ACTA UNIVERSITATIS APULENSIS No 11/2006

Proceedings of the *International Conference on Theory and Application of Mathematics and Informatics ICTAMI 2005* - Alba Iulia, Romania

# GLOBAL DISCLOSURE RISK MEASURES AND $K$-ANONYMITY PROPERTY FOR MICRODATA

Traian Marius Truţă,
Farshad Fotouhi and Daniel Barth-Jones

Abstract. In today's world, governmental, public, and private institutions systematically release data which describes individual entities (commonly referred as microdata). Those institutions are increasingly concerned with possible misuses of the data that might lead to disclosure of confidential information. Moreover, confidentiality regulation requires that privacy of individuals represented in the released data must be protected. To protect the identity of individual entities from the microdata a large number of disclosure control methods have been proposed in the literature (such as sampling, simulation, data swapping, microaggregation, etc.). To compare different approaches to achieve data protection, various disclosure risk measures have been proposed in the literature. We introduced in our earlier papers a customized global disclosure risk measure that varied between a minimal and maximal value. In the mean time, Samarati and Sweeney have introduced a property, called $k$-anonymity, which must be satisfied by a microdata to guarantee the protection of individual entities [Samarati 2001, Sweeney 2002a]. In this paper we describe our disclosure risk measures, the k-anonymity property, and then we compare their advantages and disadvantages. The global disclosure risk measures offer more information about the level of protection and they can be customized based on the specific privacy requirements for a given microdata. On the other end, k-anonymity property can be obtained automatically with efficient algorithms, while the usage of the global disclosure risk measures still involves human intervention.

## 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 [Trottini 2003]. Moreover, confidentiality regulation requires that privacy of individuals represented in the released data must be protected. Some regulations that enforce the privacy of individuals are: the *US Health Insurance Portability and Accountability Act (HIPAA)* [HIPAA 2002], the *Canadian Standard Association's Model Code for the Protection of Personal Information* [Rotenberg 2000] and the *Australian Privacy Amendment Act 2000* [APA 2000].

*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 individuals in the data, and, thereby, disclosing information about these individuals [Bethlehem 1990, Tendick 1994].

There are two types of disclosures, namely, identity disclosure and attribute disclosure. *Identity disclosure* refers to identification of an entity (person, institution) and *attribute disclosure* occurs when the intruder finds out something new about the target entity [Lambert 1993]. We notice that identity disclosure does not imply attribute disclosure. It may happen that the intruder does not find anything new when he identifies an entity. Also we can have attribute disclosure without identity disclosure.

We refer to initial data as *microdata*. *Microdata* represents a series of records, each record containing information on an individual unit such as a person, a firm, an institution, etc [Willemborg et al. 2001]. 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.). At present, microdata is released for use by the third party after the data owner has masked the data to limit the possibility of disclosure. Typically, names and other identifying information are removed from the microdata before being released for research use. We will call the final microdata *masked* or *released microdata* [Dalenius et al. 1982]. We will use the term *initial microdata* for microdata where no disclosure control methods were applied.

The major goal of disclosure control for microdata is to protect the confidentiality of the individuals. Several disclosure control techniques were proposed in the literature (for a good survey see [Willemborg et al. 2001]). To

increase the level of protection, several methods are usually applied in succession in the disclosure control process. Some of the most used disclosure control techniques are: *global recoding* (also known as *generalization*) [McGuckin et al. 1990, Sweeney 2001b], *microaggregation* [Domingo-Ferrer et al. 2002], *sampling* [Skinner et al. 1994], *suppression* [Samarati 2001b, Little 1993], and *data swapping* [Dalenius et al. 1982, Reiss 1984].

## 2. Global disclosure risk measures

Global Disclosure Risk Measures

*Disclosure risk* is the risk that a given form of disclosure will be encountered if masked microdata is released [Chen et al. 1998]. *Information loss* is the quantity of information in the initial microdata, which does not occur in masked microdata because of disclosure control methods [Willemborg et al. 2001]. When protecting the confidentiality of individuals, the data owner must satisfy the two conflicting requirements: protecting confidentiality for the entities from the initial microdata and maintaining analytic properties in the masked microdata [Kim et al. 2001]. The ultimate goal is minimizing disclosure risk so as to comply with existing regulations, while simultaneously minimizing information loss for statistical inference [Fellegi 1972]. Since fully optimal minimization of both measures is not possible (decreasing disclosure risk will usually lead to increase information loss and vice versa), the data owner must select a compromise between disclosure risk and information loss values [Duncan et al. 2001].

As we mentioned earlier, masked microdata is already de-identified, that is direct identifiers like name, address etc. are removed. However, it is possible for an intruder to link a record from masked microdata to some additional data he/she may know and to be able to disclose information about a particular target. Because each record represents a particular entity, we may consider defining disclosure risk at the record level. As mentioned before, for a non-empty masked microdata, disclosure risk is not zero. The goal of any disclosure control method is to minimize the risk of disclosure. One way to achieve this is to consider a threshold value for disclosure risk. The risk for masked microdata should be less or equal of the threshold value. In this case the microdata is considered safe, and if the disclosure risk is greater then the threshold value the microdata is considered unsafe.

Considerable research on disclosure risk assessment [Adam et al. 1989, Benedetti et al. 1998, Bilen et al. 1992, Denning et al. 1979, Elliot 2000, Fuller 1993, Greenberg et al. 1992, Lambert 1993, Paass 1988, Skinner et al. 2002, Spruill 1983] has resulted in a variety of proposed disclosure risk models, but all researchers are unanimous in the conclusion that the risk of disclosure 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 [Steel et al. 2001]. The selected attributes are called keys in disclosure avoidance literature [Willemborg et al. 2001]. 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 log-normal [Fienberg et al. 1998], Poisson-Gamma [Bethlehem et al. 1990], Dirichlet-multinomial [Takemura 1999], and negative-binomial [Chen et al. 1998]. Greenberg and Zayatz have proposed a procedure that is not dependent on a parametric statistical distribution [Greenberg et al. 1992]. Other approach was proposed by Lambert who defined disclosure risk as matter of perception [Lambert 1993]. *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 [Lambert 1993].

Recent work can be categorized into two directions: individual and global disclosure risk. Benedetti and Franconi introduced *individual risk methodology* [Benedetti et al. 1998]. 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. Other papers extended this approach [Benedetti et al. 1999, Polettini 2003, Di Consigolio et al. 2003]. *Global disclosure risk* is defined in terms of the expected number of identifications in the released microdata. Elliot and Skinner define disclosure risk measures as the proportion of correct matches amongst those records in the population that match a sample unique masked microdata record [Elliot 2000, Skinner et al. 2002]. Other global disclosure risk measures have been proposed in [Polettini 2003, Dobra et al. 2003].

In the area of disclosure control for microdata we introduced three global disclosure risk measures. Our disclosure risk measures compute the global disclosure risk for the released data and are not linked to a target individual [Polletini 2003]. . We proposed as initial metrics, 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.

For simplicity in this paper we will describe in detail our global disclosure risk for microdata, different disclosure control methods and a general disclosure risk formulation were presented in other papers [Truta et al. 2003a, Truta et al. 2003b, Truta et al. 2004a, Truta et al. 2004b].

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 data owner. Usually, the owner specifies a minimum size for a group. More formally, let be $X = \{x_1, x_2, \ldots, x_n\}$ 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 $P = \{C_1, C_2, \ldots, C_{m(P)}\}$ of $X$ is a partition where the size of group $C_i$, $1 \leq i \leq m(P)$ 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

$$\bar{x}_{c_i} = \frac{1}{|C_i|} \cdot \sum_{x_l \in C_i} x_l \tag{1}$$

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

$$mim_{p \in P_k} \sum_{i=1}^{m(P)} \sum_{x_j \in C_i} |x_j - \bar{x}_{c_i}| \tag{2}$$

where m(P) is not part of the input [Oganian et al 2001].

To quantify disclosure risk is a very difficult task. Usually, an intruder uses some external information, and together with the masked microdata he can match the corresponding records and, subsequently, disclose confidential individual information. The external information is a very general concept, and, therefore, we need to make assumptions about this external knowledge in order to predict the disclosure risk. To simplify our formulation we consider the initial microdata as a set of records with values from three types of attributes:

identifier, confidential and key attributes. The attributes are split into three categories as follow:

- $I_1, I_2, ..., I_m$ are identifier attributes such as *Name* and *SSN* that can be used to identify a record. Those attributes are present only in initial microdata because express information which can lead to 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.

The first assumption we make is that the intruder does not have specific or confirmed knowledge of any confidential information. 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. Since, the owner of the data often does not have complete knowledge about the external information available to an intruder, by using this assumption; the data owner will be able to determine whether the disclosure risk is under an acceptable disclosure risk threshold value. This assumption does not reduce the generality of the problem.

The microaggregation method preserves the same number of records ($n$) in the masked microdata as in the initial microdata. We cluster the data from both microdata sets based on their key values. In the statistical disclosure control literature, such clusters are typically referred to as *equivalence classes* [Zayatz 1991] or *cells* [Chen et al. 1998]. We define the following notations for initial microdata:

- $F$ – the number of clusters;
- $Ak$ – the set of elements from the $k$-th cluster for all $k$, $1 \leq k \leq F$;
- $F_i = |\{A_k || A_k| = i, \text{ for all } k = 1, .., F\}|$ for all $i$, $1 \leq i \leq n$. $F_i$ represents the number of clusters with the same size;

- $n_i = |\{x \in A_k | |A_k| = i, \text{ for all } k = 1, .., F\}|$ for all $i$, $1 \leq i \leq n$. $n_i$ represents the number of records in clusters of size $i$.

Similar notations are defined for the masked microdata set:

- $f$ – the number of clusters with the same values for key attributes.
- $M_k$ – the set of elements from the $k$-th cluster for all $k$, $1 \leq k \leq f$.
- $f_i = |\{M_k | |M_k| = i, \text{ for all } k = 1, ..., f\}$ for all $i$, $1 \leq i \leq n$. fi represents the number of clusters with the same size.
- $t_i = |\{x \in M_k | |M_k| = i, \text{ for all } k = 1, ..., f\}$ for all $i$, $1 \leq i \leq n$. $t_i$ represents the number of records in clusters of size $i$.

To relate initial microdata to masked microdata we define a $n$ x $n$ matrix *called the classification matrix* $C$. 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 microdata. Mathematically, this definition can be expressed in the following form: For all $i = 1, ..., n$ and for all $j = 1, ..., n$; $c_{ij} = |\{x \in M_k \text{ and } x \in A_p | |M_k| = i, \text{for all } k = 1, ..., f \text{ and } |A_p| = j, \text{ for all } p = 1, ..., F\}$.

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

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

The first measure of disclosure risk, called minimal disclosure risk, is based on the percentage of unique records, which is discussed by Fienberg [Fienberg et al. 1998]. This measure is defined as:

$$DR_{\min} = \frac{c_{11}}{n} \tag{3}$$

For the second measure, called maximal disclosure risk, the distribution of the records that are not unique in both initial and masked microdata is considered. The probability of record linkage is included in this formulation:

$$DR_{\max} = \frac{\sum_{k=1}^{m} \sum_{j=1}^{k} c_{kj}}{n} \tag{4}$$

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

$$W = \begin{pmatrix} w_{11} & 0 & ... & 0 \\ w_{21} & w_{22} & ... & 0 \\ ... & ... & ... & ... \\ w_{n1} & w_{n2} & ... & w_{nn} \end{pmatrix} \tag{5}$$

with the following properties:

- $w_{jj} \geq w_{j+1j} \geq \ldots \geq w_{nj}$ for all $j$, $1 \leq j \leq n$.
- $w_{i1} \geq w_{i2} \geq \ldots \geq w_{ii}$ for all $i$, $1 \leq i \leq n$.
- $w_{j1} \geq w_{j+1,2} \geq \ldots \geq w_{n,n-j+1}$ for all $j$, $1 \leq j \leq n$.
- $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = n$.

Disclosure risk weight matrix increases the importance of unique values relative to the rest of records, and likewise, attributes a greater importance for records with double occurrences relative to records with greater frequencies, and so on. A detailed explanation of disclosure risk weight matrix can be found in [Truta et al. 2003b]. The owner of the data defines the weight matrix, and this matrix captures particularities of that specific initial microdata.

The last formula proposed for disclosure risk, called weighted disclosure risk, is:

$$DR_W = \frac{1}{n \cdot w_{11}} \sum_{k=1}^{n} \frac{1}{k} \sum_{j=1}^{k} w_{kj} \cdot c_{kj} \tag{6}$$

Please note that when $c_{11} = n$ and all other weights are 0 in disclosure risk weights matrix, $DR_w = DR_{\min}$. Also when all weights are equal with $c_{11}$ in disclosure risk weights matrix $DR_w = DR_{\max}$.

## 3. $k$-Anonymity Property

Sweeney and Samarati introduced the property called k-*anonymity* to characterize the danger of identity disclosure [Samarati 2001, Sweeney 2002a]. In their papers the term *quasi-identifier attributes* is used instead of key attributes.

The *k-anonymity property* for a masked microdata (MM) is satisfied if with respect to key attribute set (KA) if every count in the frequency set of MM with respect to KA is greater or equal to $k$ [LeFevre et al. 2005]. In Table 1, we show an example of masked microdata where 2-anonymity is satisfied.

| Age | Zip | Sex | Illness |
|---|---|---|---|
| 50 | 43102 | M | Colon Cancer |
| 30 | 43102 | F | Breast Cancer |
| 30 | 43102 | F | HIV |
| 20 | 43102 | M | Diabetes |
| 20 | 43102 | M | Diabetes |
| 50 | 43102 | M | Heart Disease |

Table 1: Patient masked microdata satisfying 2-anonymity.

In this example the set of key attributes is composed from: Age, Zip and Sex. We notice that a simple SQL statement help us check whether a relation adhere to $k$-anonymity:

SELECT COUNT(*) FROM Patient GROUP BY Sex, Zip, Age.

If the results include groups with count less than $k$, the relation Patient does not have k-anonymity property with respect to KA = {Age, Zip and Sex}.

The disclosure control methods used to achieve $k$-anonymity are most of the times generalization (also known as global recoding) and suppression.

Generalization is used with categorical attributes such as *Zip Code* and *Sex*. This technique is similar with microaggregation except it applies to categorical attributes while microaggregation is performed over continuous attributes. The domain for an attribute that needs to be generalized is extended to a *domain generalization hierarchy*, which includes all possible groups for that

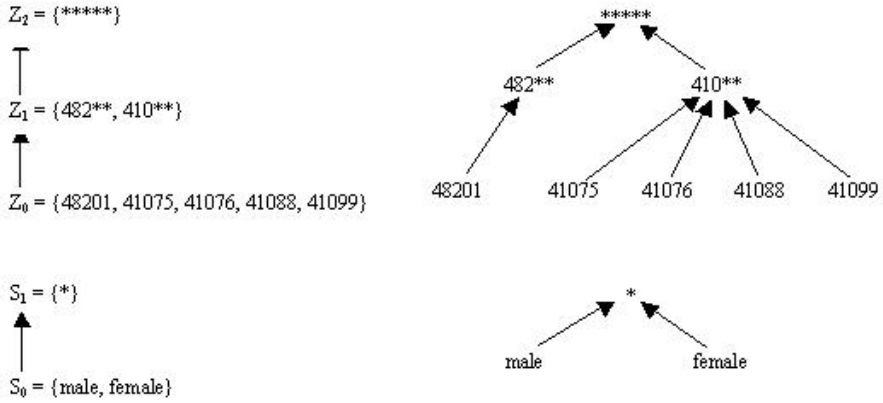


Figure 1: Examples of domain and value generalization hierarchies

specific attribute. For the attribute *Zip Code*, the domain contains all existing *Zip Codes*, while the domain generalization hierarchy contains all prefixes (without repetition) for the existing values [Samarati 2001b]. A domain generalization hierarchy is a total ordered relation between different domains that can be associated with an attribute. The values from different domains can be represented in a tree called value generalization hierarchy. We illustrate domain and value generalization hierarchy in Figure 1:

In order to apply generalization, the data owner must define the domain and value generalization hierarchies for the attributes he wants to generalize. The data owner has many choices based on the properties of each attribute. For instance, the Zip Code attribute can have a different generalization hierarchy with 6 different domains in which only one digit is removed at the time. The choice of the hierarchies (domain generalization mainly, the value generalization hierarchy usually is generated based on the chosen domain generalization hierarchy) is an important factor in the success of the masking process.

When two or more attributes are generalized the data owner can create a generalization lattice to visualize all possible combination (see Figure 2). Generalization lattices are introduced by LeFevre [LeFevre et al. 2005].

The generalization method (also called full domain generalization or global recoding) maps the entire domain of a key attribute in initial microdata to a

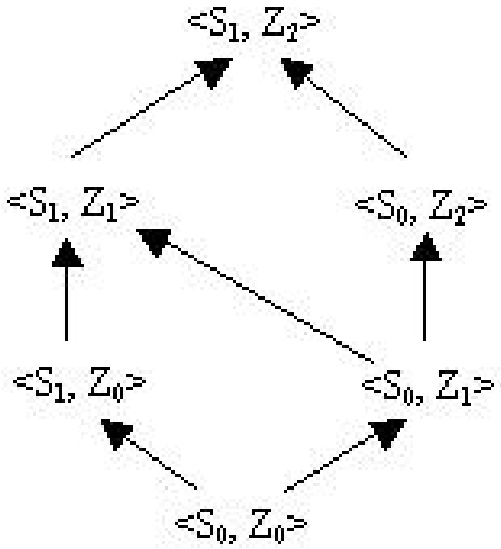


Figure 2: Generalization lattice for Zip and Sex attribute

more general domain from its domain generalization hierarchy [Samarati 2001, Sweeney 2002b].

Using only generalization every initial microdata can be transformed to a masked microdata that satisfy k-anonymity. Practical experiments have shown that the amount of generalization is too high and the resulting microdata will usually be useless. This is the reason why a second method called suppression is used to reduce the information loss created by generalization.

Suppression means to remove tuples from the microdata that destroy $k$-anonymity (have a frequency less than $k$). After each generalization we can easily compute the number of tuples that have a frequency of key attribute values less than $k$. If this number is below a choused threshold we would be better off if we remove those tuple and we avoid an extra generalization.

Using generalization and suppression we can obtain different masked microdata that satisfy $k$-anonymity property. It is easy to prove that if k-anonymity is achieved for a node $X$ in the generalization lattice, $k$-anonymity is satisfied for every node from the node $X$ to the upper level of the lattice [Samarati 2001]. From the construction of the lattice we know that on every path we

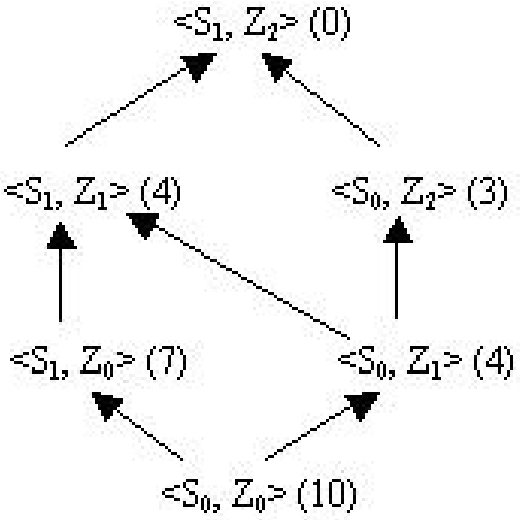


Figure 3: Example for minimal generalization with suppression threshold (TS)

lose information when we move up in the lattice. Therefore, the data owner is interested in finding the node or nodes that are closer to the bottom of the lattice. A node $X$ that satisfies $k$-anonymity and there is no other node $Y$ such that $X$ in on the path from root to $Y$ ($X$ different of $Y$) represents a minimal generalization [Samarati 2001]. The data owner wants to find one or all minimal generalization that creates $k$-anonymity property.

For the same initial microdata we can have different minimal generalization based on the threshold selected for suppression. In Figure 3, we show between the parentheses how many tuples does not satisfy $k$-anonymity for every generalization. It is easy to show that for every initial microdata and every generalization lattice, the numbers of records not satisfying $k$-anonymity decreases when the amount of generalization increases. Therefore on every path we must have increasing numbers from the upper level node to the bottom.

The Table 2 shows which node corresponds to minimal generalizations for different values of TS. We notice that the minimal generalization is not unique, but reduces considerably the number of generalizations that satisfy

k-anonymity.

| TS | $0, 1, 2$ | $3$ | $4, 5, 6$ | $7, 8, 9$ | $\geq 10$ |
|---|---|---|---|---|---|
| Node | $< S_1, Z_2 >$ | $< S_0, Z_2 >$ | $< S_0, Z_1 >$ | $< S_1, Z_0 >$and $< S_0, Z_1 >$ | $< S_0, Z_0 >$ |

Table 2: Minimal generalization for different threshold (TS) values

An important feature of $k$-anonymity is that there are several efficient algorithms which determine a minimal generalization with a suppression threshold that satisfy $k$-anonymity [Samarati 2001, LeFevre et al. 2005, Sweeney 2002b].

## 4. Comparison between Disclosure Risk and K-Anonymity

In the previous sections we presented disclosure risk measures for microaggregation and $k$-anonymity property for generalization and suppression. Although they are defined differently we could find similarities between our formulation for disclosure risk and $k$-anonymity definition. First we note that microaggregation and generalization are similar, we can consider microaggregation a form of generalization for continuous attributes. Therefore, we can talk about $k$-anonymity property for a masked microdata where microaggregation was applied. We will analyze in details this scenario with the use of illustrations. The second possibility, when we want to compute disclosure risk for a masked microdata with generalization and suppression, requires the use of the disclosure risk formulation for discrete attributes presented in a previous paper [Truta et al. 2004b].

To illustrate the disclosure risk measures and $k$-anonymity property we consider the initial microdata with three different masked microdata in Figure 4. Age and Sex are considered key attributes. In each masked microdata the microaggregation is applied for attribute Age with different sizes (2, 4 and 8). In Figure 5, we show the number of cluster with the same size and the number of records in corresponding clusters of the specified size.

| Initial Microdata | | |
|---|---|---|
| RecNo | Age | Sex |
| 1 | 8 | M |
| 2 | 10 | M |
| 3 | 19 | F |
| 4 | 23 | F |
| 5 | 37 | F |
| 6 | 43 | F |
| 7 | 68 | F |
| 8 | 72 | F |

| Initial Microdata | | |
|---|---|---|
| RecNo | Age | Sex |
| 1 | 9 | M |
| 2 | 9 | M |
| 3 | 21 | F |
| 4 | 21 | F |
| 5 | 40 | F |
| 6 | 40 | F |
| 7 | 70 | F |
| 8 | 70 | F |

| Initial Microdata | | |
|---|---|---|
| RecNo | Age | Sex |
| 1 | 15 | M |
| 2 | 15 | M |
| 3 | 15 | F |
| 4 | 15 | F |
| 5 | 55 | F |
| 6 | 55 | F |
| 7 | 55 | F |
| 8 | 55 | F |

| Initial Microdata | | |
|---|---|---|
| RecNo | Age | Sex |
| 1 | 35 | M |
| 2 | 35 | M |
| 3 | 35 | F |
| 4 | 35 | F |
| 5 | 35 | F |
| 6 | 35 | F |
| 7 | 35 | F |
| 8 | 35 | F |

Figure 4: Initial microdata and three masked microdata

| $IM$ | $n = 8$ | $n_1 = 8$ | $n_2 = \ldots = n_8 = 0$ |
|---|---|---|---|
| | $f = 8$ | $F_1 = 8$ | $F_2 = .. = F_8 = 0$ |

| $MM_1$ | $n = 8$ | $t_1 = 0$ | $t_2 = 8$ | $t_3 = \ldots = t_8 = 0$ |
|---|---|---|---|---|
| | $f = 4$ | $f_1 = 0$ | $f_2 = 4$ | $f_3 = \ldots = f_8 = 0$ |

| $MM_2$ | $n = 8$ | $t_1 = 0$ | $t_2 = 4$ | $t_3 = 0$ | $t_4 = 4$ | $t_5 = \ldots = t_8 = 0$ |
|---|---|---|---|---|---|---|
| | $f = 3$ | $f_1 = 0$ | $f_2 = 2$ | $f_3 = 0$ | $f_4 = 1$ | $f_5 = \ldots = f_8 = 0$ |

| $MM_3$ | $n = 8$ | $t_1 = 0$ | $t_2 = 2$ | $t_3 = t_4 = t_5 = 0$ | $t_6 = 6$ | $t_7 = t_8 = 0$ |
|---|---|---|---|---|---|---|
| | $f = 2$ | $f_1 = 0$ | $f_2 = 1$ | $f_3 = f_4 = f_5 = 0$ | $f_6 = 1$ | $f_7 = f_8 = 0$ |

Figure 5: Characterization of data based on cluster sizes

The next step in performing this example is to compute classification matrices for each masked microdata. Applying the definition, we obtain the matrices $C_1$, $C_2$ and $C_3$ of size 8 x 8 corresponding to $MM_1$, $MM_2$ and $MM_3$ respectively:

$$C_1 = \begin{pmatrix} 0 & 0 & ... & 0 \\ 8 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 0 \end{pmatrix} \quad C_2 = \begin{pmatrix} 0 & 0 & ... & 0 \\ 4 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ 4 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 0 \end{pmatrix}$$

$$C_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Since all records have a unique combination of key attributes in initial microdata, only first column has non-zero elements. We consider the following three weight matrices:

$$W_1 = \begin{pmatrix} 8 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 0 \end{pmatrix} \quad W_2 = \begin{pmatrix} \frac{2}{7} & 0 & ... & 0 \\ \frac{2}{7} & \frac{2}{7} & ... & 0 \\ ... & ... & ... & ... \\ \frac{2}{7} & \frac{2}{7} & ... & \frac{2}{7} \end{pmatrix}$$

$$W_3 = \begin{pmatrix} 4 & 0 & 0 & ... & 0 \\ 2 & 2 & 0 & ... & 0 \\ 0 & 0 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 0 \end{pmatrix}$$

All requirements for those matrices are considered. The owner of the data chooses the data values in the weight matrix, for which we provide three examples ($W_1$,$W_2$,and $W_3$). Such matrices instantiate the disclosure risk measure based on the data owner's privacy concerns. It is easy to notice that the first weight matrix correspond to $DR_{\min}$, the second to $DR_{\max}$. The value 2/7 is chosen due to the requirement that the sum of all weights must be equal with the number of records in the initial microdata, which is 8, and the requirement that all weights should be equal for computing maximal disclosure risk. In the case of $W_3$, records with double occurrence in masked microdata would be considered unsafe, but their weight is lowered compared with unique elements. The owner of the data considers all records with three or more occurrences safe, therefore their weight is 0. In Table 3, we show the disclosure risk values for each combination of sampling set and disclosure risk weight matrix.

| | W1 | W2 | W3 |
|---|---|---|---|
| MM1 | 0 | 0.5 | 0.25 |
| MM2 | 0 | 0.375 | 0.125 |
| MM3 | 0 | 0.25 | 0.0625 |

Table 3: Disclosure risk values

From this example we notice, as expected, that when the size parameter for microaggregation increases disclosure risk decrease.

Every combination of key attribute values occurs two or more times for each masked microdata presented above. On the other hand there is a key attribute value combination that occurs exactly twice for every microdata set. We can conclude that all three masked microdata satisfy 2-anonymity property and none of them satisfy 3-anonymity property.

By comparing the masked microdata sets from Figure 4, we note that there are some differences between them. The disclosure risk measures capture those differences by including in their formulations the frequency count of key attribute value combination (for maximal and weighted disclosure risk). We also note that minimal disclosure risk is 0 if and only if the masked microdata satisfy 2-anonymity property. More than that, a proper choice of weight matrix (with non zero values on all possible positions based on weight matrix requirements on the first k rows and zero value anywhere else) determines a

weighted disclosure risk with value 0 if and only if the masked microdata satisfy k-anonymity property. We can safely conclude that disclosure risk measure has a greater flexibility than k-anonymity and can be easily customized by the data owner based on initial microdata privacy requirements. On the other end, k-anonymity property can be obtained automatically with efficient algorithms.

## 5. Conclusions

A customizable disclosure risk measure for microdata disclosure control technique and k-anonymity property were discussed in this paper. Those disclosure risk measures and the degree of anonymity 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. We established the similarities that exist between those two methods to assess the level of protection for individuals represented in the initial microdata. The global disclosure risk measures offer more information about the level of protection and they can be customized based on the specific privacy requirements for a given microdata. On the other end, k-anonymity property can be obtained automatically with efficient algorithms, while the usage of the global disclosure risk measures still involves human intervention.

## 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, 515-556

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

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

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

[5] 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

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

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

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

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

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

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

[12] 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

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

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

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

[16] Fienberg S.E., Markov U.E. (1998), *Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data*, Journal of Official Statistics, Vol. 1, No. 4, 385-397

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

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

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

[20] Kim J.J., Winkler W.E. (2001), *Multiplicative Noise for Masking Con-*

*tinuous Data*, American Statistical Association, Proceedings of the Section on Survey Research Methods, cd-rom

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

[22] LeFevre K., DeWitt D., Ramakrishnan R. (2005), *Incognito: Efficient Full-Domain K-Anonymity*, Proceedings of the SIGMOD Conference

[23] Little R. J. A. (1993), *Statistical Analysis of Masked Data*. Journal of Official Statistics, Vol. 9, No. 2, 407 – 426

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

[25] Oganian A., Domingo-Ferrer J. (2001), *On the Complexity of Optimal Microaggregation for Statistical Disclosure Control*, Statistical Journal of the United Nations Economic Commission for Europe, Vol. 18, No. 4, 345-354

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

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

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

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

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

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

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

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

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

[35] Sweeney L. (2002a), *K-Anonymity: A Model for Protecting Privacy*, In-

ternational Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, 557-570

[36] Sweeney L. (2002b), *Achieving K-Anonymity Privacy Protection using Generalization and Suppression*, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, 571-588

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

[38] Tendick P., Matloff, N. (1994), *A modified random perturbation method for database security*, ACM Transactions on Database Systems, Volume 19, Number 1

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

[40] Truta T.M., Fotouhi F., Barth-Jones D. (2003a), *Disclosure Risk Measures for Microdata*, Proceedings of the International Conference on Scientific and Statistical Database Management, Cambridge, Ma, 15 – 22

[41] Truta T.M., Fotouhi F., Barth-Jones D. (2003b), *Privacy and Confidentiality Management for the Microaggregation Disclosure Control Method*, Proceedings of the Workshop on Privacy and Electronic Society, In Conjunction with 10th ACM CCS, Washington DC, 21 – 30

[42] Truta T.M., Fotouhi F., Barth-Jones D. (2004a), *Disclosure Risk Measures for Sampling Disclosure Control Method*, Proceedings of ACM Symposium on Applied Computing, 301-306

[43] Truta T.M., Fotouhi F., Barth-Jones D. (2004b), A*ssessing Global Disclosure Risk Measures in Masked Microdata*, Proceedings of the Workshop on Privacy and Electronic Society, In Conjunction with 11th ACM CCS, Washington DC, 85 – 93

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

[45] Zayatz L.V. (1991), *Estimation of the Number of Unique Population Elements Using a Sample*, Proc. Survey Research Methods Section, 369-373. 1991

Traian Marius Truţă
Department of Computer Science
Northern Kentucky University

Highland Heights, KY, 41099, USA
trutat1@nku.edu

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

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