Disclosure Risk Measures for the Sampling Disclosure Control Method

Traian Marius Truta
Department of Computer Science
Wayne State University
Detroit, MI 48202, USA
001-313-5776711
mtruta@cs.wayne.edu

Farshad Fotouhi
Department of Computer Science
Wayne State University
Detroit, MI 48202, USA
001-313-5775527
fotouhi@cs.wayne.edu

Daniel Barth-Jones
Center for Healthcare Effectiveness
Wayne State University
Detroit, MI 48202, USA
001-313-5778387
dbjones@med.wayne.edu

ABSTRACT
In this paper, we introduce three microdata disclosure risk measures (minimal, maximal and weighted) for sampling disclosure control method. The minimal disclosure risk measure represents the percentage of records that can be correctly identified by an intruder based on prior knowledge of key attribute values. The maximal disclosure risk measure considers the risk associated with probabilistic record linkage for records that are not unique in the masked microdata. The weighted disclosure risk measure allows the data owner to compute the risk of disclosure based on weights associated with different clusters of records. The weights allow a flexible specification of the relative importance of varying cluster sizes in probabilistic record linkage. We show that weighted disclosure risk measure is always between the values of minimal and maximal disclosure risk measures, and moreover for certain values of the weights, the weighted disclosure risk measure is equal to one of the other two measures. 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
Security, Measurement.

Keywords
Statistical Disclosure, Data Privacy, Microdata, Disclosure Risk and Sampling.

SAC'04, March 14-17, 2004, Nicosia, Cyprus.

1. INTRODUCTION
In today's world, information is available from many sources. Governmental, public, and private institutions are often required to make their data electronically available [4]. At the same time, confidentiality regulations are increasingly being developed to assure that publicly released information will not comprise the privacy of individuals or organizations represented in released data sets. 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 [8]. Under the simplified "safe harbor" provision of the HIPAA privacy rule, health care data cannot be released without removing all explicit identifiers such as names, addresses, and phone numbers, as well as other attributes, such as birth date, ZIP codes and gender that could potentially be used by a confidentiality intruder to identify a specific individual. Unfortunately, after this "safe harbor" removal step, the de-identified data has typically been stripped of much of its utility for statistical or healthcare services research. The release of data in aggregate form, which was used extensively in the past [11], frequently does not satisfy the needs of all data recipients, particularly when the user desires additional statistics or analysis than the owner of the data is prepared to provide [10]. Therefore, in order to enforce different confidentiality regulations, and at the same time to allow the users access to the data, microdata disclosure control techniques have been developed in the context of statistical databases [16].

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, thereby disclosing information about these entities [1, 13].

Microdata represents a series of records, where each record contains information on an individual entity [15]. Microdata released for use by third parties, after the data has been masked to limit the possibility of disclosure, is called *masked* or *released microdata* [3]. To avoid confusion, we will use the term *initial microdata* for microdata where no disclosure control methods were applied. Several statistical disclosure control techniques to mask the microdata have been proposed in the literature [15]. To increase confidentiality, more than one method is often applied in the disclosure control process.

In very broad terms, *disclosure risk* is the risk that a given form of disclosure will be encountered if masked microdata is released [2].

In this paper, we define three disclosure risk measures for sampling method. We justify our choices and we analyze different properties of those measures. Our disclosure risk measures compute the overall disclosure risk and are not linked to a target individual. We choose, in the beginning, two extreme measures called minimal disclosure risk (DR_{min}) and maximal disclosure risk (DR_{max}), and we then define a more general measure (D_W) based on a weight matrix. The disclosure risk measures presented in this paper are validated in our experiments.

There are many ways to define disclosure risk. Lambert defines disclosure risk as matter of perception [9]. Greenberg and Bethlehem discuss the probability of *population uniqueness* [7, 1]. This measure is related to our minimum disclosure risk measure. Other measures define disclosure risk as the proportion of sample unique records that are population unique [6, 12]. Eliot defines a new measure of disclosure risk as the proportion of correct matches amongst those records in the population, which match a sample unique masked microdata record [5]. We extend those discussions by considering 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 present sampling disclosure control method, and we define and analyze minimal, maximal and weighted disclosure risk measures for this method.

The remainder of this paper is organized as follows: Section 2 discusses disclosure risk measures for sampling method, Section 3 shows experimental results, and Section 4 gives future work in the area of disclosure control for microdata.

2. DISCLOSURE RISK MEASURES FOR SAMPLING METHOD

Sampling is the disclosure control method in which only a subset of records is released [12]. Due to the fact that statistical offices use frequently sampling (for example, elections surveys) the methods of statistical inference (e.g., proportions, ratios, means, variance, confidence intervals, regression analysis, etc.) from sample data are well established [14].

There are several methods to perform sampling over an initial set of data [14]. Since, by assumption, all the key and identifier values from the initial microdata are known by a presumptive intruder, the disclosure risk measures proposed are valid for any type of sampling. In our experiments, as well as in the illustrations, we applied *simple random sampling*, in which t distinct units are selected from the n units in the population in such a way that every possible combination of t units is equally likely to be in sample selected. The percentage of records selected in the sampling set:

$$sf = \frac{t}{n} \tag{2.1}$$

is called sampling factor.

The problem of quantifying disclosure risk is a difficult one because disclosure of confidential information usually occurs only if the intruder has some external information, and, it is usually not possible to know or anticipate this information. Therefore, we need to make assumptions about this knowledge in order to predict the disclosure risk. 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. Due to this consideration, this assumption does not reduce the generality of the problem.

Due to the sampling, the number of records (t) in the masked microdata (also called sampling set) is less or equal than the number of records (n) in initial microdata. We cluster the data from initial microdata and sampling set based on their key values. In the statistical disclosure control literature, such clusters are typically referred to as *equivalence classes* [17] or *cells* [2]. We define the following notations for initial microdata:

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

The following relations are true:

$$n_i = i \cdot F_i, \quad i = 1,.., n \tag{2.2}$$

$$\sum_{i=1}^{n} F_i = \sum_{i=1}^{n} \frac{n_i}{i} = F \tag{2.3}$$

$$\sum_{i=1}^{n} n_i = \sum_{i=1}^{n} i \cdot F_i = n \tag{2.4}$$

Similar notations are defined for the sampling set:

- S – the sampling set (i.e. the set of all records after the sampling method is applied, masked microdata);
- 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.

For sampling set, we have the following relations:

$$t_i = i \cdot f_i, \quad i = 1,.., t \tag{2.5}$$

$$\sum_{i=1}^{t} f_i = \sum_{i=1}^{t} \frac{t_i}{i} = f \tag{2.6}$$

$$\sum_{i=1}^{t} t_i = \sum_{i=1}^{t} i \cdot f_i = t \qquad (2.7)$$

To relate initial microdata to masked microdata we define the *classification matrix C*. It represents a *t x n* matrix that describes the correlation 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 $j = 1, .., n$; $c_{ij} == |\{x \in M_k$ and $x \in A_p \mid |M_k| = i$, for all $k = 1, .., f$ and $|A_p| = i$, for all $p = 1, .., F\}|$.

The classification matrix *C* has the following properties:

$$c_{ij} = 0 \text{ for all } i > j,\ i = 1, .., t;\ j = 1, .., n \qquad (2.8)$$

$$\sum_{j=1}^{n} c_{ij} = t_i \text{ for all } i = 1, .., t \qquad (2.9)$$

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

Algorithm 2.1. (Classification matrix construction)

```
Initialize each element from C with 0.
   For each element s from sampling set S do
   Count the number of occurrences of key values
     of s in sampling set S.
   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 cij by 1.
End for.
```

The first measure of disclosure risk is based on the percentage of unique records [6]. In our case we consider records that are unique in both initial microdata and sampling set. This measure is defined as:

$$DR_{min} = \frac{c_{11}}{n} \qquad (2.10)$$

This measure has some limitations. It does not consider the distribution of the records that are not unique in both initial microdata and sampling set. Since the number of occurrences of the key values for a particular record is greater in initial microdata, we consider the probability of record linkage as inversely proportional with the number of occurrences in initial microdata. This probability is included in maximal disclosure risk measure for sampling:

$$DR_{max} = \frac{\sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{\min(k,t)} c_{ik}}{n} \qquad (2.11)$$

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

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1t} & \dots & w_{1n} \\ 0 & w_{22} & \dots & w_{2t} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_{tt} & \dots & w_{tn} \end{pmatrix} \qquad (2.12)$$

with the following properties:

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

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. Because the cluster sizes for a set of key values will typically differ between the initial microdata and the sample, the first weight matrix property specifies that the weights selected should not increase across the row *i*'s as the corresponding records belonging to clusters from the initial microdata increase in size. The next two properties specify the weight constraints for records within clusters of a particular size in the initial microdata (i.e., within a particular column *j*), and specify that disclosure risk weights selected should not increase as the cluster size in the sample decreases. The following two properties guarantee that the weights assigned to groups of records do not increase as the corresponding cluster sizes increase along the matrix diagonals (i.e., as cluster sizes increase in both the initial and masked microdata). We add the last condition, which limits the sum of all weights, for expedient scaling. The combined effect of these weighting constraints allows the data owner considerable flexibility in addressing probabilistic record linkage risks, while also accounting for the potential disclosure risks posed by the data intruder's prior beliefs about confidential variables which could possibly be used to distinguish individuals.

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

$$DR_W = \frac{1}{n \cdot w_{11}} \sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{\min(k,t)} w_{ik} \cdot c_{ik} \qquad (2.13)$$

Next, we prove two lemmas for those risk measures; the first one gives a range for an arbitrary weighted disclosure risk measure for a specific initial and masked microdata, and the second one shows that any disclosure risk measure lies between 0 and 1.

Lemma 2.1.

Given sampling set *S*, and weight matrix *W,* the following relations are true: $DR_{min} \leq DR_w \leq DR_{max}$.

Proof

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

$$DR_{min} = \frac{c_{11}}{n} = \frac{1}{n \cdot w_{11}} \cdot w_{11} \cdot c_{11} \leq \frac{1}{n \cdot w_{11}} \sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{\min(k,t)} w_{ik} \cdot c_{ik} = DR_W$$

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

$$DR_W = \frac{1}{n \cdot w_{11}} \sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{\min(k,t)} w_{ik} \cdot c_{ik} = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{\min(k,t)} \frac{w_{ik}}{w_{11}} \cdot c_{ik} \leq$$

$$\frac{\sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{\min(k,t)} c_{ik}}{n} = DR_{max..}$$

q.e.d.

Lemma 2.2.

Given sampling set S, and weight matrix W, $0 \leq DR_W \leq 1$.

Proof

Using lemma 2.1 and the fact that c_{11} is greater then 0 we get: $0 \leq DR_{W.}$

Then $DR_W \leq DR_{max.} = \dfrac{\sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{\min(k,t)} c_{ik}}{n} \leq \dfrac{\sum_{k=1}^{n} \sum_{i=1}^{\min(k,t)} c_{ik}}{n} = \dfrac{t}{n} \leq 1.$

q.e.d.

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

To illustrate those disclosure risk measures we consider the initial microdata with three different sampling sets in Figure 2.1. *Age* and *Sex* are considered key attributes. In Figure 2.2., we show the number of cluster with the same size and the number of records in corresponding clusters of the specified size. The properties (2.2) to (2.7) can be easily verified.

Initial Microdata

RecNo	Age	Sex
1	10	M
2	30	M
3	20	M
4	20	F
5	10	F
6	25	F
7	20	M
8	25	F
9	10	M
10	20	M

Sampling Set 1 (S1)

RecNo	Age	Sex
1	10	M
2	30	M
3	20	M
4	20	F
5	10	F

Sampling Set 2 (S_2)

RecNo	Age	Sex
1	10	M
3	20	M
6	25	F
7	20	M
9	10	M

Sampling Set 3 (S_3)

RecNo	Age	Sex
1	10	M
3	20	M
7	20	M
9	10	M
10	20	M

Figure 2.1. Initial microdata and three sampling sets

IM	n=10	n_1=3	n_2=4	n_3=3	n_4=…=n_{10}=0
	F=6	F_1=3	F_2=2	F_3=1	F_4=..=F_{10}=0

S_1	t=5	t_1=5	t_2=…=t_5=0	
	f=5	f_1=5	f_2=…=f_5=0	

S_2	t=5	t_1=1	t_2=4	t_3=t_4=t_5=0
	f=3	f_1=1	f_2=2	f_3=f_4=f_5=0

S_3	t=5	t_1=0	t_2=2	t_3=3	t_4=t_5=0
	f=2	f_1=0	f_2=1	f_3=1	f_4=f_5=0

Figure 2.2. Characterization of data based on cluster sizes

Next, we compute classification matrices for each sampling set. We get the matrices C_1, C_2 and C_3 of size 5 x 10 corresponding to S_1, S_2 and S_3 respectively:

$$C_1 = \begin{pmatrix} 3 & 1 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix} \quad C_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 2 & 2 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0. & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix} \quad C_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 2 & 0 & 0 & \dots & 0 \\ 0 & 0 & 3 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

The properties (2.8) and (2.9) can be noticed in above matrices. We consider the following four weight matrices:

$$W_1 = \begin{pmatrix} 10 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix} \quad W_2 = \begin{pmatrix} 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & \dots & 0.4 \\ 0 & 0.4 & 0.4 & 0.4 & 0.4 & \dots & 0.4 \\ 0 & 0 & 0.4 & 0.4 & 0.4 & \dots & 0.4 \\ 0 & 0 & 0 & 0.4 & 0.4 & \dots & 0.4 \\ 0 & 0 & 0 & 0 & 0.4 & \dots & 0.4 \end{pmatrix}$$

$$W_3 = \begin{pmatrix} 6 & 2 & 0 & 0 & 0 & \dots & 0 \\ 0 & 2 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix} \quad W_4 = \begin{pmatrix} 3 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 3 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 3 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

The owner of the data chooses the data values in the weight matrix, for which we provide four examples (W_1,W_2, W_3 and W_4). Such matrices instantiate the disclosure risk measure based on the data owner's privacy concerns. From lemma 2.2, we conclude that the first weight matrix correspond to DR_{min}, the second to DR_{max}. The value 0.4 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 10, 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 initial 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 the case of W_4, the owner of the data chooses a different strategy in assessing disclosure risk. Unique records as well as records with double or triple occurrence in both initial and masked microdata are considered a threat to individual privacy, and their weight is chosen equal. Records with triple occurrence in initial microdata, but only with single or double occurrence in sampling set are considered safe, and do not count in disclosure risk computation. In the Figure 2.3, we show the disclosure risk values for each combination of sampling set and disclosure risk weight matrix.

	W_1	W_2	W_3	W_4
S_1	0.3	0.383	0.316	0.316
S_2	0	0.2166	0.05	0.116
S_3	0	0.2	0.033	0.2

Figure 2.3. Disclosure risk values

As expected from lemma 2.1, each disclosure risk value lies between minimal and maximal disclosure risk, therefore those two measures computed for a given initial microdata and sampling set provides a range for any disclosure risk value. However, if the owner of the data requires a better estimate of disclosure risk a weight matrix can be defined. There are situations when it is difficult to decide which sampling set offers more protection. From the Figure 2.2 we notice that both S_2 and S_3 offer more data protection then S_1. The challenge is to choose which sampling set minimizes disclosure risk between S_2 and S_3. The solution to this problem is strongly related to the characteristics of the data and to the risk perceived by the owner of the data due to records with more than one occurrence. Those considerations are embedded in the disclosure risk weight matrix, which provides the owner of the data with the facility to develop parameterized disclosure risk measures.

3. EXPERIMENTAL RESULTS

We used simulated medical record billing data to perform a series of tests. The data contains the following attributes: *Age*, *Sex*, *Race*, *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 two sets of key attributes: KA_1 = {*Age*, *Sex*, *Zip*} and KA_2 = {*Age*, *Race*, *Sex*}.

Then, for each combination of initial microdata and key attributes, we generated sampling sets using simple random sampling procedure with different sampling factors. Next, we computed minimal and maximal disclosure risk in each situation. Figure 3.1 and Figure 3.2 shows the results of our experiment. From Lemma 2.1, we know that every DR_W lies between DR_{min} and DR_{max}, therefore, the depiction of the extreme measures give us the range for any possible DR_W.

From the Figures 3.1 and 3.2, we notice that the number of key combination values and their distribution within initial microdata determine the values of disclosure risk. Since *Zip* attribute takes a large number of values comparing with *Race* attribute, the risk of disclosure for KA_1 should be greater than for KA_2. The proposed disclosure risk measures give the expected result. One important property of the disclosure risk results we can infer from the experiments performed is that they decrease almost linearly with the sampling factor. This result is due to the sample random sampling. When an alternative sampling method is used, such linear relationships between the sampling factors and the disclosure risk values might not hold. The advantage in the case of sample random sampling is that the owner of the data can approximate disclosure risk simply by knowing its value for the initial microdata without applying sampling method (when sampling factor is 1 all records are considered in the sampling set) and for one other level of the sampling factor in order to make the determination of the slope. The owner of the data can thus decide the trade-off between disclosure risk and information loss. When applying the sampling disclosure control method, one important tool resulting from this work on the estimation of disclosure risk is the ability to compute minimal and maximal disclosure risks and determine the disclosure risk interval, or when more precision is needed, the weighted disclosure risk must be computed for a specified weight matrix.

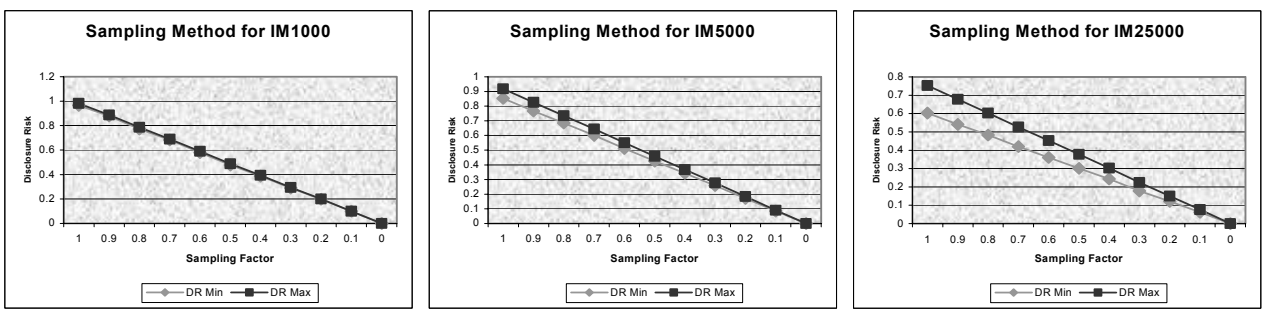

Figure 3.1. KA_1 is the set of key attributes.

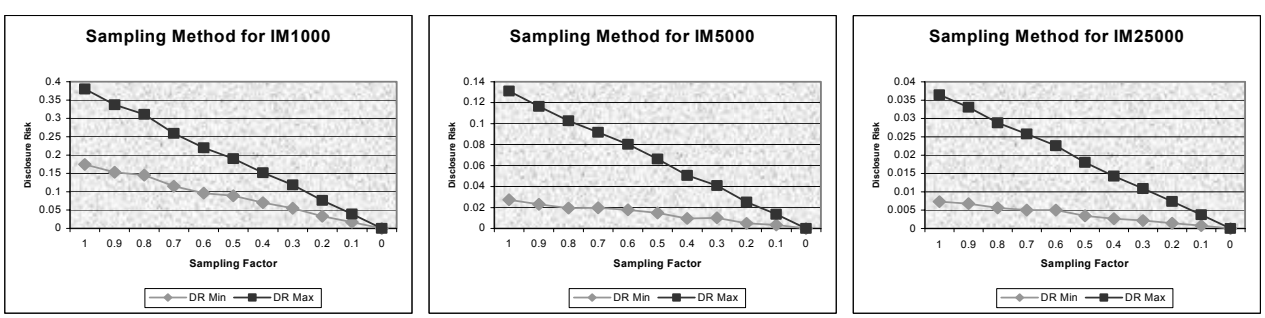

Figure 3.2. KA_2 is the set of key attributes.

4. CONCLUSIONS AND FUTURE WORK

Several disclosure risk measures for microdata were presented for sampling disclosure control method and tested using simulated data sets. We are currently pursuing appropriate disclosure risk metrics for other methods, like microaggregation, data swapping or randomization methods. The future work in this field can be divided into three areas: developing disclosure risk measures for other methods and generalizing them for combinations of methods, expressing information loss in general formulas in order to investigate the dependence between information loss and disclosure risk, and applying more than one disclosure control methods successively to minimize both disclosure risk and information loss.

5. REFERENCES

[1] 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.

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

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

[4] Duncan, G.T., Jabine, T. B., Wolf, V. A. de (1993), Private Lives and Public Policies, National Academy Press.

[5] Elliot, M. J. (2000), DIS: a New Approach to the Measurement of Statistical Disclosure Risk, International Journal of Risk management, 39 –48.

[6] Fienberg, S. E.; Markov, U. E. (1998), Confidentiality, uniqueness, and disclosure limitation for categorical data, Journal of Official Statistics, 385 - 397.

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

[8] HIPAA (2002), Health Insurance Portability and Accountability Act, http://www.hhs.gov/ocr/hipaa/privrulepd.pdf

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

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

[11] Schwartz, M. D., Denning, D. E., Denning, P. J., (1979), Linear Queries in Statistical Databases, ACM Transactions on Database Systems, Vol. 4, No. 2, 156-167.

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

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

[14] Thompson, S. K. (ed) (2002), Sampling. John Wiley & Sons.

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

[16] Willemborg L., Waal T. (ed) (1996), Statistical Disclosure Control in Practice. Springer Verlag.

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