AUTOMATIC GENERATION OF MASKED MICRODATA

by
Traian Marius Truta,
Farshad Fotouhi, Daniel Barth-Jones

Abstract. 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. Several microdata disclosure control frameworks exist in literature but they focus on specific disclosure problems. Our proposed framework attempts to define the microdata disclosure control problem more generally. In this paper we describe the architecture of a software system called AMMG (Automatic Masked Microdata Generator). The system will generate masked microdata with low disclosure risk and information loss. A general framework for microdata disclosure control is proposed for this system. Also, existing disclosure risk measures are extended by this research. Variables in the microdata are classified at two-levels, one specified by the data owner and the other indicating the knowledge states of potential data intruders. These classifications form the basis for organizing disclosure risk scenarios. The disclosure risk measure presented in this paper is validated in our illustrations.
Keywords: Statistical Disclosure, Data Privacy, Microdata, Disclosure Risk

1. Introduction

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.). Due to existing regulations in various areas, microdata should be released for use by the third party 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 *masked* or *released microdata* (Dalenius et al 1982).

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 (Bethlehem 1990, Tendick 1994).

There are two types of disclosures, namely**,** identity disclosure and attribute disclosure. *Identity disclosure* refers to 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 1991).

A third party can access masked microdata for many purposes. For example, a hospital releases the masked microdata for all its patients to a pharmaceutical company that will use this data to determine the frequency of occurrence for specific diseases. This information can then be used to perform market analysis. Many similar scenarios exist for medical data, census data, survey data, etc. In this example, an aggregate table is created from masked microdata. As useful as those aggregate tables are, the underlying microdata provides more valuable information. As a result, the demand for detailed masked microdata by public and private research communities has been increasing (McGuckin et al 1990). Therefore, today, the trend is to release masked microdata, not only masked aggregate tables, because of the flexibility it offers in extracting a great amount of information by a third party.

As seen in the following example, usually, a third party extracts different statistical characteristics from masked microdata; therefore, disclosure control is sometimes called statistical disclosure control (Willemborg et al 2001).

Figure 1.1 shows relationship between microdata (IM), masked microdata (MM), table data (T) and masked table data (MT). In the figure, the function *f* is applied to Initial Microdata (IM) to generate Masked Microdata (MMD). Function *f* is called masking function for microdata, while *f*' is collection of functions which are applied to MM to generate Masked Tables (MT). Additionally, *g* is a collection of functions that is used to create aggregate tables, while *g'* is a collection of functions that is applied to the aggregate tables to create masked aggregate tables.

In the example from Figure 1.1, we do not fully specify the masking functions; we focus on relationships between different types of data that occur in disclosure control problem. Note that dashed line in a cell means that the corresponding value was suppressed by a disclosure control technique.

In very broad terms, *disclosure risk* is the risk that a given form of disclosure will encounter 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 (Willemborg et al 2001).

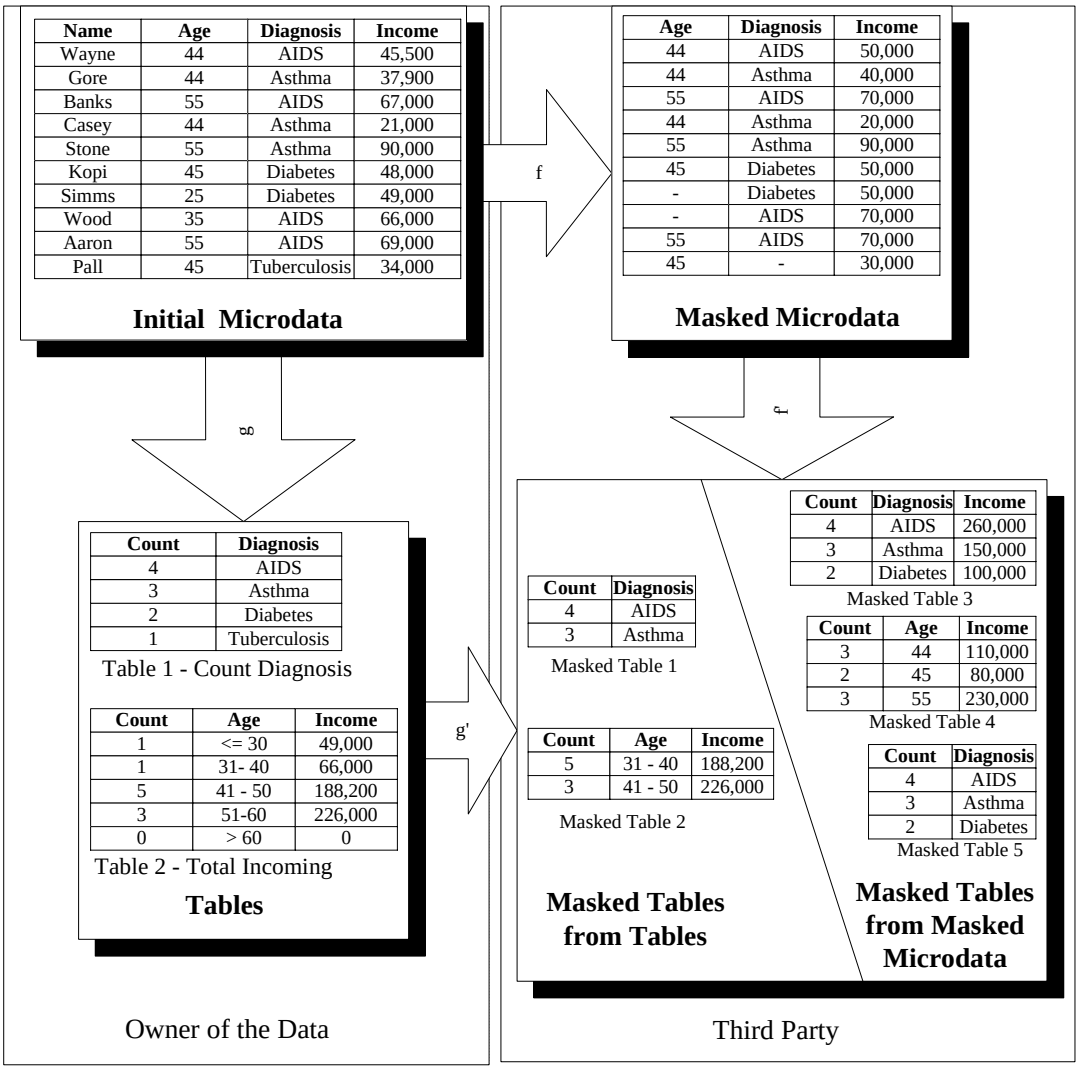

Figure 1.1 - Relationships between IM, MM, T and MT.

The problem of quantifying disclosure risk is a very difficult one because disclosure 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 in order to predict the disclosure risk. Unfortunately, the assumptions we are forced to make are sometimes not accurate with a given masked microdata set.

The masked microdata is used for statistical purposes. Therefore, it is often the case that only a subset (called sampling factor) of records from the initial microdata is released (usually random sampling is preferred). If *N* is the number of elements in initial microdata and *n* the released number of elements we call *sf = n / 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, as stated before, 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 within a given range different statistical measures is called *statistical integrity* (Fellegi 1972).

The major goal of disclosure control for microdata is to protect the confidentiality of the data. Several statistical disclosure control techniques (global recoding, local suppression, microaggregation, sampling, simulation, adding noise, rounding, post randomization method, data swapping etc.) were proposed in the literature (Adam et al 1989, Tendick et al 1994, McGuckin et al 1990, Duncan et al 1991, Dalenius et al 1982, Domingo-Ferrer et al 2002, Kim 1986, Muralidhar et al 1999). To increase confidentiality, more than one method is often applied in disclosure control process. In this paper we do not describe those methods further.

In this paper, we describe the architecture of our Automatic Masked Microdata Generator (AMMG) software system that integrates existing statistical disclosure methodologies via a general framework for microdata disclosure control. We have also generalized and adapted a disclosure risk measure for a target unit presented in the literature (Willemborg 2001). Our proposed framework attempts to define the microdata disclosure control problem more generally. Toward this end, variables in the microdata are classified at two-levels, one specified by the data owner and the other indicating the knowledge states of potential data intruders. These classifications form the basis for organizing disclosure risk scenarios. The disclosure risk measure presented in this paper is validated in our illustrations.

Other researchers have developed software products related to microdata disclosure control. One of the major products released is μ-Argus (Hundepool et al 1999). It supports a small number of disclosure control methods, and it implements disclosure risk requirements using a threshold value based on frequency of occurrence. The risk for masked microdata should be less then or equal to the threshold value. Our system extends the disclosure risk and information loss component. We also allow for adding new disclosure methods as they become available. Datafly (Sweeney 1997) is another system similar to μ-Argus. Our system differs from the existing systems in the disclosure control framework, through improvements in the disclosure risk measures,

and through the use of an open architecture for the addition of new disclosure control techniques.

The remainder of this paper is organized as follows: Section 2 contains the architecture of AMMG software, Section 3 describes Initial Microdata Analyzer component which is based on our proposed general framework for microdata disclosure control, Section 4 describes the Disclosure Risk Analyzer component which contains our generalization, and Section 5 contains future work in this area of disclosure control for microdata.

2. Architecture of the Automatic Masked Microdata Generator

In this section we will present architecture of the AMMG (Automatic Masked Microdata Generator) system. The system consists of five components namely: Data Converter, Initial Microdata Analyzer, Disclosure Method Selection, Disclosure Risk and Information Loss Analyzer, and Masked Microdata Generator.

Figure 2.1 shows the relationship among these components. *Data Converter* is a component responsible for mapping different data types and formats to a uniform initial format called initial microdata. Initial microdata represents a series of records, each record containing information on an individual unit, such as a person, or a firm. In the next section, we provide an in depth description of the initial microdata and the masked microdata.

The *Initial Microdata Analyzer* categorizes the initial microdata into three groups namely: Identifiers, Keys and Confidential attributes. Identifiers are those attributes that can easily be used to identify a record such as name and SSN. Keys correspond to those attributes that may be known by an intruder. Examples of such attributes are zip code and country. Confidential attributes are those attributes that are rarely known by an intruder, such as principle diagnosis for a patient. The Initial Microdata Analyzer allows for manual intervention for adjusting the attribute categories as needed by a user. The general framework for microdata disclosure control presented in Section three provides more understanding of this component.

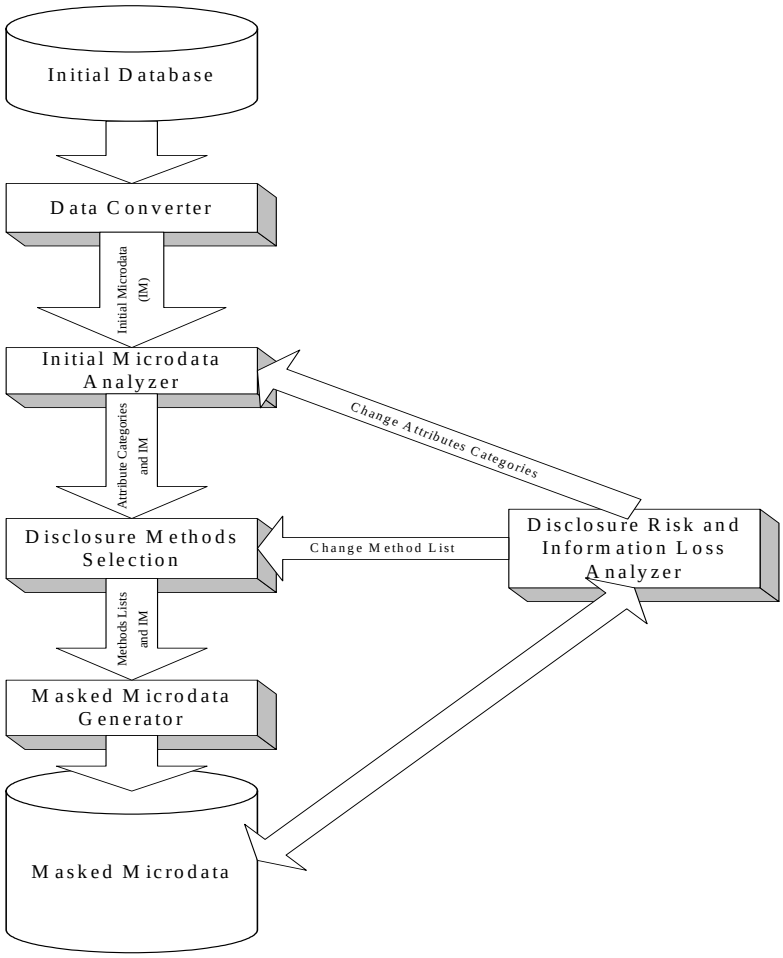

Figure 2.1 – Architecture of AMMG

Given an initial microdataset, the *Disclosure Method Selection* module allows a user to manually choose from a list of disclosure methods such as Global Recoding, Local Suppression, Rounding, and/or Data Swapping. A wizard can also determine this list automatically. If the wizard determines the list, the parameters associated with each method are then calculated automatically. Otherwise, the user can determine the parameters. The parameters associated with each method are then input by a user or by the wizard. The list of methods can contain 0 elements (the list is empty and no disclosure method is used in this situation for masking data), 1 (only one method is

used to mask data) or more than one method (the methods are applied one at the time in order to mask the data).

Masked Microdata Generator is the component that automatically, based on a given list of methods, computes the masked microdata. Masked Microdata is the output of AMMG Application, and is stored in a database. The user then has the option of saving the masked microdata for future use, such as comparing it with subsequent system output. Masked microdata has the same structure as initial microdata, except that the data is slightly modified to limit the possibility of disclosure. Typically, names and other identifying information are removed from initial microdata, and only masked microdata is released for research use.

Disclosure Risk and Information Loss Analyzer is the component that, based on initial microdata and masked microdata, estimates the value of disclosure risk and information loss. This module can also use the masked microdata to allow for a given threshold of disclosure risk and information loss to determine the optimal masked microdata among the computed ones. Section 4 describes an extension for the disclosure risk measure proposed by (Willemborg et al 2001).

Currently, we are in the process of implementing a prototype of the above system using a three-tier architecture (MySQL as a database server, Java Server Pages for project implementation and Apache Tomcat as a web server).

The system will walk the user through a series of tasks. First the user is autentificated by a user id and his password. Then, the initial microdata is selected from an existing database on mysql server and, then, the masked microdata name is chosen. In this step a project name is selected. We cannot have two projects or two masked microdata for different projects with the same name. The system will prompt the user to select other names if such a coincidence exists. The next screen shows all attributes from the initial microdata and the user can define them as identifier, key or confidential. We call this selection of attributes as Remove Identifier method. The user can, then, select one disclosure list method from the list shown on the screen. Each method requires its own setting. Currently, we have implemented only a subset of those methods. We have also implemented the management of all disclosure control methods. Each method can be deleted or updated at any time. After the desired list of methods is finalized the user generates masked microdata. He can see the masked microdata though our system or he can use it as any table in mysql system. In Apendix A we will show the current functionality of AMMG.

3. Initial Microdata Analyzer

The initial microdata consists of a set of n records with values from three types of attributes: identifier (I), confidential (S) and key attributes (K). Depending of the initial microdata, it is possible for the masked microdata not to have all three attribute types. We consider the initial microdata as a matrix with 3 partitions corresponding to

the different categories of attributes. The rows correspond to the units (individual units) and the columns represent the attributes.

$$IM = [I \mid K \mid S] \quad (3.1)$$

where

$$I = [\, i_{ij} \,] \text{ of order } n \text{ x } m \quad (3.2)$$

$$K = [\, k_{ij} \,] \text{ of order } n \text{ x } p \quad (3.3)$$

$$S = [\, s_{ij} \,] \text{ of order } n \text{ x } q \quad (3.4)$$

We labeled the attributes as follows: $I_1, I_2, \ldots, I_m, K_1, K_2, \ldots, K_p, S_1, S_2, \ldots, S_q$. For an attribute X we use the following notation for its associated domain: $\boldsymbol{D}(X)$.

Let us express the general form of the masked microdata:

$$M = [K' \mid S'] \quad (3.5)$$

where

$$K' = [\, k'_{ij} \,] \text{ of order } t \text{ x } p \quad (3.6)$$

$$S' = [\, s'_{ij} \,] \text{ of order } t \text{ x } q \quad (3.7)$$

The number of entities in the masked microdata can be different then the number from the initial microdata. The set of attributes in the masked microdata is obtained by removing identifier attributes from initial mirodata.

A record in initial microdata represents an entity. Because of the disclosure control, the identifier attributes are removed from this initial microdata, and values from key and confidential variables can be either suppressed (in this case their values are set to unknown, i.e., utilizing local suppression) or altered (if perturbative disclosure control are used). This motivates our use of the prime notation for key and confidential values in the microdata. (k_{uv} may be different of k'_{uv} and s_{uv} can also differ from s'_{uv} for some u and v).

Because of the simulation and sampling disclosure control methods, the number of records in initial microdata is, usually, not equal to the number of records in masked microdata.

We will use the following notations:

- n – the number of records in initial microdata;

- t – the number of records in masked microdata;
- r – the number of records from masked microdata with a matching correspondent in the initial microdata (one to one relationship).

Among n, t and r we have: $r \leq n$ and $r \leq t$.

We define the simulated factor as

$$fs = \frac{t - r}{t} \qquad (3.8)$$

The simulated factor represents the quantity of information simulated in masked microdata. The range for *fs* is between 0 and 1, but to preserve the validity of the microdata *fs* should be close to 0.

The factor

$$sf = \frac{r}{n} \qquad (3.9)$$

is called sampling factor. It represents the part of the initial microdata that is to be released to the public. The range for *sf* is between 0 and 1. We can have useful microdata for almost any value of the *sf*. For instance, if *sf* = 0.05 and the data is from a census with 100,000 respondents then the masked microdata will have 100,000 x 0.05 = 5,000 records. This number of records will be sufficient to preserve statistical properties of the initial microdata. It is clear that the amount of precision in the microdata is increased by the value of *sf*.

The above classification of attributes is made at the owner of the data level. We may have a similar classification at the researcher (third party) level. In this way, we can divide each record into two parts: known fields and unknown fields. This classification is made at the record level. We illustrate this with the following example. Let:

$$x_u = (i_{u1}, \ldots, i_{um}, k_{u1}, \ldots, k_{up}, s_{u1}, \ldots, s_{uq}) \text{ and}$$

$$x_v = (i_{v1}, \ldots, i_{vm}, k_{v1}, \ldots, k_{vp}, s_{v1}, \ldots, s_{vq}) \qquad (3.10)$$

be two records from initial microdata. Let us assume that both records are shown in the final microdata. Their form will be:

$$x'_u = (k'_{u1}, \ldots, k'_{up}, s'_{u1}, \ldots, s'_{uq}) \text{ and}$$

$$x'_v = (k'_{v1}, \ldots, k'_{vp}, s'_{v1}, \ldots, s'_{vq}) \qquad (3.11)$$

We pose the scenario where an intruder tries to disclose information from the above two records. The intruder has considerable external information (prior information) about the above individuals. Let ext_u and ext_v be two external information records that represent the above individuals.

$$ext_u = (i_{u,m1}, \ldots, i_{u,me}, k_{u,p1}, \ldots, k_{u,pf}, s_{u,q1}, \ldots, s_{u,qg}) \text{ and}$$

$$ext_v = (i_{v,m'1}, \ldots, i_{v,m'e'}, k_{v,p'1}, \ldots, k_{v,p'f'}, s_{v,q'1}, \ldots, s_{v,q'g'}) \quad (3.12)$$

where

$$1 \le m_1 < m_2 < \ldots < m_e \le m$$

$$1 \le p_1 < p_2 < \ldots < p_f \le p$$

$$1 \le q_1 < q_2 < \ldots < q_g \le q$$

$$1 \le m'_1 < m'_2 < \ldots < m'_{e'} \le m$$

$$1 \le p'_1 < p'_2 < \ldots < p'_{f'} \le p'$$

$$1 \le q'_1 < q'_2 < \ldots < q'_g \le q' \quad (3.13)$$

As can be seen from this example, the set of known attributes for the entities *u* and *v* are different. Therefore, the microdata (both initial and masked) can be seen as collection of entities each composed by known or unknown fields. A disclosure takes place if the intruder can use the released microdata to disclose information about unknown fields.

In an ideal scenario, the known fields will always be a subset of fields which states for identifier attributes and key attributes, however, there are situations where some confidential fields are also known fields and, therefore, more disclosure can take place. Due to this fact, it is very difficult to have an optimal disclosure control method for general cases.

4. Disclosure Risk Analyzer Component

To disclose information about individuals using masked microdata and prior information, the intruder needs to elaborate a strategy. The strategy used by an intruder for attempting a disclosure is called *disclosure scenario* (Willemborg et al 2001).

The intruder wants to obtain information about a set of individuals. We call this set of individuals the *target units*. In one disclosure scenario, those target units can be

chosen in different ways, based on prior information, or based on different characteristics in the microdata. It is important to notice that the intruder can change his target units while the disclosure scenario is in process.

A target unit should represent a known individual for the intruder. The set of all individuals in the initial microdata is $U = \{I(X_i) \mid 1 \leq i \leq n\}$. This set is called *universe*. $I(X_i)$ represents the identifier associated to the record i in the microdata. The target unit will represent a non-empty subset of U. The masked microdata is a set of de-identified records. This set is equal with $S = \{Y_i \mid 1 \leq i \leq t\}$. Since we allow simulated units in the masked microdata, let us assume that all those units represent a simulated individual labeled I_0. The set $U \cup \{I_0\}$ is labeled U_e and is called the *extended universe*.

As explained in previous sections, the masked microdata units represent elements from U_e, and, therefore, a mapping exists,

$$Id: S \rightarrow U_e \qquad (4.1)$$

where $Id(Y_i) = I(X_j)$ is the identifier belonging to the unit i in the masked microdata. The disclosure control attempts to make it more difficult for the intruder to ascertain this mapping.

Let $I(X_j)$ represent the identity of the record X_j (the values of confidential attributes), and an element from the target unit be denoted as Ext_i. (external information). Then, the set of target units will be $TU = \{Ext_i, i \in \{1, 2, 3, .., k\}\}$. We notice that Ext_i and X_j can have different values for the same attribute even if both represent the same individual.

The set of attributes from prior information, usually, is a subset of key attributes, but for a small number of individuals, some confidential values are also known.

Because the intruder does not access the initial microdata, we will use the identifiers as $I(Ext_i)$.

What we need to compute is the conditional probability of linking the identity of Ext_i to Y_j given external information and masked microdata. We can represent this probability as follows:

$$P(Id(Y_j) = I(Ext_i) \mid Ext_i, M) \qquad (4.2)$$

The above measure is similar with the one presented in (Willemborg et al 2001). We computed it in the same manner, but we use our general framework:

$$P(Id(Y_j) = I(Ext_i) \mid Ext_i, M) =$$

$$\frac{P(Ext_i, M \mid Id(Y_j) = I(Ext_i)) \cdot P(Id(Y_j) = I(Ext_i))}{P(Ext_i, M)} \qquad (4.3)$$

where

$$P(Ext_i, M) = \sum_{k=1}^{t} P(Ext_i, M \mid Id(Y_k) = I(Ext_i)) \cdot P(Id(Y_k) = I(Ext_i)) + P(M, Ext_i \mid I(Ext_i) \notin Id(S)) \cdot P(I(Ext_i) \notin Id(S)) \qquad (4.4)$$

Most of the different terms for the expressions (*4.3*) and (*4.4*) are interpreted in a similar way as in (Willemborg et al 2001) and calculated as shown in a-d below.

a) $P(I(Ext_i) \notin Id(S))$

$$P(I(Ext_i) \notin Id(S)) = 1 - P(I(Ext_i) \in Id(S)) = 1 - \pi_i, \qquad (4.5)$$

where π_i represents the inclusion probability for the $I(Ext_i)$ and the inclusion probability is approximated with the sampling factor

$$\pi_i = sf = \frac{r}{n}.$$

b) $P(Ext_i, M \mid Id(Y_j) = I(Ext_i))$

The event $Id(Y_j) = I(Ext_i)$ is independent of matrix *M*, therefore:

$$P(Ext_i, M \mid Id(Y_j) = I(Ext_i)) = P(Ext_i \mid M, Id(Y_j) = I(Ext_i)) \cdot P(M) \qquad (4.6)$$

c) $P(M, Ext_i \mid I(Ext_i) \notin Id(S))$

The events $I(Ext_i) \notin Id(S)$ and the masked microdata *M* are independent. Therefore:

$$P(M, Ext_i \mid I(Ext_i) \notin Id(S)) = P(Ext_i \mid M, I(Ext_i) \notin Id(S)) \cdot P(M) \qquad (4.7)$$

d) $P(Id(Y_k) = I(Ext_i))$

This probability depends on the sampling procedure, and, implicitly, on the inclusion probability, the simulation procedure, and the order of records in the masked

microdata. We assume that this order is given by a random permutation of all *t* elements from the masked microdata. In this way, simulated records are mixed with the identity records, and, we cannot make any distinction between those two categories. We can compute this probability as follows:

$$P(Id(Y_k) = I(Ext_i)) =$$

$$P(Id(Y_k) = I(Ext_i) \mid I(Ext_i) \in Id(S)) \cdot P(I(Ext_i) \in Id(S)) =$$

$$\frac{1}{t} \cdot \pi_i = \frac{\pi_i}{t}, \text{ for } k = 1, 2, .., t \qquad (4.8)$$

This last term is a generalization of Willemborg, et. al.'s work (2001) by including the simulation method in the interpretation of the above probability.

Substituting relations (*4.4*) to (*4.8*) into equation (*4.3*) we get:

$$P(Id(Y_j) = I(Ext_i) \mid Ext_i, M) =$$

$$\frac{P(Ext_i \mid M, Id(Y_j) = I(Ext_i)) \cdot P(M) \cdot \frac{\pi_i}{t}}{\sum_{k=1}^{t} \left(P(Ext_i \mid M, Id(Y_k) = I(Ext_i)) \cdot P(M) \cdot \frac{\pi_i}{t} \right) + P(Ext_i \mid M, I(Ext_i) \notin Id(S)) \cdot P(M) \cdot (1 - \pi_i)} \qquad (4.9)$$

In equation (4.9), we simplify *P*(*M*) and we obtain:

$$P(Id(Y_j) = I(Ext_i) \mid Ext_i, M) = \frac{\pi_i \cdot t^{-1} \cdot r_{i,j}}{\pi_i \cdot t^{-1} \cdot \sum_{k=1}^{t} r_{i,k} + (1 - \pi_i)} \qquad (4.10)$$

where

$$r_{i,k} = \frac{P(Ext_i \mid M, Id(Y_k) = I(Ext_i))}{P(Ext_i \mid M, I(Ext_i) \notin Id(S))} \qquad (4.11)$$

is known in record linkage literature as the *probability ratio* (Winkler 1995). It measures the probability of the prior information associated with the target unit given

the microdata matrix, compared with probability of prior information when it does not represent any record from the microdata.

Example:

Assume that an intruder has a masked microdata table with 20,000 records and this microdata contains three attributes: `Age`, `Zip Code` and `Income`. The first two attributes are key attributes while the last one is the confidential attribute. We make several assumptions for this microdata:

- `Age` is represented by the number of years. The entities from the masked microdata have the values for ages uniformly distributed across a range of 50 years ($A \sim U(18, 68)$, where A is the random variable which represents the values for `Age` attribute);
- `Zip Code` attribute contains only the first three digits of the real zip code. The entities from the masked microdata have the values for zip codes uniformly distributed between 0 and 999 ($Z \sim U(0, 999)$ where Z is the random variable which represents the values for `Zip Code` attribute);
- Those two random variables are independent ($Cov(A, Z) = 0$).

Let us assume that the target person is John Smith and his `age` and `zip codes` are 26 and 482. The number of possible combinations `age` and `zip codes` is 50 x 1,000 = 50,000. Therefore, we have:

$$P(Ext_i = (age = 26 \,\&\, zip = 482) \,||\, M, I(Ext_i) \notin Id(S)) = \frac{1}{50{,}000}$$

Let us assume that in the microdata we have only one record with this `age` and `zip code`:

$$P(Ext_i = (age = 26 \,\&\, zip = 482) \,|\, M, Id(Y_j) = I(Ext_i)) = 1$$

for a fixed j. For any remaining k between 1 and t ($k \neq j$) the above probability is 0. Those exact probabilities of 0 and 1 are because of no measurement errors. Therefore, we can conclude that: $r_{i,j} = 50{,}000$ and $r_{i,k} = 0$ for any $k \neq j$.

Let us assume that the population size is 200,000,000, therefore, the inclusion probability is $\pi_i = 20{,}000/200{,}000{,}000 = 1/10{,}000$

We substitute these values into (*4.10*):

$$P(Id(Y_j) = I(Ext_i) \,|\, Ext_i = (age = 26 \,\&\, zip = 482), M) = 0.024\%$$

As we can see the probability is very small, and, therefore, the disclosure risk is acceptable.

The reason for this small value of the disclosure risk is the distribution of values for key attributes for the population (in this example uniform distribution) and the number of distinct values (50,000).

We will show that, by modifying each of those two factors, the disclosure risk will be altered significantly.

Case A:

Let us assume that age attribute is not uniform distributed over the same range. We modify the uniform density function:

$$f_{Age} = \begin{cases} 0, \ldots\ldots\ldots\ldots x \notin (18,68) \\ 1/50, \ldots\ldots x \in (18,68) \end{cases}$$

to the following density function:

$$f'_{Age} = \begin{cases} 0, \ldots\ldots\ldots\ldots\ldots\ldots x \notin (18,68) \\ 1/10{,}000, \ldots\ldots\ldots\ldots x \in (18,28) \\ 9{,}999/40{,}000, \ldots\ldots x \in (28,68) \end{cases}$$

Therefore, the following probability:

$$P(Ext_i = (age = 26 \,\&\, zip = 482) \,||\, M, I(Ext_i) \notin Id(S)) = \frac{1}{10{,}000{,}000}$$

and $r_{i,j}$ = 10,000,000.

The inclusion probability and the number of elements in the microdata remain unchanged (π_i = 1/10,000; t = 20,000).

We substitute these values into (4.10):

$$P(Id(Y_j) = I(Ext_i) \mid Ext_i = (age = 26 \,\&\, zip = 482), M) = 4.98\%$$

The disclosure risk is considerably higher than when compared with the previous scenario.

Case B:

Now, let us assume that `Age` is an attribute that contains the number of years and the number of months and `Zip Code` to be an attribute with the range 0 to 99,999. We assume that both key variables are distributed uniformly in the entire population. The number of distinct values (equally likely to occur) is 600 x 100,000 = 60,000,000.

Therefore,

$$P(Ext_i = (age = 26 \,\&\, zip = 48201) \mid M, I(Ext_i) \notin Id(S)) = \frac{1}{60{,}000{,}000}$$

and $r_{i,j}$ = 60,000,000.

The disclosure risk will be:

$$P(Id(Y_j) = I(Ext_i) \mid Ext_i = (age = 26 \,\&\, zip = 48201), M) = 23.08\%$$

The disclosure risk is significant in this situation. ♦

Using formula (*4.10*) we are able to compute (or rather, estimate) the disclosure risk for a given target unit. This method, therefore, is used when we want to compute risk per unit. The second way in which we want to express disclosure risk is considering the disclosure risk for the entire microdata file. The ultimate goal is to unify those two approaches in a practical measure for disclosure risk. This final measure will be included in the final software component.

5. Future Work

Our ultimate goal is to develop all of the software components in the AMMG system. For this we need not only to implement existing results in the literature, but also to extend the results in various areas; these include: practical measures for information loss, cost functions to include both disclosure risk and information loss measures, and develop an adaptive algorithm for finding a list of disclosure control methods to be applied to any given initial microdataset.

The next step is to analyze the results of AMMG system. We will use real data sets from healthcare area, and we compare our results with existing software.

Appendix A – AMMG Interface

In this appendix we show the interface of AMMG software. The autentification part requires a user id and a password as shown in Figure A.1.

Figure A.1 – Login Page

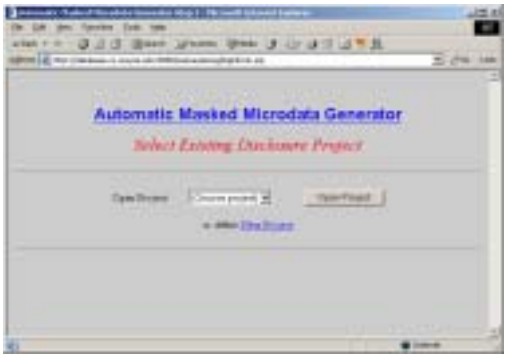

Figure A.2 – Select Disclosure Project Page

In the next screen (Figure A.2) the user can select an existing project or he can define a new project. We will choose the second option for this illustration.

The initial microdata selected (Figure A.3) called *patient* contains the following 10 records (table A.1). The user can select an initial microdata which was previously stored as a table in our project database. The database server for this project is mysql. The user gives the project name and the final microdata name.

Name	**SSN**	**Age**	**State**	**Diagnosis**	**Income**	**Billing**
John Wayne	123456789	44	MI	AIDS	45,500	1,200
Mary Gore	323232323	44	MI	Asthma	37,900	2,500
John Banks	232345656	55	MI	AIDS	67,000	3,000
Jesse Casey	333333333	44	MI	Asthma	21,000	1,000
Jack Stone	444444444	55	MI	Asthma	90,000	900
Mike Kopi	666666666	45	MI	Diabetes	48,000	750
Angela Simms	777777777	25	IN	Diabetes	49,000	1,200
Nike Wood	888888888	35	MI	AIDS	66,000	2,200

Mikhail Aaron	999999999	55	MI	AIDS	69,000	4,200
Sam Pall	100000000	45	MI	Tuberculosis	34,000	3,100

Table A.1

Figure A.3 – Project Settings

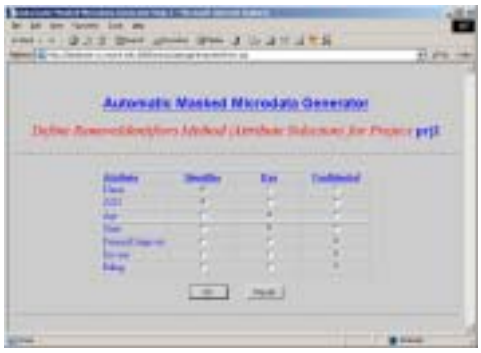

Figure A.4 – Attribute Selection

The next step is to define identifier, key and confidential attributes. We choose Name and SSN as identifier attributes, Age and State as key attributes and the remaining ones are confidential attributes (Figure A.4).

After attribute selection phase is completed the user can select several disclosure control methods to be applied in this project. For each method the user specifies its parameters. For this example we have chosen microaggregation for Age attribute (with groups of size 2) and sampling with sampling factor 0.8. At any time we can modify the list of methods, either by updating their parameters or by removing them from the project (Figure A.5).

When the masked microdata is generated, we can see the results in the next window (Figure A.6).

Figure A.5 – Methods List

Figure A.6 – Masked Microdata Table

As expected, only 8 records are in the masked microdata, but surprisingly records number 4 and 8 are unique with respect to Age attribute (we expected groups of two or more records of Age values). This result is due to the fact that methods are applied in order. The first method, microaggregation, indeed created groups with common values for Age attribute, but the sampling method eliminated two records and this is why exactly two records in the masked microdata are unpaired.

REFERENCES

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

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.

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

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.

Duncan G. T., Pearson R. W. (1991), *Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future*. Statistical Science, Vol. 6, No. 3, 219 – 239.

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

Hundepool A., Willemborg L., Wessels A., Gemerden L., Tiourine S., Hurkens C. (1999), *η-Argus User Manual*, http://www.cbs.nl.sdc.

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.

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

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

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

Sweeney L. (1997), *Guaranteeing Anonymity When Sharing Medical Data, The Datafly System*, MIT Artificial Intelligence Laboratory Working Paper, Cambridge, 344.

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

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

Winkler W. E. (1995), *Matching and Record Linkage.* in B. G. Cox et al (ed) Business Survey Methods, New York: J. Wiley, 355-384.

Authors:

Traian Marius Truta, Farshad Fotouhi - Department of computer science, Wayne State University, Detroit, MI 48202, USA, mtruta,fotouhi@cs.wayne.edu

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