Universitat Politècnica de Catalunya

Dissertation Thesis

Applied Mathematics Ph.D. programme

Security and Information Loss in the Protection of Statistical Databases

November 2002

Ph.D. Candidate: Anna Oganian

Ph.D. Advisor: Josep Domingo-Ferrer

Tutor: Carles Padró

Contents

1 Preface **7**

2 Introduction **9**

2.1 Situation and objectives 9

2.2 Structure of the thesis 13

3 State of the art **15**

3.1 Microdata protection 15

3.1.1 Notation 17

3.1.2 Perturbative methods 18

3.1.3 Nonperturbative methods 25

3.1.4 Existing performance measures for microdata 26

3.2 Tabular data protection 29

3.2.1 Data formats 30

3.2.2 Classification of tabular data 30

3.2.3 Perturbative methods for tabular data 32

3.2.4 Nonperturbative methods for tabular data 32

3.2.5 Existing performance measure for tabular data 33

4 Microdata performance measures **37**

4.1 Information loss measures 37

4.2 Disclosure risk measures for microdata 41

4.2.1 Distance-based record linkage 41

4.2.2 Probabilistic record linkage 42

4.2.3 Interval disclosure 42
4.3 Comparison of SDC methods for continuous microdata 43
4.3.1 Identification of relevant microdata masking methods . 44
4.3.2 Test data collection 46
4.3.3 Information loss metrics 47
4.3.4 Disclosure risk assessment 48
4.3.5 Empirical work 48
4.4 Conclusions of the comparison 53

5 Complexity of microaggregation **55**
5.1 Background 57
5.2 Microaggregation is NP-hard 57

6 Critique of sensitivity rules **67**
6.1 Analysis of the dominance rule 68
6.1.1 Replacing the dominance rule 72
6.2 Analysis of the p% and the pq-rules 75
6.2.1 Replacing the p% and pq-rules 78
6.3 Conclusions on sensitivity rules 80

7 *A posteriori* DR measure **83**
7.1 An *a posteriori* DR measure 84
7.2 Application to different scenarios 86
7.2.1 Cell suppression 87
7.2.2 Rounding 88
7.2.3 Table redesign 88
7.2.4 Linked tables 89
7.2.5 Minimizing disclosure risk in linked table release 91

8 Conclusions **99**
8.1 Results of this thesis 100
8.2 Future research 102

A Appendix. Probabilistic record linkage **105**
A.1 Theoretical concepts 105
A.1.1 Background 105
A.1.2 Weights 106
A.1.3 Decision procedure 106
A.1.4 Estimation of the u_i 107
A.1.5 Blocking 107
A.2 Parameter estimation methodology 108
A.2.1 Comparison configuration frequencies 108
A.2.2 The EM algorithm 109
A.2.3 Calculation of threshold values 112
A.3 Matching algorithm 113
A.3.1 Composite weight calculation 113
A.3.2 Assignment 113
A.3.3 Duplicates 115
A.3.4 File preparation 115

Our contributions **117**

Other references **119**

Chapter 1

Preface

This thesis is about security and information loss in the protection of statistical databases. Publication of statistical information by statistical offices raises the problem of preventing disclosure of confidential information about particular respondents (individuals or enterprises) without significantly damaging the utility of the data being protected.

In this way, disclosure risk and information loss are two important aspects of the protection of statistical data. Measuring them is one of the main issues of this thesis.

This report is divided according to the two types of statistical data: microdata (individual respondent records) and tabular data.

The first part of the thesis addresses issues in the field of microdata protection. One of them is a unified approach for measuring disclosure risk and information loss in microdata protection: new measures for both disclosure risk and information loss are proposed. A second issue in this part of the thesis is related to one of the most widely used microdata protection methods, namely microaggregation. The complexity analysis of microaggregation with minimal information loss is presented.

The second part of the thesis deals with tabular data protection and presents the results obtained in regard to disclosure risk measurement. Firstly, flaws of the currently used *a priori* measures (sensitivity measures which are

used prior to the application of protection methods) are highlighted. Secondly, new *a priori* and *a posteriori* measures are presented which work better than current measures.

Chapter 2

Introduction

2.1 Situation and objectives

Statistical database protection or Statistical Disclosure Control (SDC) is a part of information security which tries to prevent published statistical information (tables, individual records) from disclosing to third parties the contribution of specific respondents, who may be individuals or enterprises. Statistical data are assumed to have been collected by a statistical office through a survey or a census and the "third parties" are users outside the statistical office, *e.g.* policy makers, academic researchers and the general public.

Traditionally, statistical offices used to disseminate statistical information only in form of tables. But microdata (records which contain information about individuals) offer far greater flexibility for statistical research –especially of an exploratory nature– than tables, so there has been an increasing demand from users for such data, and statistical offices have increasingly been yielding to this demand.

In both cases, the statistical office has to be careful when providing outside users with access to such data. A statistical office typically has an obligation towards the individuals to which the data refer to make sure that it is (almost) impossible for a user to use the data to disclose confidential

information about these individuals.

SDC achieves its goal through *modification* of the data, that is, by applying techniques such as grouping of categories, suppression or alteration of values in records or table cells. In order to do so purposefully, a statistical office needs criteria to check whether data are safe enough to be released, possibly only to a particular group of users and possibly under certain conditions as to the use of the data.

Such criteria can be used during the process that transforms unsafe data into safe ones by employing SDC modification techniques. Of course, an objective in such a transformation process should be to retain as much information in the data as possible or, equivalently, to modify the data as little as possible.

How information loss and disclosure risk can be quantified is one of the most important problems in SDC and is a major theme of this thesis. As there are two types of data which may be released by statistical offices (microdata and tabular data) and their structure is quite different, this thesis is divided into two parts: the first one treats microdata protection and the second one deals with tabular data protection.

In the first part, performance measures for microdata protection are one of the main issues. Performance measures are of two kinds: information loss and disclosure risk measures. We next define what information loss and disclosure risk are:

1. *Information loss* is the loss of utility of the released data for a legitimate data user.

2. *Disclosure risk* is the risk to respondent confidentiality that the data releaser (typically a statistical agency) would experience as a consequence of releasing the table.

Strictly speaking, information loss measures depend on the data uses to be supported by the SDC-protected data. However, potential data uses are very diverse and it may be hard even to identify them all at the moment of

data release by a statistical office. It is thus desirable for the data protector to be able to measure information loss in a generic way that reflects how much harm is being inflicted to the data by a given masking method; the amount of information loss measured in this generic way should roughly correspond to the amount of information loss for a reasonable range of data uses. On the other side, when a mixture of data protection techniques is used, generic measures applicable to the various methods are needed. Unfortunately, a good deal of the information loss measures that have been proposed in the literature only make sense for specific protection methods and are thus useless for performance comparison among a wide range of methods.

On the other hand, the assessment of the performance of an SDC method cannot be limited to information loss; disclosure risk must also be measured. A method that optimizes the trade-off between both magnitudes subject to some user requirements turns out to be the best option. There is little in the literature on disclosure risk for microdata that can be used for a broad class of methods –disclosure risk measures tend to be even more method-specific than information loss measures–. So one of the aims of our research has been to devise generic performance measures that encompass disclosure risk and information loss.

Once these criteria have been specified, the stage has been set to deal with the transformation process itself: transforming unsafe data into safe data can be formulated as a collection of optimization problems whose solution is far from trivial. In case the original optimization problems turn out to be too hard to solve, one has to find heuristic methods that yield good approximate solutions in many practical instances. One of the most widely used transformation procedures is a collection of SDC methods called microaggregation, which aggregate data into small homogeneous groups and output the average value of the elements of each group. These average values are the only information being released. A lot of heuristics exist for microaggregation but, for data of dimension greater than 1, no polynomial-time algorithm has been proposed so far. Therefore, the complexity analysis

of multivariate microaggregation may be of great use because it will define the direction of the research related to that method: either try to find the exact polynomial-time algorithm or give up these attempts and concentrate on devising good microaggregation heuristics. It is shown in this thesis that microaggregation with minimal information loss is an NP-hard problem when data are of dimension greater than 1.

The second part of this thesis deals with tabular data protection. Tabular data constitute the most traditional output released by statistical agencies. Being aggregate data, one might infer that tables cannot leak information about specific respondents. As argued in [32], it turns out that table cells often do contain information on a single or very few respondents, which implies a disclosure risk for the data of those respondents. In these cases, disclosure control methods must be applied to the tables prior to their release.

Disclosure assessment for tables is usually performed *a priori*, that is, before tables are protected. The standard approach is to use a *sensitivity rule* to decide whether a table cell is sensitive and should be protected. These rules have received critiques for failing to adequately reflect the risk of disclosure, but these have been limited to numerical counterexamples for particular choices of the parameters of these rules. In this research line, we prove that, in general, these rules do not reflect disclosure risk in a proper way and we propose a new *a priori* measure which overcomes the drawbacks of current sensitivity rules.

Finally, we introduce another type of disclosure risk measures for tables: *a posteriori* disclosure risk measures, which operate *after* application of SDC methods. We argue that evaluation of the quality of the protection offered by tabular SDC methods is not well covered in the literature, the basic flaw being that disclosure risk is usually measured based on the original table *before* protecting it. However, a table cell may still be sensitive *after* application of the SDC methods. So, in order to achieve a higher level of security, disclosure risk must be measured after the application of SDC methods and, if a particular table cell is still considered sensitive, protection methods must

be applied again.

2.2 Structure of the thesis

This thesis report is organized as follows.

Chapter 3 contains the state of art of the main areas of this dissertation: microdata protection and tabular data protection. Basic definitions and properties as well as general ideas are presented. The main problems are outlined for later discussion. The rest of chapters are original contributions so this chapter will be often recalled in the rest of the thesis.

Chapter 4 deals with performance measures for microdata protection. In particular, generic information loss and disclosure risk measures are proposed. A comparison of SDC methods based on the proposed measures is presented as well.

Chapter 5 presents a complexity analysis of the microaggregation method. More precisely, the proof of the NP-hardness of optimal microaggregation is given in full detail.

Chapters 6 and 7 deal with performance measures for tabular data:

- Chapter 6 presents critiques of currently used sensitivity rules which measure disclosure risk on an *a priori* basis. A new measure based on the reciprocal of Shannon's entropy is proposed as an alternative which is free from the drawbacks identified for existing sensitivity rules.

- Chapter 7 introduces a different approach for measuring disclosure risk, namely *a posteriori* measures of disclosure risk. The application of such measures to several SDC methods is also presented in that chapter.

In the Appendix, the basic definitions and properties of record linkage techniques are recalled.

Chapter 3

State of the art

This chapter surveys the state of art of the topics which are treated in this thesis. Such background information will provide the knowledge needed to identify the weaknesses of the existing proposals and their possible solutions.

3.1 Microdata protection

Statistical disclosure control (SDC) of microdata seeks the protection of individual records that contain values of variables for a single or a few respondents, who can be persons, enterprises or other individual units.

Thus, the main goal of SDC is to modify the original data so that they can be published without giving away confidential information that can be linked to specific respondents. The challenge for SDC is to achieve this modification with minimum loss of the detail and accuracy sought by database users. Thus we can say that, any SDC method has two conflicting goals:

1. Minimize disclosure risk, that is the risk to respondent confidentiality that the data releaser (typically a statistical agency) would experience as a consequence of releasing the data.

2. Maximize data utility, that is the value of the released data to a legitimate data user.

Consequently, the performance of a statistical disclosure control method can be measured as the tradeoff it achieves between maximizing the value of the released data (minimizing information loss) and keeping disclosure risk acceptably low.

SDC methods for microdata are usually known as *masking methods*, of which there is a wide range. From the point of view of their operational principles, current masking methods fall into the following two categories:

- *Perturbative.* The microdata set is distorted before publication. In this way, unique combinations of scores in the original data set may disappear and new unique combinations may appear in the perturbed data set. Such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed data set do not significantly differ from the statistics that would be obtained on the original data set.

- *Nonperturbative.* Nonperturbative methods do not alter data; rather, they produce partial suppressions or reductions of detail on the original data set.

From the point of view of the data to which they are applied, a second classification of masking methods can be established:

- *Continuous.* This class of methods is suitable for masking data corresponding to continuous variables. A variable is considered to be *continuous* if it is numerical and arithmetic operations can be performed with it. Examples are income and age. Note that a numerical variable does not necessarily have an infinite range: for example, the range of age is finite.

- *Categorical.* This class of methods is suitable for masking data corresponding to categorical variables. A variable is considered to be *categorical* when it takes values over a finite set and standard arithmetic operations on it do not make sense. Examples are day of the week and eye color.

Our research in SDC will focus on continuous methods. Therefore, only continuous methods currently in use will be reviewed. Nonetheless, it should be noted that some of the described methods are suitable for categorical data as well.

3.1.1 Notation

We will next introduce the notation that will use in what follows. We assume that the information in a microdata file can be represented as a two-dimensional table where one dimension is the set of records *i.e.*, elements, individuals, persons) and the other is the set of variables.

The microdata file contains a value for each record-variable pair, so that it can be modeled as a function:

$$\mathrm{V{:}O} \rightarrow D(V_1) \times D(V_2) \times \cdots \times D(V_m)$$

where O denotes the set of records, V_1, $V_2, \cdots, V(m)$ denote the variables and $D(V_i)$ refers to the range of variable V_i.

Without loss of generality, the m-dimensional function V can be assumed to be of the form:

$$\mathbf{V}(O) = (V_1(O), V_2(O), \cdots, V_m(O))$$

where $V_i(\cdot) : O \rightarrow D(V_i)$ is a one-dimensional function assigning a value for variable V_i to a given record.

Using the above notation, we can restate the purpose of SDC in a more formal manner. The goal of SDC is to supply the user with a masked microdata file V' similar to the original V (*i.e.*, with low information loss) in such a way that:

1. Disclosure risk (*i.e.*, risk of identification of an individual) is low.
2. User analyses (regressions, means, etc) on V' and on V yield the same or at least similar results.

In the following two subsections, masking methods that can be used to produce V' from V will be discussed. Both perturbative and nonperturbative methods will be described.

3.1.2 Perturbative methods

Perturbative methods allow the entire microdata set to be released, although in perturbed form. That is to say, original values are distorted so that the distorted values are publishable.

Most perturbative methods reviewed below (including additive noise, data swapping, microaggregation) are special cases of matrix masking. If the original microdata set is V, then the masked microdata set V' is computed as

$$V' = AVB + C$$

where A is a record-transforming mask, B is a variable-transforming mask, and C is a displacing mask noise [25].

Additive noise

Additive noise [25, 44, 47, 65, 6, 66] consists of adding random noise with the same correlation structure as the original unmasked data. It is currently the only method that can preserve correlations. Let $v_{ij} = V_i(o_j)$ be the unmasked value of variable V_i for individual o_j. Let $e_{ij} = E_i(o_j)$ be the noise added to v_{ij}, and let $v'_{ij} = v_{ij} + e_{ij}$. Further, let $V = V_{ij}$ be the matrix having v_{ij} as elements, and similarly $E = e_{ij}$ and $V' = v'_{ij}$. It is assumed that the expected value of the noise is $E(E) = 0$ and its variance is $Var(E) = cVar(V)$ for some constant c. The variance of unmasked variables can be recovered as $Var(V')/(1 + c)$. White (*i.e.*, Gaussian) noise is most frequently used, even though it may be subject to the *bias problem* (Matloff 1986): if V_i is a continuous positive variable with a strictly decreasing density function (*i.g.*, the exponential density) to which a perturbation that is symmetrical around

0 has been added (*e.g.*, Gaussian noise), Matloff shows that

$$E(V_i|V_i' = w) < w$$

where V_i' is the perturbed version of V_i. Thus, the constant c is the only parameter that can be tuned. It alters (increases or decreases) the random noise being inoculated. The nature of additive noise makes it unsuitable for categorical data. But it is well suited for continuous data, for the following reasons:

- It makes no assumptions about the range of possible values for V_i (which may be infinite).
- The noise being added is typically continuous and with mean zero, which suits continuous original data well.
- No exact matching is possible with external files. Depending on the amount of noise added, approximate (interval) matching might be possible.

Microaggregation

Records are clustered into small aggregates or groups of size at least k. Rather than publishing an original variable V_i for a given record, the average of the values of V_i over the group to which the record belongs is published. The rationale behind microaggregation is that confidentiality rules in use permit publication of microdata sets if the records correspond to groups of k or more individuals, where no individual dominates (i.e., contributes too much to) the group and k is a threshold value. To minimize information loss, groups should be as homogeneous as possible.

Classical microaggregation according to [15] required that all groups except perhaps one be of size k; allowing groups to be of size $\geq k$ depending on the structure of data can be termed *data-oriented microaggregation* [22, 21]. Figure 3.1 illustrates the advantages of variable-sized groups.

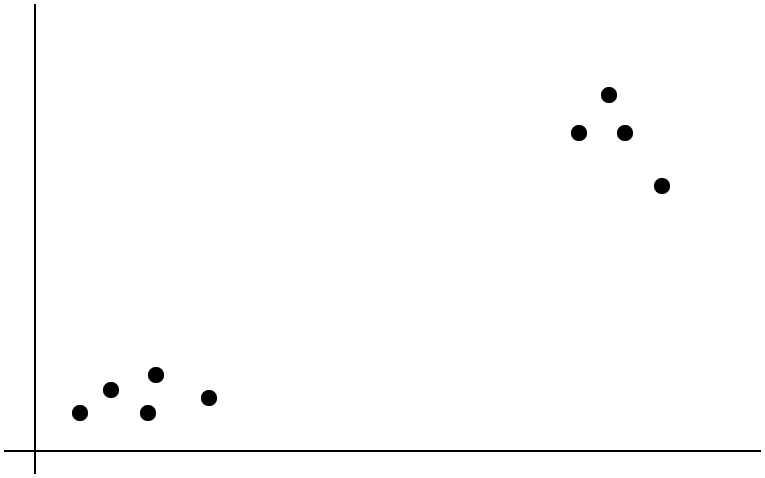


Figure 3.1: Variable-sized groups versus fixed-sized groups

If classical fixed-size microaggregation with $k = 3$ is used, we obtain a partition of the data into three groups, which looks rather unnatural for the data distribution given. If, in contrast, variable-sized groups are allowed, then the five data points on the left can be kept in a single group and the four on the right in another group; such a variable-size grouping yields more homogeneous groups, which implies lower information loss.

A number of heuristic methods have been proposed for solving the microaggregation problem. They can be divided into univariate and multivariate:

- Univariate methods deal with multivariate data sets by microaggregating one variable at a time - that is, variables are sequentially and independently microaggregated. This approach is known as individual ranking or blurring [15] and, while it causes low information loss, it can lead to rather high disclosure risks, as we showed in [2].

- Multivariate methods may use two approaches:

- Rank multivariate data by projecting them onto a single axis – for example, using the first principal component or the sum of z-scores– and then take all groups (except perhaps one) of the same size [1], [15], [14].
- Deal directly with unprojected data [22]; [49]. When working with unprojected data, one can microaggregate all variables of the data set at a time, or independently microaggregate groups of two variables at a time, three variables at a time, and so on. This yields groups of variable size $\geq k$. Heuristics to microaggregate unprojected data draw on search techniques including hierarchical clustering, genetic algorithms and tabu search.

Regarding the existence of exact algorithms which could solve the microaggregation problem in polynomial time, the situation prior to our contribution was as follows:

- A polynomial exact algorithm had been proposed only for the 1-dimensional case, *i.e.* microaggregation of univariate data [35]. This proposal is very recent and cannot be extended to higher dimensional data.

- There was no assessment of the complexity of microaggregation of multivariate data. In particular, there was no NP-hardness result which would justify the use of the aforementioned heuristic algorithms. The only available results in the literature concern the complexity of some clustering problems (see for example [7]) which are similar to microaggregation in the sense that clustering also deals with the optimal partition of elements into homogeneous groups. However, in the case of clustering, the number of groups is fixed, whereas for microaggregation it is not; besides, in clustering problems there are no constraints on the cardinality of groups, whereas in microaggregation problems groups must be of cardinality greater or equal than k.

Data distortion by probability distribution

This method, also called probability distortion [46], is suitable for both continuous and categorical variables. Three steps are needed to compute the distorted version of a confidential original data set:

1. Identification of the underlying density function of each of the confidential variables in the data set and estimation of the parameters associated with the density function.

2. Generation of a distorted series for each confidential variable from the estimated density function.

3. Mapping and replacement of the confidential series by the distorted series.

In the identification and estimation stage, the original series of the confidential variables (*e.g.* salary) is screened to determine which of a set of predetermined density functions fits the data best. Goodness of fit can be tested with the Kolmogorov-Smirnov test. An example set of predetermined density functions could include Poisson, exponential, normal, gamma, Weibull, log-normal, uniform, triangular, chi-square. If several density functions are acceptable at a given significance level, selecting the one yielding the smallest value for the Kolmogorov-Smirnov statistics is recommended. If no density in the predetermined set fits the data, the frequency-imposed distortion method can be used. With the latter method, the original series is divided into several intervals (somewhere between 8 and 20). The frequencies within the interval are counted for the original series and become a guideline to generate the distorted series. By using a uniform random number generating subroutine, a distorted series is generated until its frequencies become the same as the frequencies of the original series. If the frequencies in some intervals overflow, they are simply discarded.

Once the best-fit density function has been selected, the generation stage feeds the estimated distribution parameters to a random value-generating routine to produce the distorted series.

The final stage, mapping and replacement, is needed only if the distorted variables are to be used jointly with other nondistorted variables. Mapping consists of ranking the distorted series and the original series in the same order and replacing each element of the original series with the corresponding distorted element.

It must be stressed here that the approach described in [46] is for one variable at a time. One could imagine a generalization of the method using multivariate density functions. However, this is not a trivial undertaking. It requires multivariate ranking/mapping and can lead to very poor fits.

Resampling

Originally proposed for protecting tabular data [21, 36], resampling can also be used for microdata. Let V be an original variable in a data set with n records. Take with replacement t independent samples $X_1, \cdots, X_t$ of size n of the values of V. Independently rank each sample (using the same ranking criterion for all samples). Finally, for $j = 1$ to n, compute the j-th value v'_j of the masked variable V' as the average of the j-th ranked values in $X_1, \cdots, X_t$.

Lossy compression

The idea is to regard a numerical microdata file as an image (with records being rows, variables being columns, and values being pixels). Lossy compression –for example, JPEG (Joint Photographic Experts Group 2001)– is then used on the image, and the compressed image is interpreted as a masked microdata file. Depending on the lossy compression algorithm used, appropriate mappings between variable ranges and color scales will be needed.

Camouflage

Vector camouflage [33] is a method for giving unlimited, correct numerical responses to *ed hoc* queries to a database while not compromising confidential numerical data. No probabilistic assumptions are made, and optimization

1	K	3.7	4.4	10	D	8.0	3.4
2	L	3.8	3.4	2	L	4.5	3.2
3	N	3.0	4.8	1	H	3.0	4.8
4	M	4.5	5.0	3	M	3.7	4.4
5	L	5.0	6.0	4	N	5.0	6.0
6	H	6.0	7.5	5	L	4.5	5.0
7	H	4.5	10.0	6	F	6.7	9.5
8	F	6.7	11.0	9	C	10	7.5
9	D	8.0	9.5	7	K	3.8	11.0
10	C	10.0	3.2	8	H	6.0	10.0

Table 3.1: Example of rank swapping. Left, original file; right, rankswapped file

techniques are used to camouflage the sensitive record (exact answer) in an infinite set of records; thus providing an interval answer. The information loss is the transformation of a point answer into an interval answer. Because of its nature, this method is suitable only for continuous data.

Rank swapping

Although originally described only for ordinal variables [53], this method can be used for any numerical variable. First, values of variable V_i are ranked in ascending order. Then, each ranked value of V_i is swapped with another ranked value randomly chosen within a restricted range (*e.g.*, the rank of two swapped values cannot differ by more then p percent of the total number of records).

Example 1 *In Table 3.1, we can see an original microdata file on the left and its rankswapped version on the right. There are four variables and ten records in the original file; the second variable is alphanumeric, and the standard alphabetic order has been used to rank it. A value of* $p = 10\%$ *has been used for all variables.*

Rounding

Rounding methods replace original values of variables with rounded values. For a given variable V_i, rounded values are chosen among a set of rounding points defining a *rounding set*. In a multivariate original data set, rounding is usually performed one variable at a time (univariate rounding); however, multivariate rounding is also possible. The operating principle of rounding makes it suitable only for continuous data.

Example 2 *Assume a continuous variable V. Then we have to determine a set of rounding points $\{p_1, \cdots, p_r\}$. One possibility is to take rounding points as multiples of a base value b, that is $p_i = b * i$ for $i = 1, \cdots, r$. The set of attraction for each rounding point p_i is defined as the interval $[p_i - b/2, p_i + b/2]$, for $i = 2, \cdots, r-1$; for p_1 and p_r, respectively, the sets of attraction are $[0, p_1 + b/2]$ and $[p_r - b/2, V_{max}]$, where V_{max} is the largest possible value for variable V. Now an original value v of V is replaced with the rounding point corresponding to the set of attraction where v lies.*

3.1.3 Nonperturbative methods

Nonperturbative methods do not rely on distortion of the original data but on partial suppression or reductions of detail.

Global recoding

This method was designed first for categorical variables. For a categorical variable V_i, several categories are combined to form new (normally less specific) categories, thus resulting in a new V_i' which is a discretized version of V_i. In other words, a potentially infinite range $D(V_i)$ is mapped onto a finite range $D(V_i')$. This is a technique used in the μ-Argus SDC package [37, 38, 39].

This technique is more appropriate for categorical microdata where it helps disguising records with strange combinations of categorical variables. Global recording is used heavily by statistical offices.

Example 3 *If there is a record with 'Marital status = Widow/er' and 'Age = 17', global recording could be applied to 'Marital status' to create a broader category 'Widower or divorced', so that the probability of the above record being unique would diminish.*

Global recoding can also be used on a continuous variable, but the inherent discretization very often leads to an unaffordable loss of information. Also, arithmetical operations that were straightforward on the original V_i are no longer easy or intuitive on the discretized V_i'.

Top- and bottom-coding

Top- and bottom-codding is a special case of global recoding for use on variables that can be ranked, that is, continuous or categorical ordinal. The idea is that top values (those above a certain threshold) are lumped together to form a new category. The same is done for bottom values (those below a certain threshold). This method is implemented in the μ-Argus software [38, 39].

3.1.4 Existing performance measures for microdata

As we noted above, information loss and disclosure risk measures define the quality of the masking methods.

Information loss measures quantify the loss of informational content from the point of view of the legitimate data user whose aim is to perform some particular research on the data. Therefore, information loss clearly depends on the type of research or data use the user is interested in. Strictly speaking, statistical offices releasing protected data should determine information loss depending on the particular use for which the protected data set has been designed.

Information loss measures

If, for example, what is critical for the user is that variances of the original data set are be preserved in the protected data set, then the impact of SDC procedures may be measured as variance alteration. In particular, let $data_{Old}$ represent the original data and $data_{New}$ the protected data and suppose $\hat{\theta}(data_{Old})$ is a consistent estimator of θ if the original data were available. Suppose that $\hat{\theta}(data_{New})$ is the corresponding estimator of θ based on the protected data. The variance impact of the SDC technique may be measured by the relative variance

$$Var(\hat{\theta}(data_{New})/Var(\hat{\theta}(data_{Old})$$

If what is to be preserved in the protected data are the means of the original data set, information loss may be measured as a bias impact of SDC procedures. The bias impact may be defined as the ratio between the bias of $\hat{\theta}(data_{New})$ and the bias of $\hat{\theta}(data_{Old})$. A common assumption would be that the bias of the latter estimator may be treated as negligible, so that the bias impact may be summarized by

$$Bias(\hat{\theta}(data_{New}) = E[\hat{\theta}(data_{New})] - \theta$$

It can be seen that the information loss measures just discussed are not generic, but strongly oriented for a particular data use.

However, when a statistical office has to publish statistical data for the public at large, the potential data uses are very diverse and it may be even hard to identify them all at the moment of data release. Therefore, generic information loss measures are desirable. The amount of information loss measured in this way should roughly correspond to the amount of information loss for a reasonable range of data uses. At the same time these generic information measures would allow performance comparisons of different protection techniques.

Shannon's entropy is a general information measure. In [69], the use of

entropy in SDC is discussed in detail. We summarize it here. The basic idea of this approach can be best described by drawing an analogy between a message sent across a noisy channel and a data set modified by SDC methods. In both cases, the receiver of the perturbed message may try to reconstruct the original message. Using (supposed) knowledge of the noise process, the receiver can try to reconstruct a probability distribution over all original messages, and pick the one that is most likely. The entropy of this probability distribution over the original messages will be a measure of the information loss due to the transmission process.

However, it is important to note here that not even the authors of [69] are totally sure that entropy quantifies the information loss perceived by a user of the protected data. We agree that the entropy is not a good measure of information loss for microdata. The reason is that the user of the published microdata is generally interested in preserving the analytical validity of the data, and for this preservation information loss is best characterized by the impact on means and covariance matrix. The problem with entropy is that, although it measures the uncertainty about the original data after masked data are published, it does not reflect impacts on covariances and means so well.

Disclosure risk measures

The situation for disclosure risk is similar to what has been said for information loss, in that generic disclosure risk measures cannot be found in the literature. Previous work basically relates to disclosure risk for nonperturbative methods, in particular sampling methods in which a sample of the original data set is published. Disclosure risk is measured here as the probability that a sample unique is a population unique [28, 63]. Unless the sample size is much smaller than the population size, such a probability can be dangerously high; in that case, an intruder who locates a unique value in the released sample can be almost certain that there is a single individual in the population with that value, which is very likely to lead to that

individual's identification.

However, sampling cannot be used for continuous data, because sampling a continuous variable unperturbed for all records in the sample is most likely going to result in disclosure. The explanation is that, if variable V_i is present in an external administrative public file, unique matches with values of the corresponding published variable V_i' are very likely, because for a continuous variable (even one truncated due to digital representation) it is very unlikely that V_i takes the same value for two different records, *i.e.*, $V_i(o_1) = V_i(o_2)$ if $o_1 \neq o_2$.

If, for a continuous identifying variable, the score of a respondent is only approximately known by an attacker (as assumed in [68]), it might still make sense to use sampling to protect that variable. However, assumptions about limitations in attackers' knowledge are perilous and may prove too optimistic if good quality external administrative files are at hand.

The uniqueness property as stated above is not relevant for perturbative methods, because the released data set is a distorted version of the original one. In summary, there is little in the previous literature on disclosure risk for microdata that can be used for a broad class of perturbative methods disclosure risk assessment tends to be method-specific.

3.2 Tabular data protection

Protection of tabular data consists of the protection of aggregate data computed from survey or census responses.

Tables are central in official statistics: many survey and census data are categorical in nature, so that their representation as cross-classifications or tables is a natural reporting strategy. Tabular data being thus aggregate data, one might be tempted to think they are not supposed to contain information that can reveal the contribution of particular respondents. However, as noted in [32], in many cases table cells do contain information on a single or very few respondents, which implies a disclosure risk for the data of those

respondents. In these cases, disclosure control methods must be applied to the tables prior to their release.

Before starting the discussion of SDC methods for tables and their performance measures, we provide some important basic concepts and definitions related to tabular data. Disclosure risk and protection methods will be discussed in later sections.

3.2.1 Data formats

There are two data formats commonly used to display aggregate data: tables of frequency counts and tables of aggregate magnitude data.

Tables of frequency counts (frequency tables for short) count the number of respondents with specified characteristics. For example, a two-dimensional frequency count table may have rows corresponding to race categories and columns corresponding to age groupings. An individual table cell at the intersection of a given row and a given column would indicate the number of residents of a certain geographic area with that race and age combination.

Tables of aggregate magnitude data (magnitude tables for short) are analogous to frequency count tables in that they are defined by cross-classification of categorical variables. However, the cells contain aggregate values, over the corresponding respondents, of some quantity of interest. For example, a two-dimensional table on income defined by race and age would contain total incomes in each race-by-age cell.

3.2.2 Classification of tabular data

According to [69], tabular data (whether frequency tables or magnitude tables) can be of the following types:

- *Single tables*, obtained from microdata by aggregation. For such an aggregation it is necessary to define the spanning variables, that is, categorical variables, which are used to cross-classify the members of a population or sample (these variables define the cells in the table). Each

cell is uniquely determined by a combination of scores on the spanning variables. The content of each cell depends on the other ingredient defining the table, namely the response variable. The response variable is a numerical variable. The number of spanning variables used in the definition of the table is called the dimension of the table. Within each cell the scores on the response variable for the records that belong to that cell are added to yield the cell value or cell total.

- *Marginal tables*, obtained from a single table by aggregating over one or more spanning variables. The response variable of a marginal table derived from single table T is inherited from T. These tables are usually published along with single tables. This type of tables is considered because the marginals imply linear constraints, so they introduce dependencies in the data, a fact which must be taken into account when applying protection techniques.

- *Linked tables* is a set of tables that have been produced from a common microdata file, or alternatively, as marginals tables of a single super-table (a super-table means a many-dimensional table). The problem with linked tables is that they cannot be protected individually. Rather, their proper protection requires that they should be considered simultaneously. Also it may be the case when the released tables were produced not as one batch, but in the form of several batches of such tables at different points in time. So it is necessary to consider all the previously released batches when a new one is going to be released.

A number of disclosure control methods to protect tabular data have been proposed (see [69, 24] for a survey). As in the case of microdata, SDC methods for tabular data can be classified as perturbative and nonperturbative. The definition of these categories is the same with the only difference that table cells instead of microdata records are considered here. In the next subsections we will review SDC methods for tabular data proposed so far.

3.2.3 Perturbative methods for tabular data

Rounding

Rounding is similarly defined as in the case of microdata. So, all table cells are rounded to an integer multiple of a fixed rounding base b. Controlled rounding is a variant of rounding in which table additivity is preserved (*i.e.* rounded rows and columns still sum to their rounded marginals).

Swapping and simulation

In data swapping, units are swapped so that the table resulting from the swapped data set still preserves all k-dimensional margins of the original table. A more elaborate version of swapping was proposed in [31], whereby the original table is replaced by a random draw from the exact distribution under the log-linear model whose minimal sufficient statistics correspond to the margins of the original table. Further extensions of this idea would lead to drawing a synthetic table from the full distribution of all possible tables with the same margins of the original table.

3.2.4 Nonperturbative methods for tabular data

Cell suppression

If a table cell is deemed sensitive, then it is suppressed from the released table (primary suppression). If marginal totals or other linked tables are also to be published, then it may be necessary to remove additional table cells (secondary suppressions) to prevent primary suppressions from being computable. Secondary suppressions should be chosen in a way such that the utility of the resulting table is maximized.

Table redesign

Categories used to tabulate data are recoded into different (often more general) categories so that the resulting tabulation does not contain sensitive cells any more. A simple redesign could be to combine two rows containing sensitive cells to obtain a new row without sensitive cells.

Sampling

A table is released which is based on a sample of the units on which the original table was built.

3.2.5 Existing performance measure for tabular data

Information loss measures

Two main approaches for measuring information loss for tabular data are discussed in the literature:

- The first approach is based on weights associated with the various cells. This approach may be applied to cell suppression, rounding and table redesign. For a suppression pattern in a table, the information loss is measured by summing the weights corresponding to secondary suppressions. Depending on how weights are assigned, this can imply that information loss is the total value of secondary suppressions or the number of secondary suppressions:
  - If the weight associated to each candidate cell for secondary suppression (candidate cells are all cells remaining after suppression of sensitive cells) is proportional to the cell value, then information loss caused by secondary cell suppression is defined as the sum of the values of secondarily suppressed cells.
  - If the weight associated to each candidate cell for secondary suppression is constant (say 1), then information loss caused by secondary cell suppression is defined as the number of secondarily

suppressed cells.

So, information loss measured by this way is strictly dependent on the envisaged use of a table.

- The second approach uses entropy considerations. Entropy-based measures for information loss were discussed in [69]. Even though we will illustrate this approach in the case of cell suppression, similar considerations can be applied for other methods. Assume that, after applying suppression, we obtain a table with the suppressed cells. If we substitute variables for the suppressed cells, we can derive a set of linear equations. These equations result from the additivity in the table and its marginals. Suppose also that the table is nonnegative. Every combination of values for these variables that satisfies these constraints is a feasible solution for the suppressed values in the original table. The set of all such solutions is the solution set for this problem, which is a polyhedron. (Also in case of table redesign and controlled rounding we find solution sets of the same type, which give the feasible values for the modified cells of the original table). So, in all these cases we can calculate the feasibility intervals for each variable by solving two LP problems, namely one to determine the lower bound of this interval and one to determine the upper bound. After that, the appropriate distribution of probabilities for the members of a solution set must be chosen, which may be too difficult in some cases. Sometimes, the uniform distribution is used, which causes Shannon's entropy to be equal to the logarithm of the number of all possible values in the feasibility intervals (table cells are assumed to take values in a discrete domain: either integer values or real values with a fixed number of decimal positions).

In spite of their discussing it, the authors of [69] are not sure that the entropy-based approach is very useful to measure information loss (as in case of entropy-based measures for microdata). In practice, the approach based

on weights is the one being currently used.

Disclosure risk measures

The situation for disclosure risk for tabular data is quite different. There is a number of sensitivity rules which are used to decide whether a particular table cell can safely be released. These rules operate on an *a priori* basis: the original data are examined *before* they are protected and the rules are used to determine whether data can be released as they stand or should rather be protected.

For magnitude tables (normally related to economic data), there are two widely accepted sensitivity rules:

(n, k)-dominance In this rule, n and k are two parameters with values to be specified. A cell is called sensitive if the sum of the contributions of n or fewer respondents represents a fraction k or more of the total cell value.

pq-rule The prior-posterior rule is another rule gaining increasing acceptance. It also has two parameters p and q. It is assumed that, prior to table publication, each respondent can estimate the contribution of each other respondent to within less than q percent. A cell is considered sensitive if, posterior to the publication of the table, someone can estimate the contribution of an individual respondent to within p percent. A special case is the p%-rule: in this case, no knowledge prior to table publication is assumed, *i.e.* the pq-rule is used with $q = 100$.

For tables of counts or frequencies (normally related to demographic data), a so-called **threshold rule** is used. A cell is defined to be sensitive if the number of respondents is less than a threshold k.

The problem with these rules is that it is not 100% clear that the risk of disclosure for the contribution to the cell is always reflected in a proper way, or in other words, that inferences made from a cell declared nonsensitive will

always be less precise than inferences made from a cell declared sensitive. Chapter 6 is devoted to the discussion of this problem.

Another point we would like to raise here is that the above-mentioned rules are mainly oriented toward certain methods, in particular cell suppression (a sensitivity rule is used to decide which cells should be primarily suppressed) or table redesign. Their usefulness for other methods is less clear. This is confirmed by the survey [30]: sensitivity rules are nearly always used in conjunction with cell suppression.

Chapter 4

Performance measures for microdata protection

As stated in Chapter (2), our research line in microdata protection has been focused on performance measures of SDC microdata masking methods.

It was pointed out in Chapter (3) that, since statistical offices have to publish statistical data for general use, generic performance measures are desirable. In fact, the lack of such generic measures is the main reason of difficulties for the data protector to compare the performance of different methods and choose the most appropriate one.

In this chapter, generic performance measures are proposed for microdata and a quantitative comparison of disclosure control methods based on these measures is made.

4.1 Information loss measures

The approach used here to derive generic information loss measures is based on assessing how different the masked data set is from the original data set. We will say there is little information loss if the analytical structure of the masked data set is very similar to the structure of the original data set. In fact, the motivation for preserving the structure of the data set is to ensure

that the masked data set will be analytically valid and interesting.

According to [70] we have that:

- A microdata set is *analytically valid* if the following is approximately preserved (some conditions apply only for continuous variables):
  1. Means and covariances on a small set of subdomains
  2. A few margins
  3. At least one distributional characteristic
- A microdata file is *analytically interesting* if six variables on important subdomains are provided that can be validly analyzed.

More precise conditions of analytical validity and analytical interest cannot be stated without taking specific data uses into account. As imprecise as it may be, the above definition of analytical validity does shed some light on what preserving the data set structure means. We can actually try several complementary ways to assess the preservation of the structure of the original data set:

- Compare the data in the original and the masked data sets. The more similar the SDC method to the identity function, the less impact (but the higher the disclosure risk!).
- Compare some statistics computed on the original and the masked data sets. Little information loss should translate to little differences between statistics.
- Analyze the behavior of the particular SDC method used, in order to measure its impact on the structure of the original data set.

Now assume a microdata set with n individuals (records) $I_1, I_2, \cdots, I_n$ and p continuous variables $Z_1, Z_2, \cdots, Z_p$. Let X be the matrix representing the original microdata set (rows are records and columns are variables). Let X' be the matrix representing the masked microdata set. The following tools are useful to characterize the information contained in the data set:

- Covariance matrices V (on X) and V' (on X'). These can be taken as lower-triangular.
- Correlation matrices R and R' (also lower-triangular).
- Correlation matrices RF and RF' between the p variables and the p factors $PC_1, \cdots, PC_p$ obtained through principal components analysis.
- Commonality between each of the p variables and the first principal component PC_1 (or other PC_i's). Commonality is the percent of each variable that is explained by PC_1 (or PC_i). Let C be the vector of commonalities for X and C' the corresponding vector for X'.
- Factor score coefficient matrices F and F'. Matrix F contains the factors that should multiply each variable in X to obtain its projection on each principal component. F' is the corresponding matrix for X'.

There does not seem to be a single quantitative measure which completely reflects those structural differences. Therefore, we propose to measure information loss through the discrepancies between matrices X, V, R, RF, C and F obtained on the original data and the corresponding X', V', R', RF', C' and F' obtained on the masked data set.

Matrix discrepancy can be measured in at least three ways:

Mean square error Sum of squared componentwise differences between pairs of matrices, divided by the number of cells in either matrix.

Mean absolute error Sum of absolute componentwise differences between pairs of matrices, divided by the number of cells in either matrix.

Mean variation Sum of absolute percent variation of components in the matrix computed on masked data with respect to components in the matrix computed on original data, divided by the number of cells in either matrix. This approach has the advantage of not being affected by scale changes of variables.

	Mean square error	Mean abs. error	Mean variation
$X - X'$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
$V - V'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
$R - R'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$
$RF - RF'$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$
$C - C'$	$\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$	$\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$	$\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{\lvert c_i\rvert}}{p}$
$F - F'$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$

Table 4.1: Information loss measures for continuous microdata

Table 4.1 summarizes the measures proposed. In this table, p is the number of variables, n the number of records, and components of matrices are represented by the corresponding lowercase letters (*e.g.* x_{ij} is a component of matrix X). Regarding $X - X'$ measures, it makes also sense to compute those on the averages of variables rather than on all data (call this variant $\bar{X} - \bar{X}'$. For $V - V'$ measures, it would also be sensible to use them to compare only the variances of the variables, *i.e.* to compare the diagonals of the covariance matrices rather than the whole matrices (call this variant $S - S'$).

The set of measures in Table 4.1 may be used when it is necessary to assess to what extent a method damages the informational content of the data being protected. However, the assessment of the quality of an SDC method cannot be limited to information loss; disclosure risk is another magnitude that should be measured. Definitely, both factors should be taken into account when making decisions on the suitability of a particular method for a certain data set. In the next section, the information loss measures defined above are combined with an empirical assessment of disclosure risk to compare a set of SDC methods.

4.2 Disclosure risk measures for microdata

While information loss can be quantified using analytical measures (either generic or data use specific), it is far more difficult to assess disclosure risk in a way that is both analytical and applicable to all methods. For this reason, the approach we adopt here for disclosure risk evaluation is an empirical one, based on re-identification experiments carried out using record-linkage algorithms.

We briefly describe below two approaches to record linkage which yield empirical disclosure risk measures. An analytical measure based on interval disclosure is also described.

4.2.1 Distance-based record linkage

This approach to record linkage is described in [55] for the specific case of microaggregation masking and using the Euclidean distance. However, it can be generalized for any perturbative method provided that a distance between the original and the masked value can be defined. As in any record linkage context, it is assumed that an intruder has an external data set containing as key variables the same variables present in the released masked data set. The intruder is assumed to try to link the masked data set with the external data set.

Linkage then proceeds by computing the distances between records in the original and the masked data sets. The distances used are standardized to avoid scaling problems. For each record in the masked data set, the distance to every record in the original data set is computed. Then the "nearest" and "second nearest" records in the original data set are considered. A record in the masked data set is labeled as "linked" when the nearest record in the original data set has the same record number is the corresponding original record). A record in the masked data set is labeled as "linked to 2nd nearest" when the second nearest record in the original data set has the same record number. In all other cases, a record in the masked data set is labeled as "not

linked". The percent of "linked" and "linked to 2nd nearest" is a measure of disclosure risk. The Euclidean method above requires rescaling variables as well as an assumption on the weight of variables when computing a distance: for instance, in the proposal of [55], all variables have the same weight.

4.2.2 Probabilistic record linkage

In [41], a probabilistic record linkage method was described and illustrated on the 1985 Census of Tampa, Florida. The matching algorithm uses the linear sum assignment model to "pair" records in the two files to be matched (the original file and the masked file in our case). The percent of correctly paired records is a measure of disclosure risk.

Although less simple than the Euclidean method described in the previous section, this approach is attractive because it does not assume rescaling or weighting variables and requires the user only to provide two probabilities as input: one is an upper bound of the probability of a false match, and the other an upper bound of the probability of false non-match. This method is described in the appendix of this thesis.

We have used the U.S. Census Bureau implementation of probabilistic record linkage provided by William Winkler ([67, 70]) in our experimentation.

4.2.3 Interval disclosure

For a record in the masked data set, compute rank intervals as follows: each variable is independently ranked and a rank interval is defined around the value the variable takes on each record; the ranks of values within the interval for a variable around record r should differ less than $p\%$ of the total number of records and the rank in the center of the interval should correspond to the value of the variable in record r. Then the proportion of original values which fall into the interval centered around their corresponding masked value is a measure of disclosure risk. A 100% proportion means that an attacker is completely sure that the original value lies in the interval around the masked

value (interval disclosure).

4.3 Comparison of SDC methods for continuous microdata

The methodology we use to compare SDC methods is as follows:

- **Identification of relevant masking methods**. A set of widely used SDC masking methods for microdata protection were identified and implemented.

- **Test data collection**. We obtained test data from publicly available microdata files. This guarantees public-domain reproducibility of the experiments reported here. The price paid is that the data we start from are not original, but have already undergone some amount of disclosure protection.

- **Information loss and disclosure risk assessment**. These measures were introduced above.

- **Empirical work**. Experiments are conducted in view of obtaining t-uples of the form $(method, parms, risk, loss)$, where $parms$ are the input parameters to method, $risk$ is the percent of re-identified records in the test data set and $loss$ is the information loss. Given $risk$, it is possible to find $method$ and $parms$ such that $risk(loss, method(parms))$ is minimal (at least over the set of available t-uples). Also, given $loss$, it should be possible to find $method$ and $parms$ such that $risk(loss, method(parms))$ is minimal (at least over the set of available t-uples).

The details of the steps of this methodology are described in the following sections.

4.3.1 Identification of relevant microdata masking methods

As it was stated in the Chapter (3) SDC methods can be classified as perturbative and nonperturbative. Perturbative methods distort records before release, which allows release of the whole population microdata set. Nonperturbative methods do not alter data but partially suppress or reduce the detail of the original data set. We consider only perturbative methods for continuous data because reidentification is very easy for unperturbed continuous variables. The reason is that, if a continuous variable V is left unperturbed in the masked file and is present in an external administrative public file, unique matches are very likely, because for a continuous variable (even one truncated due to digital representation) V_i is not likely to take the same value for two different records (*i.e.*, it is unlikely that $V_i(o_1) = V_i(o_2)$ if $o_1 \neq o_2$. Thus, distortion turns out to be the only effective way to protect continuous microdata.

The subset of perturbative methods considered is:

- *Additive noise* (`Noise`p for short, where p is a parameter). Gaussian noise is added to the original data to get the masked data. If the standard deviation of the original variable is s, noise is generated using a $N(0, ps)$. Values of p considered in the experiments below are 0.01, 0.02, 0.04, 0.06, 0.08 up to 0.2 with 0.02 increments.

- *Data distortion by probability distribution* (abbreviated `Distr` [46]). For each variable in the original microdata set, the best fitting distribution is found; then the fitted distribution is used to generate the masked dataset. There are no parameters. Crystal Ball software [12] is used to find the best fitting distribution.

- *Resampling.* If n is the number of records in the dataset, take with replacement t independent samples $X_1, \cdots X_t$, of size n of the values of an original variable V_i. Independently rank each sample (using the

same ranking criterion for all samples). Build the masked variable V_i' by taking as first value the average of the first values of the samples, as second value the average of the second values, and so on. Resampling is tested for $t = 1$ (`Resamp`1 for short) and $t = 3$ (`Resamp`3 for short).

- *Microaggregation.* Records are clustered into small aggregates or groups of size at least k. Rather than publishing a variable for a given individual, the average of the values of the variable over the group to which the individual belongs is published. Variants of microaggregation considered include individual ranking (abbr. `MicIR`k), microaggregation on projected data using z-scores projection (abbr. `MicZ`k) and principal components projection (abbr. `MicPCP`k), and microaggregation on unprojected multivariate data considering two variables at a time (abbr. `Mic2mul`k), three variables at a time (abbr. `Mic3mul`k), four variables at a time (abbr. `Mic4mul`k), or all variables at a time (abbr. `Micmul`k). Values of k between 3 and 10 have been considered.

- *Lossy compression* (abbr. `JPEG`q, where q is a parameter). This method is new and we propose it for continuous data. The idea of this method is to regard a numerical microdata file as an image (with rows being records and columns being variables). Lossy compression, and more specifically the JPEG algorithm (Joint Photo graphic Experts Group 2001), is then used on the image, and the compressed image is interpreted as a masked microdata file. Depending on the lossy compression algorithm used, appropriate mappings between variable ranges and color scales will be needed. The JPEG quality q is taken as a parameter with values from 5 percent up to 100 percent with 5 percent increments.

- *Rank swapping* (abbr. `Rank`p, where p is a parameter). Although originally described only for ordinal variables, this method can be used for any numerical variable [53]. First, values of variable V_i are ranked in ascending order; then each ranked value of V_i is swapped with another ranked value randomly chosen within a restricted range (*e.g.*, the rank

of two swapped values cannot differ by more than p percent of the total number of records). We consider values of p from 1 to 20.

4.3.2 Test data collection

A microdata set was constructed using the Data Extraction System (DES) of the U.S. Census Bureau (`http://www.census.gov/DES/www/welcome.html`). From the available data sources, the Current Population Survey corresponding to 1995 was chosen; specifically, the file group "March Questionnaire Supplement - Person Data Files". Variables and records were selected as follows:

- *Variable selection* Our continuous variable selection was based on the requirement that the values of each variable span a wide range. Thirteen continuous variables were selected: $AFNLWGT$ (Final weight), AGI (Adjusted gross income), $EMCONTRB$ (Employer contribution for health insurance), $ERNVAL$ (Business or farm net earnings), $FEDTAX$ (Federal income tax liability), $FICA$ (Social security retirement payroll reduction), $INTVAL$ (Amount of interest income), $PEARNVAL$ (Total person earnings), $POTHVAL$ (Total other persons income), $PTOTVAL$ (Total person income), $STATETAX$ (State income tax liability), $TAXINC$ (Taxable income amount), $WSALVAL$ (Amount: Total wage and salary).

- *Record selection.* Our selection of 1080 records was based on the number of repeated values for each variable. This number should be low in principle, as one would not expect repeated values for a continuous variable. However, there were repetitions in the data set.

- The resulting data set had three properties that were important to our work:

1. The number of records was fewer than 1200, which allowed repeated experimentation with the probabilistic record linkage software in reasonable time.
2. Seven variables had no repeated values: $FEDTAX$, $AFNLWGT$, AGI, $EMCONTRB$, $PTOTVAL$, $TAXINC$, $STATETAX$. Because absence of repeated values is a distinguishing feature of really continuous variables, these seven were chosen as key variables for record linkage.
3. 1080 (the number of records) is the largest integer less than 1200 which is a multiple of 5, 8 and 9. Thus, when the microaggregation SDC method is used, the data set can be microaggregated with minimal group sizes $k = 3$, 4, 5, 6, 8, 9 and 10 so that all groups have exactly size k.

We used the resulting data to carry out the empirical work described in Section 4.3.5 below.

4.3.3 Information loss metrics

Let X and X' be the matrices representing the original and the masked data sets, respectively. Let V and R be the covariance matrix and the correlation matrix of X; let $\bar{X}$ be the vector of variable averages for X and let S be the diagonal of V. Define V', R',$\bar{X}'$ and S' analogously from X'.

The Information Loss (IL) is computed by averaging the mean variations of $X - X'$, $\bar{X} - \bar{X}'$, $V - V'$, $S - S'$ and the mean absolute error of $R - R'$; the resulting average is multiplied by 100. According to the formulae given in Section 4.1 we obtain:

$$IL = 100 * \left(\frac{\sum_{j=1}^{p} \sum_{i=1}^{n} \frac{|x_{ij} - x'_{ij}|}{|x_{ij}|}}{np} + \frac{\sum_{j=1}^{p} \frac{|\bar{x_j} - \bar{x'_j}|}{|\bar{x_j}|}}{p} + \frac{\sum_{j=1}^{p} \sum_{1 \le i \le j} \frac{|v_{ij} - v'_{ij}|}{|v_{ij}|}}{\frac{(p+1)p}{2}} + \right.$$
$$\left. + \frac{\sum_{j=1}^{p} \frac{|v_{ij} - v'_{ij}|}{|v_{ij}|}}{p} + \frac{\sum_{j=1}^{p} \sum_{1 \le i \le j} |r_{ij} - r'_{ij}|}{\frac{(p-1)p}{2}} \right) / 5$$

Term 2 is analogous to term 1 but only with respect to the averages of variables. Term 4 is analogous to term 3 but only with respect to variances of variables (not covariances).

4.3.4 Disclosure risk assessment

Distance Linkage Disclosure risk (DLD) is the average percentage of linked records using distance-based record linkage; the average is computed over the number of key variables that the intruder is assumed to know (we have considered knowledge of anywhere from one to seven variables). Similarly, the Probabilistic Linkage Disclosure risk (PLD) is the average percent of correctly paired records using probabilistic linkage. Interval Disclosure risk (ID) is the average percent of original values falling in the intervals around their corresponding masked values (averages have been computed over all parameter values, *i.e.* 1 percent to 10 percent with 1 percent increments).

4.3.5 Empirical work

We chose the following microdata masking methods for comparison: additive noise, data distortion by probability distribution, resampling, microaggregation, lossy compression and rank swapping. These methods are the most currently used by statistical offices.

Table 4.2 contains a ranking of these methods (the parameter values described in Section 3.1 of Chapter 3 were tried for each method). The table contains columns specifying IL, DLD, PLD and ID and also an overall score constructed as follows:

$$Score = 0.5(IL) + 0.125(DLD) + 0.125(PLD) + 0.25(ID)$$

The rationale of the above weighting is to give equal weight to information loss (0.5) and to disclosure risk. The 0.5 weight of disclosure risk is equally divided among ID (0.25) and record linkage. The 0.25 weight of record linkage is equally divided among both approaches to record linkage. The

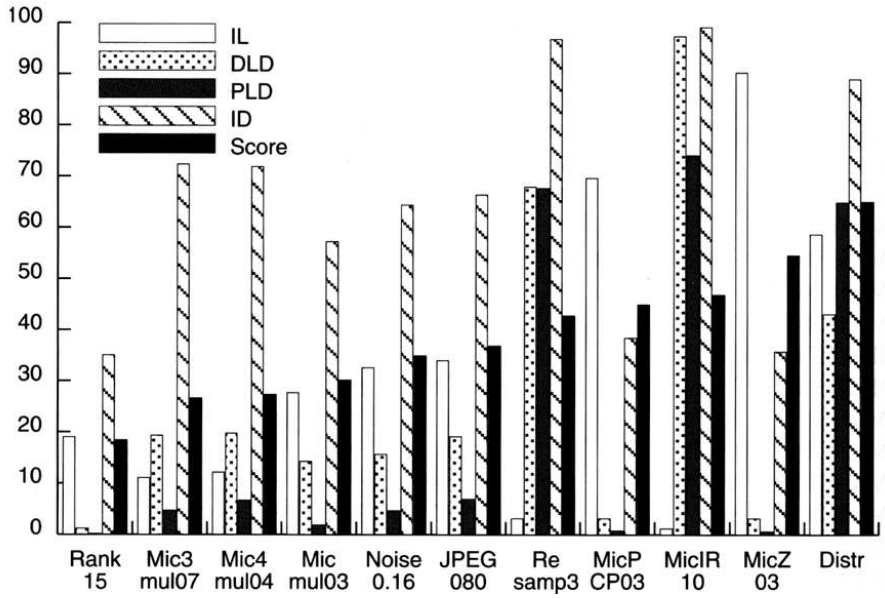


Figure 4.1: Comparison of SDC methods for continuous microdata with best parameter choice

correlation between DLD and PLD is actually 0.962, so both approaches are very similar. The (IL, DLD), (IL, PLD) and (IL, ID) correlations are -0.605, -0.551 and -0.807; thus, the lower the information loss, the higher the disclosure risk, as one would expect. The $IL - Rank$, $DLD - Rank$, $PLD - Rank$ and $ID - Rank$ columns contain the ranking of each method with respect to IL, DLD, PLD and ID; the lower the rank, the better a method performs (*i.e.* lower information loss and disclosure risk).

Figure 4.1 gives a comparison of SDC methods for continuous microdata when the best parameter choice is made. Rank swapping with parameter 15 (`Rank15`) stands out as the best performer. For multivariate microaggregation, taking three variables at a time and groups of size at least 7 (`Mic3mul07`) turns out to be the best parameter choice.

Table 4.2: Comparison results for continuous microdata

Method	IL	DLD	PLD	ID	Score	IL-Rank	DLD-Rank	PLD-Rank	ID-Rank
Rank15	19.01	1.19	0.15	35.05	18.44	53	6	7	21
Rank19	22.95	0.93	0.08	28.04	18.61	59	2	2	2
Rank16	20.91	1.39	0.11	32.18	18.69	56	8	5	16
Rank13	16.77	2.17	0.12	40.35	18.76	48	12	6	28
Rank14	19.72	1.92	0.07	37.00	19.36	55	10	1	25
Rank11	14.32	2.43	0.25	47.81	19.45	44	13	14	39
Rank12	16.37	2.50	0.25	43.73	19.46	47	14	11	35
Rank20	25.81	0.69	0.09	26.83	19.71	64	1	3	1
Rank18	25.74	0.95	0.09	29.25	20.31	63	4	4	6
Rank10	13.37	3.90	0.38	53.17	20.51	41	24	17	45
Rank17	25.12	1.52	0.20	30.95	20.51	61	9	9	10
Rank09	11.66	5.01	0.52	57.58	20.91	38	37	29	49
Rank08	11.60	6.07	0.85	63.37	22.51	37	39	39	56
Rank07	9.25	7.51	1.08	68.71	22.87	30	41	43	63
Rank06	7.87	9.02	2.79	73.80	23.86	26	43	56	71
Mic3mul07	11.06	19.34	4.70	72.34	26.62	36	68	65	69
Rank05	6.78	16.8	13.60	78.89	26.91	22	58	70	77
Mic3mul09	13.46	19.22	3.44	69.91	27.04	42	67	60	65
Mic3mul10	14.84	17.99	3.44	68.61	27.25	46	64	59	62
Mic4mul04	12.14	19.76	6.67	71.85	27.33	39	69	68	68
Mic4mul05	14.50	17.43	5.45	69.09	27.39	45	61	66	64
Mic3mul08	13.51	20.81	4.15	70.68	27.54	43	71	63	66
Mic4mul08	18.89	17.78	3.35	62.84	27.80	52	62	58	55
Mic3mul06	10.24	20.41	13.90	74.00	27.91	33	70	71	72
Mic4mul07	19.36	17.10	2.08	64.41	28.18	54	60	53	58
Mic4mul06	17.91	17.82	3.98	66.41	28.28	50	63	62	60
Mic4mul09	21.35	15.93	2.00	61.66	28.33	58	57	52	54
Mic4mul10	22.98	16.85	2.37	60.56	29.03	60	59	55	51
Mic3mul05	9.73	23.78	18.29	76.59	29.27	31	76	73	74
Mic3mul04	7.45	23.49	22.75	79.14	29.29	24	75	75	79
Mic4mul03	10.69	22.88	16.69	76.89	29.51	35	74	72	75
Rank04	5.90	22.77	22.78	84.12	29.67	20	73	76	86
Micmul03	27.67	14.26	1.88	57.23	30.16	65	54	50	47

Table 4.3: Comparison results for continuous microdata (continuation)

Method	IL	DLD	PLD	ID	Score	IL-Rank	DLD-Rank	PLD-Rank	ID-Rank
Micmul04	31.74	13.72	1.38	52.44	30.86	67	53	48	44
Mic3mul03	6.29	29.70	29.06	82.95	31.23	21	79	80	85
Micmul05	35.12	11.73	1.14	48.43	31.27	70	46	44	41
Micmul07	37.68	13.20	1.20	43.46	31.50	72	52	45	34
Micmul06	38.77	13.00	1.22	45.76	32.60	73	50	46	37
Micmul08	41.53	13.12	0.99	42.66	33.19	75	51	42	32
Rank03	5.07	31.73	36.92	89.53	33.50	18	80	83	93
Mic2mul10	10.68	49.38	27.29	77.43	34.28	34	86	78	76
Micmul10	44.69	14.66	0.50	40.41	34.34	76	55	27	29
Noise0.16	32.56	15.65	4.66	64.39	34.91	68	56	64	57
Micmul09	45.98	12.82	0.85	40.99	34.95	79	49	40	30
Mic2mul09	9.93	51.03	33.04	78.94	35.21	32	87	81	78
Mic2mul08	8.55	54.31	33.70	79.77	35.22	27	88	82	80
Mic2mul07	7.53	54.72	37.41	81.40	35.63	25	89	84	83
Noise0.12	25.24	22.21	22.39	71.58	36.09	62	72	74	67
Noise0.1	21.14	27.70	29.03	75.20	36.46	57	78	79	73
Mic2mul06	7.03	56.38	42.00	82.89	36.54	23	90	86	84
JPEG080	33.97	19.13	6.93	66.35	36.83	69	65	69	59
Noise0.14	35.13	19.21	6.24	67.62	37.65	71	66	67	61
Noise0.18	41.12	11.96	3.52	60.95	37.73	74	47	61	52
Noise0.08	17.43	36.06	39.76	79.84	38.15	49	82	85	81
Rank02	2.90	47.26	57.47	94.56	38.18	11	85	90	96
JPEG070	44.92	9.66	2.34	57.28	38.28	77	44	54	48
Noise0.2	45.97	10.01	0.97	57.63	38.77	78	45	41	50
Mic2mul05	5.88	58.97	56.84	85.40	38.77	19	92	89	88
JPEG085	29.47	23.85	24.48	72.80	38.98	66	77	77	70
Mic2mul04	4.90	61.53	60.69	87.26	39.54	17	94	91	89
JPEG090	18.17	35.37	46.98	80.87	39.60	51	81	87	82
Noise0.06	13.03	45.54	56.22	84.16	40.28	40	84	88	87
Mic2mul03	3.28	66.97	64.79	90.51	40.74	15	95	92	94
Noise0.04	8.93	58.51	65.28	88.95	42.18	28	91	94	90
JPEG075	50.45	12.67	2.90	61.27	42.49	80	48	57	53
JPEG095	9.06	60.11	66.56	89.23	42.67	29	93	96	92

Table 4.4: Comparison results for continuous microdata (continuation)

Method	IL	DLD	PLD	ID	Score	IL-Rank	DLD-Rank	PLD-Rank	ID-Rank
Resamp3	3.15	67.90	67.63	96.81	42.72	14	96	97	97
Rank01	2.34	69.19	66.35	99.54	43.00	9	97	95	106
JPEG065	57.77	7.02	1.90	53.87	43.47	81	40	51	46
Noise0.02	4.24	77.34	71.32	94.42	44.31	16	99	98	95
Resamp1	3.11	75.42	71.85	98.36	44.56	13	98	99	99
MicPCP03	69.62	3.16	0.77	38.41	44.90	84	17	38	26
JPEG055	63.70	5.57	1.26	49.70	45.13	83	38	47	42
Noise0.01	2.57	85.19	74.13	97.03	45.46	10	100	103	98
JPEG100	3.06	87.14	73.03	99.14	46.34	12	101	100	101
MicIR10	1.19	97.37	74.07	99.12	46.81	8	102	102	100
MicIR08	1.03	97.84	74.07	99.29	46.83	6	108	101	103
MicIR09	1.14	97.96	74.40	99.24	46.93	7	109	104	102
MicIR06	0.87	97.66	75.28	99.51	46.93	5	106	105	105
MicIR05	0.69	97.58	75.99	99.58	46.94	3	104	106	107
MicIR03	0.45	97.39	78.96	99.79	47.22	1	103	107	109
MicIR04	0.64	97.63	79.78	99.67	47.41	2	105	108	108
MicIR07	0.81	97.79	88.06	99.42	48.49	4	107	109	104
MicPCP04	78.84	3.43	0.62	36.00	48.92	87	19	32	23
JPEG050	73.20	4.26	0.67	47.96	49.21	86	31	36	40
JPEG060	71.24	7.66	1.52	51.71	49.69	85	42	49	43
MicPCP05	82.55	3.94	0.69	34.10	50.38	88	25	37	20
MicPCP07	89.28	4.02	0.62	32.56	53.36	91	27	33	17
MicPCP09	90.78	4.54	0.25	31.40	53.84	94	34	12	13
MicPCP06	90.26	3.37	0.50	33.42	53.97	93	18	26	19
MicZ03	90.25	3.16	0.61	35.71	54.52	92	16	31	22
JPEG035	88.80	3.65	0.44	43.20	55.71	90	20	23	33
JPEG045	87.55	4.15	0.67	46.78	56.07	89	30	35	38
MicZ04	94.94	3.70	0.53	33.04	56.26	96	21	30	18
MicPCP08	96.93	3.97	0.34	32.04	57.02	97	26	16	14
MicPCP10	97.82	4.13	0.46	31.19	57.28	98	29	24	11
JPEG040	90.99	3.72	0.66	44.98	57.29	95	22	34	36
MicZ07	102.87	4.27	0.38	30.53	59.65	99	32	20	9
MicZ06	103.92	3.88	0.41	30.43	60.10	100	23	21	8

Table 4.5: Comparison results for continuous microdata (continuation)

Method	IL	DLD	PLD	ID	Score	IL-Rank	DLD-Rank	PLD-Rank	ID-Rank
MicZ05	104.06	4.03	0.42	31.30	60.41	101	28	22	12
MicZ08	107.92	4.55	0.52	29.60	61.99	102	35	28	7
MicZ10	109.79	4.83	0.38	28.20	62.59	103	36	18	3
MicZ09	110.91	4.35	0.38	28.36	63.14	105	33	19	4
Distr	58.62	43.05	64.88	88.98	65.04	82	83	93	91
JPEG030	110.48	3.02	0.48	41.79	66.12	104	15	25	31
JPEG025	155.15	2.13	0.25	38.76	87.56	106	11	13	27
JPEG020	164.91	1.36	0.29	36.11	91.69	107	7	15	24
JPEG015	202.66	1.10	0.15	32.06	109.50	108	5	8	15
JPEG010	269.38	0.93	0.22	28.44	141.94	109	3	10	5

4.4 Conclusions of the comparison

There is a rich array of methods for microdata disclosure control. The most widely used SDC methods for continuous microdata have been identified and described in this chapter. A score combining information loss and disclosure risk has been proposed for comparison of the identified methods.

Regarding methods for masking continuous microdata (Table 4.2) `Distr`, `MicZ`k, `MicIR`k, `MicPCP`k and resampling score around or above 50 for all tried k. But publishing the original data without masking yields score 50 ($IL = 0$ and $DLD = PLD = ID = 100$), so these are not recomended. Multivariate microaggregation on unprojected data, proposed in [22], is the only form of microaggregation scoring well. Taking three variables at a time seems the best strategy but, even in this case, microaggregation is second to rank swapping. Rank swapping outperforms the best microaggregation and is the best performer (for p around 15 percent). However, being a stochastic method, it is not reproducible and this may lead to disclosure in on-line databases allowing repeated queries. Therefore, microaggregation on unprojected data can still be a good option in such cases. Lossy compression (JPEG) is not

excellent but is promising; being a new proposal, there is room for improvement (such as using other compression algorithms). It is worth noting that random noise masking as implemented to carry out this experimentation is a simple algorithm based on univariate Gaussian noise generation; there are refinements using multivariate Gaussian noise [44, 66] which should perform much better.

Chapter 5

Complexity analysis of optimal microaggregation

In this chapter, we will concentrate on assessing the complexity of microaggregation. According to Chapter 4, microaggregation compares favorably with other masking methods and, as mentioned in Chapter 3, it is one of the most frequently used methods for numerical microdata protection. This technique was proposed by Eurostat, is also being adopted by the United Nations and is the second most frequently used technique in transition countries [34, 48, 30].

In Chapter 3, we noted the fact that a lot of heuristic methods were proposed in order to solve microaggregation approximately. The only existing polynomial exact algorithm [35] is limited to univariate data. This algorithm is of little practical use as univariate microaggregation only makes sense in the following cases:

Univariate data set This is the case where the data set consists of a single variable, which is rather rare in statistical data.

Individual ranking When using individual ranking microaggregation, variables in multivariate data sets are sequentially microaggregated one at a time. However, in [2] it was shown that this type of microaggregation should not be used as it has too high a disclosure risk.

Univariate projection Multivariate data are projected on a one-dimensional space. This approach suffers from too high an information loss.

So, what is needed in practice is multivariate microaggregation. Proving NP-hardness of optimal multivariate microaggregation would explain the non-existence of multivariate polynomial exact algorithms and would justify the use of heuristics. This is precisely what will be done in this chapter.

Let us recall the microaggregation method. This method belongs to the substitution/perturbation approach. Individual records are clustered into small aggregates or groups of size greater than or equal to a given security parameter k (usual values for k are 3 or 4). Rather than publishing the value of a variable V_i for a given record, what is published is the average of the values of V_i over the group to which the record belongs. Therefore, *aggregates should be as homogeneous as possible to minimize information loss.* More formally, let us represent a microdata set with n records and d attributes as $X = \{x_1, x_2, ..., x_n\}$, where $x_i \in \mathbb{R}^d$ and let the security parameter k be a positive integer. A k-partition $P = C_1, C_2, ..., C_{m(P)}$ of X is a partition where the size of each group C_i, $1 \leq i \leq m(P)$ is at least k. Let $\mathbb{P}_k$ be the set of all k-partitions of X. Optimal microaggregation consists of finding a k-partition such that the sum of squared Euclidean distances from each x_i to the centroid $\overline{x}_{C_i} = \frac{1}{|C_i|} \sum_{x_l \in C_i} x_l$ is minimized[1], where C_i is the group to which x_i belongs.

Formally, the problem is

$$\min_{P \in P_k} \sum_{i=1}^{m(P)} \sum_{x_j \in C_i} \|x_j - \overline{x}_{C_i}\|^2 \tag{5.1}$$

where $\|.\|$ denotes Euclidean norm and $m(P)$ is *not* part of the input.

Section 5.1 gives background that will be useful for the proof. Section 5.2 contains the proof itself.

[1] In microaggregation, the sum of squared Euclidean distances is considered as a measure of information loss.

5.1 Background

Let $G = (V, E)$ be a graph, where $V = (v_1, v_2, ..., v_n)$ is a finite set of vertices and E is a set of unordered pairs of vertices, the edges of G.

Definition 1 *A* rectilinear planar layout of a graph $G = (V, E)$ *is a representation of G in a plane, such that its vertices are mapped to horizontal line segments and the edges to vertical line segments, with all endpoints of segments at positive integer coordinates. Two horizontal segments are connected by a vertical segment if and only if the corresponding vertices are adjacent in the graph.*

Proposition 1 (Rosenstiehl/Tarjan [58]) *Given a planar graph $G = (V, E)$, a rectilinear planar layout of G can be computed in time polynomial in the size of G. Moreover, the height and the width of the layout are both linear in the size of G. Without loss of generality, we may assume that all horizontal segments are at different (integer) heights.*

Proposition 2 (Domingo-Ferrer/Mateo [22]) *In an optimal solution of the microaggregation problem the groups have size greater than or equal to k and less than $2k$.*

Proposition 3 (Edwards/Cavalli-Sforza [27]) *The sum of squares of a group of points is equal to the sum of all distances between pairs of entities of this group divided by its cardinality:*

$$SSE = \sum_{x_j \in C_i} \|x_j - \overline{x}_{C_i}\|^2 = \frac{1}{|C_i|} \sum_{x_k, x_l \in C_i} \|x_k - x_l\|^2$$

5.2 Microaggregation is NP-hard

Theorem 1 *The microaggregation problem (5.1) is NP-hard.*

Proof : First we will show that microaggregation is NP-hard in the plane ($d = 2$), and then we generalize this result for an d-dimensional Euclidean space with $d > 2$.

In [56] Pferschy, Rudolf and Woeginger have shown the NP-completeness of "Clustering into Triangles", that is, given a set P of points in a plane, $|P| = 3p$, partition them into p triangles such that the sum of perimeters of all triangles is minimized. We will use their same construction to prove NP-hardness for microaggregation.

The Planar Exact Cover Problem by 3-Sets (Planar X3C), a special version of the Exact Cover Problem that was shown to be NP-complete by Dyer and Frieze [26] will be transformed to the microaggregation problem with $k = 3$ and n a multiple of 3, say $n = 3p$, where p is a positive integer. Planar X3C is defined as follows:

PLANAR EXACT COVER BY 3-SETS (Planar X3C)

Instance:

A set Q with $|Q| = 3q$; a set T of triples from $Q \times Q \times Q$ such that (i) every element of Q occurs in at most three triples and such that (ii) the induced graph is planar. (This induced graph G is defined as the graph containing a vertex for every element of Q and for every triple in T. There is an edge connecting a triple to an element if and only if the element is a member of the triple.

Clearly, G is bipartite with vertex bipartition Q and T).

Question:

Does there exist a subset of q triples in T which contains all the elements of Q?

Hence, let Q and $T \subset Q \times Q \times Q$ constitute an instance of planar X3C. We will construct a point set $S(Q,T)$, where $|S(Q,T)|$ is multiple of 3, that allows a partition into groups C_i, with $|C_i| \geq k = 3$ and SSE minimal if and

only if the planar X3C instance has a solution.

In a first step, a rectilinear planar layout for the underlying undirected graph G is computed according to Proposition 1. Then we multiply all coordinates by a factor of a large positive integer (for example 1000) in order to ensure that points on distinct horizontal (vertical) segments are sufficiently far away from each other. Next, we define the point set $S(Q,T)$. This construction is based on the triangle ∇_0 with side lengths 3,4,5 that will allow all points in $S(Q,T)$ to be kept at integer coordinates. For every element of Q, there is a so-called element point in $P(Q,T)$ and for every triple in T there are three so-called triple points forming a so-called triple triangle in $S(Q,T)$. The triple triangles are that basic triangles ∇_0 with sides 3,4,5 such that the two sides of lengths 3 and 4 are axes-parallel. The element points and the triple triangles are placed somewhere at the corresponding line segments in the rectilinear layout (because of the multiplication, there is ample space to place them).

In the next step, we consider some triple $T_i = (q_1, q_2, q_3)$ in T and the corresponding three triple points t_1, t_2, t_3 that form a triangle ∇_0. For $1 \leq i \leq 3$, the point t_i is connected to the point q_i by a chain of diamonds as it is shown in Figure 5.1. A diamond consists of two copies of ∇_0 that are glued together either by their sides of length 5 (rectangular diamond) or by their sides of length 3 (triangular diamond). All diamonds are placed in such a way that the two shorter sides of the triangles are axes-parallel. No two rectangular diamonds occur consecutively in a chain. The chains of diamonds (roughly) follow the line segments corresponding to the two vertices t_i and q_i and to the connecting edge in the graph G.

In such a construction, two things must be taken into consideration:

a) We do not want two distinct chains to come very close to each other. Again, because of the multiplication in the beginning, there are sufficiently many degrees of freedom to route the chains far away from each other. Since every element occurs in at most three triples, it is also possible to keep the chains sufficiently far from each other if they meet

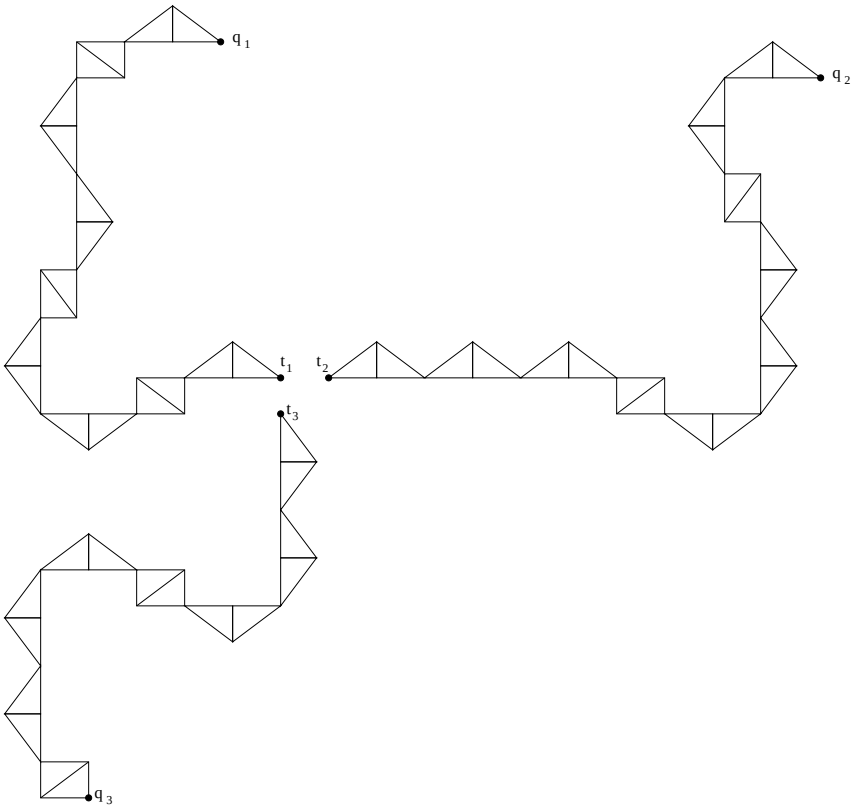


Figure 5.1: Chains of diamonds connecting t_i to q_i.

in an element point (Figure 5.2). The same holds for triple points.

b) The other problem is that it is not *a priori* clear that triple and element points can indeed be connected by such chains of diamonds: triangular shaped diamonds shift the path by ± 8 units in x-direction and 0 units in y-direction (or 0 units in x-direction and ± 8 units in y- direction); rectangular shaped diamonds shift the path by a vector of (± 3, ± 4) or (± 4, ± 3). Once more, the multiplication in the beginning of the construction removes this problem. The main idea is to use only a small number of rectangular shaped diamonds in order to reach the correct remainders for the shift if divided by 8, and to use many triangular shaped diamonds for the long distances. As the distances are

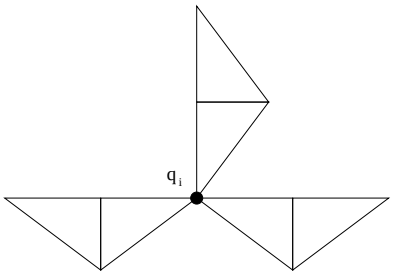


Figure 5.2: Placement of diamonds around an element point q_i.

sufficiently large, we can mix rectangular shaped and triangular shaped diamonds without ending up in problem (a).

Connecting every triple triangle to its corresponding three element points completes the construction of the point set $S(Q,T)$. It is easy to see that this can be performed in polynomial time and $|S(Q,T)|$ is divisible by 3, say $|S(Q,T)| = 3p$ (see [56]). Now we have to prove that $S(Q,T)$ can be partitioned into groups C_i , such that $|C_i| \geq 3$ and SSE minimal if and only if the Planar X3C has a solution.

Due to Proposition 2, we can reduce our search only to groups of size 3, 4 and 5. So, first we will find the minimal SSE of a group of 3, 4 and 5 points in such a structure.

1. Groups of three points. By construction, no three points in $S(Q,T)$ are at a the distance less than 3, 4 and 5 from each other, so using Proposition 3, min SSE of a group of three points is 50/3.

2. Groups of four points. All possible groups of four points with their relative SSE are presented in Figure 5.3. (We have considered only the groups of points in the order they come in the chain without skipping any point, as it is obvious that otherwise the SSE of the resulting group would not be minimal). So we see that the group of 4 points that form a rectangular diamond has minimal $SSE = 25$.

3. Groups of five points. The groups of 5 points are presented in Figure 5.4. As in the case of 4 points, we have considered only points in consecutive order. So the minimal SSE of the group of 5 points is 52.4.

We claim that (i) a partition of $S(Q,T)$ has the minimal SSE if all the groups are of size 3, and (ii) a partition into triangles with this minimal SSE is possible if and only if planar X3C allows a solution.
First we prove the second claim and then the first.

ii) As we have noted above the minimal SSE of a group of 3 points is 50/3, when the points form a triangle with sides 3,4,5. So the minimal SSE that a partition into groups all of size 3 may have is $(50/3)p$, where p is an integer that represents the number of groups.
(if) We claim that the subset T' of T that contains all triples for which the corresponding triple points form a cluster constitutes a solution to the planar X3C.
Consider some element point $q_i \in Q$. The corresponding element point is contained in exactly one cluster, and this cluster belongs to a chain of diamonds. In this chain of diamonds, every other triangle must form a cluster; therefore, the corresponding point t_i on the other end of the chain cannot be covered by any triangle cluster in the chain, and the corresponding three triangle points must form another cluster.
The clusters in the other (one or two) chains going away from q_i cover the corresponding triangle points, and these triangle points cannot form a cluster in the clustering. This way we may assign to each $q_i \in Q$ a unique triple in T. On the other hand, if we assign one q_i to some triple, the other two elements in this triple must be assigned to this triple, too. Clearly, this yields a solution to the X3C.

(only if) Now assume that the planar X3C has a solution $T' \subseteq T$. We construct a clustering into triangles as follows. All triangles corresponding to triples in T' are clusters, all triangles in $T \backslash T'$ are not. This

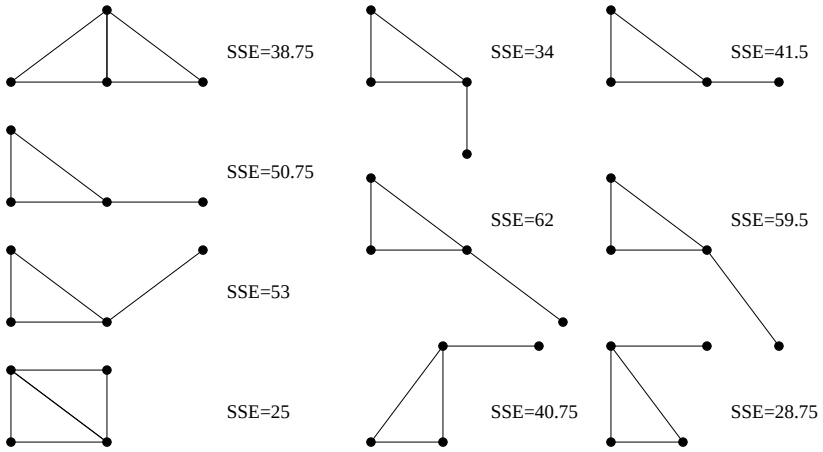


Figure 5.3: Clusters of 4 points.

completely determines which triangles on the chains have to be used in the clustering. Since T' is a solution of X3C, every element point is in exactly one cluster. So the SSE of the total construction with all the groups of size 3 is:

$$SSETOT_3 = \sum_{T_i \in T'} \sum_{j=1}^{3} (50/3) N_{ij} + (50/3)|T'| + \sum_{T_i \in T \setminus T'} \sum_{j=1}^{3} (50/3) N_{ij} =$$

$$= (50/3)(q + \sum_{i=1}^{|T|} \sum_{j=1}^{3} N_{ij}), \qquad (5.2)$$

where N_{ij} is the number of diamonds in the chain j of the triple T_i, where $1 \leq j \leq 3$ and $1 \leq i \leq |T|$.

i) Now we will show that a partition of $S(Q,T)$ has minimal SSE if all groups are of size 3.
Suppose, for the sake of contradiction, that a partition into groups of size 3,4,5 exists with $SSETOT \leq SSETOT3$. Suppose that there are x groups of size 3, y groups of size 4 and z groups of size 5, and $y \neq 0$ or $z \neq 0$ in such a partition. As noted above, the minimal SSE of a group

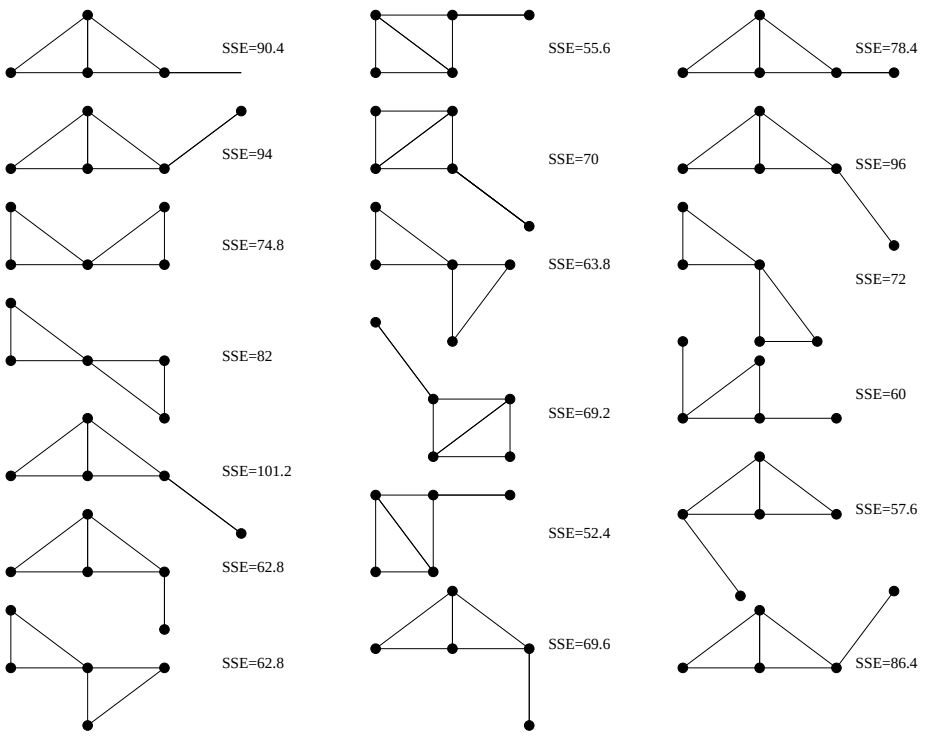


Figure 5.4: Clusters of 5 points.

of 4 points is 25, and of a group of 5 points is 52.4 and $|S(Q,T)| = 3p$ in our construction, so we have:

$$\begin{cases} 3x + 4y + 5z = 3p \\ (50/3)x + 25y + 52.4z \leq (50/3)p \end{cases}$$

$$\begin{cases} p = x + (4/3)y + (5/3)z \\ p \geq x + 1.5y + 3.144z \end{cases}$$

$(4/3)y + (5/3)z \geq 1.5y + 3.144z$ - We have reached a contradiction.

Thus, we have shown that, if a partition of $S(Q,T)$ into groups C_i has the minimal SSE (equal to $SSETOT_3$ in Expression (5.2)), then all groups are of size three. And a partition of $S(Q,T)$ into groups of size three with minimal SSE can be obtained if and only if Planar X3C

has a solution. So the partition into the groups C_i with $|C_i| \geq 3$ has a minimal SSE if and only if Planar X3C has a solution. Thus, an NP-complete problem (Planar X3C) is no harder than the microaggregation problem. Therefore, microaggregation is NP-hard in a plane.

NP-hardness of microaggregation in a d-dimensional Euclidean space follows from the NP-hardness of the same problem in the plane. (By induction, if the problem is NP- hard in a $(d-1)$-dimensional space we can construct an instance of the problem in a d-dimensional space by adding a zero d-th coordinate to all points. This instance of the microaggregation problem in dimension d has a solution if and only if the initial $d-1$-dimensional problem has a solution. So d-dimensional microaggregation is NP-hard for $d > 2$ as well). QED

In summary, we have been able to prove in this chapter that microaggregation is an NP-hard problem for dimension greater than or equal to two. It must be pointed out that our result is different from seemingly similar NP-hardness results in the literature [4, 7, 42, 52, 51, 61, 64]. Most of such results deal with objective functions different from Expression (5.1); even when a sum-of-squares objective function is considered, all known results concentrate on clustering problems consisting of dividing a set of points into a fixed number of clusters, and this without putting cardinality constraints on the groups.

Chapter 6

Critique of sensitivity rules used for table protection

In this chapter our study turns to the theme of tabular data protection. As mentioned in Chapter 2, our research line in tabular data protection has been focused on disclosure risk measures. In Chapter 3 we described currently used sensitivity rules to measure disclosure risk prior to publication of the table.

In this chapter we will prove through general counterexamples that releasing a cell declared non-sensitive by the main sensitivity rules in use ((n,k)-dominance, p%-rule, pq-rule) can actually imply higher disclosure risk than releasing a cell declared sensitive. We also propose an alternative sensitivity rule which measures the concentration of contributions and is free from the aforementioned contradictions.

Section 6.1 discusses flaws related the dominance rule and proposes a new sensitivity rule as a replacement. Section 6.2 extends the critique to the p% and the pq rules; the new sensitivity rule described in Section 6.1 can also be used as a replacement. Some conclusions are given in Section 6.3.

6.1 Analysis of the dominance rule

Remember that the (n, k) dominance rule defines a cell as sensitive if the sum of the contributions of n or fewer respondents represents a fraction k or more of the total cell value. According to [34, 48, 30], the (n, k) dominance rule is the most popular one for magnitude tables, followed by the p%-rule and the pq-rule. Yet, it is significant to note that the U.S. Census Bureau switched in 1992 from the dominance rule to the p%-rule, and the German Statistisches Bundesamt did the same in 2001.

The dominance rule has received critiques for failing to adequately reflect the risk of disclosure, but these have been limited to numerical counterexamples for particular choices of n and k. The following is a counterexample from [57] for the particular case $n = 1$:

Example 4 (Robertson and Ethier, 2002) *In the dominance rule, let* $n = 1$ *and* $k = 0.6$ *(60%). Then a cell with value 100 and contributions 59, 40, 1 is declared* not *sensitive, while a cell with value 100 and contributions 61, 20, 19 would be declared sensitive. Assume now that the second largest respondent of both cells knows the total 100 and is interested in estimating the contribution of the largest respondent. Then, for the* $(59, 40, 1)$ *cell, she removes her contribution and gets an upper bound* $100 - 40 = 60$ *for the largest contribution. For* $(61, 20, 19)$ *the upper bound she gets is* $100 - 20 = 80$*, much farther from the real largest contribution. So the cell declared non-sensitive by the rule allows better inferences than the cell declared sensitive!*

We generalize below the critique in the above example for any values of n and k. Assume a cell X in a table takes a value x which is formed by N respondent contributions $x_1, \cdots, x_N$. Equivalently,

$$x = x_1 + x_2 + \cdots + x_N$$

The dominance rule declares X to be sensitive if a few contributions (n or less) add up to a substantial fraction of x (k or more).

In order to construct a non-sensitive cell, we need the following result.

Lemma 1 *For any integer n and $k \in (0,1]$, there exists an integer N and $r \in [0,1)$ such that*

$$f(r) = \frac{r^n - 1}{r^N - 1} = k \tag{6.1}$$

Proof: For fixed n and N, with $n+1 < N$, the function

$$f(r) = \frac{r^n - 1}{r^N - 1}$$

bijectively maps the interval $[0,1)$ onto the interval $(n/N, 1]$. The lower bound of the image interval is determined as

$$\lim_{r \to 1} f(r) = \frac{n}{N}$$

Thus, the lemma holds if we take N large enough so that $(n)/N < k$.
QED

The non-sensitive cell is now constructed as follows.

Construction 1 (Non-sensitive cell X_{ns})

1. *Take $r \in [0,1)$ and N as defined in Lemma 1. Then it holds that*

$$k = \frac{r^n - 1}{r^N - 1} = \sum_{i=1}^{n} \frac{r^i (r-1)}{r^{N+1} - r} \tag{6.2}$$

2. *Let*

$$R_i := \frac{r^i (r-1)}{r^{N+1} - r} \tag{6.3}$$

3. *Consider a cell X_{ns} whose N relative contributions are*

$$x_i/x = \begin{cases} R_i & \text{for } i = 1, \cdots, n-1 \text{ and } i = n+2, \cdots, N \\ R_n - \varepsilon & \text{for } i = n \\ R_{n+1} + \varepsilon & \text{for } i = n+1 \end{cases}$$

where $\varepsilon := (R_n - R_{n+1})/3$. *With this choice of* ε, *one still has* $x_n/x > x_{n+1}/x$.

4. *According to Expression (6.2), the* n *largest relative contributions* $x_1/x, \cdots, x_n/x$ *add to* $k - \varepsilon$. *Therefore, there is no subset of* n *contributions adding to* k, *so* X_{ns} *is clearly* not *sensitive according to the dominance rule.*

The sensitive cell is constructed as follows.

Construction 2 (Sensitive cell X_s)

1. *Take the same values* N, n *and* k *used in Construction 1.*

2. *Consider a cell* X_s *whose relative contributions are*

$$x_i/x = \begin{cases} R_i & \text{for } i = 1, \cdots, n \\ (1-k)/(N-n) & \text{for } i = n+1, \cdots, N \end{cases}$$

3. *By construction, the sum of the* n *relative contributions* $x_1/x, \cdots, x_n/x$ *is* k. *Thus,* X_s *is declared sensitive by the* (n, k)*-dominance rule.*

We next show that the cell declared non-sensitive by the dominance rule can yield a closer upper bound for the largest contribution than the cell declared sensitive.

Theorem 2 *Assume a coalition of the* n *second largest contributors want to upper-bound the largest contribution. If* N *is large enough, the coalition gets a closer upper bound for the case of* X_{ns} *than for the case of* X_s.

Proof. For both X_s and X_{ns}, the n second largest contributors know that the largest contribution is upper-bounded by the total x minus their own contributions, that is

$$x_1 \leq x - (x_2 + x_3 + \cdots + x_{n+1}) \tag{6.4}$$

The distance between x_1 and the upper bound (6.4) is exactly the sum of the $N-n-1$ smallest contributions. We next show that, for large enough N, this sum is smaller for X_{ns} than for X_s (since x_1 is the same for both cells, this is equivalent to showing that the upper bound on the largest contribution is proportionally closer for X_{ns}). Now, both X_{ns} and X_s total to x, so we can use in what follows relative contributions rather than absolute contributions for both cells. For X_{ns}, the $N-n$ smallest relative contributions add to $1-k+\varepsilon$ by construction; therefore, the $N-n-1$ smallest relative contributions add to $1-k+\varepsilon$ minus the $n+1$-th largest relative contribution

$$(1-k)+\varepsilon-(R_{n+1}+\varepsilon)=(1-k)-\frac{kr^n(r-1)}{r^n-1} \tag{6.5}$$

To obtain the last term of Expression (6.5), we have used that $k=(r^n-1)/(r^N-1)$. On the other hand, for X_s, the sum of the $N-n-1$ smallest relative contributions is

$$\frac{(N-n-1)(1-k)}{N-n}=(1-k)-\frac{1-k}{N-n} \tag{6.6}$$

If $N\to\infty$, Expression (6.6) approaches $1-k$. On the other hand, if we let $N\to\infty$, then r is such that $k=1-r^n$ (according to Lemma 1). In this case, Expression (6.5) becomes

$$(1-k)-\frac{k(1-k)(r-1)}{1-k-1}=r(1-k)<(1-k)$$

Thus, N can be taken large enough so that Expression (6.6) is larger than Expression (6.5), which causes the theorem to hold. QED.

The following example illustrates that N does not need to be very large for Theorem 2 to hold.

Example 5 *For $n=1$, $k=0.369$ and $N=3$, we have $r=0.9$. The largest relative contribution is $x_1/x=0.369$ for both X_s and X_{ns}. The tail of the $N-n-1=1$ smallest relative contribution is 0.298893 for X_{ns} and 0.315498 for X_s.*

6.1.1 Replacing the dominance rule

Theorem 2 highlights a major flaw in the dominance rule. Note that n can be as small as 1 and, in that case, *a single cell contributor* (the second largest) can, without any help, get more precise estimates on the largest contribution for a cell declared non-sensitive than for a cell declared sensitive.

The explanation of the flaw is that the dominance rule does not capture well the concentration of contributions to a table cell (even if it defines sensitivity as concentration in a rough way). That is the reason why a cell declared non-sensitive by the dominance rule can have a smaller tail (sum of small contributions) than a cell declared sensitive, which results in better bounds on the largest contribution.

As an alternative sensitivity rule, it could seem natural to directly check whether the sum of a number of smallest contributions to the cell is below some threshold. The problem is how many smallest contributions should be considered. Note that, for X_{ns} and X_s, the (flawed) dominance rule actually amounts to checking whether the $N-n$ smallest contributions add to $1-k$ or less. If we consider only the $N-n-1$ smallest contributions, then X'_{ns}, X'_s could be constructed such that a result analogous to Theorem 2 can be proven for a coalition of the $n+1$ (rather than n) second largest contributors. The same argument applies to any number of smallest contributions that can be chosen.

A better sensitivity rule could be based on actually measuring the concentration of contributions. Evenness of the distribution of contributions can be measured by the entropy of the relative contributions as follows:

$$H(X) = -\sum_{i=1}^{N} (x_i/x) \log_2 (x_i/x) \tag{6.7}$$

Expression (6.7) is maximal when all relative contributions are the same, *i.e.* when $x_i/x = 1/N$ for all i. Therefore, the higher $H(X)$, the less concentration in the contributions to cell X and the less sensitive should the cell be. We next compute $H(X_{ns})$ and $H(X_s)$:

- The entropy of the relative contributions of X_{ns} is

$$H(X_{ns}) = -\sum_{i=1}^{n-1} R_i \log_2(R_i) - (R_n - \varepsilon)\log_2(R_n - \varepsilon)$$

$$-(R_{n+1} + \varepsilon)\log_2(R_{n+1} + \varepsilon) - \sum_{i=n+2}^{N} R_i \log_2(R_i) \qquad (6.8)$$

- The entropy of the relative contributions of X_s is

$$H(X_s) = -\sum_{i=1}^{n} R_i \log_2(R_i) - (1-k)\log_2(\frac{1-k}{N-n}) \qquad (6.9)$$

The theorem below shows that concentration can be higher for a cell declared non-sensitive by the dominance rule than for a cell declared sensitive, that is, $H(X_{ns}) < H(X_s)$. Thus, if the dominance rule is replaced by entropy as a sensitivity measure, X_{ns} is declared as more sensitive than X_s for large enough N, which is consistent with Theorem 2.

Theorem 3 *Let n and k be the parameters of the dominance rule. Let N be the number of contributors to cells X_s and X_{ns}. If N is sufficiently large, then $H(X_{ns}) < H(X_s)$ holds.*

Proof (Theorem 3): To compare Expressions (6.8) and (6.9) we must check whether

$$-(R_n - \varepsilon)\log_2(R_n - \varepsilon) - (R_{n+1} + \varepsilon)\log_2(R_{n+1} + \varepsilon) - \sum_{i=n+2}^{N} R_i \log_2(R_i) \stackrel{?}{<}$$

$$\stackrel{?}{<} -R_n \log_2(R_n) - (1-k)\log_2(\frac{1-k}{N-n})$$

The above inequality holds if the following inequality holds

$$R_n \log_2(\frac{1}{R_n - \varepsilon}) + \varepsilon \log_2(\frac{R_n - \varepsilon}{R_{n+1} + \varepsilon}) + \sum_{i=n+1}^{N} R_i \log_2(\frac{1}{R_i}) \stackrel{?}{<}$$

$$\stackrel{?}{<} R_n \log_2(\frac{1}{R_n}) + (1-k)\log_2(\frac{N-n}{1-k}) \tag{6.10}$$

Now, all terms on both sides of Inequality (6.10) are positive (the arguments of logs being all greater than one). However, as N increases, the second term on the right-hand side increases more rapidly than the third term on the left-hand side (increasing N on the left-hand side results only in extending the tail of a geometric series). Thus, for large enough N, the theorem holds. QED

The following example illustrates Theorem 3 for the same parameters used in Example 5.

Example 6 *For $n = 1$, $k = 0.369$ and $N = 3$, we have $r = 0.9$ and*

$$H(X_{ns}) = 1.58088 < H(X_s) = 1.5809$$

Thus, entropy gives results consistent with Example 5.

If entropy is to be used as a sensitivity rule, it is desirable to bound its range to a fixed interval, for example $[0, 1]$. A known property of the entropy of a variable taking one of N values is that $0 \leq H(X) \leq \log_2 N$. Thus, an entropy-based sensitivity rule could look like:

Algorithm 1 (Entropy-based sensitivity(t))

1. *Let parameter $t \in [0, 1]$ be a sensitivity threshold.*

2. *Given a cell X with N contributors, compute $H(X)$ using Expression (6.7).*

3. *If $H(X)/\log_2 N < t$ declare X as sensitive; otherwise, declare X as non-sensitive.*

6.2 Analysis of the $p\%$ and the pq-rules

In the pq-rule it is assumed that, prior to table publication, each respondent can estimate the contribution of each other respondent to within less than q percent. A cell is considered sensitive if, posterior to the publication of the table, someone can estimate the contribution of an individual respondent to within p percent. A special case is the $p\%$-rule: in this case, no knowledge prior to table publication is assumed, *i.e.* the pq-rule is used with $q = 100$. For the sake of simplicity, we will focus the discussion on the $p\%$-rule. Analysis of the pq-rule is analogous. Note, that the formulation of the $p\%$-rule is in practice more specific than the definition given above. In its most commonly accepted form (see [32]), the $p\%$-rule declares a cell to be sensitive if the second largest contributor can estimate the largest contribution x_1 to within a proportion p of x_1 (for convenience, we will take p as a proportion rather than as a percent), that is, when

$$x - x_2 - x_1 \leq px_1 \tag{6.11}$$

where x_2 is the second largest contribution and x is the cell total.

Let us now construct a cell X_{ns} which is non-sensitive according to the $p\%$-rule.

Construction 3 (Non-sensitive cell X_{ns})

1. *Take* $r \in [0, 1)$ *such that*

$$g(r) := \frac{r^N - r^2}{r - 1} = p \tag{6.12}$$

For $N \geq 3$ *and* $r \in [0, 1)$, g *is an increasing function whose image includes the whole interval* $[0, 1)$. *So, for any* $p \in [0, 1)$, *we can find* r *such that* $g(r) = p$.

2. *Now take the cell total to be*

$$x := \frac{r^{N+1} - r}{r - 1} \tag{6.13}$$

and consider a cell X_{ns} whose contributions are

$$x_i = \begin{cases} r & for\ i = 1 \\ r^2 - \varepsilon & for\ i = 2 \\ r^3 + \varepsilon & for\ i = 3 \\ r^i & for\ i = 4, \cdots, N \end{cases}$$

where $\varepsilon := (r^2 - r^3)/3$. With this choice of ε, one still has $x_2 > x_3$.

3. *From the first condition in Expression (6.12),*

$$rg(r) = r^3 + r^4 + \cdots + r^N = x - r - r^2 = pr \tag{6.14}$$

so that

$$x - x_1 - x_2 = x - r - r^2 + \varepsilon > px_1 = pr$$

Thus, the cell is clearly non-sensitive according to the p%-rule.

A sensitive cell can be constructed as follows.

Construction 4 (Sensitive cell X_s)

1. *Take the same values p, r, N and x used in Construction 3.*

2. *Consider a cell X_s whose contributions are*

$$x_i = \begin{cases} r & for\ i = 1 \\ r^2 & for\ i = 2 \\ pr/(N-2) & for\ i = 3, \cdots, N \end{cases}$$

Note that the contributions to X_s add to x, because

$$\sum_{i=1}^{N} x_i = r + r^2 + pr = x \tag{6.15}$$

where the last equality is obtained using Equation (6.14).

3. *From Equation (6.15),*

$$x - x_2 - x_1 = x - r - r^2 = pr = px_1$$

Thus, X_s is declared sensitive by the p%-rule.

We next show that the cell declared non-sensitive by the p%-rule can yield a closer upper bound for the largest contribution than the cell declared sensitive.

Theorem 4 *Let p be the parameter of the p%-rule. Assume a coalition of the two second largest contributors want to upper-bound the largest contribution. For any p, if N is taken large enough, then the coalition gets a proportionally closer upper bound for the case of X_{ns} than for the case of X_s.*

Proof. For both X_s and X_{ns}, the two second largest contributors know that the largest contribution is upper-bounded by the total x minus their own contributions, that is

$$x_1 \leq x - x_2 - x_3 \tag{6.16}$$

Now, the distance between x_1 and the upper bound (6.16) is exactly the sum of the $N-3$ smallest contributions. We next show that, for large enough N, this sum is smaller for X_{ns} than for X_s (since x_1 is the same for both cells, this is equivalent to showing that the upper bound on the largest contribution is proportionally closer for X_{ns}). For X_{ns}, the $N-3$ smallest contributions add to

$$\sum_{i=4}^{N} x_i = x - r - (r^2 - \varepsilon) - (r^3 + \varepsilon) = pr - r^3 < pr \tag{6.17}$$

For X_s, the $N-3$ smallest contributions add to

$$\sum_{i=4}^{N} x_i = pr\frac{N-3}{N-2} \tag{6.18}$$

As $N \to \infty$, Expression (6.18) approaches pr and becomes thus larger than Expression (6.17). Thus, the theorem holds for large enough N. QED.

The following example illustrates that N does not need to be very large for Theorem 4 to hold.

Example 7 *For $p = 0.4375$ and $N = 5$, we have $r = 0.5$. The largest contribution is $x_1 = 0.5$ for both X_s and X_{ns}. The tail of the $N-3$ smallest relative contributions is* 0.09677 *for X_{ns} and* 0.145833 *for X_s.*

6.2.1 Replacing the p% and pq-rules

The flaw pointed out by Theorem 4 is due to the fact that neither the p%-rule nor the pq-rule reflect the concentration of contributions in a proper way. Like it happened with the dominance rule, a cell X_{ns} declared non-sensitive can have a smaller tail than a cell X_s declared sensitive; this causes X_{ns} cell to yield a better bound for the largest contribution. However, the situation is a bit safer here than for the dominance rule: with the p%-rule, no single contributor can estimate the largest contribution better for a non-sensitive cell than for a sensitive cell (at least a coalition of two contributors is needed, which may be unrealistic sometimes).

A check whether the sum of a number of smallest contributions is below some threshold is not a good alternative sensitivity rule for the p% rule for the same arguments it was not a good replacement for the dominance rule. Note that the (flawed) p%-rule actually consists of checking whether the $N-2$ smallest contributions are below px_1. For any number $N-t$ of smallest contributions considered, one could construct cells X'_{ns}, X'_s such that a result analogous to Theorem 4 holds for a coalition of the t second largest contributors.

We show below that the entropy of the relative contributions is a sensitivity measure which, for large enough N, declares X_{ns} more sensitive than X_s. We first compute the entropies of both cells:

- The entropy of the relative contributions of X_{ns} is

$$H(X_{ns}) = -(r/x)\log_2(r/x) - ((r^2-\varepsilon)/x)\log_2((r^2-\varepsilon)/x)$$

$$-((r^3+\varepsilon)/x)\log_2((r^3+\varepsilon)/x) - \sum_{i=4}^{N}(r^i/x)\log_2(r^i/x) \qquad (6.19)$$

- The entropy of the relative contributions of X_s is

$$H(X_s) = -\frac{r}{x}\log_2(\frac{r}{x}) - \frac{r^2}{x}\log_2(\frac{r^2}{x}) - \frac{pr}{x}\log_2(\frac{pr}{x(N-2)}) \qquad (6.20)$$

Theorem 5 *Let the proportion p be the parameter of the p%-rule. Let N be the number of contributors to cells X_s and X_{ns}. For any p, if N is taken large enough, then $H(X_{ns}) < H(X_s)$ holds.*

Proof (Theorem 5): To compare Expressions (6.19) and (6.20), we can simplify computations by replacing x with 1, which amounts to replacing relative contributions by absolute contributions (since both X_s and X_{ns} add to x, this does not affect the comparison). Our goal is to check whether

$$-r\log_2(r) - (r^2-\varepsilon)\log_2(r^2-\varepsilon) - (r^3+\varepsilon)\log_2(r^3+\varepsilon) - \sum_{i=4}^{N} r^i\log_2(r^i) \stackrel{?}{<}$$

$$\stackrel{?}{<} -r\log_2(r) - r^2\log_2(r^2) - pr\log_2(\frac{pr}{N-2})$$

The above

$$r^2\log_2(\frac{1}{r^2-\varepsilon}) + \varepsilon\log_2(\frac{r^2-\varepsilon}{r^3+\varepsilon})\sum_{i=3}^{N} r^i\log_2(\frac{1}{r^i}) \stackrel{?}{<}$$

$$\stackrel{?}{<} r^2\log_2(\frac{1}{r^2}) + pr\log_2(\frac{N-2}{pr}) \qquad (6.21)$$

Now, all terms on both sides of Inequality (6.21) are positive (the arguments of logs being all greater than one). However, as N increases, the second term on the right-hand side increases more rapidly than the third term on the left-hand side (increasing N on the left-hand side results only in extending the tail of a geometric series). Thus, for large enough N, the theorem holds. QED

Thus, if the p%-rule is replaced by entropy as a sensitivity measure, X_{ns} is declared as more sensitive than X_s for large enough N, which is consistent with Theorem 4. The following example illustrates Theorem 5 for the same parameters used in Example 7.

Example 8 *For* $p = 0.4375$ *and* $N = 5$, *we have* $r = 0.5$ *and*

$$H(X_{ns}) = 1.82108 < H(X_s) = 1.83946$$

Thus, we propose again to use the entropy-based sensitivity rule given by Algorithm 1 to replace the p%-rule. To replace the pq-rule, prior knowledge by the intruder must be taken into account, so one might think of substituting conditional entropy for plain entropy in Algorithm 1 (entropy would be conditional to the prior knowledge by the intruder).

6.3 Conclusions on sensitivity rules

We have constructed general counterexamples that show that none of the most widely accepted sensitivity rules for tabular protection ((n, k) dominance, p% and —by extension— pq) captures disclosure risk in a proper way when coalitions of cell contributors are a realistic scenario. Further, for any value of k, the $(1, k)$-dominance rule is problematic even without coalitions: the second largest cell contributor can compute without help a bound on the largest contribution which might be more accurate for a cell labeled as non-sensitive than for a cell labeled as sensitive. The explanation of the above misbehaviors is that the rules in use fail to adequately reflect

the concentration of contributions.

Possible and non-exclusive solutions when contributors should be regarded as potential intruders include:

- Replace sensitivity rules in use by a new entropy-based sensitivity rule (given by Algorithm 1). This new rule measures the concentration of relative contributions and has been shown to work well for the counterexamples proposed.

- Complement *a priori* risk assessment provided by sensitivity rules with *a posteriori* assessment. This will be the theme of the following chapter.

Chapter 7

A posteriori disclosure risk measure for tabular data based on conditional entropy

In the previous chapter we discussed the weaknesses of existing *a priori* disclosure risk measures and we proposed a new *a priori* sensitivity rule which works better than the previous ones. Remember that *a priori* measures decide whether the cell is sensitive before applying protection methods. So, *a priori* measures do not actually measure the disclosure risk incurred once a particular table is released. That is why using only *a priori* measures may be insufficient for the table protection.

In this chapter, we propose to complement *a priori* risk assessment with *a posteriori* risk assessment, which takes protected information into account. The proposed measure applies to a broad class of disclosure protection methods and is computable in practice.

In section 7.1 a new *a posteriori* measure of disclosure risk is discussed.

Section 7.2 describes an application of the proposed *a posteriori* disclosure risk measure to different table protection methods, both for simple tables and for linked tables. In the case of cell suppression methods, the proposed measure turns out to be useful to detect suppression patterns which offer

more protection than others.

7.1 An *a posteriori* disclosure risk measure based on conditional entropy

We have mentioned above that using only *a priori* measures may be insufficient for table protection. Now we want to illustrate this on the following examples.

Example 9 *Suppose, that the person or entity who wants to guess secret information about how much a particular respondent contributed to some cell of the table is someone who also contributed to that cell. So he obviously knows his own contribution to that cell; he also may know some additional information, for example, how many respondents have contributed to that cell, who they are, etc. This internal intruder is in a better position than an outsider to estimate the contribution of his interest. This kind of information is not taken into account by* a priori *measures. According to these, the disclosure risk is the same for all types of intruders and that is not true.*

The information held by an intruder does not only depend on her being internal or external; it clearly depends also on what information has previously been published and on how that information has been protected.

Example 10 *Assume we have an n-dimensional table whose cells are deemed sensitive, and therefore cannot be released. Only some 2-dimensional (or $(n-i)$-dimensional) tables are released, which have been obtained as projections of the n-dimensional table. Due to their origin, the released tables are linked tables, so the uncertainty about a cell value in the n-dimensional table is conditional to the particular tables released so far.*

Therefore, we propose to complement *a priori* risk assessment provided by sensitivity rules with *a posteriori* risk assessment. The latter is performed

after data have been protected and takes protected data into account to compute bounds on cells labeled as sensitive by a sensitivity rule.

Our proposal for *a posteriori* measure is to use the reciprocal of Shannon's conditional entropy [62] to express the disclosure risk in a natural and unified manner.

Entropy-based measures were already discussed in [69] for computing information loss at the table level, but not for computing disclosure risk. However, the authors of [69] do not believe entropy is a practical information loss measure. We support their opinion with the following example.

Example 11 *Assume we use rounding with integer base b to protect a table. The entropy-based information loss measure defined in [69] is the reciprocal of the number of original tables whose rounded version matches the published rounded table (*i.e. *the number of original tables "compatible" with the published one). The number of compatible tables depends on the rounding base b, but is independent on how close the published rounded values are to the original values. Thus, the entropy-based information loss measure is the same when the original table exactly corresponds to the rounded table (which happens when all cell values in the original table are multiples of b) and when all differences between corresponding cell values in the original and rounded tables are close to b/2. This does not seem to adequately reflect the utility of the published data.*

In [24], the reciprocal of Shannon's entropy (not conditional entropy) was suggested as measure of disclosure risk at the cell level. What was not clear there is how to compute the probabilities, that is, what distribution should be chosen. In fact, as we noted above, the particular distribution for an intruder depends on the knowledge of that intruder.

The above discussion suggests that the most natural *a posteriori* measure for disclosure risk is the reciprocal of conditional entropy

$$DR(X) = 1/H(X|Y = y) = 1/(-\sum_{x} p(x/y) log_2 p(x/y)) \qquad (7.1)$$

where X is a variable representing an original cell and Y is a variable representing the intruder's knowledge (which is supposed to be equal to some specific value y). The intuitive notion behind Expression (7.1) is that, the more uncertainty about the value of the original cell X (which depends on the constraints $Y = y$), the less disclosure risk (and conversely).

There are two practical problems in computing Expression (7.1):

1. Finding the set $S_y(X)$ of possible values of X given the constraints y.
2. Estimating the probabilities $p(x|y)$, *i.e.* the probability of the cell X being x given that Y is y.

As noted by [69] when discussing entropy-based information loss measures, taking the uniform probability distribution over the set $S_y(X)$ can make sense for some disclosure control methods. Using the uniform distribution, Expression (7.1) is simplified to

$$DR_{unif}(X) = 1/\log_2 m(S_y(X)) \tag{7.2}$$

where $m(S_y(X))$ is the number of possible values of the cell in $S_y(X)$.

Note 1 (On $m(S_y(X))$) *We assume in what follows that table cells take values in a discrete domain: either integer values or real values with a fixed number of decimal positions. This is the usual case in published statistical tables: count tables consist of integer values and magnitude tables consist of either integer values or real values with limited precision. Thus the set $S_y(X)$ of possible values is enumerable and it makes sense to speak of $m(S_y(X))$ as the number of cell values in $S_y(X)$.*

7.2 Application to several table protection scenarios

We show in this Section how to compute Expression (7.2) for several disclosure control methods for tables; the case of linked tables will also be

Table 7.1: A table with suppressed cells

Economic activity	Size class					Total
	4	5	6	7	8	
2,3	80	253	54	0	0	387
4	641	3694	2062	746	0	7143
5	592	x_1	329	x_2	1440	3898
6	57	x_3	946	x_4	2027	4281
7	78	0	890	1719	1743	4430
Total	1148	4353	4281	4847	5210	20139

discussed.

7.2.1 Cell suppression

The disclosure risk computation for cell suppression is illustrated by extending an example provided in [69]. Let Table 7.1 be a table from which four cells x_1, x_2, x_3 and x_4 have been suppressed. Assume that the suppressed values are integer.

According to the definition given in Section 7.1, the disclosure risk for each suppressed cell is the reciprocal of one of the following conditional entropies:

$$
\begin{aligned}
H(x_1|x_1 + x_2 = 1537, x_1 + x_3 = 406, x_i \geq 0) \\
H(x_2|x_1 + x_2 = 1537, x_2 + x_4 = 2382, x_i \geq 0) \\
H(x_3|x_1 + x_3 = 406, x_3 + x_4 = 1251, x_i \geq 0) \\
H(x_4|x_2 + x_4 = 2382, x_3 + x_4 = 1251, x_i \geq 0)
\end{aligned}
$$

Expressions (7.3) contain constraints y_i for each suppressed cell x_i which allow $m(S_{y_i}(x_i))$ to be computed by solving two linear programming (LP)

Table 7.2: A table with two rows combined

Economic activity	Size class					Total
	4	5	6	7	8	
2,3	80	253	54	0	0	387
4	641	3694	2062	746	0	7143
5,6	649	406	1275	2382	3467	8179
7	78	0	890	1719	1743	4430
Total	1148	4353	4281	4847	5210	20139

problems (one maximization and one minimization) and subtracting the solutions. In the case of Table 7.1, minimizations and maximizations bound every cell as follows: $0 \leq x_1 \leq 406$, $1131 \leq x_2 \leq 1537$, $0 \leq x_3 \leq 406$ and $845 \leq x_4 \leq 1251$. By subtracting the bounds we obtain $m(S_{y_i}(x_i)) = 407$ for $i = 1, 2, 3, 4$. Using Expression (7.2), we can compute $DR_{unif}(x_i) = 1/\log_2 407 = 0.115$ for every cell.

7.2.2 Rounding

When the table is protected by rounding, the cell entropy conditional to the rounded table depends on the rounding base b. In a rounded table without marginals, if the value of a cell x_i' is $n_i b$ (*i.e.* n_i times the rounding base), then we know that the original cell x_i must lie in the interval $I_i = [(n_i - 1/2)b, (n_i + 1/2)b)$. Thus, $DR_{unif}(x_i) = 1/log_2 m(I_i)$, where $m(I_i)$ is the number of possible cell values in I_i (keep in mind that cell values are either integer or with a fixed number of decimal positions).

7.2.3 Table redesign

This case is very similar to cell suppression. Imagine that the sensitive cells in Table 7.1 are protected by combining rows with $Economic_activity = 5$ or 6. This yields Table 7.2.

Let us label the six cells in the original row with $Economic_activity = 5$ as x_1 through x_6 and the six cells in $Economic_activity = 6$ as x_7 through x_{12} (x_6 is the marginal of the first row and x_{12} is the marginal of the second row). Then the following equalities hold:

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 + x_5 - x_6 &= 0 \\ x_7 + x_8 + x_9 + x_{10} + x_{11} - x_{12} &= 0 \\ x_1 + x_7 &= 649 \\ x_2 + x_8 &= 406 \\ x_3 + x_9 &= 1275 \\ x_4 + x_{10} &= 2382 \\ x_5 + x_{11} &= 3467 \\ x_6 + x_{12} &= 8179 \\ x_i &\geq 0 \text{ for } i = 1, \cdots, 12 \end{aligned} \tag{7.3}$$

From the above, $m(S_{y_i}(x_i))$ and $DR_{unif}(x_i)$ are computed in a way analogous to the case of cell suppression.

7.2.4 Linked tables

We will show the application of conditional entropy as *a posteriori* disclosure risk measure for linked tables with an example.

Let us consider the three-dimensional table ASR formed by cells $z_{a_i s_j r_k}$, where each cell denotes the total turnover of businesses with activity a_i and size s_j in region r_k. Assume that table ASR is not released because every cell in it is considered sensitive. Instead of ASR, some of the following tables obtained by bidimensional projection are released: $AS = \{z_{a_i s_j}\}$, which breaks down turnover by activity and business size, $AR = \{z_{a_i r_k}\}$, which breaks down turnover by activity and region, and $SR = \{z_{s_j r_k}\}$, which breaks down turnover by size and region. Assume three scenarios: 1) only AS is released; 2) AS and AR are released; 3) AS, AR and SR are released. The disclosure

risk of cell $z_{a_i s_j r_k}$ in each scenario can be expressed as:

$$DR_{unif}(z_{a_i s_j r_k}|AS) = 1/H(z_{a_i s_j r_k}|z_{a_i s_j} = \sum_k z_{a_i s_j r_k}) \tag{7.4}$$

$$DR_{unif}(z_{a_i s_j r_k}|AS, AR)$$
$$= 1/H(z_{a_i s_j r_k}|z_{a_i s_j} = \sum_k z_{a_i s_j r_k}, z_{a_i r_k} = \sum_j z_{a_i s_j r_k}) \tag{7.5}$$

$$DR_{unif}(z_{a_i s_j r_k}|AS, AR, SR)$$
$$= 1/H(z_{a_i s_j r_k}|z_{a_i s_j} = \sum_k z_{a_i s_j r_k}, z_{a_i r_k} = \sum_j z_{a_i s_j r_k}, z_{s_j r_k} = \sum_i z_{a_i s_j r_k}) \tag{7.6}$$

The released tables impose constraints on the possible cell values of the table ASR. Such constraints actually determine the simplexes $S_{AS}(z_{a_i s_j r_k})$, $S_{AS,AR}(z_{a_i s_j r_k})$ or $S_{AS,AR,SR}(z_{a_i s_j r_k})$ where $z_{a_i s_j r_k}$ should lie. By solving one LP maximization and one LP minimization for each $z_{a_i s_j r_k}$, an interval where the cell lies can be determined. Then, the cell disclosure risk is computed using Expression (7.2). If a cell is too closely bounded, then its disclosure risk is too high and disclosure control methods must be used.

When the disclosure control method chosen is cell suppression, it is important to notice that linked tables have the property that there are sets of linearly dependent constraints, so that one of the constraints in each such set may be suppressed without decreasing the rank of the whole constraint system. This will influence the quality of the protection offered by different suppression patterns: the best pattern is the one decreasing most the system rank, which results in more degrees of freedom, and thus more cell entropy and lower disclosure risk. We show this with in the following section.

Table 7.3: Constraint matrix imposed by Table AS. Here $i \in \{1, \cdots, 3\}$, $j \in \{1, \cdots, 4\}$, $k \in \{1, \cdots, 3\}$.

1	1	1																	$z_{a_1 s_1}$
			1	1	1														$z_{a_1 s_2}$
						1	1	1											$z_{a_1 s_3}$
									1	1	1								$z_{a_1 s_4}$
																			$\cdots$
																1	1	1	$z_{a_3 s_4}$

7.2.5 Minimizing disclosure risk in linked table release

For the sake of concreteness, we will resume the example of three linked tables used in the previous section. We want to estimate the disclosure risk of cells in ASR depending on the released tables. Assume that $i \in \{1, \ldots n\}$, $j \in \{1, \ldots m\}$ and $k \in \{1, \ldots t\}$.

One table released out of three linked tables

If only the table AS is released, the constraint matrix is shown in Table 7.3, where there is a column for each cell $z_{a_i s_j r_k}$. The matrix rank is nm, as all equations are linearly independent. Every choice for secondary suppressions causes the same rank decrease and consequently has an equivalent impact on the disclosure risk. Therefore, there is no room for optimization (unless there are specific additional constraints specified by the data protector).

Two tables released out of three linked tables

If tables AS and AR are released, the constraint matrix is shown in Table 7.4, where there is a column for each cell $z_{a_i s_j r_k}$. This matrix consists of n submatrices of size $(t+m) \times mt$ with rank $(t+m-1)$, that is, one constraint in each submatrix is a linear combination of the remaining constraints in the

submatrix. That is, using Gaussian elimination we have:

$$\sum_{k=1}^{t} z_{a_i r_k} - \sum_{j=1}^{m} z_{a_i s_j} = 0 \text{ for } 1 \leq i \leq n \tag{7.7}$$

Therefore, n constraints can be suppressed and the matrix rank will not change, nor will change $H(z_{a_i s_j r_k}|AS, AR)$ nor the disclosure risk. Note that only one row per submatrix can be suppressed for disclosure risk to stay unchanged, which, in terms of tables AS and AR, means only one cell per two corresponding rows in AS and AR (*e.g.* for the submatrix related to Expression (7.7), the two rows are those in AS and AR with subscript a_i).

From the above discussion, we can state the following proposition:

Proposition 4 *When two out three linked tables are released, the entropy increase and the disclosure risk decrease are maximized if the suppression patterns are chosen so that the secondary suppressions are in the same columns and rows for both released tables.*

Proof : Assume that tables AS and AR are released. Now assume that $z_{a_i s_j r_k}$ in supertable ASR has a high disclosure risk which makes it necessary to increase its entropy. So, if a cell suppression is used, a natural option is to suppress the cells in the released tables which refer to $z_{a_i s_j r_k}$. These suppressions will be called primary suppressions and will be $z_{a_i s_j}$ and $z_{a_i r_k}$ in the tables AS and AR, respectively. The suppression of these two cells will decrease the rank by 1 (see the discussion above).

Secondary suppressions will be the following:

- Two cells, say $z_{a_i s_l}$ and $z_{a_f s_j}$ in the table AS in order to prevent $z_{a_i s_j}$ from being computable (in what follows, we will say - "to protect $z_{a_i s_j}$") , which decrease the rank by 1,
- The cell $z_{a_f s_l}$ in AS to protect $z_{a_f s_j}$, which decreases the rank by 1.
- Two cells in the table AR: one in the row a_i and other in the column r_k. If $z_{a_i r_v}$ is the cell in row a_i, its suppression decreases the rank by

Table 7.4: Constraint matrix imposed by tables AS and AR. Here $i \in \{1, \cdots, 3\}$, $j \in \{1, \cdots, 4\}$, $k \in \{1, \cdots, 3\}$.

1	1	1																	$z_{a_1 s_1}$
1			1			1			1										$z_{a_1 r_1}$
	1			1			1			1									$z_{a_1 r_2}$
		1			1			1			1								$z_{a_1 r_3}$
			1	1	1														$z_{a_1 s_2}$
						1	1	1											$z_{a_1 s_3}$
									1	1	1								$z_{a_1 s_4}$
													$\cdots$					$\cdots$	
																1	1	1	$z_{a_3 s_4}$

1. When we choose the candidate for the suppression in column r_k, we should take into account that, if we choose the cell in a row other than a_f, this will not decrease the rank (because the rows where we have already suppressed cells are a_i and a_f). So, in order to increase the entropy and to decrease disclosure risk, we have to choose row a_f, that is we have to suppress the cell $z_{a_f r_k}$.

- Finally, in order to protect $z_{a_f r_k}$, we have to suppress $z_{a_f r_v}$, which will decrease the rank by 1.

So, from the above argument, we have that choosing the same rows for secondary suppressions in both tables will decrease the rank by 1 more than if we chose different rows. Therefore, this is the strategy to maximize rank decrease, which is equivalent to to maximizing entropy increase and disclosure risk decrease. QED

All three linked tables released

Now let us consider the case when three tables AS, AR, SR are released. Again, using Gaussian elimination, we obtain several sets of constraints which

are linearly dependent (exactly one constraint in each set is a linear combination of the rest). The linear combinations are as follows:

$$\sum_{k=1}^{t} z_{a_i r_k} - \sum_{j=1}^{m} z_{a_i s_j} = 0 \text{ for } 1 \leq i \leq n \tag{7.8}$$

$$\sum_{k=1}^{t} z_{s_j r_k} - \sum_{i=1}^{n} z_{a_i s_j} = 0 \text{ for } 1 \leq j \leq s-1 \tag{7.9}$$

$$\sum_{j=1}^{m} z_{s_j r_k} - \sum_{i=1}^{n} z_{a_i r_k} = 0 \text{ for } 1 \leq k \leq t-1 \tag{7.10}$$

$$z_{s_m r_t} - \sum_{j=1}^{m-1} \sum_{k=1}^{t-1} z_{s_j r_k} + \sum_{i=1}^{n} \sum_{j=1}^{m-1} z_{a_i s_j} - \sum_{i=1}^{n} z_{a_i r_t} = 0 \tag{7.11}$$

Furthermore, up to $n+m+t-1$ cells can be suppressed without changing the entropy nor the disclosure risk. But this is heavily dependent on which cells are suppressed. A suppression pattern of maximal size which does not change the rank of the system may be the following: $z_{a_1 s_m}$, $z_{a_2 s_m}$, $\cdots$, $z_{a_n s_m}$, $z_{s_1 r_t}$, $z_{s_2 r_t}$, $\cdots$, $z_{s_{m-1} r_t}$, $z_{s_m r_1}$, $z_{s_m r_2}$, $\cdots$, $z_{s_m r_t}$. Now,assume that the cell $z_{a_1 s_1 r_1}$ has a high disclosure risk which makes it necessary to increase its entropy. If cell suppression is used, a natural option is to suppress every cell in the three tables which refers to $z_{a_1 s_1 r_1}$. These suppressions will be called primary suppressions and will be $z_{a_1 s_1}$ from Table AS, $z_{a_1 r_1}$ from Table AR and $z_{s_1 r_1}$ from Table SR. Note that with these suppressions the rank of the system will decrease by 1. If $z_{a_1 s_1}$ is suppressed, the rank does not decrease because, by Expression (7.9), the suppressed cell is a linear combination

$$z_{a_1 s_1} = \sum_{k=1}^{t} z_{s_1 r_k} - \sum_{i \neq 1} z_{a_i s_1} \tag{7.12}$$

If $z_{a_1 r_1}$ is suppressed next, the rank does not change either. By Expression (7.10), we can express the suppressed cell as a linear combination of cells

which have not yet been suppressed:

$$z_{a_1 r_1} = \sum_{j=1}^{m} z_{s_j r_1} - \sum_{i \neq 1} z_{a_i r_1} \tag{7.13}$$

If $z_{s_1 r_1}$ is our third suppression, then the rank will decrease by 1, because that cell appears in Equations (7.12) and (7.13) and it is easy to see that there is no way to use the above equations to express that cell as a combination of the cells which have not yet been suppressed.

If the table is released with marginals, then a set of secondary suppressions is required to prevent primary suppressions from being computable. At this moment, it is important to choose the necessary strategy for secondary suppressions because the rank of the system and consequently the cell entropy will vary depending on what secondary suppressions are made. Let us analyze what happens with the secondary suppressions corresponding to each primary suppression:

1. Assume that, to protect $z_{a_1 s_1}$, we choose as secondary suppressions $z_{a_1 s_3}$, $z_{a_3 s_1}$ and $z_{a_3 s_3}$. Suppressing $z_{a_1 s_3}$ does not change the rank, because by Equation (7.9) we have

$$z_{a_1 s_3} = \sum_{k=1}^{t} z_{s_3 r_k} - \sum_{i \neq 1} z_{a_i s_3} \tag{7.14}$$

   where the cells on the right-hand side have not yet been suppressed. Suppressing $z_{a_3 s_1}$ does not change the rank either, because by Equation (7.8):

$$z_{a_3 s_1} = \sum_{k=1}^{t} z_{a_3 r_k} - \sum_{j \neq 1} z_{a_3 s_j} \tag{7.15}$$

   Suppression of $z_{a_3 s_3}$ causes the rank to decrease by 1, because there is no manner to express the suppressed cell as combination of other cells which have not yet been suppressed: if Equation (7.9) is used, $z_{a_1 s_3}$ is necessary but has been suppressed already; if Equation (7.8) is used,

then $z_{a_3 s_1}$ is necessary which have been suppressed; for a similar reason we cannot use Equation (7.11). Note also that, once the suppression process has started, Equation (7.11) is not very useful to obtain linear combinations, because it depends on nearly all cells.

2. Now assume that, to protect $z_{a_1 r_1}$, we choose as secondary suppressions $z_{a_1 r_3}$, $z_{a_2 r_1}$ and $z_{a_2 r_3}$. When $z_{a_1 r_3}$ is suppressed, the rank does not change, because by Equation (7.10)

$$z_{a_1 r_3} = \sum_{j=1}^{m} z_{s_j r_3} - \sum_{i \neq 1} z_{a_i r_3} \tag{7.16}$$

Suppressing $z_{a_2 r_1}$ does not change the rank either, because by Equation (7.8):

$$z_{a_2 r_1} = \sum_{j=1}^{m} z_{a_2 s_j} - \sum_{k \neq 1} z_{a_2 r_k} \tag{7.17}$$

Note, that if we now chose to suppress $z_{a_3 r_1}$, the rank would decrease by 1, because there is no manner to express it as a combination of other cells not yet suppressed (if Equation (7.10) was used, $z_{a_1 r_1}$ would be necessary and, if Equation (7.8) was used, then $z_{a_3 s_1}$ would be necessary). So, *choosing for this suppression any row in Table AR other than row 3 (which was used in Table AS) does not decrease the rank and, consequently, adds hardly any protection.* Finally, using similar arguments, it is not difficult to see that suppression of $z_{a_2 r_3}$ decreases the rank by 1 (this suppression is inevitable in order to protect $z_{a_2 r_1}$ and $z_{a_1 r_3}$).

3. As to the third primary suppression, assume that, to protect $z_{s_1 r_1}$, we choose as secondary suppressions $z_{s_1 r_2}$, $z_{s_4 r_1}$ and $z_{s_4 r_2}$. Suppressing $z_{s_1 r_2}$ does not change the rank, because, by Equation (7.10):

$$z_{s_1 r_2} = \sum_{i=1}^{n} z_{a_i r_2} - \sum_{j \neq 1} z_{s_j r_2} \tag{7.18}$$

Note that, if we chose $z_{s_1r_3}$, the rank would decrease by 1. *So, choosing for this suppression any column other than column 3 (which was used for Table AR) does not decrease the rank and adds virtually no protection.* Suppressing $z_{s_4r_1}$ does not change the rank either, because, by Equation (7.9):

$$z_{s_4r_1} = \sum_{i=1}^{n} z_{a_is_4} - \sum_{k\neq 1} z_{s_4r_k} \tag{7.19}$$

If we chose $z_{s_3r_1}$ instead of $z_{s_4r_1}$, then the rank would decrease by 1. *So, choosing for this suppression any row other than row 3 (which was used for Table AS when $z_{a_3s_1}$ was suppressed) does not decrease the rank and adds no real protection.* Finally, we have to suppress $z_{s_4r_2}$, which causes the rank to decrease by 1.

We have performed 12 suppressions altogether (primary and secondary), which decrease the rank of the system by 4. From the above, it is clear that the strategy we followed was to choose the suppression pattern which minimizes the decrease of the system rank and consequently minimizes the increase of entropy (and of protection!).

From the previous discussion, we can infer the following general result, whose proof is analogous to the proof of Proposition 4:

Proposition 5 *The entropy increase and the disclosure risk decrease in three linked tables are maximized if the suppression patterns are chosen so that the secondary suppressions are in the same columns and rows for all three tables.*

A weaker necessary condition for maximizing the decrease of disclosure risk in the case of three tables is as follows:

Proposition 6 *The entropy increase and the disclosure risk decrease in three linked tables are maximized if the suppression patterns are chosen so that the secondary suppressions are in the same row for tables with the same first variable — e.g. AS, AR—, the same column for tables with the same second variable* —e.g. *AR, SR—, and the same row for tables which share a variable in a different position* —e.g. *AS, SR—).*

Using such optimal patterns we can decrease the rank of the system of 3 linked tables by 3 more units (up to an overall rank decrease of 7.

Disclosure risk by internal intruders

Finally, a point we have to take into account here is that disclosure risk is different for different users. Let us imagine that, when solving one LP maximization and one LP minimization for $z_{a_i s_j r_k}$, we find that $995 \leq z_{a_j s_j r_k} \leq 1004$. Then, if company A is the second largest contributor to this cell with a turnover of, say, 400, then company A knows that the largest contributor (company B) has a turnover between 401 and 604. Thus, company A is able to estimate the turnover of company B within 50% of its value. However, the uncertainty of an external intruder about the turnover of company B is roughly 200% of its value: the external intruder only knows that the turnover of the largest contributor is between $\epsilon > 0$ and 1004. Therefore, for an internal intruder (respondent contributing to the cell), the measure of disclosure risk $1/H(z_{a_i s_j r_k}|\text{released tables})$ should be replaced by:

$$DR(X) = 1/H(z_{a_i s_j r_k}|\text{released tables}, \text{intruder's contribution}) \qquad (7.20)$$

.

Chapter 8

Conclusions

In this thesis, we have covered different aspects in the field of statistical database protection. Most of our attention has been devoted to performance issues of the methods used to protect statistical data. In particular, the proposals concerning information loss and disclosure risk measures represent one of the important parts of the thesis. In a background chapter, we have surveyed the state of the art of each topic by describing the main existing proposals, their properties and their weaknesses. Then our contributions are presented. Following the division of statistical data into microdata and tabular data, our results are subdivided into those concerning microdata and those concerning tabular data.

In the field of microdata protection, both information loss and disclosure risk measures were treated. Information loss measures were studied from the point of view of the legitimate user, whose interest is that the value of the released data be maximal. Disclosure risk measures for microdata were studied from the point of view of the respondent who supplies information about himself to the statistical office. Since this information may contain confidential data, the respondent's interest is that the supplied data cannot be linked with her/him. Generic information loss and disclosure risk measures were proposed instead of method-specific or data use-specific measures. Method-independent measures are important to allow comparison of several

protection methods. Data use-independent information loss measures may not be fully satisfactory for the data user, but at least they provide some guidance to the data protector when the planned uses of the released microdata are not known at the moment of protecting them.

Based on the proposed information loss and disclosure risk measures, a score has been constructed and used to rank the performance of the main SDC methods for microdata.

After comparing method performance, we have presented a complexity analysis of one of the best-performing and most used protection methods for numerical microdata, namely microaggregation. This analysis proves the NP-hardness of optimal microaggregation in the multivariate case, which justifies the use of heuristic approaches to implement that method.

In the field of tabular data protection, most of the attention has been devoted to the security of respondent contributions. In particular, sensitivity rules currently used for *a priori* disclosure risk assessment have been analyzed and their shortcomings have been highlighted. A new sensitivity rule has been proposed as a better alternative. In addition to that, we have proposed *a posteriori* disclosure risk assessment for tabular data as a complement to sensitivity rules. In other words, we have proposed to measure the risk of disclosure not only before the application of disclosure control methods but also after the application of these methods, in order to guarantee a sufficient level of security to the cell contributions. An *a posteriori* measure based on conditional Shannon's entropy has been proposed and we have shown how it can be applied to different disclosure control methods. For the case of linked tables, guidance has been given on how to choose the best protection strategy based on our *a posteriori* measure.

8.1 Results of this thesis

The following results have been presented in this thesis:

- For microdata, generic information loss measures have been proposed

instead of method specific measures. The proposed measures characterize information contained in the data set. These measures are based on the discrepancy between several statistics in the original and the masked microdata sets. These statistics include averages, covariances, correlations, correlations between variables and principal components, commonality between variables and first principal components, and factor score coefficient matrices. In addition, discrepancies between original data themselves and the corresponding masked data are also considered. Discrepancies can be quantified in three different ways: as a mean square error, as a mean absolute error and as a mean variation.

- Also for microdata, generic measures for disclosure risk based on record linkage techniques have been proposed. As in the case of information loss, these measures allow different SDC techniques to be compared and may be applicable to several different methods.

- Using a subset of the proposed measures for information loss and disclosure risk, a generic score for SDC methods for numerical microdata has been constructed. This score is an overall measure which takes into account both the impact of applying of protection methods to the data and the risk of disclosure. The most used disclosure protection methods were ranked by this score. As a result, rank swapping turned out to be the best method with relatively low information loss and disclosure risk, closely followed by multivariate microaggregation. To our best knowledge, this is the first comparison of SDC methods for microdata proposed so far.

- In Chapter 5, the complexity analysis of microaggregation has been presented. It has been shown that optimal microaggregation is NP-hard. This result justifies the current use of heuristic approaches.

- In the field of tabular data protection, we have presented a critique of current sensitivity rules used for *a priori* disclosure risk assessment.

This critique reveals the flaws of sensitivity rules. In particular, it has been shown that, for any of these rules, a closer upper bound on the contribution of the largest respondent may be reached for a cell declared non-sensitive than for a cell declared sensitive. To overcome the detected shortcomings, a new approach in measuring disclosure risk for tables was proposed which is based on the entropy of the relative contributions to a cell.

- Finally, for tabular data, we propose to complement *a priori* disclosure risk assessment with *a posteriori* disclosure risk assessment. If the disclosure risk measured on the protected table is still high, then protection methods must be applied again until the desired level of security is attained. As an *a posteriori* disclosure risk measure, the reciprocal of the conditional entropy, conditioned to the knowledge of the intruder, has been proposed. In Chapter 7, we have shown how this measure can be applied to different SDC methods like cell suppression, rounding, table redesign and also to a set of linked tables. In case of using cell suppression in linked tables, this measure turns out to be useful to detect suppression patterns which offer more protection than others do. The strategy of choosing the best candidates for secondary suppressions has been outlined in the last chapter.

8.2 Future research

In microdata protection, future research will be devoted to

- Designing polynomial algorithms for exact optimal univariate microaggregation which are more efficient than the shortest-path approach described in [35]. Although polynomial, this latter approach is not practical to protect large files (*e.g.* one million records) on a standard PC.

- Looking at alternative formulations for microaggregation, namely those based on minimizing the intra-group gaps rather than the intra-group

sum of squares (see [59, 60]). Assess the complexity and performance of these alternative formulations and design polynomial algorithms whenever possible.

- Develop new record linkage methods for disclosure risk assessment when the intruder's external data set and the protected data set do not have common variables.

- Extend microaggregation for categorical data, following the research line initiated in [23].

In the field of tabular data, a major challenge is to take the intruder's side knowledge into account when performing *a posteriori* disclosure risk assessment. In fact, several disclosure risk scenarios should be considered and compared. Some examples of side information which define different disclosure scenarios are as follows:

- The intruder knows the identity of the largest contributor to a sensitive cell.

- The intruder knows the identities and the ranking of all contributors to a sensitive cell.

- The intruder has a fuzzy knowledge of the ranking of contributors to a sensitive cell.

- The intruder is a contributor to a sensitive cell and may collude with other contributors who reveal their contributions to him.

- etc.

Appendix A

Appendix. Probabilistic record linkage

We recall in this appendix the basic operational principles of the probabilistic record linkage method described in [41]. In Section A.1, concepts necessary to understand record-linkage techniques are presented. Section A.2 describes the parameter estimation methodology and Section A.3 presents the matching algorithm.

A.1 Theoretical concepts

A.1.1 Background

Consider two computer files, A and B, consisting of records taken from a population. Each file consists of a number of fields, or "components", and a number of records, or "observations". Typically, each observation corresponds to a member of the population and the fields are attributes identifying the individual observation, such as name, address, age, and sex. The objective of the record linkage or matching process is to identify and link the observations on each file that correspond to the same individual. The records are taken to contain no unique identifiers that would make the matching operation trivial. That is, the individual fields are all subject to error. We can

define two disjoint sets M and U formed from the cross-product of A with B, the set $A \times B$. A record pair is a member of set M, if that pair represents a true match. Otherwise, it is a member of U. The record-linkage process attempts to classify each record pair as belonging to either M or U.

A.1.2 Weights

The components (fields) in common between the two files are useful for matching. Not all components, however, contain an equal amount of information, and error rates vary. For example, a field such as sex only has two value states and consequently could not impart enough information to identify a match uniquely. Conversely, a field such as surname imparts much more information, but it may frequently be reported or transcribed (keyed) incorrectly.

Weights are used to measure the contribution of each field to the probability of making an accurate classification. In [54], the concept of weights based on probabilities of chance agreement of component value states was discussed. In [29], these concepts were extended into a more rigorous mathematical treatment of the record-linkage process. Their definition of weights takes into account the error probabilities for each field by using a log-likelihood ratio. Let $m_i = Pr\{\text{component } i \text{ agrees}|r \in M\}$ and $u_i = Pr\{\text{component } i \text{ agrees}|r \in U\}$ for all record pairs r. If, for a given record pair, component i agrees (matches), then the weight for component i is $w_i = \log_2(m_i/u_i)$. If component i disagrees, then the weight $w_i = \log_2((1 - m_i)/(1 - u_i))$.

A.1.3 Decision procedure

For any record pair, a composite weight can be computed by summing the individual component weights. Since $m_i > u_i$ in most cases, fields that agree make a positive contribution to this sum, whereas fields that disagree make a negative contribution. A most significant concept advanced by [29] is an optimal decision procedure for record linkage. For this procedure, three states

are defined. A record pair is classified as a match if the composite weight is above a threshold value, a nonmatch if the composite weight is below another threshold value, and an undecided situation if the composite weight is between these two thresholds. The threshold values can be calculated (see Section A.2.3) given the acceptable probability of false matches (the probability that a record pair is classified as a match when the records do not represent the same individual) and the probability of false nonmatches.

A.1.4 Estimation of the u_i

Values for the m_i and the u_i probabilities must be estimated for each pair of files to be matched. Estimating the u_i (the probability that a component agrees given U) is simplified by the fact that the cardinality of the set U (denoted by $|U|$) is generally much greater than that of M. For two files, both of equal size, F, $|M| = pF$, where p is the proportion of matched pairs, and $|U| = F^2 - pF$. Consequently, estimates for the u probabilities can be obtained by ignoring the contribution from M and considering only the probability of chance agreement of the component i. Usually this can be estimated from a sample of pairs rather than from all pairs. Estimating the m_i probabilities (the probability that a component agrees given M) is more difficult. Conditioning on M presupposes an a priori knowledge o correctly matched pairs. This could be obtained by a prelinked sample of the population. If such a sample were obtained clerically, much expense would be involved and the error rates for the clerical operation might be too high to permit accurate parameter estimation. One solution is blocking and using a latent trait model.

A.1.5 Blocking

For files of average size $|A \times B|$ is too great to consider all possible record pairs. Since there are many more record pairs in U than in M and 2^n possible comparison configurations involving n fields, drawing record pairs at random

would require a sample size approaching all record pairs (for typical applications) to obtain sufficient information about the relatively rare M cases. The two files can be partitioned into mutually exclusive and exhaustive blocks designed to increase the proportion of matched pairs observed while decreasing the number of record pairs to compare. Comparisons are restricted to record pairs within each block. Consequently, blocking is important for the actual matching and for parameter estimation activities. Blocking is generally implemented by means of sorting the two files on one or more variables. For example, if both files were sorted by zip code, the pairs to be compared would only be drawn from those records where zip codes agree. Record pairs disagreeing on zip code would not be considered and hence would be automatically classified as nonmatches (elements of U). To be effective at enriching the M cases, such blocking variables must contain a large number of value states that are fairly uniformly distributed and such variables must have a low probability of reporting error (i.e. a high weight). Blocking is a trade-off between computation cost (examining too many record pairs) and false nonmatch rates (classifying record pairs as nonmatches because the records are not members of the same block). Multiple-pass matching techniques using independent blocking variables for each run can minimize the effect of errors in a set of blocking variables. R.P. Kelley has developed an algorithm that may assist in choosing the best blocking scheme in light of these trade-offs (see [43]).

A.2 Parameter estimation methodology

A.2.1 Comparison configuration frequencies

This section discusses the methodology used to estimate the m_i probabilities. Information about the comparison configurations observed is provided by the matching software itself, which tabulates frequencies for all 2^n possible patterns of agreement and disagreement on n fields. To increase the

proportion of matched pairs examined, these tabulations are performed using just those record pairs in which both observations come from the same block. Fortunately, the m_i probabilities may reasonably be expected to be independent of the blocking schemes chosen as long as errors in the blocking variables do not exclude an excessive number of matched records from the tabulations. This is consistent, however, with the goal of choosing blocking variables with low reporting error rates. The independence of the m_i, probabilities to choices in blocking and the effect of errors in the blocking variables to the final estimates are yet to be determined. Given frequencies for all possible agreements and disagreements, the m_i probabilities can be estimated using any of several procedures. The EM algorithm described here is the most effective of those developed and tested.

A.2.2 The EM algorithm

Given n fields and a sample of N record pairs drawn from $A \times B$, let $\gamma_i^j = 1$ if field i agrees for record pair j, let $\gamma_i^j = 0$ if field i disagrees for record pair j, for $i = 1, \cdots n$ and $j = 1, \cdots, N$. Further, let γ^j be the vector of ones and zeros showing field agreements and disagreements for the j-th pair in the sample, and let γ be the vector containing all of the γ^j. The m_i and u_i, probabilities can be defined as $m_i = Pr\{\gamma_i^j = 1 | r_j \in M\}$ and $u_i = Pr\{\gamma_i^j = 1 | r_j \in U\}$ for a randomly selected record pair r_j and $i = 1, 2 \cdots, n$. Define p as the proportion of matched pairs equal to $|M|/|M \cup U|$. The elements of $M \cup U$ (i.e., all record pairs r_j are distributed according to a finite mixture with the unknown parameters $\Phi = (m, u, p)$. We will use an EM algorithm to estimate these parameters; in particular, the m vector is of the greatest interest. Notation is consistent with that used in [16].

Let x be the complete data vector equal to $< \gamma, g >$, where $g_j = (1, 0)$ iff $r_j \in M$ and $g_i = (0, 1)$ iff $r_j \in U$. Then, the complete data log-likelihood is

$$\ln f(x|\Phi) = \sum_{j=1}^{N} g_j(\ln Pr\{\gamma^j|M\}, \ln Pr\{\gamma^j|U\})^T + \\ + \sum_{j=1}^{N} g_j * (\ln p, \ln(1-p))^T$$

Although it is quite reasonable to expect the empirical frequencies to belie independence, since errors in one component will often induce errors in another, such departures will not likely disturb the ordering by composite weights of the 2^n configurations. Consequently, we will assume an independence model:

$$Pr\{\gamma^j|M\} = \prod_{i=1}^{n} m_i^{\gamma_i^j}(1-m_i)^{1-\gamma_i^j} \tag{A.1}$$

and

$$Pr\{\gamma^j|U\} = \prod_{i=1}^{n} u_i^{\gamma_i^j}(1-u_i)^{1-\gamma_i^j} \tag{A.2}$$

Our application of the EM algorithm begins with estimates of the unknown parameters $(\hat{m}, \hat{u}, \hat{p})$ and consists of iterative applications of the expectation (E) and maximization (M) steps until the desired precision is obtained. The algorithm is not particularly sensitive to starting values, and the initial $\hat{m}$ values can be guessed. It is important, however, that the $\hat{m}$ values be greater than their corresponding $\hat{u}$ values. For Tampa, 0.9 was used for all of the initial $\hat{m}$. Estimation of $\hat{u}$ is discussed in Section A.1.4. For the E step, replace g_j with $(\hat{g}_m(\gamma^j), \hat{g}_u(\gamma^j))$, where

$$\hat{g}_m(\gamma^j) = \frac{\hat{p}\prod_{i=1}^{n} \hat{m}_i^{\gamma_i^j}(1-\hat{m}_i)^{1-\gamma_i^j}}{\hat{p}\prod_{i=1}^{n} \hat{m}_i^{\gamma_i^j}(1-\hat{m}_i)^{1-\gamma_i^j} + (1-\hat{p})\prod_{i=1}^{n} \hat{u}_i^{\gamma_i^j}(1-\hat{u}_i)^{1-\gamma_i^j}}$$

$\hat{g}_u(\gamma^j)$ can be derived similarly. For the M the complete data log-likelihood can be separated into three maximization problems. Setting the partial

derivatives equal to 0 and solving for $\hat{m}_i$, yields

$$\hat{m}_i = \sum_{j=1}^{N} [\hat{g}_m(\gamma^j)\gamma_i^j] / \sum_{j=1}^{N} [\hat{g}_m(\gamma^j)]$$

Further, the matrix of second partial derivatives can be shown to be negative-definite. In practice, we store frequency counts, $f(\gamma^j)$ for each of the possible 2^n patterns of γ^j. These counts are obtained as follows: each file is partitioned into blocks by means of the blocking variables. For each block, all record pairs in the block are examined. For each record pair, the comparison vector γ^j is computed and 1 is added to the frequency count for that particular configuration. There are 2^n such counters. The counters are not reset after a block is processed, but represent the number of observations of each configuration over all blocks. Both estimation and actual matching are accomplished using the same blocks. The EM algorithm is run once using these frequencies. The E step computes $\hat{g}_m(\gamma^j)$ for each of the 2^n patterns. This can be done without examining the individual observations, since the frequencies are a sufficient statistic for the M step. By replacing the individual observations with the frequencies, we obtain

$$\hat{m}_i = \sum_{j=1}^{2^n} [\hat{g}_m(\gamma^j)\gamma_i^j f(\gamma^j)] / \sum_{j=1}^{2^n} [\hat{g}_m(\gamma^j) f(\gamma^j)]$$

The arguments for the u probabilities are similar. Finally, the proportion of matched pairs p can be estimated by

$$\hat{p} = \sum_{j=1}^{N} [\hat{g}_m(\gamma^j)]/N = \sum_{j=1}^{2^n} [\hat{g}_m(\gamma^j) f(\gamma^j)] / \sum_{i=1}^{2^n} f(\gamma^j)$$

It must be remembered that the frequencies used for the p, m, and u estimates were obtained from record pairs within blocks and represent an accumulation over all blocks. Since blocking greatly reduces the number of nonmatched pairs observed and because blocking selects record pairs that

are likely to match, the u probability estimates obtained using blocked data will be biased. Consequently, the u probabilities must be computed directly on unblocked data, as explained in Section A.1.4, and the EM algorithm is only used to compute the m probabilities, where blocking enriches the number of matched pairs observed while avoiding comparisons on relatively large numbers of unmatched pairs. The EM algorithm is highly stable and the least sensitive to the starting values of any of the methods studied. The algorithm is very simple to implement, and the probabilities will always be within bounds.

A.2.3 Calculation of threshold values

An algorithm in the matcher determines the threshold weights as follows. There are 2^n possible configurations of agreement and disagreement of n components. These configurations can be ordered by the composite weight (the sum of the individual weights, w_i, for each component). After ordering the composite weights, the sum of $Pr(|M)$ and 1 minus the sum of $Pr(|U)$ can be calculated [see (A.1) and (A.2), Section A.2.2]. The maximum weight for a nonmatch decision is the weight of the configuration where the sum of $Pr(|M)$ does not exceed the desired probability that a matched pair should be classified as unmatched. The minimum weight for a match decision is the weight of the configuration where 1 minus the sum of $Pr(|U)$ does not exceed the desired probability that an unmatched pair should be classified as matched. Weights between these two thresholds are undecided cases. For applications such as census matching, with approximately 10 components, this technique is computationally feasible. Unpublished experimentation has been performed by sampling component configurations for problems having many components, but the large number of cells makes it difficult to obtain a sufficient number of observations in each cell, so sampling error is not an overpowering factor.

A.3 Matching algorithm

This section describes the basic operation of the matcher. Before matching, fields such as house address should be separated into components and spellings should be standardized. Both files must be sorted by the blocking variables.

A.3.1 Composite weight calculation

The matcher processes one block at a time, building a matrix (C) containing the composite weights for all pairs within the block being processed. The composite weights are computed by summing the individual weights for agreement or disagreement on each field (see Section A.1.2). The simple agreement/disagreement dichotomy modelled by the theory is too simplistic for noncategorical fields. For example, character strings are compared using an information-theoretic character comparison algorithm that provides for random insertion, deletion, replacement, and transposition of characters. The weight assigned for such comparisons is prorated according to a measure of similarity between character fields (see [40]). If two character fields match exactly, the full weight for agreement is assigned to the comparison. If they disagree slightly, however, it would be wrong to assign the disagreement weight. Consequently, the weight assigned for the comparison will be somewhat less than the full agreement weight. Similarly, weights for integer or continuous variables such as age can be prorated proportionally to the ratio of the difference and the minimum of the two values being compared (delta percent). For example, if age disagrees by one year in an 80-year-old man, it is less serious a mismatch than for a 1-year-old baby.

A.3.2 Assignment

After the matrix containing the composite weights for all pairs within the block is constructed (C_{ij} in the following), the records can be paired up (assigned). One record on file A can be assigned to one and only one record on

file B, and vice versa. We wish to choose an assignment scheme that maximizes the sum of the composite weights of the assigned record pairs. This is a degenerate transportation problem known as the linear sum assignment problem, which can be solved by a simple method requiring only addition and subtraction. The use of such a linear programming model to provide the assignments represents an advance over previous ad hoc assignment methods. The problem can be formulated as follows: Maximize

$$Z = \sum_{i=1}^{k} \sum_{j=1}^{k} C_{ij} X_{ij}$$

subject to

$$\sum_{j=1}^{k} X_{ij} = 1, \quad i = 1, 2, \cdots, k$$

and

$$\sum_{i=1}^{k} X_{ij} = 1, \quad i = 1, 2, \cdots, k$$

where C_{ij}, is the cost (weight) of matching record i on file A with record j on file B, X_{ij} is an indicator variable that is 1 if record i is assigned to record j and 0 if i is not assigned to j, k_a is the number of records in the block being processed from file A, k_b is the number of records in the block being processed from file B, and $k = maximum(k_a, k_b)$. If $k_a \neq k_b$, the matrix is made square (with dimension k) by inserting entries whose values are large negative numbers (less than any possible composite weight). This prevents these entries from being assigned. An excellent discussion of the theory of assignment problems can be found in [10]. Once an optimal assignment vector is obtained, an assigned pair can be classified as a match if the composite weight is greater than the Fellegi-Sunter threshold value. After all assigned pairs in the block are processed, the records for the next block can be read.

A.3.3 Duplicates

Duplicates can be detected by examining each, row or column of the assignment matrix. If more than one entry is above the cutoff threshold, then there is a possibility of a duplicate. Two similar records on both files would probably be two separate individuals (a father and his son, for example), but two similar records on only one file would probably be a duplicate.

A.3.4 File preparation

To match any file, free-form information must be standardized. This is especially true of fields such as street address and person name. The components of the name should be separated into individual fields (given name, middle initial, and surname). This is much more effective and accurate than trying to match an entire name as a single character string. For street address, the various components of the address should be placed in individual fields and the spellings of common abbreviations (such as BD, $BLVD$) should be standardized. Punctuation should be removed from the fields. The technique of SOUNDEX encoding [45] is a method of transforming a persons name into some code that tends to bring together all variants of the same name. For example, $Smith$ and $Smythe$ would both be coded as $S530$. Surname is often an important blocking variable. To maximize the chance that similarly spelled surnames reside in the same block, the SOUNDEX system can be used to code the names, and the SOUNDEX code can be used as a blocking variable. There are better encoding schemes than SOUNDEX, but SOUNDEX with relatively few states and poor discrimination helps ensure that misspelled names receive the same code. SOUNDEX is not recommended for matching nonblocking variables, since nonphonetic errors result in different codes and different names may receive the same code.

Our contributions

[1] J. Domingo-Ferrer, A. Oganian and V.Torra, "Information-theoretic disclosure risk measures in statistical disclosure control of tabular data", in *Proceedings of the 14th International Conference on Scientific and Statistical Database Management (SSDBM'2002)*, (J. Kennedy, ed.), Los Alamitos CA: IEEE Computer Society, 2002, pp. 227-231.

[2] J.Domingo-Ferrer, J.M. Mateo-Sanz, A.Oganian and A.Torres, "On the security of microaggregation with individual ranking: analitical attacks", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5 (to appear, 2002).

[3] J. Domingo-Ferrer and A. Oganian, "Medidas basadas en teoría de la información en el secreto estadistico de datos tabulares", in *Actas de la VII Reunión Española sobre Criptología y Seguridad de la Información*, eds. S. González and C. Martínez, Oviedo: Servicio de Publicaciones de la Universidad de Oviedo, 2002, pp. 207-218.

[4] A. Oganian and J. Domingo-Ferrer, "On the complexity of optimal microaggregation for statistical disclosure control", *Statistical Journal of the United Nations Economic Commission for Europe*, vol. 18, no. 4, 2001. Preliminary version in *2nd Joint UNECE Work Session on SDC*, Skopje, Macedonia, March 2001.

[5] J. Domingo-Ferrer, V.Torra, J.M. Mateo-Sanz and A. Oganian, "Disclosure risk assesment in statistical data protection", in *International Conference on Computational and Applied Mathematics-ICCAM'2002*, 2002.

Journal version submitted to the *Journal of Computational and Applied Mathematics*.

Other references

[1] N. Anwar, *Micro-aggregation - the small aggregates method*, Research Report, Luxembourg: Eurostat, 1993.

[2] B. C. Arnold, N. Balakrishnan and H. N. Nagaraja, *A First Course in Order Statistics*. New York: Wiley, 1993.

[3] Y. Baeyens and D. Defays, "Estimation of variance loss following microaggregation by the individual ranking method", in *Proceedings of Statistical Data Protection'98*, Luxembourg: Office for Official Publications of the European Communities, pp. 101-108, 1999.

[4] E. Boros and P.L. Hammer, "On clustering problems with connected optima in Euclidean spaces", *Discrete Mathematics*, vol. 75, pp. 81-88, 1989.

[5] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, New York: Wiley, 1992.

[6] R. Brand, "Microdata protection through noise", *Lecture Notes in Computer Science*, vol. 2316, pp. 97-116. Vol. *Inference Control in Statistical Databases*, (J. Domingo-Ferrer, ed.), Berlin: Springer, 2002.

[7] P. Brucker, "On the complexity of clustering problems", in *Optimization and Operations Research*, (R. Hehn, B. Korte and W. Oettli, eds.), Berlin: Springer-Verlag, pp.45-54, 1977.

[8] R. E. Burkard and U. Derigs, "Assignment and matching problems: solution methods with FORTRAN-programs", in *Lecture Notes in Economics and Mathematical Systems*, vol. 184, pp. 1-11, New York: Springer-Verlag, 1981.

[9] L. Buzzigoli and A. Giusti, "An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals", in *Proceedings of the Conference on Statistical Data Protection*, Luxemburg: Eurostat, pp. 131-147, 1999.

[10] L. Cooper and D.Steinberg *Methods and Applications of Linear Programming*, Philadelphia: W.B.Saunders, 1974.

[11] L. H. Cox, "Disclosure risk for tabular economic data", in *Confidentiality, Disclosure and Data Access*, (P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, eds.), Amsterdam: North-Holland, pp. 167-183, 2001.

[12] Crystal Ball, 2002, `http://www.cbpro.com`

[13] R. A. Dandekar, J. Domingo-Ferrer and F. Sebé, "LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection", *Lecture Notes in Computer Science*, vol. 2316, pp. 175-186. Vol. *Inference Control in Statistical Databases*, (J. Domingo-Ferrer, ed.), Berlin: Springer, 2002.

[14] D. Defays and N. Anwar, "Micro-aggregation: a generic method", in *Proceedings of the 2nd International Symposium on Statistical Confidentiality*, Luxembourg: Office for Official Publications of the European Communities, pp. 69-78, 1995.

[15] D. Defays and P. Nanopoulos, "Panels of enterprises and confidentiality: the small aggregates method", in *Proceedings of the 92 Symposium on Design and Analysis of Longitudinal Surveys*, Ottawa: Statistics Canada, pp. 195-204, 1993.

[16] A. P. Dempster, N. M. Laird and D. B. Rubin "Maximum likelihood from incomplete data via the EM Algorithm", *Journal of the Royal Statistical Society*, vol. 39, pp. 1-38, 1977.

[17] D. E. Denning, *Cryptography and Data Security*, Reading, MA: Addison-Wesley, 1982.

[18] J. Domingo-Ferrer, J. M. Mateo-Sanz and V. Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk", *Proc. of ETK-NTTS'2001.* Luxembourg: Eurostat, pp. 807-825, 2001.

[19] J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure and Data Access*, (P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, eds.), Amsterdam: North-Holland, pp. 111-133, 2001.

[20] J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure and Data Access*, (P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, eds.), Amsterdam: North-Holland, pp. 111-133, 2001.

[21] J. Domingo-Ferrer and J. M. Mateo-Sanz, "On resampling for statistical confidentiality in contingency tables", *Computers and Mathematics with Applications*, vol. 38, pp. 13-32, 1999.

[22] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control", *IEEE Transactions on Knowledge and Data Engineering,* vol. 14, pp. 189-201, 2002.

[23] J. Domingo-Ferrer and V. Torra, "Towards fuzzy c-means based microaggregation", in *Soft Methods in Probability, Statistics and Data Analysis*, (P. Grzegorzewski, O. Hryniewicz and M. A. Gil, eds.), Series "Advances in Soft Computing" Heidelberg: Physica-Verlag, 2002, pp. 289-294.

[24] G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman and S. F. Roehrig, "Disclosure limitation methods and information loss for tabular data", in *Confidentiality, Disclosure and Data Access*, (P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, eds.), Amsterdam: North-Holland, pp. 135-166, 2001.

[25] G.T. Duncan and R.W. Pearson, "Enhancing access to microdata while protecting confidentiality: prospects for the future", *Statistical Science*, vol. 6, pp. 219-39, 1991.

[26] M. E. Dyer and A. M. Frieze, "Planar 3DM is NP-complete", *Journal of Algorithms*, vol. 7, pp. 174-184, 1986.

[27] A. W. F. Edwards and L. L. Cavalli-Sforza, "A method for cluster analysis", *Biometrics*, vol. 21, pp. 362-375, 1965.

[28] M. J. Elliot, C. J. Skinner and A. Dale, "Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk", *Research in Official Statistics*, vol. 1, no. 2, pp.53-67, 1998.

[29] I. P. Fellegi and A. B. Sunter, "A theory for record linkage", *Journal of the American Statistical Association*, vol. 64, pp. 1183-1210, 1969.

[30] F. Felsö, J. Theeuwes and G. G. Wagner, "Disclosure limitation methods in use: Results of a survey", in *Confidentiality, Disclosure and Data Access*, (P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, eds.), Amsterdam: North-Holland, pp. 17-42, 2001.

[31] S. E. Fienberg, U. E. Makov and R. J. Steele, "Disclosure limitation using perturbation and related methods for categorical data", *Journal of Official Statistics*, vol. 14, pp. 485-512, 1998.

[32] S. Gießing, "Nonperturbative disclosure control methods for tabular data", in *Confidentiality, Disclosure and Data Access*, (P. Doyle, J. Lane,

J. Theeuwes and L. Zayatz, eds.), Amsterdam: North-Holland, pp. 185-213, 2001.

[33] R. Gopal, P. Goes and R. Garfinkel, "Confidentiality via Camouflage: the CVC approach to database query", in *Statistical Data Protection*, (J. Domingo-Ferrer, ed.), Luxembourg: Office for Official Publications of the European Communities, pp. 19-28, 1998.

[34] J. Holvast, "Statistical dissemination, confidentiality and disclosure", in *Proceedings of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality*. Luxembourg: Eurostat, pp. 191-207, 1999.

[35] S. L. Hansen and S. Mukherjee, "A polynomial algorithm for univariate optimal microaggregation", *IEEE Trans. on Knowledge and Data Engineering*, 2002 (to appear).

[36] G.R. Heer, "A Bootstrap procedure to preserve statistical confidentiality in contingency tables", in *Proceedings of the International Seminar on Statistical Confidentiality*, (D. Lievesley, ed.), Luxembourg: Office for official publications of the european communities, pp.261-271, 1993

[37] A. Hundepool, A. van de Wetering, L. Franconi, A. Capobianchi, P.-P. de Wolf, *μ-Argus User's Manual Version 3.1.* Voorburg: Statistics Netherlands, Jan. 2002.

[38] A. Hundepool and L. Willenborg, "ARGUS: Software from the SDC Project", in *Proceedings of Joint UNECE-Eurostat Work Session on Statistical Data Confidentiality* Luxemburg: UNECE-Eurostat, pp. 87-98, 1999.

[39] A. Hundepool, "The CASC project", *Lecture Notes in Computer Science*, vol. 2316, pp. 172-180. Vol. *Inference Control in Statistical Databases*, (J. Domingo-Ferrer, ed.), Berlin: Springer, 2002.

[40] M. A. Jaro, "UNIMATCH: a record linkage system user's manual", Washington, DC: U.S. Bureau of the Census, 1978.

[41] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, vol. 84, pp. 414-420, 1989.

[42] D.S. Johnson, "The NP-completeness column: an ongoing guide", *Journal of Algorithms*, vol. 7, pp. 174-184, 1986.

[43] R.P. Kelly, "Blocking considerations for record linkage under conditions of uncertainty", in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 602-605, 1984.

[44] J.J. Kim, "A method for limiting disclosure in microdata based on random noise and transformation", in *Proceedings of the ASA Section on Survey Research Methodology*, Alexandria VA:American Statistical Association, pp.303-308, 1986.

[45] D.E. Knuth, *The Art of Computer Programming, Volume 3: Sorting and Searching*, Reading MA: Addison-Wesley, 1973.

[46] C. K. Liew, U. J. Choi and C. J. Liew, "Data distortion by probability distribution", *ACM Transactions on Database Systems*, vol. 10, pp. 395-411, 1985.

[47] R.J.A. Little, "Statistical Analysis of Masked Data", *Journal of Official Statistics*, 9, pp. 407-426, 1993.

[48] T. Luige and J. Meliskova, "Confidentiality practices in the transition countries", in *Proceedings of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality*. Luxembourg: Eurostat, pp. 287-319, 1999.

[49] J. M. Mateo-Sanz and J. Domingo-Ferrer, "A method for data-oriented multivariate microaggregation", in *Proceedings of Statistical Data Protection'98*. Luxembourg: Office for Official Publications of the European Communities, pp. 89-99, 1999.

[50] J. M. Mateo-Sanz and J. Domingo-Ferrer, "A comparative study of microaggregation methods", *Qüestiió*, vol. 22, no. 3, pp. 511-526, 1998.

[51] N. Megiddo and K.J. Supowit, "On the complexity of some common geometric location problems", *SIAM Journal of Computing*, vol. 13, pp. 182-196, 1984.

[52] N. Megiddo and A.Tamir, "On the complexity of locating linear facilities in the plane", *Operations Research Letters*, vol. 13, pp. 194-197, 1982.

[53] R. Moore, "Controlled data swapping techniques for masking public use microdata sets", U. S. Bureau of the Census, 1996 (unpublished manuscript).

[54] H. B. Newcombe and J. M. Kennedy, "Record linkage", *Communications of the ACM*, vol. 5, pp. 563-566, 1962.

[55] D. Pagliuca and G. Seri, *Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey*, Esprit SDC Project, Deliverable MI-3/D2, 1999.

[56] U. Pferschy, R. Rudolf, G.J.Woeginger, "Some geometric clustering problems", *Nordic Journal of Computing*, vol. 1, no. 2, pp. 246-263, 1994.

[57] D. Robertson and R. Ethier, "Cell suppression: Theory and experience", *Lecture Notes in Computer Science*, vol. 2316, pp. 9-21. Vol. *Inference Control in Statistical Databases*,(J. Domingo-Ferrer, ed.), Berlin: Springer, 2002.

[58] P. Rosentiehl and R.E. Tarjan, "Rectilinear planar layout of planar graphs and bipolar orientations", *Discrete Computational Geometry*, vol. 1, pp. 343-353, 1986.

[59] G. Sande, "Methods for data-directed microaggregation in one or more dimensions", in *Federal Committee on Statistical Methodology Research*

Conference, Arlington VA, Nov. 14-16, 2001.
`http://www.fcsm.gov/01_papers/index.html`

[60] G. Sande, "Exact and approximate methods for data-directed microaggregation in one or more dimensions", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5 (to appear, 2002).

[61] T. Schreiber, "Clustering for data reduction and approximation", *Computer Graphics Geometry*, vol. 2, no. 2, 2000. `http://eos.wdcb.ru/cgg/1/1.htm`

[62] C. E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, pp. 379-423, 623-656, July and Oct. 1948.

[63] C. Skinner, C. Marsh, S. Openshaw and C. Wymer, "Disclosure control for census microdata" in *Journal of Official Statistics*, 10, pp. 31-51, 1994

[64] K.J. Supowit, *Topics in Computational Geometry*, Ph. D. Thesis, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Report UIUCDCS-R-81-1062, 1981

[65] G. Sullivan and W.A. Fuller, "The use of measurement error to avoid disclosure", in *Proceedings of the ASA Section on Survey Research Methodology*, Alexandria, VA: American Statistical Association, pp. 802-807, 1989.

[66] P. Tendik and N. Matloff, "A modified random perturbation method for database security", *ACM Trans. on Database Systems*, vol. 19, no. 1, pp. 47-63, 1994.

[67] U.S. Bureau of the Census, "Record Linkage Software: User Documentation". Available from U.S. Bureau of the Census, 2000.

[68] L. Willenborg and T. De Waal *Statistical Disclosure Control in Practice*, Lecture notes in statistics 111, New York: Springer-Verlag, 1996.

[69] L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*. New York: Springer-Verlag, 2001.

[70] W. Winkler, "Re-identification methods for evaluating the confidentiality of analytically valid microdata", in *Statistical Data Protection*, Luxembourg: Office for Official Publications of the European Communities, 1999. Journal version in *Research in Official Statistics*, vol. 1, no. 2, pp. 50-69, 1998.