*: 6 attributes, first 2 attributes with $k = 8$, next 2 attributes with $k = 5$, and last 2 with $k = 3$.

Note that in our experiments we apply different protection degrees to different attributes of the same file. The values used range between 2 to 8, i.e., values between the lowest protection value and a good protection degree in accordance to [Domingo-Ferrer and Torra, 2001b]. This is especially interesting when variables have different sensitivity. We have used the web application [ppdm.iiia.csic.es, 2009], which is based on [Domingo-Ferrer and Torra, 2001b], to compute standard scores to evaluate all the protected datasets. These scores are computed by means of a combination of information loss and disclosure risk values, so the best protection method is the one that optimizes the trade-off between the information loss and the disclosure risk, see Section 2.5 for more details. Table 5.9 shows the average record linkage ($AvRL(\%)$), the probabilistic information loss ($PIL(\%)$) and the overall score ($Score(\%)$) for all the protected files. Recall that the lower the score, the better. Therefore, the best score is achieved by the *M5-38* file, though the other files have a very similar score. If we focus on the average record linkage evaluation, we see that *M6-853* and *M5-38* are the files with less disclosure risk, while *M4-33* is the file with higher disclosure risk. Note, that *M4-33* is the only file with a uniform protection for all its attributes. That is, both subgroups of two attributes are microaggregated setting the protection degree, k, to 3.

Unlike the previous section, these experiments were performed on a workstation with two processors Xeon E5-2609 (4 cores) at 2.4GHz and 16GB of memory DDR3. From this workstation we used 6 cores and 16GB of memory and for these experiments we used a newer version of the solver IBM ILOG CLPEX[IBM ILOG CPLEX, 2010a], the version 12.6. This new version allow us to quantify each execution in *ticks*, a deterministic way to quantify the time, independently of the system load. According to the CPLEX V12.6 documentation [IBM ILOG CPLEX, 2010b] the length of a deterministic tick may vary by platform. Nevertheless, ticks are normally consistent measures for a given platform (combination of hardware and software) carrying the same load.

	$AvRL(\%)$	$PIL(\%)$	$Score(\%)$
M4-33	42.127	23.85	32.99
M4-28	33.47	28.40	30.94
M4-82	32.37	31.80	32.09
M5-38	26.01	31.92	28.96
M6-385	35.42	36.91	36.16
M6-853	30.65	37.76	34.06

Table 5.9: Evaluation of the protected datasets.

Mic2

	Finis Terrae		Workstation		
k	$Avg(time)$	$\sigma(time)$	$Avg(time)$	$\sigma(time)$	$Avg(ticks)$
3	2.86	0.27	395.99	2.51	$1,409,159.45$
10	17.11	13.80	383.98	3.42	$1,361,895.89$
20	5639.93	6759.68	502.76	106.32	$1,340,086.89$

Table 5.10: Time (in seconds) comparison between Finis Terrae (CESGA) and the Workstation used to perform the experiments. Additionally, the ticks are provides. All values are the average of the 10 data set variations for each k value of *Mic2* protection.

In order to provide a basis for time comparison between the execution times of the Finis Terrae and the workstation we have done some computations in both machines. In particular, in Table 5.10 we provide a time comparison between both machines for three problems: microaggregation (*Mic2*) for $k = \{3, 10, 20\}$, which were solved with the supervised learning approach using the weighted mean (d^2WM). Additionally, Figure 5.7 shows the variability of $ticks/second$ through all 10 variations of *Mic2* for $k = \{3, 10, 20\}$.

The proposed supervised metric learning for record linkage is performed for all these six protected data files. This time we have evaluated each masked dataset with the supervised learning approach with respect to all proposed parameterized aggregators and its variations. First, we have considered the weighted mean, the OWA operator and their two corresponding variations, described in Section 5.2.1. The first variation considers non-negative weights (d^2WM and d^2OWA) and the second variation, more restrictive, considers just positive weights (d^2WM_m and d^2OWA_m). These more restrictive variations make that both functions satisfy the identity of indiscernibles metric's property (property *(ii)* from Definition 2.10). Second, we have considered the three outlined optimization problems using a symmetric bilinear function described in Section 5.2.2. The first variation of the problem, d^2SB_{NC}, does not consider any addition restriction with respect to the general problem and so solves the problem formalized by Equations (5.13), (5.14) and (5.15). On the contrary, the second variation, d^2SB, considers and extra constrain, Equation (5.16) in order to force the function being positive for all the data examples. The third varia-

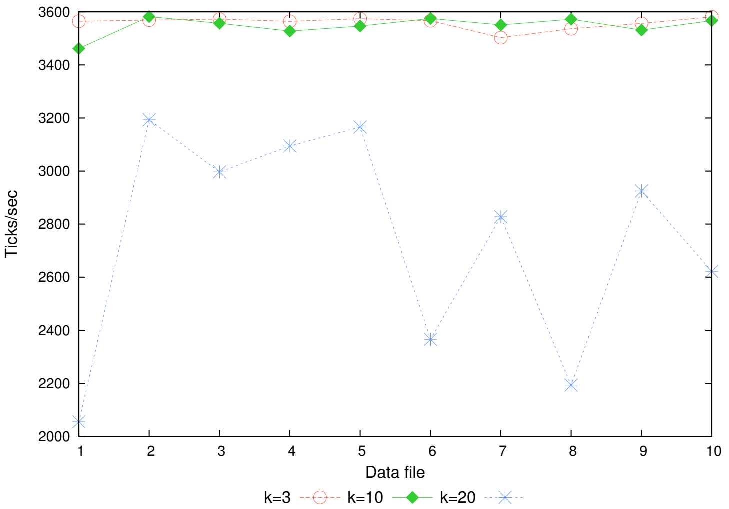


Figure 5.7: Ticks/second through all 10 variations of *Mic2* for $k = \{3, 10, 20\}$.

tion, d^2SB_{PD}, first executes the problem d^2SB_{NC} and then by means of using the Higham's algorithm [Higham, 2002], computes the nearest positive definite matrix from the one found by the optimization problem. Finally, we evaluate the optimization problem using the Choquet integral and as was described in Section 5.2.3 we have considered two variations. The first variation, d^2CI is defined by Equations (5.23), (5.24), (5.25), (5.26). The second variation, d^2CI_m considers an additional constraint to the problem, Equation (5.27), which ensures that the fuzzy measure is submodular and so it can be considered a metric.

	M4-33	*M4-28*	*M4-82*	*M5-38*	*M6-385*	*M6-853*
d^2AM	84.00	68.50	71.00	39.75	78.00	84.75
d^2MD	94.00	90.00	92.75	88.25	98.50	98.00
d^2WM	95.50	93.00	94.25	90.50	99.25	98.75
d^2WM_m	95.50	93.00	94.25	90.50	99.25	98.75
d^2OWA	89.00	87.50	88.25	67.50	94.50	95.50
d^2OWA_m	89.00	87.50	88.25	67.50	94.50	95.50
d^2CI	95.75	93.75	94.25	91.25	**99.75**	99.25
d^2CI_m	95.75	93.75	94.25	90.50	99.50	98.75
d^2SB_{NC}	**96.75**	**94.5**	**95.25**	**92.25**	**99.75**	**99.50**
d^2SB	**96.75**	**94.5**	**95.25**	**92.25**	**99.75**	**99.50**
d^2SB_{PD}	–	–	–	–	–	99.25

Table 5.11: Percentage of the number of correct re-identifications.

Table 5.11 shows the percentage of correct linkages using two non-supervised learning approaches and the supervised approaches listed above. The non-supervised learning approaches are the standard record linkage method, d^2AM (Definition 5.1, Section 5.2) and the Mahalanobis distance, d^2MD (Definition 2.12, Section 2.2.1). The values in the table are the percentages determining the correctly identified records from the total, so a percentage of 100% means that all records were re-identified.

A general behaviour showed in Table 5.11 is that all evaluated methods perform worse for *M4-28* and *M5-38* data files. However, in Table 5.9 in which according to an average of all risk measures (see Section 2.5.2) *M6-853* and *M5-38* were the two protected data files with a lower disclosure risk score. Note that although *M6-853* is considered the second best protected file, $AvRL(\%) = 30.65$, according to Table 5.11 it is the worst protected file.

Before tackling the results obtained by the presented supervised approaches we focus on the non-supervised approaches. The most noticeable fact between the standard distance-based record linkage (d^2AM) and Mahalanobis distance (d^2MD) is the improvement achieved by the latter method, which in average achieves about 22.6% more correct re-identifications and for the protected file *M5-38* achieves a maximum improvement of 48.5%. This improvement and ease computation of Mahalanobis distance make that d^2MD should be strongly considered for the disclosure risk assessment of protected datasets. However, as it is also shown in Table 5.11, these results can still be overcome by the presented optimization approaches.

The results obtained by the proposed supervised approaches show that almost for all the protected files the optimization problem with respect to the symmetric bilinear function achieves the larger number of correct matches, slightly followed by the Choquet integral by a maximum difference of exactly 1% (4 correct matches less) for *Mic3-44* and *Mic5-38* protections. Improvements obtained by the Choquet integral are also slightly followed by the ones obtained by the weighted mean, which have a maximum difference of 0.75% (3 correct matches less) for *Mic4-28* protection. With respect to results obtained by both variations of the OWA operator, they are far from the previous commented improvements in the re-identification and although they are much better than the standard distance-based record linkage, their results are even worse than the ones obtained by the Mahalanobis distance. Finally, from these results we can conclude that from the supervised learning approaches the symmetric bilinear, the Choquet integral and the weighted mean are the best methods. Comparing them in those protected files where the standard record linkage approach achieve the maximum (*M6-853*) and the minimum (*Mic5-38*) number of re-identifications, we obtained an improvement of 14.76% (by d^2SB), 14.51% (by d^2CI) and 14.01% (by d^2WM) for the *M6-853* file and improvement of 52.5% (by d^2SB), 51.5% (by d^2CI) and 50.751% (by d^2WM) for the *M5-38* file. However, to evaluate all approaches it is also important to bear in mind the problem complexity and its computing time, factor that we analyze below, in Table 5.12.

Now, we focus on the comparison between the number of correct matches between each approach (d^2WM, d^2OWA and d^2CI) and its variation, which are noted with the subscript m (d^2WM_m, d^2OWA_m and d^2CI_m). As we explained before, these variations introduce additional constraints to the original problems in order to fulfill one or more metric properties. However, despite of having new constraints, the percentage of re-identifications achieved by each variation is slightly lower than the original approach. This might happen due to reasons. The first one is because a different combination of parameter's values can lead to the same number of correct matches. The second reason is because the solution of both approaches could be exactly the same. This is the case of the weighted mean and OWA approaches, the parameters obtained by the solver in d^2WM and d^2WM_m, and d^2OWA and d^2OWA_m are the same.

For the symmetric bilinear function let us underline that all matrices obtained by d^2SB and d^2SB_{NC} satisfy the positive definiteness property, except for the last dataset (*M6-853*), which in either of the two approaches this property was not satisfied. The Higham's algorithm was applied to the matrix obtained by d^2SB_{NC} achieving a new positive definite matrix with which we obtained a similar number of correct matches than the d^2SB_{NC} approach. Specifically, with the matrix obtained by the Higham's algorithm we obtained one correct match less than the matrix obtained by d^2SB_{NC}. Recall that when the matrix obtained by the Higham's algorithm is positive definite and so all metric properties are satisfied. Note that using the symmetric bilinear function with the matrix obtained by the presented approach achieve better results than using the covariance matrix computed from the data.

	M4-33	*M4-28*	*M4-82*
d^2WM	617,614.58	600,189.76	567,472.96
d^2WM_m	21,638.74	31,454.2	33,023.73
d^2OWA	607,176.78	602,025.03	565,354.11
d^2OWA_m	4,722.36	5,483.27	4,859.99
d^2CI	**2,249,344.22**	**2,231,926.16**	**2,211,749.13**
d^2CI_m	2,246,930.92	2,222,339.29	2,123,553.72
d^2SB_{NC}	27,426.56	62,730.44	1,282,687.88
d^2SB	100,814.72	125,815.04	26,969.76

	M5-38	*M6-385*	*M6-853*
d^2WM	703,767.41	1,138,007.58	1,115,389.45
d^2WM_m	141,784.01	116,248.7	111,432.83
d^2OWA	**36,756,689.64**	1,185,783.26	1,133,943
d^2OWA_m	16,206,672.71	11,599.79	3,384.76
d^2CI	5,541,439.93	11,092,013.06	10,996,719.87
d^2CI_m	4,914,697.73	**11,129,653.95**	**11,069,537.49**
d^2SB_{NC}	–	6,103.3	5,260.63
d^2SB	2,239,318.59	13,213.44	12,235.56

Table 5.12: Computation deterministic time comparison (in ticks).

The time taken to learn the optimal weights and solve the problem for each dataset and learning approach can be seen in Table 5.12. The higher ticks' values were highlighted in bold. As it can be appreciated the Choquet integral is, in general, the approach computationally more expensive and complex. This is due to the number of constraints required in the optimization problem. This makes the symmetric bilinear function very effective. Results are similar to the Choquet integral but execution time is significantly smaller. Note that the table does not include ticks for the d^2SB_{NC} for the file *M5-38*, this is because the memory needed to solve the problem was higher than the 16GB of the workstation machine and this problem has been solved in the Finis Terrae machine (CESGA).

5.4.4 Heuristic Supervised Approach for Record linkage

Seeing how expensive is to run the supervised learning approach for record linkage when the Choquet integral is considered we proposed a heuristic approach in Section 5.3.2. In this section we compare this heuristic algorithm ($HRLA$) and the Choquet integral optimization algorithm (d^2CI) over different protected files.

This comparison is divided in two parts to tackle the optimization problem. In the first part we have focused on the percentages' comparison, in terms of the number of correct linkages and also the required times taken from both approaches. In the second part we have focused on the overfitting problem, testing both approaches with a small set for training and a big set for test.

In this first part, as in Section 5.4.3, the Choquet optimization problem was solved using the IBM ILOG CPLEX solver [IBM ILOG CPLEX, 2010a], (version 12.6) installed in the workstation. The heuristic approach $HRLA$ was completely programmed in the R statistical software and performed in a personal computer.

In order to evaluate this first experimental part we have considered two protected files used in Section 5.4.3, *M4-28* and *M5-38*. Both files were spotted as the ones with higher protection level.

	Dataset	d^2AM	d^2CI	$HRLA$
% Re-identifications	*M4-28*	68.50	93.75	91.75
	M5-38	39.75	91.25	86.75
Computational Time	*M4-28*	-	25 minutes	20 minutes
	M5-38	-	4.2 hours	20 minutes

Table 5.13: Percentage of re-identifications and computational time.

Table 5.13 shows the percentage of re-identifications and the consumed time of both approaches (d^2CI and $HRLA$). It is clear that both supervised approaches have obtained better results than the arithmetic mean (d^2AM). However, if we make a comparison between them, we can see that the $HRLA$ has an error between 2% and 5% respect to the optimum value, achieved by d^2CI. Recall that the $HRLA$ is initialized with an equilibrium fuzzy measure. Therefore, in the first iteration the $HRLA$ is at least as good as the Euclidean distance

(d^2AM). It is worth mention that, since $HRLA$ is an algorithm that finds the local minimum of a function, the results shown in that table correspond to the average of ten runs with the same configuration. The time required by the $HRLA$ to achieve similar results than d^2CI is much lower than the optimization algorithm. However, we have to remember that the time's factor of the $HRLA$ approach could be different depending on the learning rate and the number of iterations which are parameters of the algorithm set up in its initialization.

Note that we have evaluated the time in seconds instead of ticks, this is because the evaluation by ticks is exclusively given by the IBM ILOG CPLEX software.

In this second part of the evaluation, as in Section 5.4.2, to solve the Choquet optimization problem we used the IBM ILOG CPLEX solver [IBM ILOG CPLEX, 2010a], (version 12.1) installed in Finis Terrae (CESGA). While, the proposed heuristic approach $HRLA$ was completely programmed in the R statistical software and performed in a personal computer.

In this experiments part, we have anonymized the whole original file (Census) by means of four different protection methods with several degrees of protection. All these protection methods are enumerated below.

- **Rank Swapping**, (Section 2.4.2). We have considered values of $p = \{4, 5, 12, 15, 20\}$.
- **Additive Noise**, (Section 2.4.3). We have considered values of $p = \{1, 12, 16\}$.
- **Microaggregation**, (Section 2.4).
  - Z-scores projection (*MicZ*). $k = 3$
  - Multivariate microaggregation:
    - *Mic3*: microaggregation is applied in groups of three consecutive attributes independently. That is, two groups of three attributes and a single attribute. We have considered the following parameterizations $k = \{3, 5, 9\}$
    - *Mic4*: microaggregation is applied in groups of four consecutive attributes independently. That is, one group of four attributes and a group of tree attributes. We have considered the following parameterizations $k = \{4, 5, 8\}$

We suppose that the attacker has a prior knowledge, so, a linkage of 200 records between the original and the protected files (labeled training set) could be made. Then, using a supervised approach the set of Choquet integral coefficients are learned to re-identify the rest of records (880 records), i.e., the test set.

Table 5.14 shows the results obtained by d^2AM, d^2CI and $HRLA$ for the training and test partitions. In addition, it shows the times consumed by d^2CI to learn the respective fuzzy measure for each training set. The times consumed to learn the parameters with $HRLA$ are not shown because they were manually set

	d^2AM		d^2CI			$HRLA$	
Dataset	Train	Test	Time	Train	Test	Train	Test
RS-20	14.00	2.61	–	–	–	14.50	2.73
RS-15	24.50	9.89	–	–	–	26.00	8.98
RS-12	43.50	17.50	–	–	–	44.50	17.73
RS-5	94.00	78.86	**4min**	97.5	77.61	94.50	79.20
RS-4	95.50	85.23	**9sec**	100.00	80.91	97.00	85.11
Mic3-9	83.00	60.23	18min	89.50	57.16	83.00	60.11
Mic3-5	91.00	77.39	**1.5min**	96.50	74.66	93.00	76.93
Mic3-8	82.50	65.00	**5min**	91.00	62.95	83.00	65.11
Mic4-4	84.50	61.48	**2min**	88.00	58.52	84.50	61.70
Mic4-8	70.00	37.27	**13min**	75.50	35.68	70.00	37.16
Mic4-5	80.00	52.50	37min	85.00	50.45	80.00	52.50
Micz-3	0.00	0.23	**3sec**	0.00	0.11	0.00	0.23
Noise-16	87.00	70.11	1day	92.50	67.50	87.00	70.11
Noise-12	92.00	86.59	22min	97.00	80.57	93.00	86.82
Noise-1	100.00	100.00	**4sec**	100.00	99.66	100.00	100.00

Table 5.14: Percentage of re-identifications and time consumed.

to 14 minutes for all test sets. Besides, recall that the standard distance-based record linkage does not require any learning step, so it has not training times. Comment that the hyphens indicate that the corresponding computation was not finished, because it needed more than 300 hours (time limitation imposed by CESGA).

In the evaluation of the training process we have considered the times needed to learn the parameters and the percentage of correct re-identifications. As expected the highest re-identification results were achieved by d^2CI, achieving a maximum improvement of 8.5% and 8% when compared with d^2AM and $HRLA$, respectively. $HRLA$ obtained similar re-identification percentages than the standard distance-based record linkage (d^2AM), achieving a maximum improvement of a 2%. With respect to the times consumed to learn the parameters, $HRLA$ was manually set to 14 minutes, while the times need by d^2CI are very variable, depending on the protection method used. They range from few seconds to several hours. In Table 5.14 we highlighted in bold all d^2CI times lower than $HRLA$ times, i.e. 14 minutes.

In the test process the heuristic algorithm for record linkage ($HRLA$) has achieved the best re-identifications' percentages in 10 out of the 15 cases. It has achieved an improvement of at most 6% when compared with the optimization problem, this is a clear indicator of overfitting. Nevertheless, $HRLA$ has achieved similar re-identification results than d^2AM. This is due to the fact that $HRLA$ is initialized with the equilibrated weights and they were slightly changed by this algorithm. Although all the protection processes are different, they mainly rely on the addition of noise to each variable, so a distance function as the Euclidean distance can clearly re-identify some of the records, obviously always depending on the amount of noise added, that is the protection degree applied for the method.

5.4.5 Identification of Key-Attributes

By finding the optimal parameters that give us the maximum number of re-identifications we also obtain information regarding the relevance of each attribute or attribute interactions. That is, the attributes or combinations of them with the highest values are those that have more weight/relevance for the linkage process.

This means that we can establish a direct correspondence between the weights associated to each attribute with its disclosure risk, providing thus a disclosure risk estimation for each individual attribute. For example, an attribute with a high weight has a greater disclosure risk. Statistical agencies can then apply specific protection methods to concrete attributes if their individual disclosure risk is greater than expected in relation to the other attributes of the data set, that is, performing non-uniform protections.

As a first example, we consider the case of 7 attributes and 400 randomly selected records from original Census data set protected with additive noise. Unlike in previous additive noise tests, in this case we used different protection parameters for each attribute: attribute V_1 with $p = 1$, V_2 with $p = 2$, V_3 with $p = 3$, and so on. Table 5.15 shows for each attribute, the weights obtained with the d^2WM, and the parameter p of the additive noise used to protect the attribute.

Variable	V_1	V_2	V_3	V_4	V_5	V_6	V_7
p	1	2	3	4	5	6	7
Weight	0.97970	0.01484	0.00500	0.0	0.0	0.00046	0.0

Table 5.15: Weights identifying key-attributes for a file protected with additive noise, where each variable is protected with different values for the parameter p.

As expected, V_1 is the attribute with a clear higher weight since it is the variable with lower perturbation, and thus, the one that provides better information for the record linkage. In order to further analyze these results we have considered the re-identification using single variables. That is, we test the distance-based record linkage using only one variable each time. The results shown in Table 5.16, show that the re-identification percentages of each variable separately relate to the weights previously obtained. It is also interesting to note that single-variable record linkage obtains very poor re-identification results as compared to the record linkage with all 7 variables, which gives a 100% of correct matches.

This approach to identify key records can be compared to the Special Uniques Identification problem [Elliot and Manning, 2001, Elliot et al., 2002], which identifies records with a high risk of re-identification in a microdata file. In our case, we do not identify the risky records, but the risky variables.

Variable	V_1	V_2	V_3	V_4	V_5	V_6	V_7
p	1	2	3	4	5	6	7
Re-identification	29.5%	14.75%	10.5%	6.75%	7%	4.25%	4%

Table 5.16: Re-identification percentages using single variables for a file protected with additive noise, with different values of p for each variable.

In the case of the symmetric bilinear function, we have compared the covariance matrices used in the Mahalanobis distance (d^2MD) and the inverses of the weighting matrices obtained by the supervised approach using the symmetric bilinear function d^2SB_{NC} for the first five datasets and the matrix obtained by d^2SB_{PD} for the last case (because it needed an extra step to be positive definite). These are supposed to be similar than the covariance matrices or a scaled variation of those. However, when we compare both matrices by means of the mean square error (Equation 5.29), the results show that both matrices are different. See Table 5.17.

$$MSE(V, V') = \frac{\sum_{j=1}^{n} \sum_{1 \leq i \leq j} (v_{ij} - v'_{ij})^2}{\frac{n(n+1)}{2}} \tag{5.29}$$

	Mean square error
M4-33	18.49
M4-28	48.75
M4-82	$2,784.81$
M5-38	7.26
M6-385	15.91×10^6
M6-853	12.77×10^{16}

Table 5.17: Mean square error between covariance matrices and the positive definite matrices obtained.

Finally, we evaluate the relevance of the information provided by the fuzzy measure computing the supervised learning approach using the Choquet integral. This is the most interesting approach to spot relevant information for the re-identification process, since the fuzzy measure provides the relevance information of all single attributes and also all possible attribute combinations.

Once we solve the optimization problem with d^2CI for a concrete dataset, we can reconstruct the original fuzzy measure from the Möbius transform obtained using the Zeta transform, see Definition 2.5.

Table 5.18 shows the fuzzy measure obtained for *M4-28* and Figure 5.8 shows the lattice representation of this fuzzy measure. Each subset of variables $A \subseteq V$ is represented by the dyadic representation of its index k as described in Section 5.2.3, Example 10. We can see that the first variable provides a high degree of information for the re-identification. Note that all subsets which include it has a weight greater than 0.999. This variable has been protected with $k = 2$,

k	μ_k	k	μ_k
0000	0.000000	1000	0.999168
0001	0.154253	1001	0.999799
0010	0.003651	1010	0.999800
0011	0.207612	1011	0.999899
0100	0.039292	1100	0.999268
0101	0.154353	1101	0.999899
0110	0.096721	1110	0.999900
0111	0.207712	1111	0.999999

Table 5.18: Fuzzy measure for *M4-28*.

which preserves more information (is less distorted) than the last two variables (recall that they are protected with $k = 8$). It is also interesting to note that the highest weight of a two element subset is the one which includes the first and third variables. Each of these variables correspond to different protected blocks (one with $k = 2$ and the other with $k = 8$). So our approach is useful to detect that to combine variables with complementary information is useful in re-identification.

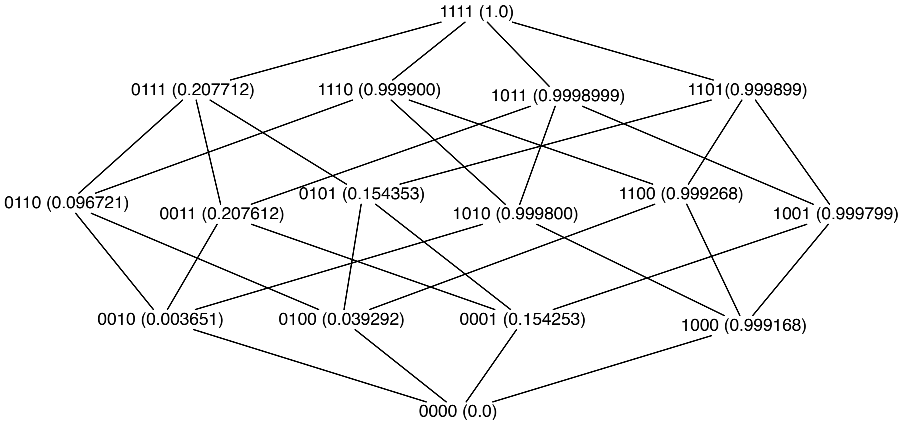


Figure 5.8: Fuzzy measure lattice for *M4-28*. Dyadic representation of the set and measures in brackets.

It is also interesting to observe the case of the fuzzy measure for the files with 6 variables. Table 5.19 shows the fuzzy measure for *M6-853*, and Figure 5.9 the lattice representation of the measure for all subsets with a weight $\mu_k \geq 0.1$.

Note, for example that all the sets of four elements (leaves from last row in Figure 5.9) include at least one element of each block of variables. That is, one element of the variables microaggregated with $k = 8$, one with $k = 5$, and one with $k = 3$. As stated before all these variables provide complementary

k	μ_k	k	μ_k
000000	0.0000000000	010000	0.0004214525
000001	0.0007378150	010001	0.0008378150
000010	0.0000000000	010010	0.0220367576
000011	0.0220367576	010011	0.0221367576
000100	0.0000000000	010100	0.0640252746
000101	0.0640252746	010101	0.0641252746
000110	0.0640252746	010110	0.0641252746
000111	0.0641252746	010111	0.8247279668
001000	0.0019057155	011000	0.0127378590
001001	0.0160771077	011001	0.0213887564
001010	0.0220367576	011010	0.0221367576
001011	0.0221367576	011011	0.0222367576
001100	0.0020057155	011100	0.0641252746
001101	0.0641252746	011101	0.0642252746
001110	0.0641252746	011110	0.8247279668
001111	0.0642252746	011111	0.8248279668
100000	0.0081683003	110000	0.0221158311
100001	0.0221158311	110001	0.0222158311
100010	0.0221158311	110010	0.0222158311
100011	0.0222158311	110011	0.0424704875
100100	0.0082683003	110100	0.0641252746
100101	0.0641252746	110101	0.0642252746
100110	0.0641252746	110110	0.9998000000
100111	0.9998000000	110111	0.9999000000
101000	0.0082683003	111000	0.0222158311
101001	0.0222158311	111001	0.0223158311
101010	0.0423704875	111010	0.0424704875
101011	0.0424704875	111011	0.0425704875
101100	0.0083683003	111100	0.0642252746
101101	0.0642252746	111101	0.0643252746
101110	0.9998000000	111110	0.9999000000
101111	0.9999000000	111111	1.0000000000

Table 5.19: Fuzzy measure for *M6-853*.

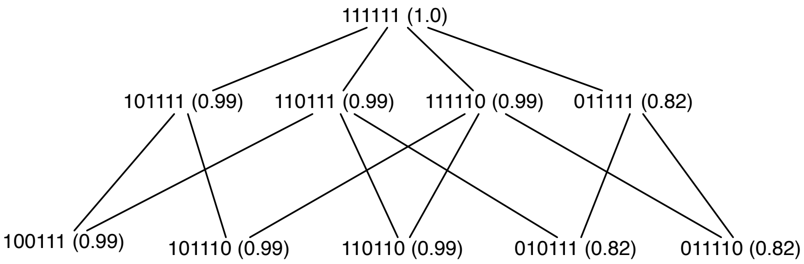


Figure 5.9: Partial fuzzy measure lattice for *M6-853* including all measures with values larger than 0.1.

information, which helps in the linkage process.

We also show in Table 5.20 the weights obtained for the same dataset if we compute the weighted mean distance d^2WM. In this case the most important variable seems be the 5th one. It comes as no surprise that this variable is present in all the measures from Figure 5.9. Note also that measures for sets which differ in the presence of the second and last variables are approximately the same. These variables do not seem to provide useful information for the record linkage.

k	*weight*
100000	0.016809573957189
010000	0.00198841786482128
001000	0.00452923777074791
000100	0.138812880222131
000010	0.835523953314578
000001	0.00233593687053289

Table 5.20: Weight vector for *M6-853*, when using d^2WM.

5.5 Conclusions

In this chapter we have introduced a supervised learning approach for record linkage as well as a set of parameterized functions that improve the current record linkage techniques. Although, the proposed approach is focused on solving a data privacy problem, it is easily extrapolated to other similar data matching scenarios. Along this chapter, we have introduced a formalization of a general re-identification problem to improve the current measures to evaluate the risk of sensitive information disclosure of a protected dataset. This problem is modeled as a mixed integer linear optimization problem.

We have proposed a set of four parameterized functions that can be used in the general supervised learning problem. They are the following: weighted mean, ordered weighted averaging operator, a symmetric bilinear function and the Choquet integral. All of them are used with the same purpose but as we saw they have different characteristics and complexities. We have formulated the different optimization problems and discussed their approach, the set of constraints and the number of parameters that have to be learned for each function by the optimization problem. Moreover, for each function we have studied different problem variations consisting of adding new constraints in order to make them satisfy as many metric/distance properties as possible.

Table 5.11 shows the results in terms of percentages of the number of correct matches between original records and protected records for all the proposed methods. It is easy to observe the high improvement achieved by all supervised approaches compared to the standard distance-based record linkage. However, from those approaches OWA is the one that get the worse results. Although these results overcome the standard method, OWA is not able to improve the results of the Mahalanobis distance. Additionally, this table also shows the percentage of correct links achieved by all the previous functions variations. Recall that these variations satisfy more metric properties than the original functions. When they are compared with the original functions, we saw a slight diminishment in the number of correct re-identifications. Nonetheless, since these parameterized function variations have more constraints than the original ones, the space to find a feasible solution is reduced. Hence, the solver leads to a result with fewer re-identifications but at the same time as the space is smaller it seems that it does not need as much time as the original function. See Table 5.12.

Motivated by the high computing times required to find the fuzzy measures by the Choquet integral we proposed a heuristic algorithm, which we compared to the optimization problem solved by the commercial solver, CPLEX. The first experimental part was focused on the comparison of methods in terms of time and number of correct re-identifications. The results obtained solving the problem with the IBM's solver guarantees the convergence to the optimal solution, but it requires more time, whereas the time required by the heuristic approach to solve the same problem remains low and stable. Regarding the results in this comparison the heuristic method obtained an error rate from 2% to 5%.

In the second part of the experiments, both methods were compared in order to study and cope with the other ways to solve the optimization problems faster and with less overfitting. This was motivated by the fact that it is not always possible to analyze the whole dataset, for instance, when the data practitioner wants a preliminary risk estimation. That is, he could evaluate the risk of a sample extracted from the dataset with which he is able to get a less accurate but faster risk estimation. The results suggest that the problems solved by the IBM's solver suffers from overfitting. The parameters learned in the training process give an accurate linkage information when they are also evaluated on the training data. However, when they are tested to evaluate the test data, they give worse results than the parameters learned by the heuristic approach.

Finally, we can conclude that if we have to perform an exhaustive disclosure risk evaluation and we have enough time and computational resources it is highly recommended solving the problem by using the CPLEX solver. In this way, the solver gives us the optimal combination of weights in order to obtain the maximum number of correct linkages. Moreover, with these weights we are able to perform a more efficient study with which we can spot which are the weaker attributes or sets of attributes that can lead to a re-identification problem. Thus, the protection practitioner could apply stronger masking methods to those attributes. Otherwise, if the resources needed are not available we could use the heuristic approach, that provides a good approximation to the optimal solution. That is, we recommend solving the problem with global optimization mechanisms to protection practitioners, whereas heuristic methods are more suitable to get fast but less accurate risk evaluations or to test possible record linkage attacks.

Chapter 6

Conclusions and Future Directions

In this last chapter we review all the work that have been presented along the previous chapters. We start summarizing all contributions. Then, we draw some conclusions about the work developed in this thesis. Finally, we show some open research lines to future development.

6.1 Summary of Contributions

In this dissertation we have contributed to different aspects within the field of data privacy. On the one hand, we have focused on the development of new mechanisms to protect sensitive textual data so, it can be released without disclosing any sensitive information related to individuals or topics. As we saw in Chapter 2 document protection is divided in two areas, depending on intended use of the data: the *declassification of documents* and the *privacy preserving text mining.* We provided several contributions on both areas. On the other hand, we also focused on the study of advanced techniques for microdata *disclosure risk assessment.* We summarize all of these contributions below.

In **Chapter 3** we presented some work related to *declassification technologies.* We developed a semi-automatic technique for confidential document sanitization. This technique was defined in two anonymization steps: One concerning the anonymization of entities' identifiers and the second one concerning to the anonymization of parts of the texts containing sensitive content. In addition, we presented a query-based method, which combines some information retrieval metrics to provide an evaluation of sanitized documents. This evaluation is performed in terms of the disclosure risk and the information that has been lost in the sanitization process. Finally, we test the semi-automatic sanitization method with some real classified documents published by WikiLeaks. According to the measures introduced, the results show a relatively low loss of information and a high reduction of the risk.

Privacy preserving text mining contributions were addressed in **Chapter 4**. Through classical documents' collections representations, vector space models, we established a relationship between classical privacy preserving and unstructured textual data protection. We have defined two anonymization methods for vector space models. They are two modifications of the multivariate microaggregation in which the privacy is ensured because of the satisfaction of the k-anonymity principle. On the one hand, in Section 4.3 we described the first variation, the Spherical microaggregation. Since vector space models usually are high-dimensional and very sparse data, we described and formalized an algorithm based on the microaggregation to exploit these characteristics. The main differences between the classical and the spherical microaggregation are on the distance and aggregation functions used. We have introduced a new partition and aggregation operations relying on the cosine distance instead of the Euclidean distance, which is usually employed in the classical microaggregation. Whereas, the formalization of the partition step was relatively straightforward, for the aggregation formalization we have provided a proof. We proved that the representative of the clusters is the one that minimizes the sum of distances between all the members of the cluster and the representative. The main idea behind this approach is to merge the k most similar documents in order to make them indistinguishable within the merged documents. An exhaustive testing has been carried out to evaluate the performance of the method. On the other hand, in Section 4.5 we proposed a second microaggregation variation, the Semantic microaggregation. Unlike the previous method, this exploits the semantic meanings of words. This approach relies on a generic lexical database such as WordNet, which provides tools to define distances and relations between words. As before, the partition and aggregation steps were modified according to semantic distances and aggregations. Hence, the cluster-based protection exploits the data words semantic relations by using specific distances and aggregations functions. This fact solves some problems presented by vector space models such as synonymity relations.

Finally, in **Chapter 5** we presented *advanced techniques for disclosure risk assessment*. This is a contribution in the area of record linkage as a disclosure risk evaluation method. In this chapter we described a supervised learning method for distance-based record linkage. This was introduced as a general approach and was formalized as an optimization problem. Specifically, it is a Mixed Integer Linear Programming (MILP) problem. Its goal is to find the parameters that maximize the number of correct re-identifications between two databases. Thus, its key point is the parameterized function used. We have proposed a set of four different parameterized functions. They are the weighted mean, the ordered weighted averaging (OWA) function, a symmetric bilinear function and the Choquet integral. We conducted a comprehensive testing analysis in which all such methods achieve a larger number of correct re-identifications than the standard distance-based record linkage techniques. We also showed the relevance of the analysis of the functions' parameters for protection practitioners.

6.2 Conclusions

In this thesis we have covered different aspects of data privacy. Most of our efforts were focused on the development of new technologies for documents' protection such as document declassification and protection methods which preserve text mining analysis. Additionally, we have focused on the research of more complex methods to evaluate the disclosure risk of a protected microdata set, so they could provide much more information than an estimation value of how difficult would be for an intruder to obtain sensitive information.

Methods for document declassification are a must-have in an open and transparent democracy. Sanitization methods are a helpful tool for enterprises and governments. We have provided tools to declassify documents in a faster and efficient way. Therefore, providing specific tools to automate sanitization we uphold the freedom of information while the national security is not jeopardised and additionally, unauthorized releases of confidential information such as the one performed by WikiLeaks could be avoided. Moreover, benchmarking methods to evaluate the utility and risk of sanitized methods are completely necessary to compare different sanitization techniques or spot possible mistakes done by automatic or manually sanitizations.

Disciplines such as Privacy Preserving Data Mining (PPDM) and Statistical Disclosure Control (SDC) put many efforts on the design and development of protection methods. One of the main privacy principles that has been proposed is k-anonymity, which is a privacy plus for any protection methods that can ensure it. E.g., microaggregation. For this reason we developed new protection methods for unstructured textual datasets based on microaggregation. One exploits the sparsity of the data and the other exploits the semantic relations of word meanings. The experiments concerning to the first approach were very satisfactory. All protected datasets lost a small amount of information. Specially, this could be noted in the classification evaluations and the specific information loss evaluations. The sum of errors evaluation show that the maximum information loss was 50% when the higher considered protection degree was applied. Moreover, with respect to the classification evaluations we obtained a lost of information up to 15% and 20% for each respective dataset. However, lower protection degrees are enough to keep the data safe, so values from 3 to 10 could be enough to protect most of the problems. These protection degrees are also recommended for the Semantic microaggregation approach.

A lot of research has been done in data privacy related to the evaluation of the possibilities an intruder has to get new sensitive information. That is, the disclosure risk assessment. In this thesis we have continued this research line and we have done a two-fold contribution. On the one hand, we demonstrated that by means of using supervised learning methods for disclosure risk assessment we were able to obtain better disclosure risk estimates and so, we improved the current distance-based methods significantly . On the other hand, we explained how useful could be using such supervised learning methods for the disclosure risk assessment, not only to obtain more accurate evaluations, but also it is important to analyze the information provided for such methods, that is, the learned parameters. Analyzing the resultant parameters' values we were able to spot which are the attributes or combinations of attributes that can lead to a security breach. Since intruders exploit whatever weakness they detect to achieve their goal, the study of this information could be considered a relevant countermeasure to detect these weaknesses. Therefore, protection practitioners can use this information to improve the dataset protection. Obviously, these supervised learning techniques can also be used by intruders to spot data weaknesses.

6.3 Future Directions

In this section we present some open research directions for each of the presented contributions. We summarize them below.

In this thesis we presented two approaches for the protection of textual unstructured data. One is used for document declassification, while the other is used to anonymize vector space models.

We have presented a new method for *document declassification*, however for a complete declassification this process still needs the help of an expert. Therefore, the need of fully-automate declassification methods is still open. We give a couple of ideas for further improvements in this topic. On the one hand, supervised or semi-supervised learning methods could be applied to tagged examples. An example could be the consideration of automatic text summarization techniques. This could be used to create text summaries from the whole document taking into account a set of given constraints. These constraints will state what should be considered sensitive information. On the other hand, techniques such as term generalization could be embedded to an automatic sanitization method. That is, despite of removing some concepts they could be generalized in order to reduce the information loss. Some preliminary results in this line are described in [Abril et al., 2011]. Consequently, mechanisms to evaluate the quality of sanitized documents should also be considered for further improvements. These mechanisms will be also used to make comparisons between different declassification methods. It would be interesting to apply learning processes that could be used to find the best overall descriptive equations.

Then, we contributed in the development of new *protection methods* for textual data that preserve text mining. They consist of a protection for vector representation of unstructured textual data, i.e. vector space models. Many improvements can be done in order to reduce loss of information, some ideas could be extracted from the text mining state-of-the-art. For instance, vector space models could be extended with word correlation matrices and so, distances will take into account these correlations and will be more accurate. In the case of the Semantic microaggregation, a straightforward improvement could consist of using a bunch of lexical databases, such as WordNet or other more specific databases like UMLS. Additionally, it would be interesting the study of different distances and aggregation operators to improve the microaggregation methods we have presented, as well as other anonymization techniques.

Finally, we present some future directions concerning to the contributions on the *advanced methods for disclosure risk assessment*. One interesting direction is to improve the results of the symmetric bilinear function by formalizing and solving the problem with semi-definite programming and compare the results with the heuristic approaches presented. Another research line is to address the overfitting and time computing problems of the record linkage approach based on the Choquet integral. Our proposal is to develop heuristic approaches that exploit the information provided by the fuzzy measures. That is, we are interested in executing the supervised learning approach based on the Choquet integral in small data partitions. The advantage is that the execution time for each partition will be lower than the execution for the whole dataset. However, each obtained fuzzy measure will be an optimal solution for each data partition, but not for the whole dataset. In this line, we are interested in defining distances between fuzzy measures in order to find possible relations between them and the number of re-identifications achieved. Additionally, we want to define a function that allows us to aggregate a set of fuzzy measures. Thus, we could generate a fuzzy measure that aggregates the information contained in the set of fuzzy measures obtained from computing the supervised learning approach in the small data partitions and so, it should achieve a higher number of re-identifications in the whole dataset than any other aggregated fuzzy measure. In addition, we are interested in the study of protection methods that take into account supervised learning methods for disclosure risk to obtain better protection datasets. That is, protection methods that iteratively could evaluate the risk and perform the necessary transformations into the data based on the obtained parameter values.

Contributions

[1] Abril, D., Navarro-Arribas, G., and Torra, V. (2010). Towards privacy preserving information retrieval through semantic microaggregation. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 296–299.

[2] Abril, D., Navarro-Arribas, G., and Torra, V. (2010). Towards semantic microaggregation of categorical data for confidential documents. In Torra, V., Narukawa, Y., and Daumas, M., editors, *Modeling Decisions for Artificial Intelligence*, volume 6408 of *Lecture Notes in Computer Science*, pages 266–276. Springer Berlin Heidelberg.

[3] Abril, D., Navarro-Arribas, G., and Torra, V. (2011). On the declassification of confidential documents. In Torra, V., Narakawa, Y., Yin, J., and Long, J., editors, *Modeling Decision for Artificial Intelligence*, volume 6820 of *Lecture Notes in Computer Science*, pages 235–246. Springer Berlin Heidelberg.

[4] Abril, D., Navarro-Arribas, G., and Torra, V. (2012). Choquet integral for record linkage. *Annals of Operations Research*, 195(1):97–110.
(**SCI Index 2012: 1.029**).

[5] Abril, D., Navarro-Arribas, G., and Torra, V. (2012). Improving record linkage with supervised learning for disclosure risk assessment. *Information Fusion*, 13(4):274 – 284. Information Fusion in the Context of Data Privacy.
(**SCI Index 2012: 2.262**).

[6] Abril, D., Navarro-Arribas, G., and Torra, V. (2013). Towards a private vector space model for confidential documents. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 944–945, New York, NY, USA. ACM.

[7] Abril, D., Navarro-Arribas, G., and Torra, V. (2013). Vector space model anonymization. In *Artificial Intelligence Research and Development: Proceedings of the 16th International Conference of the Catalan Association for Artificial Intelligence*, volume 256, pages 141–150. IOS Press.

[8] Abril, D., Navarro-Arribas, G., and Torra, V. (2014). Aprendizaje supervisado para el enlace de registros a travs de la media ponderada. In *Actas*

de la XIII Reunin Española sobre Criptologa y Seguridad de la Informacin (RECSI 2014). Forthcoming.

[9] Abril, D., Navarro-Arribas, G., and Torra, V. (2015). Spherical microaggregation: anonymizing sparse vector spaces. *Computers & Security*, 49(0): 28 – 44.
(**SCI Index 2014: 1.172**).

[10] Abril, D., Navarro-Arribas, G., and Torra, V. (2015). Supervised learning using a symmetric bilinear form for record linkage. *Information Fusion*. Information Fusion in the Context of Data Privacy. In press.
(**SCI Index 2014: 3.472**).

[9] Abril, D., Torra, V., and Narukawa, Y. (2014). Comparing fuzzy measures through their Möbius transform. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1 – 6, Salamanca, Spain. IEEE.

[10] Abril, D., Torra, V., and Navarro-arribas, G. (2011). Supervised Learning Using Mahalanobis Distance for Record Linkage. In Bernard De Baets, Radko Mesiar, L. T., editor, *Proc. of 6th International Summer School on Aggregation Operators (AGOP)*, pages 223–228, Benevento, Italy. Lulu.

[11] Murillo, J., Abril, D., and Torra, V. (2012). Heuristic supervised approach for record linkage. In Torra, V., Narukawa, Y., Lpez, B., and Villaret, M., editors, *Modeling Decisions for Artificial Intelligence*, volume 7647 of *Lecture Notes in Computer Science*, pages 210–221. Springer Berlin Heidelberg.

[12] Navarro-Arribas, G., Abril, D., and Torra, V. (2014). Dynamic anonymous index for confidential data. In Garcia-Alfaro, J., Lioudakis, G., Cuppens-Boulahia, N., Foley, S., and Fitzgerald, W. M., editors, *Data Privacy Management and Autonomous Spontaneous Security*, Lecture Notes in Computer Science, pages 362–368. Springer Berlin Heidelberg.

[13] Nettleton, D. and Abril, D. (2012). Document sanitization: Measuring search engine information loss and risk of disclosure for the wikileaks cables. In Domingo-Ferrer, J. and Tinnirello, I., editors, *Privacy in Statistical Databases*, volume 7556 of *Lecture Notes in Computer Science*, pages 308–321. Springer Berlin Heidelberg.

[14] Qureshi, M. A., Younus, A., Abril, D., ORiordan, C., and Pasi, G. (2013). Cirgdisco at replab2013 filtering task: Use of wikipedias graph structure for entity name disambiguation in tweets. In *CLEF (Online Working Notes/Labs/Workshop)*.

[15] Torra, V., Abril, D., and Navarro-arribas, G. (2011). Fuzzy methods for database protection. In *European Society for Fuzzy Logic and Technology (EUSFLAT)*, volume 1, pages 439 – 443, Aix-Les-Bains, France. Atlantis Press.

[16] Torra, V., Navarro-Arribas, G., and Abril, D. (2010). On the applications of aggregation operators in data privacy. In Huynh, V.-N., Nakamori, Y., Lawry, J., and Inuiguchi, M., editors, *Integrated Uncertainty Management and Applications*, volume 68 of *Advances in Intelligent and Soft Computing*, pages 479–488. Springer Berlin Heidelberg.

[17] Torra, V., Navarro-Arribas, G., and Abril, D. (2010). Supervised learning for record linkage through weighted means and owa operators. *Control and Cybernetics*, Vol. 39, no 4:1011–1026.
(**SCI Index 2010: 0.3**).

[18] Abril, D., Navarro-Arribas, G., and Torra, V. (2014). Data privacy with R. In Navarro-Arribas, G. and Torra, V., editors, *Advanced Research in Data Privacy*, volume 567 of *Studies in Computational Intelligence*, pages 63–82. Springer International Publishing.

[19] Nettleton, D. and Abril, D. (2014). An information retrieval approach to document sanitization. In Navarro-Arribas, G. and Torra, V., editors, *Advanced Research in Data Privacy*, volume 567 of *Studies in Computational Intelligence*, pages 151–166. Springer International Publishing.

Other References

[Abril et al., 2011] Abril, D., Navarro-Arribas, G., and Torra, V. (2011). On the declassification of confidential documents. In Torra, V., Narakawa, Y., Yin, J., and Long, J., editors, *Modeling Decision for Artificial Intelligence*, volume 6820 of *Lecture Notes in Computer Science*, pages 235–246. Springer Berlin Heidelberg.

[Aggarwal, 2005] Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 901–909. VLDB Endowment.

[Agrawal and Srikant, 2000] Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 439–450. ACM Press.

[Anandan et al., 2012] Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P., and Si, L. (2012). t-plausibility: Generalizing words to desensitize text. *Trans. Data Privacy*, 5(3):505–534.

[Baeza-Yates and Ribeiro-Neto, 2011] Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition) (ACM Press Books).* Addison-Wesley Professional, 2 edition.

[Barbaro and Zeller, 2006] Barbaro, M. and Zeller, T. J. (2006). A face is exposed for AOL searcher no. 4417749. *The New York Times.*

[Batini and Scannapieco, 2006] Batini, C. and Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications).* Springer-Verlag New York, Inc.

[Beliakov et al., 2011] Beliakov, G., James, S., and Li, G. (2011). Learning choquet-integral-based metrics for semisupervised clustering. *Fuzzy Systems, IEEE Transactions on*, 19(3):562–574.

[Bellet et al., 2013] Bellet, A., Amaury, H., and Marc, S. (2013). A survey on metric learning for feature vectors and structured data. *CoRR*, abs/1306.6709.

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls): Integrating biomedical terminology. http://www.nlm.nih.gov/research/umls.

[Brand, 2002] Brand, R. (2002). Microdata protection through noise addition. In Domingo-Ferrer, J., editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 97–116. Springer Berlin Heidelberg.

[Brand et al., 2002] Brand, R., Domingo-Ferrer, J., and Mateo-Sanz, J. (2002). Reference datasets to test and compare sdc methods for protection of numerical microdata. *Technical report, European Project IST-2000-25069 CASC.*

[Canada, 2010] Canada, S. (2010). Record linkage at statistics canada. http://http://www.statcan.gc.ca/record-enregistrement/index-eng.htm.

[Census Bureau, 1995] Census Bureau, U. (1995). Data extraction system. http://www.census.gov/.

[CESGA, 2010] CESGA (2010). Centro de supercomputación de galicia. http://www.cesga.es.

[Chakaravarthy et al., 2008] Chakaravarthy, V. T., Gupta, H., Roy, P., and Mohania, M. K. (2008). Efficient techniques for document sanitization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 843–852, New York, NY, USA. ACM.

[Choquet, 1953] Choquet, G. (1953). Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295.

[Chow et al., 2011] Chow, R., Staddon, J., and Oberst, I. (2011). Method and apparatus for facilitating document sanitization. US Patent App. 12/610,840.

[Christen, 2012] Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *Knowledge and Data Engineering, IEEE Transactions on*, 24(9):1537–1555.

[Colledge, 1995] Colledge, M. (1995). Frames and business registers: An overview. business survey methods. *Wiley Series in Probability and Statistics.*

[Cox and Cox, 1994] Cox, T. F. and Cox, M. A. A. (1994). *Multidimensional Scaling.* Chapman & Hall/CRC Monographs on Statistics & Applied Probab. Chapman and Hall.

[Cumby and Ghani, 2011] Cumby, C. and Ghani, R. (2011). A machine learning based system for semi-automatically redacting documents. In Shapiro, D. G. and Fromherz, M. P. J., editors, *Twenty-Third Innovative Applications of Artificial Intelligence*, pages 1628–1635.

[Dalenius and Reiss, 1982] Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6(1):73 – 85.

[Dantzig, 1951] Dantzig, G. (1951). *Maximization of a Linear Function of Variables Subject to Linear Inequalities*, pages 339–347. T. C. Koopmans, Ed.

[DARPA, 2010] DARPA (2010). New technologies to support declassification. Request for Information (RFI) Defense Advanced Research Projects Agency. Solicitation Number: DARPA-SN-10-73.

[data.gov, 2010] data.gov (2010). Usa government.

[data.gov.uk, 2010] data.gov.uk (2010). Uk government.

[Defays and Nanopoulos, 1993] Defays, D. and Nanopoulos, P. (1993). Panels of enterprises and confidentiality: The small aggregates method. In *Proc. of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204. Statistics Canada.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38.

[Deza and Deza, 2009] Deza, M. and Deza, E. (2009). *Encyclopedia of distances.* Springer Verlag.

[Dhillon and Modha, 2001] Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, 42(1-2):143–175.

[DOE/OSTI, 2014] DOE/OSTI (2014). U.s department of energy. Deparment of energy researches use of advanced computing for document declassificatin. http://www.osti.gov/opennet.

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

[Domingo-Ferrer et al., 2001] Domingo-Ferrer, J., Mateo-sanz, J. M., and Torra, V. (2001). Comparing sdc methods for microdata on the basis of information loss and disclosure. In *Proceedings of ETK-NTTS 2001, Luxemburg: Eurostat*, pages 807–826. Eurostat.

[Domingo-Ferrer et al., 2004] Domingo-Ferrer, J., Seb, F., and Castell-Roca, J. (2004). On the security of noise addition for privacy in statistical databases. In Domingo-Ferrer, J. and Torra, V., editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 149–161. Springer Berlin Heidelberg.

[Domingo-Ferrer and Torra, 2001a] Domingo-Ferrer, J. and Torra, V. (2001a). *Disclosure Control Methods and Information Loss for Microdata*, pages 91–110. Elsevier.

[Domingo-Ferrer and Torra, 2001b] Domingo-Ferrer, J. and Torra, V. (2001b). *A quantitative comparison of disclosure control methods for microdata*, pages 111–133. Elsevier.

[Domingo-Ferrer and Torra, 2002] Domingo-Ferrer, J. and Torra, V. (2002). Validating distance-based record linkage with probabilistic record linkage. In Escrig, M., Toledo, F., and Golobardes, E., editors, *Topics in Artificial Intelligence*, volume 2504 of *Lecture Notes in Computer Science*, pages 207–215. Springer Berlin Heidelberg.

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

[Domingo-Ferrer et al., 2006] Domingo-Ferrer, J., Torra, V., Mateo-Sanz, J. M. i. S., and Sebe, F. (2006). Empirical disclosure risk assessment of the ipso synthetic data generators. In *Monographs in Official Statistics, UNECE/Eurostat Work Session on Statistical Data Confidentiality*, pages 227–238.

[Duda et al., 2012] Duda, R., Hart, P., and Stork, D. (2012). *Pattern Classification.* John Wiley and Sons.

[Duncan et al., 2001] Duncan, G. T., Keller-mcnulty, S. A., and Stokes, S. L. (2001). Disclosure risk vs. data utility: The r-u confidentiality map. Technical report, Chance.

[Dunn, 1946] Dunn, H. (1946). Record linkage. *American Journal of Public Health*, 36(12):1412–1416.

[Elliot and Manning, 2001] Elliot, M. J. and Manning, A. M. (2001). The identification of special uniques. In *Proceedings of GSS Methodology Conference.*

[Elliot et al., 2002] Elliot, M. J., Manning, A. M., and Ford, R. W. (2002). A computational algorithm for handling the special uniques problem. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):493–509.

[Elmagarmid et al., 2007] Elmagarmid, A., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16.

[E.O. 13526, 2009] E.O. 13526 (2009). Executive order 13526, of the us administration. Classified National Security Information, Section 1.4, points (a) to (h). http://www.whitehouse.gov/the-press-office/ex-ecutive-order-c lassified-national-security-information.

[Erola et al., 2010] Erola, A., Castell-Roca, J., Navarro-Arribas, G., and Torra, V. (2010). Semantic microaggregation for the anonymization of query logs. In Domingo-Ferrer, J. and Magkos, E., editors, *Privacy in Statistical Databases*, volume 6344 of *Lecture Notes in Computer Science*, pages 127–137. Springer Berlin Heidelberg.

[Fellegi and Sunter, 1969] Fellegi, I. and Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

[Friedman, 1994] Friedman, J. H. (1994). Flexible metric nearest neighbor classification. Technical report, Department of Statistics, Stanford University.

[Fukunaga, 1990] Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Computer science and scientific computing. Elsevier Science.

[Ghinita et al., 2008] Ghinita, G., Tao, Y., and Kalnis, P. (2008). On the anonymization of sparse high-dimensional data. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 715–724.

[Goldberger et al., 2004] Goldberger, J., Roweis, S., Hinton, G., and Salakhutdinov, R. (2004). Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press.

[Grabisch, 1995] Grabisch, M. (1995). A new algorithm for identifying fuzzy measures and its application to pattern recognition. In *Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., Proceedings of 1995 IEEE Int*, volume 1, pages 145–150.

[Halkidi et al., 2008] Halkidi, M., Gunopulos, D., Vazirgiannis, M., Kumar, N., and Domeniconi, C. (2008). A clustering framework based on subjective and objective validity criteria. *ACM Trans. Knowl. Discov. Data*, 1(4):4:1–4:25.

[Hartley, 1958] Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, 14(2):174–194.

[Hastie and Tibshirani, 1996] Hastie, T. and Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(6):607–616.

[Herzog et al., 2007] Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data quality and record linkage techniques.* Springer.

[Higham, 2002] Higham, N. (2002). Computing the nearest correlation matrix a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343.

[Hong et al., 2011] Hong, T.-P., Lin, C.-W., Yang, K.-T., and Wang, S.-L. (2011). A heuristic data-sanitization approach based on tf-idf. In Mehrotra, K., Mohan, C., Oh, J., Varshney, P., and Ali, M., editors, *Modern Approaches in Applied Intelligence*, volume 6703 of *Lecture Notes in Computer Science*, pages 156–164. Springer Berlin Heidelberg.

[ppdm.iiia.csic.es, 2009] ppdm.iiia.csic.es (2009). (PPDM) Privacy Preserving Data Mining.

[http://www.pingar.com, 2014] http://www.pingar.com (2014). Pingar entity extraction software.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

[IBM ILOG CPLEX, 2010a] IBM ILOG CPLEX, I. (2010a). High-performance mathematical programming engine. international business machines corp. http://www-01.ibm.com/software/integration/optimization/cplex/.

[IBM ILOG CPLEX, 2010b] IBM ILOG CPLEX, I. (2010b). Ibm ilog cplex optimization studio v12.6.0 documentation. http://pic.dhe.ibm.com/infocenter/cosinfoc/v12r6/index.jsp.

[Jaro, 1989] Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

[Jero, 2012] Jero, L. (2012). On leveraging gpus for security: discussing k-anonymity and pattern matching. Master's thesis, Università degli Studi Roma.

[Johnson, 1970] Johnson, C. R. (1970). Positive definite matrices. *The American Mathematical Monthly*, 77(3):pp. 259–264.

[Karmarkar, 1984] Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395.

[Lasko and Vinterbo, 2010] Lasko, T. and Vinterbo, S. (2010). Spectral anonymization of data. *Knowledge and Data Engineering, IEEE Transactions on*, 22(3):437–446.

[Laszlo and Mukherjee, 2005] Laszlo, M. and Mukherjee, S. (2005). Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Trans. on Knowl. and Data Eng.*, 17(7):902–911.

[Liu and Wang, 2013] Liu, J. and Wang, K. (2013). Anonymizing bag-valued sparse data by semantic similarity-based clustering. *Knowledge and Information Systems*, 35(2):435–461.

[Luenberger and Ye, 2008] Luenberger, D. and Ye, Y. (2008). *Linear and Nonlinear Programming.* International Series in Operations Research & Management Science. Springer.

[Machanavajjhala et al., 2007] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1).

[Mahalanobis, 1936] Mahalanobis, P. C. (1936). On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55.

[Manning et al., 2008a] Manning, C. D., Raghavan, P., and Schütze, H. (2008a). *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA.

[Manning et al., 2008b] Manning, C. D., Raghavan, P., and Schütze, H. (2008b). *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA.

[Martínez et al., 2012a] Martínez, S., Sánchez, D., and Valls, A. (2012a). Semantic adaptive microaggregation of categorical microdata. *Computupers and Security*, 31(5):653–672.

[Martínez et al., 2012b] Martínez, S., Valls, A., and SáNchez, D. (2012b). Semantically-grounded construction of centroids for datasets with textual attributes. *Knowledge-Based Systems*, 35:160–172.

[Mateo-Sanz et al., 2005] Mateo-Sanz, J. M., Domingo-Ferrer, J., and Sebé, F. (2005). Probabilistic information loss measures in confidentiality protection of continuous microdata. *Data Mining and Knowledge Discovery*, 11(2):181–193.

[Mccallum and Wellner, 2003] Mccallum, A. and Wellner, B. (2003). Object consolodation by graph partitioning with a conditionally-trained distance metric. *In Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 19–24.

[McLachlan and Krishnan, 1997] McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics).* Wiley-Interscience.

[Meystre et al., 2010] Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., and Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):70+.

[Miller, 1995] Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

[Mitchell, 2002] Mitchell, J. E. (2002). Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of Applied Optimization*, pages 65–77.

[Moore, 1996] Moore, R. (1996). Controlled data-swapping techniques for masking public use microdata sets. *Statistical Research Division Report Series*, pages 96–04.

[Narayanan and Shmatikov, 2008] Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125.

[Narukawa, 2007] Narukawa, Y. (2007). Distances defined by choquet integral. In *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, pages 1–6. IEEE.

[Navarro-Arribas et al., 2012] Navarro-Arribas, G., Torra, V., Erola, A., and Castellí-Roca, J. (2012). User k-anonymity for privacy preserving data mining of query logs. *Inf. Process. Manage.*, 48(3):476–487.

[Neamatullah et al., 2008] Neamatullah, I., Douglass, M., Lehman, L.-w. H., Reisner, A., Villarroel, M., Long, W., Szolovits, P., Moody, G. B., Mark, R. G., and Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC MEDICAL INFORMATICS AND DECISION MAKING*, 8(32).

[Newcombe and Kennedy, 1962] Newcombe, H. B. and Kennedy, J. M. (1962). Record linkage: making maximum use of the discriminating power of identifying information. *Commun. ACM*, 5(11):563–566.

[Newcombe et al., 1959] Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic linkage of vital records. *Science*, 130:954–959.

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

[OKF, 2013] OKF (2013). Open data index. https://index.okfn.org/country.

[Pagliuca and Seri, 1999] Pagliuca, D. and Seri, G. (1999). Some results of individual ranking method on the system of enterprise acounts annual survey. *Esprit SDC Project, Delivrable MI-3/D2.*

[Pham and Yan, 1999] Pham, T. D. and Yan, H. (1999). Color image segmentation using fuzzy integral and mountain clustering. *Fuzzy Sets and Systems*, 107(2):121 – 130.

[Porter, 1980] Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

[Rutkin, 2013] Rutkin, A. (2013). How your facebook profile reveals more about your personality than you know.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

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

[Samelin et al., 2012] Samelin, K., Phls, H., Bilzhause, A., Posegga, J., and de Meer, H. (2012). Redactable signatures for independent removal of structure and content. In Ryan, M., Smyth, B., and Wang, G., editors, *Information Security Practice and Experience*, volume 7232 of *Lecture Notes in Computer Science*, pages 17–33. Springer Berlin Heidelberg.

[Schrijver, 1986] Schrijver, A. (1986). *Theory of Linear and Integer Programming.* John Wiley & Sons, Inc., New York, NY, USA.

[Schultz and Joachims, 2004] Schultz, M. and Joachims, T. (2004). Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems 16.* MIT Press.

[Simonite, 2013] Simonite, T. (2013). Ads could soon know if youre an introvert (on twitter).

[Strehl et al., 2000] Strehl, A., Strehl, E., Ghosh, J., and Mooney, R. (2000). Impact of similarity measures on web-page clustering. In *In Workshop on Artificial Intelligence for Web Search (AAAI) 2000*, pages 58–64. AAAI.

[Sugeno, 1974] Sugeno, M. (1974). *Theory of Fuzzy Integrals and its Applications.* PhD thesis, Tokyo Institute of Technology, Tokyo, Japan.

[Sun and Chen, 2011] Sun, S. and Chen, Q. (2011). Hierarchical distance metric learning for large margin nearest neighbor classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(07):1073–1087.

[Sweeney, 2002] Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570.

[Tan et al., 2005] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition).* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Tenenbaum et al., 2000] Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

[Torra, 2004] Torra, V. (2004). Microaggregation for categorical variables: A median based approach. In Domingo-Ferrer, J. and Torra, V., editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 162–174. Springer Berlin Heidelberg.

[Torra, 2008] Torra, V. (2008). Constrained microaggregation: Adding constraints for data editing. *Transactions on Data Privacy*, 1:86–104.

[Torra, 2010] Torra, V. (2010). Privacy in data mining. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 687–716. Springer US.

[Torra et al., 2006] Torra, V., Abowd, J., and Domingo-Ferrer, J. (2006). Using mahalanobis distance-based record linkage for disclosure risk assessment. In Domingo-Ferrer, J. and Franconi, L., editors, *Privacy in Statistical Databases*, volume 4302 of *Lecture Notes in Computer Science*, pages 233–242. Springer Berlin Heidelberg.

[Torra and Domingo-Ferrer, 2003] Torra, V. and Domingo-Ferrer, J. (2003). Record linkage methods for multidatabase data mining. In Torra, V., editor, *Information Fusion in Data Mining*, volume 123 of *Studies in Fuzziness and Soft Computing*, pages 101–132. Springer Berlin Heidelberg.

[Torra and Narukawa, 2007] Torra, V. and Narukawa, Y. (2007). *Modeling Decisions: Information Fusion and Aggregation Operators*. Springer.

[Torra and Narukawa, 2012] Torra, V. and Narukawa, Y. (2012). On a comparison between mahalanobis distance and choquet integral: The choquetmahalanobis operator. *Information Sciences*, 190(0):56 – 63.

[Weinberger et al., 2006] Weinberger, K. Q., Blitzer, J., and Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 1473–1480. MIT Press.

[Wikileaks, 2010] Wikileaks (2010). Cable repository, http://www.cablegatesearch.net.

[Willenborg and Waal, 2001] Willenborg, L. and Waal, T. (2001). *Elements of Statistical Disclosure Control*. Springer-Verlag.

[Winkler, 2003] Winkler, W. E. (2003). Data cleaning methods. *Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[Winkler, 2004] Winkler, W. E. (2004). Re-identification methods for masked microdata. In Domingo-Ferrer, J. and Torra, V., editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 216–230, Heidelberg, Berlin. Springer Berlin Heidelberg.

[WordNet, 2010] WordNet (2010). Princeton university "about wordnet." princeton university. http://wordnet.princeton.edu.

[Wu and Palmer, 1994] Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Xing et al., 2003] Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. (2003). Distance metric learning, with application to clustering with side-information. In *Advances in neural information processing systems*, pages 505–512. MIT Press.

[Yager, 1988] Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *Systems, Man and Cybernetics, IEEE Transactions on*, 18(1):183–190.

[Yager, 1993] Yager, R. R. (1993). Families of owa operators. *Fuzzy Sets Syst.*, 59(2):125–148.

[Yahoo! News, 2010] Yahoo! News (2010). Top 10 revelations from wiki leaks cables, http://news.yahoo.com/blogs/lookout/top-10-revelations-wikileaks-cables.html.

[Yancey et al., 2002] Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In *Inference Control in Statistical Databases, From Theory to Practice*, volume 2316, pages 135–152, London, UK. Springer-Verlag.

[Yao, 1982] Yao, A. C. (1982). Protocols for secure computations. In *SFCS '82: Proceedings of the 23rd Annual Symposium on Foundations of Computer Science*, pages 160–164, Washington, DC, USA. IEEE Computer Society.