Disclosure Risk Measures for Microdata

Traian Marius Truta[1], Farshad Fotouhi[2], Daniel Barth-Jones[3]
Department of Computer Science (1, 2) and Center for Healthcare Effectiveness (3)
Wayne State University, Detroit, MI, 48202, USA
mtruta@cs.wayne.edu, fotouhi@cs.wayne.edu, dbjones@med.wayne.edu

Abstract

In this paper, we define several disclosure risk measures for microdata. We will analyze disclosure risk based on the disclosure control techniques applied to initial microdata. Disclosure Control is the discipline concerned with the modification of data containing confidential information about individual entities, such as persons, households, businesses, etc. in order to prevent third parties working with these data from recognizing entities in the data and thereby disclosing information about these entities. In very broad terms, disclosure risk is the risk that a given form of disclosure will occur if a masked microdataset is released. Microdata represents a series of records, each record containing information on an individual unit. The disclosure risk measures presented in this paper are validated in our experiments.

1. Introduction

Microdata represents a series of records, where each record contains information on an individual unit such as a person, a firm, an institution, etc. [17]. Microdata can be represented as a single data matrix where the rows correspond to the units (individual units) and the columns to the attributes (as name, address, income, sex, etc.). Due to existing regulations in various areas, microdata can be released for use by the third party only after the owner of the data has masked the data to limit the possibility of disclosure. Typically, names and other identifying information are removed from original records before being released for research use. We will call the final microdata as *masked* or *released microdata* [3].

Disclosure Control is the discipline concerned with the modification of data containing confidential information about individual entities, such as persons, households, businesses, etc., in order to prevent third parties working with these data from recognizing entities in the data and thereby disclosing information about these entities [2, 16].

In very broad terms, *disclosure risk* is the risk that a given form of disclosure will be encountered if a masked microdata is released. *Information loss* is the quantity of information, which exists in the initial microdata and because of disclosure control methods does not occur in masked microdata [17].

The problem of quantifying disclosure risk is a difficult one because disclosure of confidential information usually occurs only if the intruder has some external information and the owner of the data cannot possibly know or anticipate this information. Therefore, we need to make assumptions about this knowledge to predict the disclosure risk. Unfortunately, the assumptions we are forced to make are sometimes not accurate with a given masked microdataset.

The masked microdata is used for statistical purposes. Therefore, it is often the case that only a subset (called sampling set) of records from the initial microdata is released (Usually random sampling is employed.). If n is the number of elements in initial microdata and t is the released number of elements, we call $sf = t / n$ the sampling factor. Appling this method of sampling reduces the number of records and reduces disclosure risk. This method also increases information loss; one might initially be tempted to conclude that the information loss is at least $1 - sf$. In actuality, the loss will likely be much smaller, because, this masked microdata is used for statistical purposes, and, therefore, it is important to consider different statistical measures in expressing information loss (mean, variance, standard deviation). Moreover, the masked microdata can be considered useful only if those statistical measures are sufficiently precise. This property, of preserving different statistical measures within a given range, is called *statistical integrity* [7].

The major goal of disclosure control for microdata is to protect the confidentiality of the identities of individuals from the data. Several statistical disclosure control techniques such as global recoding [16, 13], local suppression [12], microaggregation [4], sampling [15], simulation [1], adding noise [9], rounding [14], post randomization method [10], data swapping [3] etc. were

proposed in the literature (For an excellent survey of all those methods see [17].). To increase confidentiality, more than one method is often applied in the disclosure control process.

In this paper, we define a set of disclosure risk measures based on combinations of particular methods. We justify our choices and we analyze different properties of those measures. Our disclosure risk measures compute the overall disclosure risk and are not linked to a target individual. We choose, in the beginning, two extreme measures called minimal disclosure risk (DR_{min}) and maximal disclosure risk (DR_{max}), and we then define a more general measure (D_w) based on a weight matrix. The disclosure risk measures presented in this paper are validated in our experiments. We have implemented those measures, and we have executed different experiments on simulated medical data.

There are many ways to define disclosure risk. Lambert defines disclosure risk and harm as matter of perception [11]. Those two measure the perception of an intruder. Also, in this paper, two types of disclosure, namely**,** identity disclosure and attribute disclosure, are presented. *Identity disclosure* refers to the identification of an entity (such as a person or an institution) and *attribute disclosure* refers to an intruder finding out something new about the target entity [11]. The results presented are based on probabilities that quantify the perception of the intruder.

Willemborg presents a different approach [17]. The risk per record is estimated using various assumptions. This work is more of a theoretical interest.

The most common approach deals with population unique [8, 15]. Greenberg and Bethleem discuss the probability of "population uniqueness"[8, 2]. We extended this measure to minimum disclosure risk measure, which we present later. Other measures define disclosure risk as proportion of sample unique records that are population unique [7, 15]. Eliot defines a new measure of disclosure risk as the proportion of correct matches amongst those records in the population, which match a sample unique masked microdata record [5]. We extend those discussions and present a more practical approach. We define a framework for microdata disclosure control and we make assumptions about the external information known by a presumptive intruder. Then, we discuss remove identifier, sampling and microaggregation methods and we define disclosure risk measures for any combination of those disclosure control methods applied in succession to a microdataset.

The remainder of this paper is organized as follows: Section 2 describes the framework for microdata disclosure control, Section 3 discusses disclosure risk measure for remove identifiers method, Section 4 discusses disclosure risk measures where sampling and microaggregation are applied in succession to the initial microdata, Section 5 shows experimental results, and Section 6 gives future work in the area of disclosure control for microdata.

2. General framework for microdata

The initial microdata consists of a set of n records with values from three types of attributes: identifier (I), confidential (S) and key (K) attributes. $I_1, I_2, .., I_m$ are identifier attributes such as *Name* and *SSN* that can be used to identify a record. Those attributes are present only in the initial microdata because they express information, which can lead to the identification of a specific entity. $K_1, K_2, ..., K_p$ are key attributes such as *Zip Code* and *Age* that may be known by an intruder. Key attributes are present in masked microdata as well as in the initial microdata. $S_1, S_2, ..., S_q$ are confidential attributes such as *Principal Diagnosis* and *Annual Income* that are rarely known by an intruder. Confidential attributes are present in masked microdata as well as in the initial microdata.

We represent the initial microdata as a matrix (IM) with 3 partitions that correspond to different categories of attributes. The rows represent the entities (individual entities) and the columns represent the attributes. Therefore:

$$IM = \left[I \quad | \quad K \quad | \quad S \right] \qquad (2.1)$$

where $I = [\, i_{ij} \,]$ of order $n \times m$, $K = [\, k_{ij} \,]$ of order $n \times p$, and $S = [\, s_{ij} \,]$ of order $n \times q$.

The general form of the masked microdata (M) is:

$$M = \left[K' \quad | \quad S' \right] \qquad (2.2)$$

where $K' = [\, k'_{ij} \,]$ of order $t \times p$, and $S = [\, s'_{ij} \,]$ of order $t \times q$.

The number of records in the masked microdata (t) differs from the number of records in initial microdata (n) due to disclosure control methods such as sampling and simulation. The corresponding attribute values may also differ due to perturbative methods (such as global recoding, microaggregation, data swapping and so on) used in disclosure control process (this is why we use prime notation).

Let r be the number of records of masked microdata with a corresponding record in initial microdata. Therefore, the following relations are always true $r \leq n$ and $r \leq t$.

We define the simulated factor:

$$fs = \frac{t - r}{t} \qquad (2.3)$$

which represents the quantity of information simulated in masked microdata

The sampling factor sf is defined as:

$$sf = \frac{r}{n} \qquad (2.4)$$

which represents the percentage of records from the initial microdata that are released to the public.

So far, the classification of attributes was made based on the ownership view. We have a similar classification based on the researcher view. In this way, we can divide each record into two parts: known fields and unknown fields. This classification is at the record level. More detailed discussion on this topic can be found in [17].

In an ideal scenario, the known fields will be a subset of identifier and key attributes, but, unfortunately, there are situations when some confidential fields are also known fields and, therefore, more disclosure can take place. Due to this fact, it is very difficult to have a disclosure control method for general case. We can make assumptions about intruder possible external knowledge and, based on those assumptions, we will define several disclosure risk measures for various methods.

The first assumption we make is that the intruder does not know any confidential information. Still the intruder may have a perception of some confidential information. For instance, if the intruder has to choose the income for a physician between two possible values, $20,000 and $100,000, the intruder would probably guess the right.

The second assumption is that an intruder knows all the key and identifier values for population. To identify individuals from masked microdata the intruder will execute a record linkage operation between external information dataset and masked microdata.

3. Disclosure risk measures for remove identifiers method

Based on our previous assumptions, we consider the following external information available to an intruder:

$$Ext = [I \mid K] \qquad (3.1)$$

Due to the fact that masked microdata is obtained by simply removing the identifiers attributes we will call it *initial masked microdata (IMM).*

We note that initial masked microdata is defined as a projection on key and confidential attributes of initial microdata:

$$IMM = \Pi_{K,S}(IM) \qquad (3.2)$$

Disclosure risk measures the percentage of records that are correctly linked by an intruder knowing IMM and Ext.

We make the assumption that the key attributes are discrete. We cluster the data from IMM based on their key values. Therefore, in each cluster we will include records with the same values for their key attributes. We define the following:

- n – the number of entities in the population.
- F – the number of clusters.
- A_k – the set of elements from the k-th cluster for all k, $1 \le k \le F$.
- $F_i = |\{A_k \mid |A_k| = i, \text{ for all } k = 1, .., F\}|$ for all i, $1 \le i \le n$. F_i represents the number of clusters with the same length.
- $n_i = |\{x \in A_k \mid |A_k| = i, \text{ for all } k = 1, .., F\}|$ for all i, $1 \le i \le n$. n_i represents the number of records in clusters of length i.

Based on the above notations we have the following relations:

$$n_i = i \cdot F_i, \quad i=1,.., n \qquad (3.3)$$

$$\sum_{i=1}^{n} F_i = \sum_{i=1}^{n} \frac{n_i}{i} = F \qquad (3.4)$$

$$\sum_{i=1}^{n} n_i = \sum_{i=1}^{n} i \cdot F_i = n \qquad (3.5)$$

The first measure of disclosure risk is based on the percentage of unique records, which is discussed by Fienberg [7]. This is represented by:

$$DR_{min} = \frac{n_1}{n} \qquad (3.6)$$

Since we made the assumption that an intruder has knowledge about identifier and key values, this measure represents the percentage of records from the population that can be correctly de-identified by the intruder. This is a minimal disclosure risk value. Any other measure we may define should be greater then this minimal disclosure risk.

This measure has its limits. It does not consider the distribution of the records that are not unique. For example, when population size is 100, if $n_2 = 100$ or $n_{100} = 100$ then the value for DR_{min} will be 0 for both cases. However, in the case of $n_2 = 100$, the probability of an intruder de-identify one record is 0.5, and for case $n_{100} = 100$ the probability is 0.01. Due to this limitation for minimal disclosure risk, we introduce a new disclosure risk measure, which consider a non-unique population. This measure is defined as:

$$DR_{max} = \sum_{i=1}^{n} \frac{1}{i} \cdot \frac{n_i}{n} = \frac{F}{n} \qquad (3.7)$$

When the number of distinct values increases, the disclosure risk also increases. This measure is the maximum value for measuring disclosure risk. Only when more external information, i.e. confidential information, is available the value of disclosure risk can be greater then DR_{max}.

This measure has its limitations too. For example, when population size is 100, if $n_{10} = 100$ or $n_1 = 9$, $n_{91} = 91$ then the value for DR_{max} will be 0.1 for both cases. Those two situations are not equivalent. In the first case disclosure risk is intuitively lower than in the second case. Therefore, we introduce a weight system to increase the importance of unique values over the rest of records, and

the importance of records with double occurrence for key values over the records with more then double occurrence and so on. This constitutes the basic idea for our third measure.

We define a disclosure risk weight vector $w = (w_1, w_2, \ldots, w_N)$ with the following properties:

a) $w_i \in \boldsymbol{R}_+$ for all $i = 1, \ldots, n$.

b) $w_i \geq w_j$ for all $i \leq j$, $i,j = 1, \ldots, n$.

c) $\sum_{i=1}^{n} w_i = n$.

Using the disclosure risk weight vector, we define a new disclosure risk measure as follows:

$$DR_w = \frac{1}{n \cdot w_1} \sum_{i=1}^{n} w_i \cdot F_i \qquad (3.8)$$

Lemma 3.1.

For every disclosure risk weights vector w, the following relations are true:
$DR_{min} \leq DR_w \leq DR_{max}$.

Proof

To show $DR_{min} \leq DR_w$ we have:

$$DR_{min} = \frac{n_1}{n} = \frac{1}{n \cdot w_1} \cdot w_1 \cdot n_1 = \frac{1}{n \cdot w_1} \cdot w_1 \cdot F_1 \leq \frac{1}{n \cdot w_1} \sum_{i=1}^{n} w_i \cdot F_i = DR_{w.}$$

To show $DR_w \leq DR_{max}$ we have:

$$DR_w = \frac{1}{n \cdot w_1} \sum_{i=1}^{n} w_i \cdot F_i = \frac{1}{n} \sum_{i=1}^{n} \frac{w_i}{w_1} \cdot F_i \leq \frac{1}{n} \sum_{i=1}^{n} 1 \cdot F_i = DR_{max..}$$

Lemma 3.2.

For every disclosure risk weight vector w, $0 \leq DR_w \leq 1$.

Proof

Using lemma 3.1 and the fact that n_1 and F are numbers between 0 and n we get: $0 \leq DR_w \leq 1$.

q.e.d.

Please note that when $w = (n, 0, 0, \ldots, 0)$ disclosure risk measure $DR_w = \frac{1}{n \cdot w_1} \sum_{i=1}^{n} w_i \cdot F_i = \frac{F_1}{n} = \frac{N_1}{n} = DR_{min.}$ and when disclosure risk weight vector is $w = (1, 1, 1, \ldots, 1)$ disclosure risk measures $DR_w = \frac{1}{n \cdot w_1} \sum_{i=1}^{n} w_i \cdot F_i = \frac{1}{n} \sum_{i=1}^{n} 1 \cdot F_i = DR_{max..}$

To illustrate those measures, we consider several examples of initial masked microdata with characteristics described in Table 3.1-3.4**.** Given a vector w, we compute disclosure risk for all those initial masked microdata using DR_w (see Table 3.5).

$n = 100$	$n_1 = 10$	$n_2 = 20$	$n_3 = 30$	$n_4 = 40$	$n_4 = \ldots n_{100} = 0$
$F = 40$	$F_1 = 10$	$F_2 = 10$	$F_3 = 10$	$F_4 = 10$	$F_4 = \ldots F_{100} = 0$

Table 3.1. – Initial masked microdata A

$n = 100$	$n_1 = 5$	$n_2 = 40$	$n_3 = 15$	$n_4 = 40$	$n_4 = \ldots n_{100} = 0$
$F = 40$	$F_1 = 5$	$F_2 = 20$	$F_3 = 5$	$F_4 = 10$	$F_4 = \ldots F_{100} = 0$

Table 3.2. – Initial masked microdata B

$n = 100$	$n_1 = 9$	$n_2 = \ldots = n_{90} = 0$	$n_{91} = 91$	$n_{92} = \ldots n_{100} = 0$
$F = 10$	$F_1 = 9$	$F_2 = \ldots = F_{90} = 0$	$F_{91} = 1$	$F_{92} = \ldots F_{100} = 0$

Table 3.3. – Initial masked microdata C

$n = 100$	$n_1 = \ldots = n_9 = 0$	$n_{10} = 10$	$n_{11} = \ldots n_{100} = 0$
$F = 10$	$F_1 = \ldots = F_9 = 0$	$F_{10} = 10$	$F_{11} = \ldots F_{100} = 0$

Table 3.4. – Initial masked microdata D

W	IMM A	IMM B	IMM C	IMM D
(n, 0, 0, …, 0)	10%	5%	9%	0%
(1, 1, 1, …, 1)	40%	40%	10%	10%
(2n/3, n/3, 0, …, 0)	15%	15%	9%	0%
(n/2, n/3, n/6, 0, …, 0)	20%	20%	9%	0%
(n/2, n/4, n/6, n/12, 0, …, 0)	20%	18.33%	9%	0%
(n/3, n/4, n/4, n/6, 0, …, 0)	30%	28.33%	9%	0%

Table 3.5. – Disclosure risk examples for remove identifiers method

Please note that $DR_w = DR_{min.}$ when $w = (n, 0, 0, ..., 0)$ and $DR_w = DR_{max}$ when disclosure risk weight vector is $w = (1, 1, 1, ..., 1)$. As one can observe from the above table, using only the remove identifiers method disclosure risk is usually high. Therefore, at least one more disclosure control method needs to be applied to reduce disclosure risk. Disclosure risk computations when exactly one method is performed after remove identifiers step are presented in [18].

4. Disclosure risk measures for sampling and microaggregation methods

After the remove identifiers method is applied, in order to protect the confidentiality of the entities, usually, more than one disclosure control method is applied. In this section, we analyze disclosure risk measures when both sampling and microaggregation are applied, in any order, to the same initial microdataset.

Sampling is the disclosure control method in which only a subset of records is released [15].

Microaggregation is a disclosure technique applicable to quantitative attributes. It can be applied to a single attribute (univariate microaggregation) at a time, or to a group of attributes (multivariate microaggregation). We will briefly discuss the univariate case.

The idea behind this method is to sort the records from the initial microdata with respect to an attribute A, create groups of consecutive values, replace those values by the group average. How the groups are formed is up to the owner of the data. Usually, the owner specifies a minimum size for a group. More formally, let be $X = \{x_1, x_2, ..., x_n\}$ where x_i is the value of attribute A for record i and let k be the minimum size of a group. A k-partition $P = \{C_1, C_2, ..., C_{m(P)}\}$ of X is a partition where the size of group C_i, $1 \le i \le m(P)$ is at least k. Let P_k be the set of all k-partitions of X. Optimal microaggregation consists of finding a k-partition such that the sum of distances from each x_i to the average value for each partition

$$\bar{x}_{C_i} = \frac{1}{|C_i|} \cdot \sum_{x_l \in C_i} x_l \qquad (4.1)$$

is minimized, where C_i is the group to which x_i belongs. Formally, the problem is:

$$min_{P \in P_k} \sum_{i=1}^{m(P)} \sum_{x_j \in C_i} | x_j - \bar{x}_{C_i} | \qquad (4.2)$$

where $m(P)$ is not part of the input [4].

We use the following notations:

- n, F, A_k, n_i and F_i, for all $i = 1, .., n$ – have the same meaning as in the previous section.
- f – the number of clusters with the same values for key attributes in MM.
- We cluster all records from MM based on their key values. M_k – the set of elements from the k-th cluster for all k, $1 \le k \le f$.
- $f_i = | \{M_k \mid |M_k| = i, for\ all\ k = 1, .., f\} |$ for all i, $1 \le i \le n$. f_i represents the number of clusters with the same length.
- $t_i = | \{x \in M_k \mid |M_k| = i, for\ all\ k = 1, .., f\} |$ for all i, $1 \le i \le n$. t_i represents the number of records in clusters of length i.
- C – *the classification matrix*. . This $t \times n$ matrix represents the correlation between masked microdata and initial masked microdata. Each element of C, c_{ij}, represents the number of records that appears in clusters of size i in the masked microdata and appeared in clusters of size j in the initial masked microdata. Mathematically, this definition can be expressed in the following form: For all $i = 1, .., t$ and for all $j = 1, .., n$; $c_{ij} = | \{x \in M_k$ and $x \in A_p \mid |M_k| = i,$ for all $k = 1, .., f$ and $|A_p| = j,$ for all $p = 1, .., F\} |$.

Relations (3.1), (3.2) and (3.3) are holding. We have the following extra relations:

$$t_i = i \cdot f_i, \quad i=1,.., t \qquad (4.3)$$

$$\sum_{i=1}^{t} f_i = \sum_{i=1}^{t} \frac{t_i}{i} = f \qquad (4.4)$$

$$\sum_{i=1}^{t} t_i = \sum_{i=1}^{t} i \cdot f_i = t \qquad (4.5)$$

$$\sum_{j=1}^{n} c_{ij} = t_i \text{ for all } i = 1, .., t \qquad (4.6)$$

$$\sum_{i=1}^{t} \sum_{j=1}^{n} c_{ij} = t \qquad (4.7)$$

The following algorithm describes how to calculate elements of C, the classification matrix.

Algorithm 4.1. (Classification matrix construction)
Initialize each element from $\mathbf{C}$ *with* $\mathbf{0}$.
For each element $\mathbf{s}$ *from masked microdata* $\mathbf{MM}$ *do*
Count the number of occurrences of key values of $\mathbf{s}$ *in masked microdata* $\mathbf{MM}$.*Let* $\mathbf{i}$ *be this number.*
Count the number of occurrences of key values of $\mathbf{s}$ *in initial microdata* $\mathbf{IM}$.*Let* $\mathbf{j}$ *be this number.*
Increment $\mathbf{c_{ij}}$ *by 1.*
End for.

Now, we define three disclosure risk measures similar to the previous sections. The first two are introduced below:

$$DR_{min} = \frac{c_{11}}{n} \quad (4.8)$$

$$DR_{max} = \frac{\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} c_{ik} + \sum_{j=1}^{k-1} c_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} c_{ik}}{n} \quad (4.9)$$

DR_{min} represents the percentage of records from the population that the intruders can de-identify because c_{11} represents the number of records unique in the initial microdata as well as in the masked microdata. This is the minimal disclosure risk value. DR_{max} takes in consideration the probability of correct linking for non-unique records.

For the third measure we define a disclosure risk weight matrix, W, as follows:

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1t} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2t} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ w_{t1} & w_{t2} & \dots & w_{tt} & \dots. & w_{tn} \end{pmatrix} \quad (4.10)$$

with the following properties:

- $w_{jj} \geq w_{jj+1} \geq \dots \geq w_{jn}$ for all j, $1 \leq j \leq n$
- $w_{1j} \leq w_{2j} \leq \dots \leq w_{jj}$ for all j, $1 \leq j \leq t$
- $w_{1j} \leq w_{2j} \leq \dots \leq w_{tj}$ for all j, $t+1 \leq j \leq n$
- $w_{1j} \geq w_{2,j+1} \geq \dots \geq w_{t,j+t}$ for all j, $1 \leq j \leq n - t$
- $w_{1j} \geq w_{2,j+1} \geq \dots \geq w_{n-j+1,n}$ for all j, $n - t < j < n$
- $w_{jj} \geq w_{j+1j} \geq \dots \geq w_{nj}$ for all j, $1 \leq j \leq n$
- $w_{j1} \leq w_{j2} \leq \dots \leq w_{jj}$ for all j, $1 \leq j \leq n$
- $w_{j1} \geq w_{j+1,2} \leq \dots \leq w_{n,n-j+1}$ for all j, $1 \leq j < n$
- $\sum_{i=1}^{t}\sum_{j=1}^{n} w_{ij} = n$

The last formula proposed for disclosure risk is:

$$DR_W = \frac{1}{n \cdot w_{11}}\left(\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} w_{ik} \cdot c_{ik} + \sum_{j=1}^{k-1} w_{kj} \cdot c_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} w_{ik} \cdot c_{ik}\right) \quad (4.11)$$

By a proper choice for a disclosure risk weight matrix, the owner of the data will be able to obtain a better approximation for disclosure risk based on the data characteristics.

Lemma 4.1.

For every disclosure risk weights matrix W the following relations are true:

$DR_{min} \leq DR_W \leq DR_{max}$

Proof

To show $DR_{min} \leq DR_W$ we have:

$$DR_{min} = \frac{c_{11}}{n} = \frac{1}{n \cdot w_{11}} \cdot w_{11} \cdot c_{11} \leq \frac{1}{n \cdot w_{11}} \cdot w_{11} \cdot c_{11} + \frac{1}{n \cdot w_{11}}\left(\sum_{k=2}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} w_{ik} \cdot c_{ik} + \sum_{j=1}^{k-1} w_{kj} \cdot c_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} w_{ik} \cdot c_{ik}\right)$$

$= DR_W$.

To show $DR_W \leq DR_{max}$ we have:

$$DR_W = \frac{1}{n \cdot w_{11}}\left(\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} w_{ik} \cdot c_{ik} + \sum_{j=1}^{k-1} w_{kj} \cdot c_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} w_{ik} \cdot c_{ik}\right) =$$

$$\frac{1}{n}\left(\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} \frac{w_{ik}}{w_{11}} \cdot c_{ik} + \sum_{j=1}^{k-1} \frac{w_{kj}}{w_{11}} \cdot c_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} \frac{w_{ik}}{w_{11}} \cdot c_{ik}\right)$$

$$\leq \frac{\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} c_{ik} + \sum_{j=1}^{k-1} c_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} c_{ik}}{n} = DR_{max}.$$

q.e.d.

Lemma 4.2.

For every disclosure risk weights matrix W, $0 \leq DR_W \leq 1$.

Proof

Using lemma 4.1 and the fact that c_{11} is greater then 0 we get: $0 \leq DR_W$.

Then $DR_W \leq DR_{max.} =$

$$\frac{\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} c_{ik} + \sum_{j=1}^{k-1} c_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} c_{ik}}{n} \leq \frac{\sum_{i=1}^{t}\sum_{j=1}^{n} c_{ij}}{n} = \frac{t}{n} \leq 1.$$

q.e.d.

Please note that when $c_{11} = n$ and all other weights are 0 in disclosure risk weights matrix DR_W is equal with DR_{min}. Also when all weights are equal ($c_{ij} = 1 / t$ for all i, $1 \leq i \leq t$ and for all j, $1 \leq j \leq n$) in disclosure risk weights matrix DR_W is equal with DR_{max}.

5. Experimental results

We used simulated medical record billing data to perform a series of tests. The data contains the following attributes: *Age*, *Race*, *Age_Cat* (in five years increments), *Zip* and *Amnt_Billed*. In our experiment, we used three sets of initial microdata; one with size 50 (called *IM50*), one with size 500 (*IM500*), and the last one with size 5000 (*IM5000*), all with the same set of attributes. For each initial microdata we considered four sets of key attributes (please see Appendix 1 for details regarding those attributes):

- KA_1 = {*AGE, RACE, SEX, ZIP*}

- KA_2 = {*AGE, RACE, SEX*}
- KA_3 = {*AGE_CAT, RACE, SEX, ZIP*}
- KA_4 = {*AGE_CAT, RACE, SEX*}

Then, for each of those 12 different scenarios, we applied various series of disclosure control methods, and we computed minimal and maximal disclosure risk. In this paper, we present a few scenarios, which combine both sampling and microaggregation. Figure 5.1 shows disclosure risk variations for sampling followed by microaggregation for *Age* attribute.

We notice that microaggregation is effective for *Age* values when the group size is large. The reason is the initial grouping of the *Age* attribute values from the initial microdata. Similar results were obtained when we first applied microaggregation for *Age* attribute and then sampling, still there are some differences. Those differences are depicted in Figure 5.2.

Both minimal disclosure risk and maximal disclosure risk are lowered when sampling is applied first followed by microaggregation. This is true for any sampling factor as well as the one presented in Figure 5.2. The reason for this result is due to the fact that by applying microaggregation after the sampling, the group size for *Age* attribute is at least equal with microaggregation parameter. However when microaggregation is applied first, due to the sampling, the group size may be less than microaggregation parameter.

6. Conclusions and Future Work

In this paper, several disclosure risk measures were presented. We implemented those results and executed a series of tests over simulated sets of data. From the experiments, we drew one conclusion about the order of applying more than one disclosure control method for an initial microdata: the sampling followed by microaggregation performs better with respect to disclosure risk computations than vice versa.

The disclosure risk weight matrix should capture the specifics of the data and the goals of the data owner. The data owner can change priorities between different clusters of records. More testing must be done to develop automated techniques for choosing the disclosure risk weight matrix. In this paper, we were able to determine the range for disclosure risk for any weight matrix, and this range is independent of the disclosure risk weight matrix chosen. Most of the time by using the interval between minimal and maximal disclosure risk, the data owner can decide if the data is protected against disclosure. When more accuracy is needed, the data owner must choose the weight matrix for disclosure risk computations.

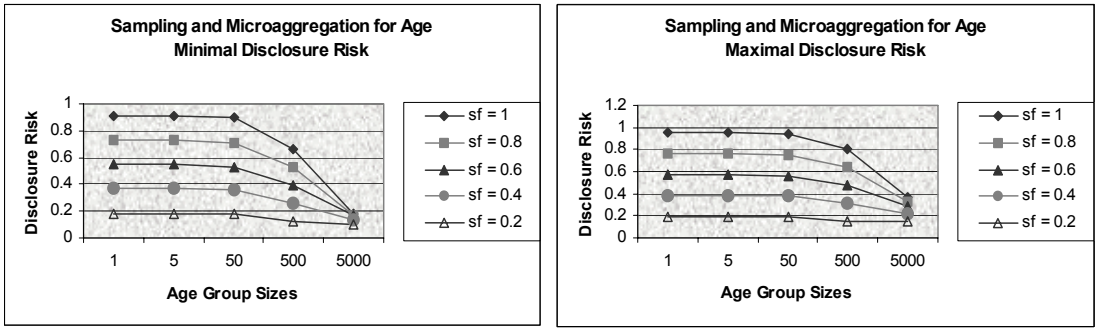

Figure 5.1. – Sampling, followed by microaggregation for Age when IM5000 and KA_1 are used.

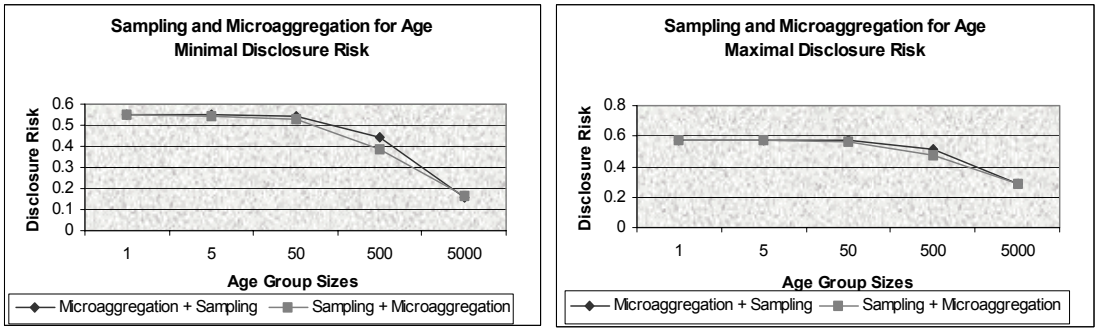

Figure 5.2. – Sampling and microaggregation for Age when IM5000 and KA_1 are used.

The future work in this field can be divided into three areas. One is to develop disclosure risk measures for other methods and to generalize them further, the second one is to express information loss in general formulas, to study the dependence between information loss and disclosure risk, and the last one is to find patterns of applying successively more than one disclosure control methods for a given initial microdata to minimize both disclosure risk and information loss. Practical experiments that will use data intrusion simulation [5] techniques must be finally performed for various data sets in order to validate the results.

References:

[1] Adam N. R., Wortmann J. C. (1989), *Security Control Methods for Statistical Databases: A Comparative Study*. ACM Computing Surveys, Vol. 21, No. 4.

[2] Bethlehem J. G., Keller W. J., Pannekoek J. (1990), *Disclosure Control of Microdata*. Journal of the American Statistical Association, Vol. 85, Issue 409, 38-45.

[3] Dalenius T., Reiss S. P. (1982), *Data-Swapping: A Technique for Disclosure Control*. Journal of Statistical Planning and Inference 6, 73-85.

[4] Domingo-Ferrer J., Mateo-Sanz J. (2002), *Practical Data-Oriented Microaggregation for Statistical Disclosure Control*. IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 1, 189-201.

[5] Elliot, M. J. (2000), *DIS: a new approach to the measurement of statistical disclosure risk,* International Journal of Risk Management, 39 –48.

[6] Fellegi I. P. (1972), *On the Question of Statistical Confidentiality*. Journal of the American Statistical Association, Volume 67, Issue 337, 7-18.

[7] Fienberg, S. E.; Markov, U. E. (1998), *Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data,* Journal of Official Statistics, 385 - 397.

[8] Greenberg, B.; Zayatz, L. (1992), *Strategies for Measuring Risk in Public Use Microdata Files*, Statistica Neerlandica, 33 – 48.

[9] Kim J. J. (1986), *A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation*. American Statistical Association, Proceedings of the Section on Survey Research Methods, 303-308.

[10] Kooiman, P.; Willemborg, L.; Gouweleeuw, J. (1997), *PRAM: A Method for Disclosure Limitation for Microdata,* Report, Department of Statistical Methods, Statistical Netherlands, Voorburg.

[11] Lambert D. (1993), *Measures of Disclosure Risk and Harm*. Journal of Official Statistics, Vol. 9, 313-331.

[12] Little, R. J. A. (1993), *Statistical Analysis of Masked Data,* Journal of Official Statistics, Vol. 9, 407-426.

[13] McGuckin R. H., Nguyen S. V. (1990), *Public Use Microdata: Disclosure and Usefulness*. Journal of Economic and Social Measurement, Vol. 16, 19 – 39.

[14] Muralidhar K., Sarathy R. (1999), *Security of Random Data Perturbation Methods*, ACM Transactions on Database Systems, Vol. 24, No. 4, 487-493.

[15] Skinner, C. J.; Marsh, C.; Openshaw, S.; Wymer, C. (1994), *Disclosure control for census microdata*, Journal of Official Statistics, 31-51.

[16] Tendick P., Matloff, N. (1994), *A Modified Random Perturbation Method for Database Security.* ACM Transactions on Database Systems, Volume 19, Number 1.

[17] Willemborg L., Waal T. (ed) (2001), *Elements of Statistical Disclosure Control*. Springer Verlag.

Appendix 1 – Number of Distinct Values for Key Attributes

Attribute	IM50	IM500	IM5000
SEX	2	2	3
RACE	4	7	7
AGE	38	88	102
AGE_CAT	16	18	19
ZIP	44	291	771
$\mathbf{KA_1}$ = {AGE, RACE, SEX, ZIP}	50	498	4775
$\mathbf{KA_2}$ = {AGE, RACE, SEX}	48	273	650
$\mathbf{KA_3}$ = {AGE_CAT, RACE, SEX, ZIP}	50	493	4314
$\mathbf{KA_4}$ = {AGE_CAT, RACE, SEX}	39	105	167