International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:10, 2010

On Identity Disclosure Risk Measurement for Shared Microdata

M. N. Huda, S. Yamada, N. Sonehara

Abstract**—**Probability-based identity disclosure risk measurement may give the same overall risk for different anonymization strategy of the same dataset. Some entities in the anonymous dataset may have higher identification risks than the others. Individuals are more concerned about higher risks than the average and are more interested to know if they have a possibility of being under higher risk. A notation of overall risk in the above measurement method doesn't indicate whether some of the involved entities have higher identity disclosure risk than the others. In this paper, we have introduced an identity disclosure risk measurement method that not only implies overall risk, but also indicates whether some of the members have higher risk than the others. The proposed method quantifies the overall risk based on the individual risk values, the percentage of the records that have a risk value higher than the average and how larger the higher risk values are compared to the average. We have analyzed the disclosure risks for different disclosure control techniques applied to original microdata and present the results.

Keywords**—**Anonymization, microdata, disclosure risk, privacy.

I. INTRODUCTION

MANY organizations hold a huge amount of data containing information on individual unit such as a person, a company, an institution, etc. These data, called microdata [1], hold valuable information which can be unfolded only through statistical analysis on the data. Therefore, often parts of these data are shared with third parties for research purposes, data analysis or application testing to discover the important information in them. On the other hand, shared microdata containing personal information can become a major source of privacy violation if appropriate privacy protection technologies are not applied on the data before sharing with others.

Three types of attributes of microdata are under consideration in the context of this paper: (i) *quasi-identifiers (K),* (ii) *sensitive attributes (S)* and (iii) *identifiers (I)*. Quasi-identifiers are the attributes that in combination can be used to identify an individual. E.g., country, postal code, gender, age, date of birth, etc. Sensitive attributes contain sensitive information about an entity. E.g., salary, diseases, political views, etc. Identifiers are the attributes that explicitly identify individuals. E.g., Social Security Number, passport number, complete name etc.

Third parties may get microdata from data collectors for data mining and research. Some organizations may also sell microdata to commercial data brokers [2][3]. The shared microdata may contain quasi-identifier attributes and sensitive information (e.g. disease names) of entities. Privacy protection technologies re-arrange or modify the data in such a way that the subject individuals cannot be identified from the data. Such modification, known as data masking or anonymization, is more often than not a legal requirement [2]. Typically, names and other identifying information are removed from original records before being shared with others. Third parties usually utilize microdata through statistical analysis.

Though the identification attributes are removed from the data before sharing with third parties, there still remains the possibility of identification threats in the anonymous data through quasi-identifiers. A malicious user, who has access to the shared microdata, can obtain publicly available identification databases (e.g., voter lists) containing the same quasi-identifier attributes as in the shared microdata and can match them with the shared records to potentially re-identify and thus reveal sensitive information about the entities.

Sensitive Attrib (S)	**Quasi-Identifier (K)**		
Disease	*Zip Code*	*Gen*	*Y. of Birth*
Cancer	2314	M	1959
Strep	2314	F	1955
Diarrhea	2342	M	1959
Gallstones	2319	F	1966
Gastric Ulcer	2323	F	1987
Pneumonia	2314	M	1975
Flu	2321	F	1975
Meningitis	2324	M	1967
Diabetes	2310	M	1961
Allergies	2337	M	1974
Cancer	2345	F	1978
Cancer	2328	F	1961

(a) Shared microdata

Matching

Quasi-Identifier (K)			**Identifier (I)**
Zip Code	*Gen*	*Y. of Birth*	*Name*
2314	M	1959	John Smith
2432	M	1962	Alan Smith
2314	F	1965	Alice Brown
2342	M	1959	Hercules Green
4249	F	1955	Marie Kirkpatrick
4723	M	1978	Albert Blackwell
4249	F	1975	Gill Stringer
2342	M	1964	Douglas Henry
2314	M	1975	Bill Nash
4237	F	1942	Alicia Fred
2323	F	1982	Leslie Hall
2321	F	1973	Freda Shields
4723	F	1959	Beverly McCulsky

(b) Identification database (publicly available)

Fig. 1. Example of some shared microdata (hospital record) and identification database (e.g., voter list). Record linking operation can determine who was diagnosed for what disease.

Fig. 1 illustrates such matching and identification using

M. N. Huda, S. Yamada and N. Sonehara are with the National Institute of Informatics, Tokyo, Japan (e-mail: {huda, shigeki, sonehara}@nii.ac.jp).

publicly available identification databases. The malicious user can identify and associate sensitive information when the same quasi-identifiers are present in both the shared microdata and the identification database. For example, by matching quasi-identifiers between the two data tables, the user can find out that "John Smith" has been diagnosed for "Cancer".

The original data holder (data owner) seeks to limit the risk that malicious users, hereafter called intruders, are able to identify sampled units in the shared data. In order to protect linking records between the released data and records from external databases, several statistical disclosure control techniques such as global recoding [5][6], local suppression [7], microaggregation [8], sampling [9], simulation [10], adding noise [11], rounding [12], post randomization method [13], data swapping [14] etc. were proposed in literature Reference [1] presents a survey of all those methods. To increase confidentiality, more than one method is often applied in the disclosure control process. We will call the final microdata as masked or shared microdata [14].

The data owner can assess the strength of an anonymization by estimating the disclosure risk of the data. In very broad terms, disclosure risk is the risk that a given form of disclosure will be encountered if a masked microdata is released. Two types of disclosure, namely**,** identity disclosure and attribute disclosure are discussed in literature. 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 [15]. In this paper, we focus on identity disclosure.

Identity disclosure risk assessment metrics have been developed by many researchers based on mainly: (1) number/ percentage of unique records and (2) probability of identification. Because of the privacy acts, it has been a common practice that microdata are shared with third parties after anonymization, which usually results in "few or no unique records" in the anonymous microdata. Thus, the risk measurements based on the percentage of unique records are no longer in common use. Instead, probability-based risk measurement methods have been well accepted because of its considerations of both unique records and non-unique records in the assessment.

A. Motivations

The overall risk (or the risk for the whole shared microdata) can be expressed with a single numerical value that is calculated as the average of the risks of all entities. However, in these measures, the same overall risk can be resulted from different anonymization of the same entities that create different group sizes of the records. The risk value calculated from a uniform grouping (risk distribution), where every entity has the same risk, can be the same as the risk value calculated from a non-uniform risk distribution (where some entities have higher risks than the others). Thus, a notation of overall risk in the existing probability-based metrics doesn't indicate whether some of the involved entities have higher risks than the others. So, existing metrics do not fully capture the risk distributions among the entities. For example, an overall risk value of 20 x 0.4 /20 = 0.4 calculated from a risk distribution of "0.4 risks for all 20 entities" will not create the same feeling in the involved people as in the case of the same overall risk value of (0.1x10 + 0.7x10)/20=0.4 calculated from a risk distribution of "0.1 risks for 10 entities and 0.7 risks for the rest 10 entities". Individuals are more concerned about higher risks and are more interested to know if they have a possibility of being under higher risk than the average. Also, the data owner is interested to know if the applied anonymization is at its optimal point for the same overall risk. Thus, we need a metric that not only expresses the overall risk, but also conveys information about the risk distribution among the members.

B. Our Contributions

In this paper, first we analyze all of the possible identification risk distributions among the members for a fixed average risk of a given microdata. From the analysis, we formulate the distribution of identification risk that is most unbalanced among the members and the distribution of identification risk that is most balanced among the members. Then we describe the algorithm of our proposed method for measuring disclosure risk that not only indicates the average risk value but also captures whether some of the members have higher risk than the others. The proposed method quantifies the overall risk by considering (1) the individual risk values, (2) the percentage of the records that goes above the average and (3) how larger the higher risk values are compared to the average. A notion of overall risk value in our proposed metric will help (i) the involved individual to feel the real risk including the average of low and high risks and the possibility of being in higher risk compared to the rest, and (ii) data releasing authorities to realize whether their anonymization is optimized for a given average risk. The data releasing authorities can decide whether their anonymous data is in the best possible state for a fixed average risk. The proposed method is applicable in all areas involving data privacy: from the customer's shopping behavior data of a shopping mall, to the network user's activity data of an organization, to the patient's health records of a hospital. Finally, we evaluate our method by using simulated data and compare with existing measurement schemes.

The remainder of this paper is organized as follows: Section II illustrates related disclosure risk measurement methods proposed previously by other authors. Section III briefly describes a framework for microdata disclosure control and makes assumptions about the external information known by a presumptive intruder. Section IV describes and analyzes probability-based risk measurement methods. Section V describes the proposed identity disclosure risk measurement method and also presents the algorithm. Section VI illustrates experimental results created with simulated data and Section VII concludes the paper.

II. PREVIOUS WORKS

Considerable research on disclosure risk assessment [1][9] [10][15]-[42] has resulted in a variety of proposed disclosure risk assessment methods. Many of the proposed disclosure risk measures are some function of the number of population or sample unique. These include, among others the measures proposed in [18] - [25]. One of the most intuitive ways is to count the number of unique records with respect to a limited set of attributes [43], called "keys" in disclosure avoidance literature [1]. References [16] and [9] define disclosure risk as a proportion of sample unique records that are population unique. Reference [17] 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.

Uniqueness is relevant because population unique generally have higher risks of identification disclosure than non-unique ones. Indeed, it has been suggested that uniqueness is a necessary condition for identification [21]. Substantial work has been done on estimating the number of population unique from a sample of data when the population follows a particular distribution such as Poisson-Gamma [19], Dirichlet-multinomial [44], and negative-binomial [22]. The paper in reference [20] has proposed a procedure that is not dependent on a parametric statistical distribution. Reference [15] defines disclosure risk as matter of perception. While useful, population uniqueness does not account for the nature of the information possessed by the intruder. For example, when the intruder knows a particular target is in the sample and knows values of that target's record, the intruder can identify the target when it is a sample unique, even if it is not a population unique.

Other authors have proposed that a linking attempt is carried out between released records with target records, either by direct matching using external databases [26][27] or indirect matching using the existing database [28][45][46] [38]. In both approaches, the organization essentially mimics the behavior of an intruder trying to match released records to target records. Recent publications [28]-[36] assess risk based on probability of linking the released records with external data by considering the attackers' knowledge and taking possible external data sources into account.

These approaches address many of the shortcomings of relying on population or sample unique. They can permit one to account for varying degrees of intruder knowledge, are equally appropriate for continuous and categorical data, and can be applied to assess the effects of statistical disclosure limitation techniques.

III. GENERAL FRAMEWORK FOR MICRODATA

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.) [18]. The shared microdata consist of a set of *n* records with values from two types of attributes: Quasi-identifier (K) and Sensitive (S) attributes. Identifier attributes, such as Name and SSN, can be used to identify a record and thus are removed from the microdata before sharing with third parties. Quasi-identifier attributes such as Zip Code and Age, cannot identify a record alone, but may do so when several quasi-identifiers are combined and linked with external data sources containing identifiers. Privacy concern arises only if the microdata contains sensitive information about the related entities. Sensitive attributes, such as disease name, must be protected from being associated with individual identity through a disclosure control mechanism.

We represent the shared microdata (SM) as a matrix with 2 partitions that correspond to two categories of attributes: quasi-identifiers (K) and sensitive attributes (S). Each row corresponds to an individual entity and the columns represent one or more quasi-identifier attributes (that can be used to identify individual) and one or more sensitive attributes. Therefore:

$$SM = [K \mid S] \tag{1}$$

where,

$K=K_1, K_2, \ldots, K_p$ i.e., $K=[k_{ij}]$ of order $n \times p$

$S= S_1, S_2, \ldots, S_q$ *i.e.,* $S=[s_{ij}]$ of order $n \times q$.

Due to applied disclosure control methods, such as sampling and simulation, the number of records in the shared microdata (n) differs from the number of records in initial microdata (N). The corresponding attribute values may also differ due to perturbative methods (such as global recoding, microaggregation, data swapping and so on) used in disclosure control processes.

Disclosure of confidential information usually occurs if the intruder has some related external information which is difficult to know or anticipate. Therefore, we need to make assumptions about this knowledge to calculate the disclosure risk. The assumption we make about the intruder is that an intruder has the quasi-identifiers along with the confidential information (from the shared microdata) and identifiers along with the quasi-identifiers (from external data sources) values for population. Our assumption also considers that the key attributes are discrete; the identification database has all of the data records corresponding to the entities in the shared microdata. Though the population unique and the sample unique may not be equal, we consider sample uniqueness for measuring identity disclosure.

Based on our previous assumptions, external information available to an intruder is:

$$Ext = [K \mid I] \tag{2}$$

We note that shared microdata can be expressed as a projection on quasi-identifiers and confidential attributes of initial microdata:

$$SM = \Pi_{K,S}(IM) \tag{3}$$

We group the data from shared microdata based on their quasi-identifier values. All of the quasi-identifiers are combined and considered as a single combination attribute. Therefore, in each group we will include records with the same values for their quasi-identifier attributes. We define the following:

N – the number of entities in the population
n – the number of entities in the shared microdata
G – the number of groups
A_k – the set of elements from the *k*-th group $\forall k$, $1 \le k \le$ G.
$|A_k| = i$, *for all k* = 1, ..,*G*, $\forall i$, $1 \le i \le$ n.
G_i – the number of groups with the group size *i*.
n_i – the number of records in all of the groups of size *i*.
Thus, we have the following relations:

$$n_i = i \times G_i, \forall i = 1 \ldots n \tag{4}$$

$$G = \sum_{i=1}^{n} G_i = \sum_{i=1}^{n} \frac{n_i}{i} \tag{5}$$

$$n = \sum_{i=1}^{n} n_i = \sum_{i=1}^{n} i \times G_i \tag{6}$$

Let us consider an example sample data consisting of nine data records with different quasi-identifier values as shown in Fig. 2. The "Rec No" column was added for the convenience in referring to a specific record with its corresponding number and it is not a part of the microdata. The data are grouped into different groups based on the quasi-identifier values (K). The records with the same quasi-identifier values are grouped together. Different variable values are indicated on the right side of Fig. 2. By A_1={1} we mean that the record no. 1 is in the first group.

SM

Rec No	*K*	*S*
1	a	Cancer
2	b	Strep
3	c	Diarrhea
4	c	Gallstones
5	b	Gastric Ulcer
6	d	Pneumonia
7	e	Flu
8	e	Meningitis
9	e	Diabetes

Variable values (for the left side data)
n= 9, G= 5
G_1= 2:
A_1 = {1},
A_2 = {6}
G_2= 2:
A_3 = {2, 5}
A_4 = {3, 4}
G_3= 1:
A_5 = {7, 8, 9}

Fig. 2 Example sample records and their grouping according to the microdata framework.

There are 5 unique values in the quasi-identifiers. So, G=5. There are 2 groups with single records; G_1=2; A_1 = {(a, Cancer)}, A_2= {(d, Pneumonia)}. There are 2 groups with double records; G_2=2; A_3 = {(b, Strep), (b, Gastric Ulcer)}, A_4= {(c, Diarrhea), (c, Gallstones)}. There is 1 group with 3 records; G_3=1, A_5 = {(e, Flue), (e, Meningitis), (e, Diabetes)}.

IV. Unique and Probability-Based Risk

The first measure of disclosure risk is based on the percentage of unique records in the population and can be called as the minimum disclosure risk.

$$DR_{\min} = \frac{n_1}{N} \tag{7}$$

When the sample unique is equal to the population unique,

$$DR_{\min} = \frac{n_1}{n} \tag{8}$$

Since we made the assumption that an intruder has knowledge about identifier and key values, this measure represents the percentage of records from the sample that can be correctly re-identified by the intruder. This is a minimal disclosure risk value.

This measure has its limits. It does not consider the distribution of the records that are not unique. Thus, another method considers the disclosure risk based on the probability of identification risks for all of the data records. In this method, the identity disclosure risk of all records in a group *k* ($1 \le k \le$ G) of group size *i* is

$$DR_{pr}(i) = \frac{1}{i} = P_i \tag{9}$$

where, P_i is the probability of identification of an entity in the population. The overall disclosure risk for the population can be computed by taking the average of the risks of all of the records. Thus, the overall disclosure risk in the probability-based metric becomes,

$$DR_{pr} = \frac{1}{n}\sum_{i=1}^{n}(n_i \times P_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{n_i}{i} = \frac{1}{n}\sum_{i=1}^{n} G_i = \frac{G}{n} \tag{10}$$

From the above equation, we see that for a fixed number of records (n) the overall risk depends upon the number of groups or unique values in the quasi-identifier attributes. For a fixed number of records, more groups results in more disclosure risk.

A. Analysis of probability-based risk

The average risk measurement does not reflect whether some records are in higher identification risk than the others. As long as the total number of groups is fixed for a fixed number of records the overall risk in the probability-based metric is the same. However, different group sizes (and their probabilities) are possible for a fixed total group (G) and fixed number of records (n). For example, let us consider that there are 12 data records and they are distributed into 3 groups. For the fixed value of (n=12) and (G=3), several distributions of the three group sizes are possible. Let us consider three different distributions, as shown in Fig. 3, among all possible distributions.

In the first distribution (Distribution 1), all of the three groups are of equal sizes i.e., each groups has 4 group members. Thus, the probability of identification of each member of each group is ¼. However, in the second distribution (Distribution 2), two groups are of group size 2 and one group is of group size 8. Thus, the probability of identification of each member of the first two groups is ½ and the probability of identification of each member of the third group is 1/8. In the third distribution (Distribution 3), two groups are of group size 1 and one group is of group size 10. Thus, the probability of identification of each member of the first two groups is 1 and the probability of identification of each member of the third group is 1/10.

Distribution 1		Distribution 2		Distribution 3	
G=3		G=3		G=3	
G_4=3	DR_{pr}=1/4	G_2=2, G_8=1	DR_{pr}=1/4	G_1=2, G_{10}=1	DR_{pr}=1/4

Fig. 3. Different group sizes for a fixed total number of groups (G=3) and a fixed number of records (n=12).

We can understand that a linking attack by intruder gives maximum 25% confidence in identifying all entities in Distribution 1. In Distribution 2, a linking attack would give maximum 50% confidence in identifying two entities, while in Distribution 3, a linking attack would give 100% confidence in identifying two entities. However, the overall risk, calculated by taking the average, does not differentiate between any two distributions among the above three. Thus, the average-based overall risk measure cannot express people's feelings about high risk and we need a better metric.

V. Proposed Threshold-based Disclosure Risk Measurement Method

The disclosure risk should not only reflect the average risks but also indicate how high the higher risks are compared to the average. So, in addition to taking the average, it should consider the risks of records that are above the average and their differences with the average.

For the same average risk (G/n), different distributions of risk among the members are possible. The risk distribution depends on the distribution of group sizes ($|A_1|$, $|A_2|$,…$|A_G|$). Let us consider, for example, the value of n=12 and G=3. The number 12 can be grouped into 3 groups in 9 different ways. They are (4, 4, 4), (3, 4, 5), (2, 5, 5), (2, 4, 6), (1, 5, 6), (1, 4, 7), (1, 3, 8), (1, 2, 9), (1, 1, 10). The first distribution (4, 4, 4) implies that four records have one distinct quasi-identifier value, four other records have another distinct quasi-identifier value and the rest four records have another distinct quasi-identifier value. The group sizes and hence the probabilities of identification are most balanced (equal for all) among all members. In the last distribution (1, 1, 10), two records have two distinct quasi-identifier values and ten records have another distinct quasi-identifier value. The grouping and hence the probabilities of identification are most unbalanced (1 for two members and 1/10 for ten members) among all members.

In the most balanced partitioning or grouping of *n* records into *G* groups there can be zero (if n is completely divisible by G) or more (if there is a remainder) groups with group size equal to (quotient +1) and one or more groups with group size equal to the quotient. Thus, the following equations can be deduced.

$$n_{\left\lfloor \frac{n}{G} \right\rfloor+1} = \left(\left\lfloor \frac{n}{G} \right\rfloor + 1\right) \times G_{\left\lfloor \frac{n}{G} \right\rfloor+1} = \left(\left\lfloor \frac{n}{G} \right\rfloor + 1\right) \times \left(n - \left\lfloor \frac{n}{G} \right\rfloor \times G\right) \qquad (11)$$

$$n_{\left\lfloor \frac{n}{G} \right\rfloor} = \left\lfloor \frac{n}{G} \right\rfloor \times G_{\left\lfloor \frac{n}{G} \right\rfloor} = \left\lfloor \frac{n}{G} \right\rfloor \times \left(G - \left(n - \left\lfloor \frac{n}{G} \right\rfloor \times G\right)\right) \qquad (12)$$

$$n = n_{\left\lfloor \frac{n}{G} \right\rfloor+1} + n_{\left\lfloor \frac{n}{G} \right\rfloor} \qquad (13)$$

In the most unbalanced partitioning of *n* into *G* groups there can be (G-1) groups with group size 1 and one group with group size (n-(G-1)).

$$n_1 = (G-1) \times G_1 \qquad (14)$$

$$n_{n-G+1} = 1 \times G_{n-G+1} \qquad (15)$$

$$n = n_1 + n_{n=G+1} \qquad (16)$$

When we take the average as the overall risk, the overall risk value solely depends upon the number of groups. Different distribution for the same total number of groups does not have any impact on the overall risk value. However, the maximum identification probabilities in different distributions for a fixed value of G and n vary and people are more concerned with higher risks than lower risks. So, the grouping having higher risk records should get a higher risk value and the grouping having evenly distributed records should get a lower risk value. We define the overall risk of a given dataset as:

$$DR_{thr} = \frac{1}{n}\left(G + \frac{m}{m+1} \times \frac{|A_k| \times \sum_{k=1}^{G} (P_k - P_{(b)})}{\sum_{k=1}^{G} (P_{(u)} - P_{(b)})} \right) \qquad (17)$$

$$\forall (P_k \text{ - } \mathrm{P}_{(b)}) >= 0$$

where,

m is the number of possible partitioning of n records into G groups and can be calculated using partitioning theory,

P_k is the probability of identification of an entity in the K^{th} group of the given distribution,

$P_{(b)}$ is the probability of identification of an entity in the K^{th} group for the balanced distribution,

$P_{(u)}$ is the probability of identification of an entity in the K^{th} group for the most unbalanced distribution.

Fig. 4 briefly describes the algorithm for calculating our threshold-based identity disclosure risk.

Algorithm 1 Disclosure Risk (SM)

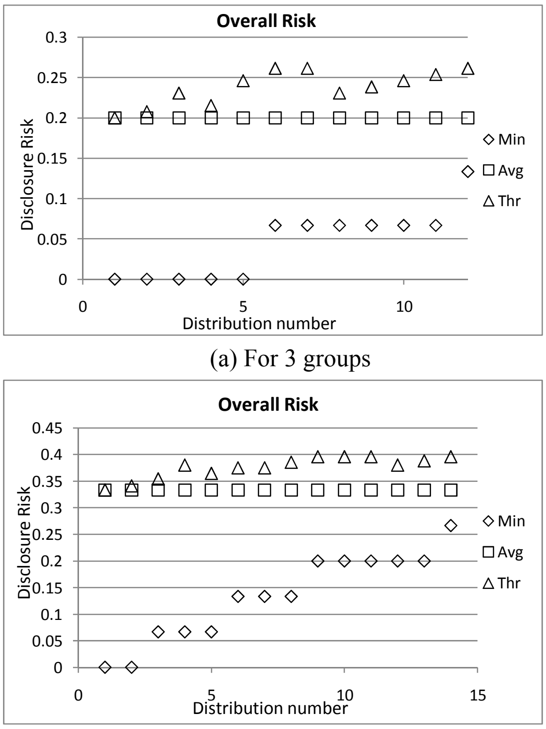
```
Count number of records $n$
Count total number of groups G
Calculate the number of possible distributions $m$ for the counted $n$, $G$
Find the group sizes $|A_k|$ $\forall(1 \le k \le G)$ and sort them in ascending order. Let
    $$P_k = \frac{1}{|A_k|}$$
Find the group sizes $|A_{(b)}|$ for the most balanced partitioning of $n$ into $G$ parts by using the equations 11, 12 and 13 and sort them ascending. Let
    $$P_{(b)} = \frac{1}{|A_{(b)}|}$$
Find the group sizes $|A_{(u)}|$ for the most unbalanced partitioning of $n$ into $G$ parts by using the equations 14, 15 and 16 and sort them ascending. Let
    $$P_{(u)} = \frac{1}{|A_{(u)}|}$$
Calculate Delta($\delta$) for $\forall(P_k - P_{(b)}) >= 0$; $(1 \le k \le G)$
    $$\delta = \frac{m}{m+1} \times \frac{1}{n} \times \frac{n_k \sum_{k=1}^{G} P_k - P_{(b)}}{\sum_{k=1}^{G} P_{(u)} - P_{(b)}}$$
Calculate the Disclosure Risk as
    $$DR_{thr} = \frac{G}{n} + \delta$$
```

Fig. 4 Algorithm for Calculating Threshold-based Disclosure Risk for a given shared data

VI. Experimental Results

We generated simulation-based shared microdata for different values of n and G. For several scenarios, we calculate and compare the disclosure risks in three metrics: unique record-based "Min" metric, probability-based "Avg" metric and our threshold-based "Thr" metric. In the graphs, we show risks for all possible distributions of n records into G groups. However, a given shared microdata will have any one of the possible distributions and thus will have one particular risk value on the graph depending on its distribution.

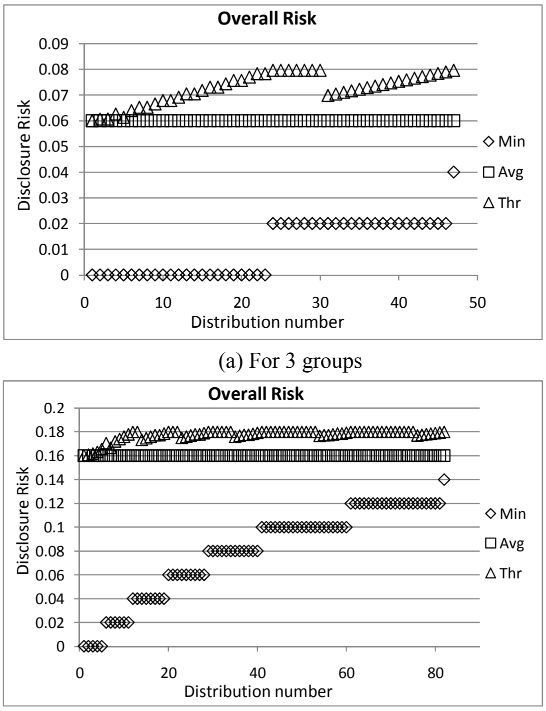
The graphs in Fig. 5, 6 and 7 show the disclosure risks in the metrics for 15 records divided into 3 and 5 groups, 50 records divided into 3 and 8 groups, and 100 records divided into 3 and 8 groups respectively. The X axis takes the risk distribution number. The value on the X axis has no relation with the overall risk measurement. This is because a given dataset will have one of the possible distribution numbers. We use this distribution number to refer to one of the possible distributions.



(a) For 3 groups

(b) For 5 groups

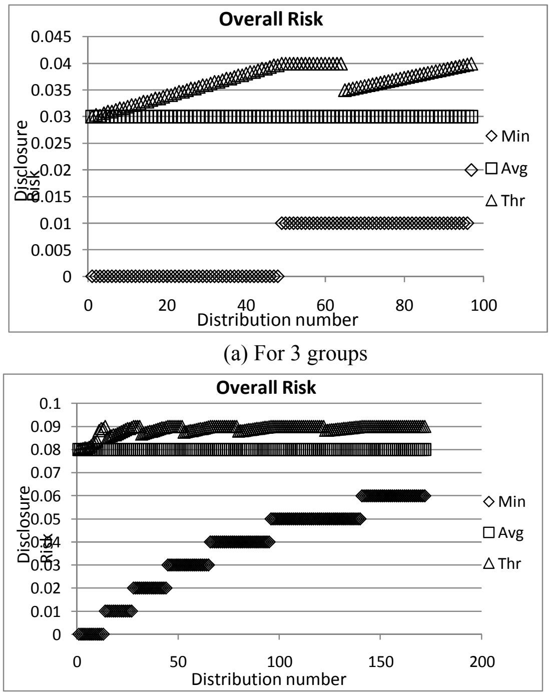
Fig. 5 Disclosure Risks for all possible distribution 15 records into 3 groups and 5 groups.



(a) For 3 groups

(b) For 8 groups

Fig. 6 Disclosure Risks for all possible distribution of 50 records into 3 groups and 8 groups.



(a) For 3 groups

(b) For 8 groups

Fig. 7 Disclosure Risks for all possible distribution of 100 records into 3 groups and 8 groups.

From the graphs, we see that the min metric gives quite a lower value of risk compared to the other two metrics but increases for certain groupings (towards un-evenly distributed group sizes). In all of the graphs, at a specific point, i.e., the most balanced distribution of group sizes, our threshold-based metric calculates the same risk value as the risk value in the probability based Avg. metric. Thus, if the given shared data is grouped in the most balanced size, the Thr. metric and the Avg. metric will calculate the same disclosure risk. However, for all other distribution of group sizes, Thr. Metric gives a higher risk value than the Avg. metric, which implies some of the entities have higher disclosure risks than others.

VII. Conclusion

In our threshold-based risk measurement, the higher the identification risk of some groups compared to other groups, the higher the overall risk will be. For the same number of groups, the risk will be the minimum and equal to the value in the Avg. metric when all of the group sizes will be equal and the risk will be the maximum and equal to when the group sizes will be most un-evenly distributed. Before releasing a set of anonymous data, an organization can evaluate the risk value for the whole dataset by taking the average. However, to evaluate how well the anonymization has been done, it needs to calculate the risk value in the Thr. metric and check whether the value is close to the maximum or minimum. If it is at the minimum, then the anonymization is at its best state and if it is at the maximum, the anonymization is at its worst state. In the Thr metric, with a risk value equal to the average indicates that an individual has an equal chance of being identified as others and a higher risk value indicates that an individual has a higher chance of being identified than others.

References

[1] L. Willemborg, T. Waal, “Elements of Statistical Disclosure Control”. Springer Verlag. 2001.

[2] P. Kosseim and K. El Emam, "Privacy Interests in Prescription Records, Part 1: Prescriber Privacy," IEEE Security and Privacy, vol. 7, pp.72-76, 2009

[3] K. El Emam and P. Kosseim, "Privacy Interests in Prescription Records, Part 2: Patient Privacy," IEEE Security and Privacy, vol. 7, pp.75-78, 2009

[4] J. Lane, P. Heus and T. Mulcahy, “Data access in a cyber world: making use of cyberinfrastructure”, Transactions on Data Privacy, 1(1), pp.2–16, 2008

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

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

[7] R. J. A. Little, “Statistical Analysis of Masked Data”, Journal of Official Statistics, Vol. 9, pp.407-426, 1993.

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

[9] C. J. Skinner, C. Marsh, S. Openshaw, and C. Wymer, “Disclosure control for census microdata”, Journal of Official Statistics, pp.31-51. 1994.

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

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

[12] K. Muralidhar and R. Sarathy, “Security of Random Data Perturbation Methods”, ACM Transactions on Database Systems, Vol. 24, No. 4, pp.487-493, 1999.

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

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

[15] D. Lambert, “Measures of Disclosure Risk and Harm”. Journal of Official Statistics, Vol. 9, pp.313-331, 1993.

[16] S. E. Fienberg, U. E. Markov, “Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data”, Journal of Official Statistics, pp385 – 397, 1998.

[17] M. J. Elliot, “DIS: a new approach to the measurement of statistical disclosure risk”, International Journal of Risk Management, pp.39 –48, 2000.

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

[19] J. G. Bethlehem, W. J., Keller, and J. Pannekoek, “Disclosure control of microdata”. Journal of the American Statistical Association., vol. 85, pp.38-45, 1990.

[20] B. Greenberg, and L. Zayatz, “Strategies for measuring risk in public use microdata files”. Statistica Neerlandica, vol. 46, pp.33-48, 1992.

[21] C.J. Skinner, C. Marsh, S. Openshaw, and C. Wymer, “Disclosure control for census microdata”. Journal of Official Statistics., vol. 10, pp.31-51, 1994.

[22] G. Chen, and S. Keller-McNulty, “Estimation of identification disclosure risk in microdata”. Journal Official Statistics., vol. 14, pp.79-95, 1998.

[23] S.E. Fienberg, and U.E. Makov, “Confidentiality, uniqueness and disclosure limitation for categorical data”, Journal Official Statistics, vol. 14, pp.385-397, 1998.

[24] S.M. Samuels, “A Bayesian, species-sampling-inspired approach to the unique problems in microdata disclosure risk assessment”. Journal Official Statistics, vol. 14, pp.373-383, 1998.

[25] M.J. Elliot, and A. Dale,” Scenarios of attack: the data intruder’s perspective on statistical disclosure risk”. Netherlands Official Statist., Spring, pp.6-10, 1999.

[26] G. Paass, “Disclosure risk and disclosure avoidance for microdata”. J.Bus.Econ.Statist., vol. 6, pp.487-500, 1988.

[27] U. Blien, H. Wirth, and M. Müller, “Disclosure risk for microdata stemming from official statistics”. *Statistica Neerlandica,*vol. 46*,* pp. 69-82, 1992.

[28] X. Xiao, Y. Tao and N. Koudas, “Transparent Anonymization: Thwarting Adversaries Who Know the Algorithm, ACM Transactions on Database Systems (TODS)”, Vol. 35, Issue 2, April 2010.

[29] V.S. Laks, Lakshmanan and T. NG Raymond and G. Ramesh, “On disclosure risk analysis of anonymized itemsets in the presence of prior knowledge”, ACM Transactions on Knowledge Discovery from Data (TKDD) Vol.2 , Issue 3 October 2008.

[30] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. VENKITASUBRAMANIAM, “l-Diversity: Privacy Beyond k-Anonymity, ACM Transactions on Knowledge Discovery from Data (TKDD)” Vol. 1 , Issue 1 , March 2007.

[31] F. K. Dankar and K. E. Emam, “A Method for Evaluating Marketer Re-identification Risk”, Proceedings of the 2010 EDBT/ICDT Workshops, Lausanne, Switzerland 2010

[32] T.M. Truta, F. Fotouhi and D. Barth-Jones, “Assessing Global Disclosure Risk in Masked Microdata”, Proceedings of the 2004 ACM workshop on Privacy in the electronic society table of contents, Washington DC, USA, pp.85 - 93, 2004

[33] T.M. Truta, F. Fotouhi and D. Barth-Jones, “Disclosure Risk Measures for the Sampling Disclosure Control Method”, Proceedings of the 2004 ACM symposium on Applied computing, Nicosia, Cyprus, pp.301 - 306, 2004.

[34] M. Bezz, “Expressing privacy metrics as one-symbol information”, Proceedings of the 2010 EDBT/ICDT Workshops, Lausanne, Switzerland, Article No.: 29, 2010

[35] J. Domingo-Ferrer and David Rebollo-Monedero, “Measuring Risk and Utility of Anonymized Data Using Information Theory”, Proceedings of the 2009 EDBT/ICDT Workshops, Saint-Petersburg, Russia Pages: 126-130, 2009

[36] T.M. Truta, F. Fotouhi and D. Barth-Jones, “Privacy and Confidentiality Management for the Microaggregation Disclosure Control Method: Disclosure Risk and Information Loss Measures”, Proceedings of the 2003 ACM workshop on Privacy in the electronic society, Washington, DC, pp.21 - 30, 2003.

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

[38] R. Benedetti, L. Franconi, “Statistical and Technological Solutions for Controlled Data Dissemination”. *Proceedings of New Techniques and Technologies for Statistics*, Vol. 1, pp. 225-232, 1998.

[39] D. E. Denning and P. J. Denning, “Data Security”. *ACM Computing Surveys*, Vol. 11, pp. 227-249, 1979.

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

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

[42] T.M. Truta, F. Fotouhi and D. Barth-Jones, “Disclosure risk measures for microdata”, Proceedings of the 15th International Conference on Scientific and Statistical Database Management, Cambridge, MA, Page: 15-22, 2003

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

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

[45] G.T. Duncan and D. Lambert, “The risk of disclosure for microdata”, *J.Bus.Econ. Statist.*, vol. **7**, pp.207-217, 1989.

[46] D. Lambert, “Measures of disclosure risk and harm”. Journal Official Statistics., vol.**9**, pp.313-331, 1993.