Assessing Global Disclosure Risk in Masked Microdata

Traian Marius Truta
Department of Mathematics and Computer Science
Northern Kentucky University
Highland Heights, KY 41076, USA
trutat1@nku.edu

Farshad Fotouhi
Department of Computer Science
Wayne State University
Detroit, MI 48202, USA
fotouhi@wayne.edu

Daniel Barth-Jones
Center for Healthcare Effectiveness
Wayne State University
Detroit, MI 48202, USA
dbjones@med.wayne.edu

ABSTRACT
In this paper, we introduce a general framework for microdata and three disclosure risk measures (minimal, maximal and weighted). We classify the attributes from a given microdata in two different ways: based on their potential identification utility and based on the order relation that exists in their domain of value. We define inversion and change factors that allow data users to quantify the magnitude of masking modification incurred for values of a key attribute. The disclosure risk measures are based on these inversion and change factors, and can be computed for any specific disclosure control method, or any combination of methods applied in succession to a given microdata. Using simulated medical data in our experiments, we show that the proposed disclosure risk measures perform as expected in real-life situations.

Categories and Subject Descriptors
K.4.1 [**Computers and Society**]: Public policy Issues – *privacy, regulation*.

General Terms
Measurement, Security.

Keywords
Statistical Disclosure, Data Privacy, Microdata, Disclosure Risk, Information Loss and Microaggregation.

1. INTRODUCTION
Governmental, public, and private institutions that systematically release data are increasingly concerned with possible misuses of their data that might lead to disclosure of confidential information [42]. Moreover, confidentiality regulation requires that privacy of individuals represented in the released data must be protected.

WPES'04, October 28, 2004, Washington, DC, USA.

In the U.S., for example, privacy regulations promulgated by the Department of Health and Human Services as part of the *Health Insurance Portability and Accountability Act (HIPAA)* went into effect in April 2003 in order to protect the confidentiality of electronic healthcare information [21]. Other countries have promulgated similar privacy regulations (for example, the *Canadian Standard Association's Model Code for the Protection of Personal Information* [33] and the *Australian Privacy Amendment Act 2000* [3]). Various privacy regulations analyzed from a database perspective are presented in [2].

Microdata represents a series of records, where each record contains information on an individual entity [45]. After microdata has been masked to limit the possibility of disclosure and released for use by third parties, it is called *masked* or *released microdata* [9]. To avoid confusion, we will use the term *initial microdata* for microdata where no disclosure control methods were applied. Data masking methods (also called statistical disclosure control techniques) such as: *sampling* [35], *global and local recoding* [27, 34, 40], *suppression and local suppression* [34, 26], *microaggregation* [13], *simulation* [1], *adding noise* [22], *randomization or perturbation methods* [28, 24, 29, 16], *data swapping* [9, 32], *substitution* [37] etc. have been extensively discussed in literature [45].

Disclosure risk is the risk that a given form of disclosure will be encountered if masked microdata is released [8]. *Information loss* is the quantity of information, which existed in the initial microdata but which does not occur in masked microdata because of disclosure control methods [45]. When protecting the confidentiality of individuals, the owner of the data must satisfy the two conflicting requirements: protecting confidentiality for the entities from the initial microdata and maintaining analytic properties in the masked microdata [23]. The ultimate goal is minimizing disclosure risk so as to comply with existing regulations, while simultaneously minimizing information loss for statistical inference [18]. Since fully optimal minimization of both measures is not possible (decreasing disclosure risk will usually lead to increase information loss and vice versa), the owner of the data must select a compromise between disclosure risk and information loss values [15].

Considerable research on disclosure risk assessment [1, 4, 7, 10, 17, 19, 20, 25, 30, 36, 38] has resulted in a variety of proposed disclosure risk models, but is unanimous in the conclusion that disclosure risk cannot be eliminated completely. Accordingly, such research has focused on limiting disclosure risks to threshold levels.

One of the most intuitive ways to measure disclosure risk for microdata is to count the number of unique records with respect to a limited set of attributes [39]. The selected attributes are called keys in disclosure avoidance literature [45]. Substantial work has been done on estimating the number of population uniques from a sample of data when the population follows a particular distribution such as Poisson-Gamma [6], Dirichlet-multinomial [40], and negative-binomial [8]. Greenberg and Zayatz have proposed a procedure that is not dependent on a parametric statistical distribution [20]. Other approach was proposed by Lambert who defined disclosure risk as matter of perception [25]. *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 [25].

Recent work can be categorized into two directions: individual and global disclosure risk. Benedetti and Franconi introduced *individual risk methodology* [4]. The risk is computed for every released entity from masked microdata. In this scenario, the individual risk for each entity is the probability of correct identification by an intruder. All records with individual risk above a fixed threshold are defined as being at risk, and disclosure control methods must be used to protect these records. Other papers extend this approach [5, 31, 11]. *Global disclosure risk* is defined in terms of the expected number of identifications in the released microdata. Elliot and Skinner define 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 [17, 36]. Other approaches have been proposed in [31, 12].

In this paper, we extend disclosure risk measures that has previously presented for specific disclosure control methods [43, 44] to be suitable when masked microdata is obtained from initial microdata by any combination of disclosure control methods. Our formulations for disclosure risk measures compute overall disclosure risks for given datasets and are not linked to target individuals. To develop those disclosure risk measures, we consider the probabilistic linkage as well as characteristics of the data and the level of protection desired by the data owner. We define a framework for microdata disclosure control that incorporates assumptions about the external information known by a presumptive intruder. Then, we define and analyze minimal, maximal and weighted disclosure risk measures.

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. Such attributes are commonly deleted from the initial microdata in order to prevent direct identification. $K_1, K_2, ..., K_p$ are key attributes such as *Zip Code* and *Age* that may be known by an intruder. Key attributes are commonly fields which ideally would be retained in masked microdata if possible, but which often pose potential disclosure risks. $S_1, S_2, ..., S_q$ 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 units) and the columns represent the attributes. Therefore:

$$IM = \left[I \quad | \quad K \quad | \quad S \right] \tag{2.1}$$

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

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

$$M = \left[K' \quad | \quad S' \right] \tag{2.2}$$

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

The number of records in the masked microdata (t) can differ from the number of records in initial microdata (n) due to sampling and simulation. The corresponding attribute values may also differ due to disclosure control methods that modify values (such as microaggregation, data swapping, etc.) in the disclosure control process (motivating the use of the prime notation).

We call the actions taken by the owner of the data in order to protect the initial microdata with one or more disclosure control methods the *masking process*. The masking process can alter the initial microdata in three different ways: *changing the number of records*, *changing the number of attributes* and *changing values of specific attributes*. The change in number of attributes is always used, since the removal of identifier attributes is the first step for data protection. We call this first mandatory step in the masking process the *remove identifiers method*. The other two types of changes may or may not be applied to the initial microdata. The most general scenario is when all three changes are applied to the given initial microdata.

While change in the number of records is caused by two techniques: simulation [1] and sampling [35], the change of attribute values occurs during a larger number of disclosure methods (microaggregation [13], data swapping [9], adding noise [22], etc.). The domain of values and the existence of an order relation for those values determine what kind of methods can be applied in the masking process. Based on those two characteristics, we introduce the following classification:

- The attribute A is a *Continuous Attribute* (C) if the domain of values ranges over an infinitely divisible continuum of values. The attributes *Distance* and *Length* are examples of CA.
- The attribute A is a *Discrete Ordered Attribute* (DO) if the number of possible values for the attribute A is finite and there is a total order relationship between those values. *Age* and *Income* are examples of such attributes. Most of the time the attributes in this category are numerical.
- The attribute A is a *Discrete Partial Ordered Attribute* (DPO) if the number of possible values for the attribute A is finite and there is a partial order relationship between those values. *Zip Code* with its prefixes fits into this category. Such attributes are ordinal categorical variables.
- The attribute A is a *Discrete Unordered Attribute* (DU) if the number of possible values for the attribute A is finite and there is no order relationship between those values. *Sex* and *Race* are examples of such attributes. Such attributes are categorical.

In reality, due to data representation, continuous attributes can be seen as discrete attributes with a very large number of possible values, so we group the first two types of attributes defined above into one called *Ordered Attributes* (*O*). The next two categories can be renamed to *Partial Ordered Attribute* (*PO*) and *Unordered Attribute* (*U*). Numerical attributes belong to the first category, and categorical attributes can be classified as partial ordered or unordered attributes. When a categorical attribute is ordered, the data owner can define a distance function between its elements that maps its values to a numerical domain. Therefore, without loosing generality, we can assume that all categorical attributes are either partial ordered or unordered attribute.

In order to describe the masking process, a few assumptions are needed. The first assumption we make is that the intruder does not have specific knowledge of any confidential information. Still, the intruder may have some unconfirmed suspicions about confidential information. For instance, if the intruder has to choose the income for a physician from two possible values, \$10,000 and \$100,000, the intruder would probably guess the latter. The second assumption is that an intruder knows all the key and identifier values from the initial microdata, usually through access to an external dataset. In order to identify individuals from masked microdata, the intruder will execute a record linkage operation between the external information dataset and masked microdata. This assumption maximizes the amount of external information available to an intruder. Since, disclosure risk increases when the quantity of external information increases, this second assumption guarantees that any disclosure risk value computed by one of the proposed measures is an upper bound to the disclosure risk value when the amount of external information available to an intruder is not maximal. Since, the data owner often does not have complete knowledge about the external information available to an intruder, this assumption allows the data owner to determine whether the disclosure risk is under an acceptable disclosure risk threshold value and, therefore, this assumption does not reduce the generality of the problem.

Based on the above assumptions only key attributes are subject to changes in the masking process. We consider, for simplicity of notation that the first r records from initial microdata maintain their position in the masked microdata. Since the data owner executes the masking process, he may order the records as desired without loosing generality. Therefore, x_{ij} represents the value for record x_i (for all i between 1 and r) of attribute K_j (for all j between 1 and p), and, x'_{ij} represents the corresponding value in masked microdata.

For any ordered attribute O_k we define the notion of inversion. The pair (x_{ik}, x_{jk}) is called *inversion for attribute* O_k if $x_{ik} < x_{jk}$ and $x'_{ik} < x'_{jk}$ for i,j between 1 and r. We label the total number of inversions for the attribute O_k, inv_k. By definition:

$$inv_k = |\{(x_{ik}, x_{jk}) \mid x_{ik} < x_{jk} \text{ and } x'_{ik} > x'_{jk},\ 1 \le i,j \le r\}| \qquad (2.3)$$

The maximum number of distinct pairs (x_i, x_j), when $1 \le i,j \le r$ is $r \cdot (r-1)/2$. In the extreme situation all such pairs are inversions. Therefore,

$$0 \le inv_k \le \frac{r \cdot (r-1)}{2} \qquad (2.4)$$

for any attribute O_k. By definition, inv_k is a positive number.

We define *inversion factor for attribute* O_k the minimum between 1 and the number of inversions for attribute O_k over average number of inversions. We label it with if_k. Therefore,

$$if_k = \min\left(1, \frac{4 \cdot inv_k}{r \cdot (r-1)}\right) \qquad (2.5)$$

We notice that when if_k increases for any k disclosure risk decreases due to the possibility of false matches.

For the remaining categories of attributes, we define the terms of strong and weak change for partial ordered attributes and change for unordered attributes as follows. The pair (x_{ik}, x'_{ik}) is called a *change* for attribute U_k if $x_{ik} \neq x'_{ik}$. The total number of changes for a given attribute, labeled ch_k is:

$$ch_k = |\{(x_{ik}, x'_{ik}) \mid x_{ik} \neq x'_{ik},\ 1 \le i \le r\}| \qquad (2.6)$$

We define *change factor for attribute* U_k the number of changes over maximum number of possible changes.

$$cf_k = \frac{ch_k}{r} \qquad (2.7)$$

The week and strong change are defined for partial ordered attribute. For example, in Figure 2.1, we consider *Zip Code* attribute with several values. We include in its domain of value all recoded zip codes such as *482**, *48**, or *. The values for a partial ordered attribute can be mapped to a tree, and there is always an element that is the root of the three (in our example "*"), that is comparable with all other values and is the maximum element. We label it with *max*.

A *weak change* for attribute PO_k is a pair of distinct values (x_{ik}, x'_{ik}) where exist at least one value a such as $a \ge x_{ik}$ *and* $a \ge x'_{ik}$, and a is not equal with *max*. Otherwise, the pair of distinct values (x_{ik}, x'_{ik}) is called *strong change*. Based on this definition the pair (48201, 48301) is a week change since the value 48* is greater than both 48201 and 48301 and is not equal with *. The pair (48202, 88202) is a strong change.

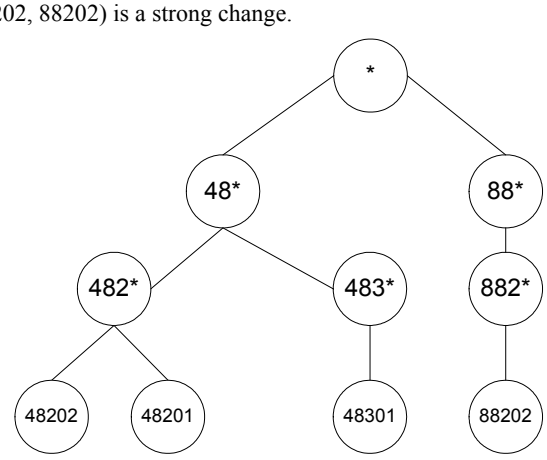

Figure 2.1. *Zip Code* Attribute Tree.

To characterize the "weakness" of the change, the data owner must define a set of rules to compare various weak changes. The *weak change factor* (labeled $wcf(x_{ik}, x'_{ik})$) will be associated to every pair of values, and its range will be the interval [0, 1]. The value 0, means that x_{ik} is equal to x'_{ik}, and the value 1 means the pair (x_{ik}, x'_{ik}) is a strong change. For *Zip Code* attribute, the data

owner can consider 5 - the number of identical digits starting with the left divided by 5 as the weak change factor for any pair of zip codes.

The change factor for attribute PO_k is defined as follows:

$$cf_k = \frac{1}{r} \cdot \sum_{i=1}^{r} wcf(x_{ik}, x'_{ik}) \qquad (2.8)$$

The inversion factor and change factor allow us to measure disclosure risk when various methods such as data swapping [9], global and local recoding [27, 34, 40] or randomization method [28, 24, 29, 16], are employed. We note that the change factor is not null when a *PO* or *U* attribute has been subject to changes in values. When the masking process modifies values from an ordered attribute, the inversion factor may still be equal to zero. Top and bottom coding [45] and microaggregation [13] are examples of disclosure control methods that keep the inversion factor as zero. We label those methods as *order preserving disclosure control methods*. We call all methods that input a non-null inversion factor as *non-order preserving methods*.

3. DISCLOSURE RISK MEASURES

We cluster the data from initial microdata and masked microdata based on their key values [43]. In the statistical disclosure control literature, such clusters are typically referred to as *equivalence classes* [Zayatz 1991] or *cells* [8].

We define the following notations for initial microdata (*IM*):

- 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 size i;
- $n_i = |\{x \in A_k \mid |A_k| = i, \text{ for all } k = 1, .., F\}|$ for all i, $1 \le i \le n$. n_i represents the number of records in clusters of size i.
- Similar notations are defined for the masked microdata (*MM*):
- f – the number of clusters with the same values for key attributes;
- B_k – the set of elements from the k-th cluster for all k, $1 \le k \le f$;
- $f_i = |\{B_k \mid |B_k| = i, \text{ for all } k = 1, .., f\}|$ for all i, $1 \le i \le t$. f_i represents the number of clusters with the size i;
- $t_i = |\{x \in B_k \mid |B_k| = i, \text{ for all } k = 1, .., f\}|$ for all i, $1 \le i \le t$. t_i represents the number of records in clusters of size i.

To relate initial microdata to masked microdata we define the *classification matrix C*. It represents a $t \times n$ matrix that describes the relationship between sampling set and initial microdata. Each element of C, c_{ij}, is equal with the total number of records that appears in clusters of size i in the sampling set, and, in clusters of size j in the initial microdata. Mathematically, this definition can be expressed in the following form: for all $i = 1, .., t$ and for all $j = 1, .., n$; $c_{ij} = |\{x \in B_k \text{ and } x \in A_p \mid |B_k| = i, \text{ for all } k = 1, .., f \text{ and } |A_p| = j, \text{ for all } p = 1, .., F\}|$.

Various relations between the defined terms are described by Truta, Fotouhi and Barth-Jones [43]. We note that this classification is based on the chosen set of key attributes. If the set of key attributes is changed, the process of computing cluster sizes and the elements from classification matrix must be restarted. The following algorithm describes how to calculate elements of the classification matrix based on a given set of key attributes:

Algorithm (Classification matrix construction)

Input: *IM – Initial Microdata*
M – Masked Microdata
K – Set of Key Attributes
Output: *C – Classification Matrix*
Initialize each element from C with 0.
*For each record from masked microdata **MM** (ordered **s′** = 1 to t) do*
Concatenate the specified key attributes to form a composite key.
*Count the number of occurrences of composite key values of **s′** in **MM**. Let **i** be this number.*
*Find the corresponding record **s** from **IM** such that **s** is masked to **s′** in **MM**.*
*Count the number of occurrences of composite key values of **s** in **IM**. Let **j** be this number.*
*Increment c_{ij} by **1**.*
End for.

The disclosure risk measures proposed in [43] can be generalized from any combination of sampling and microaggregation to any possible masked microdata in which the inversion factor for any ordered attribute is 0 and the changing factor for any unordered or partial ordered attribute is also 0. To further generalize those measures we need to include the inversion and change factor for corresponding key attributes. To simplify notations we define *inversion-change factor* (*icf*) as equal with inversion factor (*if*) for ordered attributes and equal with change factor (*cf*) for unordered and partial ordered attributes. To characterize disclosure risk when the change factor is not 0, we introduce the following intermediary disclosure risk measures:

$$DR_{min}^{int} = \left(\prod_{k=1}^{p}(1 - icf_k)\right) \cdot \frac{c_{11}}{n} \qquad (3.1)$$

$$DR_{max}^{int} = \left(\prod_{k=1}^{p}(1 - icf_k)\right) \cdot \frac{\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} c_{ik} + \sum_{j=1}^{k-1} c_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} c_{ik}}{n} \qquad (3.2)$$

$$DR_{W}^{int} = \left(\prod_{k=1}^{p}(1 - icf_k)\right) \cdot \frac{1}{n \cdot w_{11}} \cdot \left(\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} w_{ik} \cdot c_{ik} + \sum_{j=1}^{k-1} w_{kj} \cdot c_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} w_{ik} \cdot c_{ik}\right) \qquad (3.3)$$

The disclosure risk weight matrix, W, is defined as:

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1t} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2t} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ w_{t1} & w_{t2} & \dots & w_{tt} & \dots & w_{tn} \end{pmatrix} \quad (3.4)$$

with the following properties (see [43] for details):

- $w_{jj} \geq w_{jj+1} \geq \dots \geq w_{jn}$ for all j, $1 \leq j \leq n$
- $w_{1j} \leq w_{2j} \leq \dots \leq w_{tj}$ 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 intermediary disclosure risk measures are not always accurate. Let us assume that the change factor for a key attribute A is 1. This rare situation occurs when, for instance, the attribute is unordered and all values are modified. In that case the intermediary disclosure risk (minimal, maximal or weighted, for any weight matrix) will be equal to 0. In reality the intruder may notice that is something wrong with values of that key attribute and, therefore, he may not consider this key attribute in the process of disclosing confidential information from masked microdata. In that case the set of useful key attributes does not include the attribute A. Similar situations may exist for a key attribute with a non-null change factor. The disclosure risk computed considering this attribute for a specified weight matrix may be lower than the disclosure risk computed without that key attribute. In order to find the correct value for disclosure risk, we have to consider all possible subsets of key attributes, compute disclosure risk for each subset and select the maximum value.

We describe the method of computing disclosure risk measures using a binary vector v with p elements (p is the number of key attributes), $v = (v_1, v_2, \dots, v_p)$. The zero elements in this vector means that the corresponding key attributes are not considered in disclosure risk computation. Each vector component (v_i) takes either value 1 or 0, therefore the number of distinct vectors v for a fixed p is 2^p. For any v, we define its corresponding subset of key attributes K^v as follows:

$$K^v = \{K_i \mid v_i=1, i=1,2,\dots,p\} \quad (3.5)$$

To compute disclosure risk, we need to consider each possible non-empty subset K^v, and compute its corresponding classification matrix. We label the classification matrix with C^v, and its corresponding elements with c^v_{ij}. The disclosure risk measures corresponding to the vector v are depicted below:

$$DR^v_{min} = \left(\prod_{k=1}^{p}(1 - v_k \cdot icf_k)\right) \cdot \frac{c^v_{11}}{n} \quad (3.6)$$

$$DR^v_{max} = \left(\prod_{k=1}^{p}(1 - v_k \cdot icf_k)\right) \cdot \frac{\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} c^v_{ik} + \sum_{j=1}^{k-1} c^v_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} c^v_{ik}}{n} \quad (3.7)$$

$$DR^v_W = \left(\prod_{k=1}^{p}(1 - v_k \cdot icf_k)\right) \cdot \frac{1}{n \cdot w_{11}} \cdot \left(\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} w_{ik} \cdot c^v_{ik} + \sum_{j=1}^{k-1} w_{kj} \cdot c^v_{kj}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} w_{ik} \cdot c^v_{ik}\right) \quad (3.8)$$

The disclosure risk for a given weight matrix (minimal and maximal disclosure risks are obtained for particular disclosure risk weight matrices [43]) is the maximum of all disclosure risk values when all binary vectors v are considered. Therefore:

$$DR_{min} = \max_v \{DR^v_{min}\} \quad (3.9)$$

$$DR_{max} = \max_v \{DR^v_{max}\} \quad (3.10)$$

$$DR_W = \max_v \{DR^v_W\} \quad (3.11)$$

We label the vector that maximizes DR_W as $v(W)$. The subset of key attributes that maximize disclosure risk value is dependent not only of the data manipulations from the masked microdata, but also by the choice of the weights matrix. The vectors that maximizes the extreme disclosure risk measures are labeled as $v(min)$, and $v(max)$ respectively. Those three vectors may or may not be equal. The following two properties are true for any choice of disclosure risk weight matrix W (see Appendix 1 for their proof):

$$DR_{min} \leq DR_W \leq DR_{max} \quad (3.12)$$

$$0 \leq DR_W \leq 1 \quad (3.13)$$

The algorithm that generates all possible subsets of key attributes to compute disclosure risk has the exponential complexity $O(2^p)$. Fortunately, in a real initial microdata, the number of key attributes is low (usually less than 5). Moreover, the owner of the data can reduce the number of subsets to check if the inversion-change factor is either 0 or 1. It is easy to show that the key attributes corresponding to the inversion-change factor 0 will have values of 1 on their corresponding position in $v(W)$, for any disclosure risk weight matrix. Also, when the inversion-change factor is 1, their corresponding key attributes will be excluded from the search. We label the number of key attributes with icf equal with 0 as p_0, and the number of key attributes with icf equal with 1 as p_1. The number of checked combination will be reduce since the complexity becomes $O(2^{p-p0-p1})$.We are currently investigating several polynomial heuristics based on greedy algorithms, to determine the combination of key attributes that maximizes disclosure risk for a given matrix W. Our preliminary

results hold for local maximum values, and we hope to determine the global maximum.

4. EXPERIMENTAL RESULTS

We used simulated medical record billing data to perform a series of tests. After the identifier attributes are removed, the data contains the remaining attributes: *Age*, *Sex*, *Zip* and *Amount_Billed*. In our experiment, we used three sets of initial microdata; with sizes *n*=1,000 (*IM1000*), n=5,000 (*IM5000*), and *n*=25,000 (*IM25000*), all with the same set of attributes. For each initial microdata we considered the following set of key attributes: *KA* = {*Age*, *Sex*, *Zip*}. *Age* attribute is an ordered attribute, *Sex* is unordered, and *Zip* is partial ordered.

In the first masking process, we added random noise [28, 22] to *Age* attribute. The generated noise used in our experiments has a mean of 0 and a variance of cv^2, where c is a constant defined by the data owner, and v^2 is the variance for the attribute *Age.* However, more advanced noise perturbation methods are have been proposed, such as correlated noise [22], bias corrected correlated noise [41], or multiplicative noise [23]. We choose the simplest form of noise for simplicity. After the noise is added to the *Age* attribute, its values may decrease beyond 0, or surpass 100. Any such extreme values were bottom and top coded to 0 and 100.

Figure 4.1 shows the minimal and maximal disclosure risk values for all three initial microdata when the constant c varies. The inversion factor increases when the variance of the noise increase, but the rate of increase drops to 0 when the variance is too high.

In the second experiment, we sort the record based on number of occurrences of key values. We take p% records from the initial microdata with fewer occurrences of key values, and we reverse the sex of chosen records. The results when the percent of chosen records varies are depicted in Figure 4.2. In this experiment, we notice that disclosure risk decreases only when the percentage of chosen records is low. The cause of this surprisingly result is that the disclosure risk computed using only *Age* and *Zip* attributes is greater than the disclosure risk computed for all three attributes. By using disclosure risk measures, the owner of the data can avoid the mistake of overprotecting one specific attribute without any benefit in terms of disclosure risk.

In the last experiment, we apply local recoding [40] to attribute *Zip*. We sort the records from all three initial microdata based on number of occurrences for *Zip* attribute. When a zip code has the number of occurrences less than a fixed threshold t, the last two digits of that zip code are suppressed. When t is larger then the maximum number of occurrences for a zip code, all zip values will be reduced to three digits. Minimal and maximal disclosure risk values for various values of t are showed in Figure 4.3. Both minimal and maximal disclosure risks have the largest decrease in this last experiment. This occurs because the *Zip* attribute has the largest number of distinct values, compared with the other two keys attributes. In each case, when the threshold t is greater than the maximum number of occurrences for zip values, all zip values have been reduced to three digits, and the disclosure risk becomes constant.

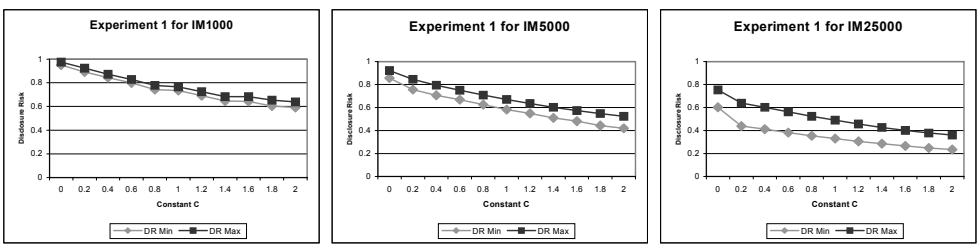

Figure 4.1. Random Noise added to *Age* Attribute.

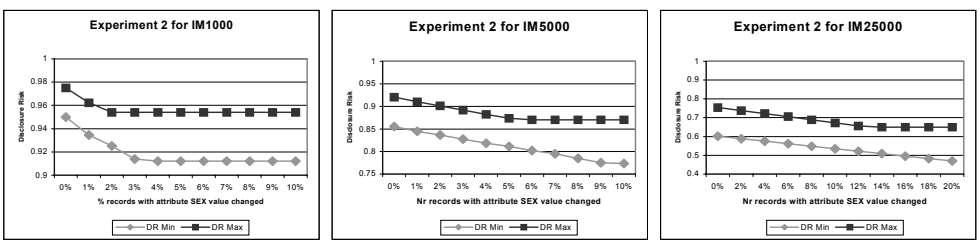

Figure 4.2. Change of the *Sex* attribute value for *p%* records.

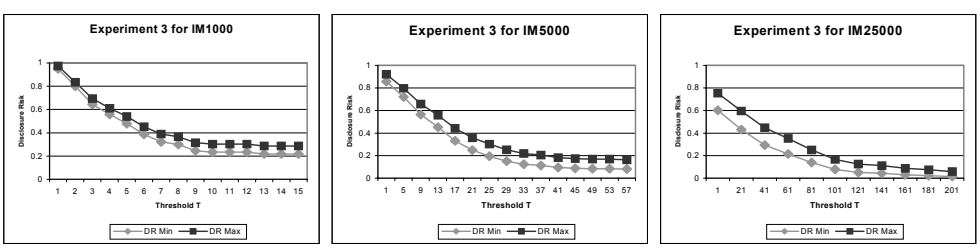

Figure 4.3. Local Recoding for *Zip* attribute.

Comparing the results from these experiments, a strong relationship between the decrease of the disclosure risk value and the number of distinct values in initial microdata for the masked key attribute can be seen. Based on these observations, data owners would typically benefit from starting the masking process by applying to the initial microdata disclosure control methods (such as global recording or microaggregation) that alter key attributes with large numbers of distinct values.

We have presented only a few experiments, but the generality of proposed disclosure risk measures makes possible experiments where several disclosure control methods are applied in succession to the initial microdata. The next experimental step is to include information loss computation (as those proposed in [14]), and to analyze the relationship between disclosure risk and information loss values.

5. CONCLUSIONS AND FUTURE WORK

A general framework for microdata disclosure control and a customizable disclosure risk measure were proposed in this paper. The boundaries for any disclosure risk measure were established between minimal and maximal disclosure risk. Those measures can be computed for any masking process, and they may become an important decision factor for the owner of the data in selecting which disclosure control methods he should apply to a given initial microdata. The experiments we performed showed that usually the proposed disclosure risk measures decreases when the data is modified more and more from its original form. There are situations when the owner of the data modifies more than is necessary some key attributes, but those kind of potential problems can be eliminated using the proposed disclosure risk measures. An important finding is that disclosure risk decrease substantially when a key attribute with a large number of distinct values is masked. Methods that hold good results are microaggregation [43] for ordered attributes and, as seen in our experiments, global and local recoding for partial ordered attributes.

The next step in this research is to compute information loss and to analyze its relation with disclosure risk. The disclosure risk versus information loss diagrams improve the understanding of the problem and help the owner of the data to choose the disclosure risk methods as well as the parameters for each method.

The final goal is to investigate those diagrams for different masking processes, and to derive patterns of applying more than one disclosure control method to a specific initial microdata to minimize both information loss and disclosure risk.

6. REFERENCES

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

[2] Agrawal, R., Kiernan, J., Srikant, R., Xu, Y. Hippocratic Databases. *Proc. of the 28th Int'l Conference on Very Large Databases*, Hong Kong, China, 2002.

[3] APA. The Australian Privacy Amendment (Private Sector) Act, 2000, *Available online at http://www.privacy.gov.au/publications/npps01.html*.

[4] Benedetti, R., Franconi, L. Statistical and Technological Solutions for Controlled Data Dissemination. *Pre-proceedings of New Techniques and Technologies for Statistics*, Vol. 1, 1998, 225-232.

[5] Benedetti, R., Franconi, L., Piersimoni, F. Per-record Risk of Disclosure in Dependent Data. *Proceedings of the Conference on Statistical Data Protection*, 1999.

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

[7] Bilen, U., Wirth, H., Muller, M. Disclosure Risk for Microdata Stemming from Official Statistics. *Statistica Neerlandica*, Vol. 46, 1992, 69-82.

[8] Chen, G., Keller-McNulty, S. Estimation of Deidentification Disclosure Risk in Microdata. *Journal of Official Statistics*, Vol. 14, No. 1, 1998, 79-95.

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

[10] Denning, D. E., Denning, P. J. Data Security. *ACM Computing Surveys*, Vol. 11, 1979, 227-249.

[11] Di Consigolio, L., Franconi, L., Seri, G. Assessing the Risk of Disclosure: An Experiment. *Joint ECE/EUROSTAT Work Session on Data Confidentiality*, Luxembourg, 2003.

[12] Dobra, A., Fienberg, S. E., Trottini, M. Assessing the Risk of Disclosure of Confidential Categorical Data. *Bayesian Statistics*, Vol. 7, Oxford University Press, 2003, 125-144.

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

[14] Domingo-Ferrer, J., Mateo-Sanz, J., Torra, V. Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk, *Pre-proceedings of ETK-NTTS'2001,* Vol. 2, Luxembourg: Eurostat, 2001, 807-826.

[15] Duncan, G., Keller-McNulty, S., Stokes, S. Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. *Technical Report LA-UR-01-6428*, Statistical Sciences Group, Los Alamos National Laboratory, 2001.

[16] Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J. Privacy Preserving Mining of Association Rules. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, 217-228.

[17] Elliot, M. J. DIS: A New Approach to the Measurement of Statistical Disclosure Risk. *International Journal of Risk Management*, 2000, 39 –48.

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

[19] Fuller, W. A. Masking Procedure for Microdata Disclosure Limitation. *Journal of Official Statistics*, Vol. 9, 1993, 383-406.

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

[21] HIPAA. Health Insurance Portability and Accountability Act. 2002, *Available online at http://www.hhs.gov/ocr/hipaa.*

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

[23] Kim, J. J., Winkler, W. E. Multiplicative Noise for Masking Continuous Data. American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 2001, cd-rom.

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

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

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

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

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

[29] Muralidhar, K., Sarathy, R. Recent Advances in Perturbation Methods. *Joint ECE/EUROSTAT Work Session on Data Confidentiality*, Luxembourg, 2003.

[30] Paass, G. Disclosure Risk and Disclosure Avoidance for Microdata. *Journal of Business and Economic Statistics*, Vol. 6, 1988, 487-500.

[31] Polettini, S. Some Remarks on the Individual Risk Methodology. *Joint ECE/EUROSTAT Work Session on Data Confidentiality*, Luxembourg, 2003.

[32] Reiss, S. P. Practical Data-Swapping: The First Steps. *ACM Transactions on Database Systems*, Vol. 9, No. 1, 1984, 20-37

[33] Rotenberg, M. (ed) The Privacy Low Sourcebook 2000: United States Law, International Law, and Recent Developments, *Electronic Privacy Information Center*, 2000.

[34] Samarati, P. Protecting Respondents Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, No. 6, 2001, 1010-1027.

[35] Skinner, C. J., Marsh, C., Openshaw, S., Wymer, C. Disclosure Control for Census Microdata. *Journal of Official Statistics*,1994, 31-51.

[36] Skinner, C. J., Elliot, M. J. A Measure of Disclosure Risk for Microdata. *Journal of the Royal Statistical Society*, Series B, Vol. 64, 2002, 855-867.

[37] Singh, A. C., Yu, F., Dunteman, G. H. MASSC: A New Data Mask for Limiting Statistical Information Loss and Disclosure. *Joint ECE/EUROSTAT Work Session on Data Confidentiality*, Luxembourg, 2003.

[38] Spruill, N. L.: The Confidentiality and Analytic Usefulness of Masked Business Microdata. *Proceedings of the American Statistical Association*, Section on Survey Research Methods, 1983, 602-613

[39] Steel, P., Sperling, J. The Impact of Multiple Geographies and Geographic Detail on Disclosure Risk: Interactions between Census Tract and ZIP Code Tabulation Geography. *Bureau of Census*, 2001.

[40] Takemura, A. Local Recoding by Maximum Weight Matching for Disclosure Control of Microdata Sets. *ITME Discussion Paper*, No.11, 1999.

[41] Tendick P., Matloff, N. A Modified Random Perturbation Method for Database Security. *ACM Transactions on Database Systems*, Vol. 19, No. 1, 1994.

[42] Trottini, M. Assesing Disclosure Risk and Data Utility: A Multiple Objectives Decision Problem. *Joint ECE/EUROSTAT Work Session on Statistical Data Confidentiality*, Luxembourg, 2003.

[43] Truta, T. M., Fotouhi, F, Barth-Jones, D. Disclosure Risk Measures for Microdata. *International Conference on Scientific and Statistical Database Management*, 2003 15 – 22.

[44] Truta, T. M., Fotouhi, F, Barth-Jones, D.: Disclosure Risk Measures for Sampling Disclosure Control Method. *Annual ACM Symposium on Applied Computing*, 2004.

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

7. APPENDIX

A. Proof of property 3.12 ($DR_{min} \le DR_W \le DR_{max}$)

A.1. We show that $DR_{min} \le DR_W$:

$$DR_W = \max_v \{DR_W^v\} = DR_W^{v(W)} \ge DR_W^{v(min)} =$$

$$\left(\prod_{k=1}^{p}(1-v(min)_k \cdot icf_k)\right) \cdot$$

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

$$=\left(\prod_{k=1}^{p}(1-v(min)_k \cdot icf_k)\right) \cdot \frac{1}{n \cdot w_{11}} \cdot$$

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

$$= DR_{min}^{v(min)} + \left(\prod_{k=1}^{p}(1-v(min)_k \cdot icf_k)\right) \cdot \frac{1}{n \cdot w_{11}} \cdot$$

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

$$\ge DR_{min}$$

A.2. We show that $DR_W \le DR_{max}$:

$$DR_{max} = \max_{v}\{DR_{max}^{v}\} = DR_{max}^{v(max)} \geq DR_{max}^{v(W)} =$$

$$\left(\prod_{k=1}^{p}(1 - v(W)_k \cdot icf_k)\right) \cdot \frac{\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} c_{ik}^{v(W)} + \sum_{j=1}^{k-1} c_{kj}^{v(W)}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} c_{ik}^{v(W)}}{n} =$$

$$\left(\prod_{k=1}^{p}(1 - v(W)_k \cdot icf_k)\right) \cdot \frac{1}{n \cdot w_{11}} \cdot \left(\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} w_{11} \cdot c_{ik}^{v(W)} + \sum_{j=1}^{k-1} w_{11} \cdot c_{kj}^{v(W)}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} w_{11} \cdot c_{ik}^{v(W)}\right) \geq$$

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

B. Proof of property 3.12 ($0 \leq DR_W \leq 1$)

B.1. We show that $0 \leq DR_W$:

$$DR_W \geq DR_{min} = \max_{v}\{DR_{min}^{v}\} = DR_{min}^{v(min)} =$$

$$\left(\prod_{k=1}^{p}(1 - v(min)_k \cdot icf_k)\right) \cdot \frac{c_{11}^{v(min)}}{n} \geq 0.$$

We used that $0 \leq icf_k \leq 1$(from definition), and, consequently, all multiplication terms are positive.

B.1. We show that $DR_W \leq 1$:

$$DR_W \leq DR_{max} = \max_{v}\{DR_{max}^{v}\} = DR_{max}^{v(max)} =$$

$$\left(\prod_{k=1}^{p}(1 - v(max)_k \cdot icf_k)\right) \cdot \frac{\sum_{k=1}^{t}\frac{1}{k}\left(\sum_{i=1}^{k} c_{ik}^{v(max)} + \sum_{j=1}^{k-1} c_{kj}^{v(max)}\right) + \sum_{k=t+1}^{n}\frac{1}{k}\sum_{i=1}^{t} c_{ik}^{v(max)}}{n} \leq$$

$$\left(\prod_{k=1}^{p}(1 - v(max)_k \cdot icf_k)\right) \cdot \frac{\sum_{i=1}^{t}\sum_{k=1}^{n} c_{ik}^{v(max)}}{n} \leq$$

$$\left(\prod_{k=1}^{p}(1 - v(max)_k \cdot icf_k)\right) \cdot \frac{t}{n} \leq \left(\prod_{k=1}^{p}(1 - v(max)_k \cdot icf_k)\right) \leq 1$$

We used that $0 \leq icf_k \leq 1$(from definition), $v(max)_k$ is 0 or 1, and, consequently, all multiplication terms are in the interval [0 , 1].