Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®

ELSEVIER Journal of Computational and Applied Mathematics 164–165 (2004) 285–293

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.com/locate/cam

Disclosure risk assessment in statistical data protection☆

Josep Domingo-Ferrer[a,*], Vicenç Torra[b]

[a] *Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain*

[b] *Institut d'Investigació en Intelligència Artificial, Campus de Bellaterra, E-08193 Bellaterra, Catalonia, Spain*

Received 8 August 2002; received in revised form 14 February 2003

Abstract

Statistical data protection, also known as statistical disclosure control, is about methods that try to prevent published statistical information (tables, individual information) from disclosing the contribution of specific respondents, who may be individuals or enterprises. In addition to keeping disclosure risk acceptably low, methods used for statistical data protection should not significantly damage the utility of the data being protected. This paper surveys different ways to assess the risk of disclosure in the protection of both individual data (called microdata) and tabular data. A noteworthy result also presented is that the most widely used rule for assessing disclosure risk in tabular data protection is flawed.

Keywords: Disclosure risk; Statistical disclosure control; Official statistics; Business statistics

1. Introduction

Statistical data confidentiality, also known as statistical disclosure control (SDC), is the discipline that seeks to modify statistical data so that they can be published by statistical offices without giving away confidential information that can be linked to specific respondents behind the data. Although released information should be as detailed as possible from the users' viewpoint, data utility (i.e. accuracy) is in conflict with respondents' privacy. In other words, a tradeoff must be reached between the user's wish that the information loss caused by SDC methods be as low as possible and the wish by respondents that the disclosure risk (risk that released records can be linked to specific respondents) be kept low as well. See [2,3,16] for details on SDC methods in use.

☆ Work partly funded by the European Union under project "CASC" IST-2000-25069.
* Corresponding author. Tel.: +34-977-558270; fax: +34-977-559710.
E-mail addresses: jdomingo@etse.urv.es (J. Domingo-Ferrer), vtorra@iiia.csic.es (V. Torra).

0377-0427/$ - see front matter
doi:10.1016/S0377-0427(03)00643-5

1.1. Contribution and plan of this paper

Statistical offices release two kinds of data through their statistical databases: microdata sets (individual respondent records) and tabular data. Any attempt to compare methods for statistical data protection should focus on two basic attributes:

(1) *Disclosure risk*: A measure of the risk to respondent confidentiality that the data releaser (typically a statistical agency) would experience as a consequence of releasing the table.
(2) *Data utility*: A measure of the value of the released table to a legitimate data user.

A first approach for measuring data utility is to take generic measures such as the reciprocal of the mean squared error between the original and released data [5,3]. While this may be useful as a crude approach, a more accurate utility assessment must necessarily take into account the specific data uses the user is interested in. Thus, strictly speaking, there is no universal data utility measure.

The situation for disclosure risk is quite different, since it does not depend on particular data uses. Therefore, a unified approach for measuring disclosure risk is reasonable and even desirable. Such a unified approach exists in the literature for tabular data, but not for microdata. The contribution of this paper is twofold: first, we propose a unified risk assessment methodology for microdata based on record linkage and, second, we highlight the shortcomings of the most commonly used rule for risk assessment in tabular data protection.

In Section 2, disclosure risk assessment for microdata is discussed. Section 3 deals with disclosure risk assessment for tabular data; in particular, it is shown that the (very popular) dominance rule used to assess disclosure risk is flawed. Conclusions are summarized in Section 4.

2. Disclosure risk assessment in microdata protection

Disclosure risk assessment for microdata is performed a posteriori, that is, after protecting the microdata. First, an original microdata set is protected, and then the risk of disclosure is computed by taking as one of the inputs the protected microdata set.

Literature on disclosure risk for microdata is basically related to nonperturbative methods based on sampling, in which the protected microdata set is obtained as a sample of the original data set. Disclosure risk here is measured as the probability that a sample unique is a population unique [6,14]. If the size of the sample is similar to the size of the whole population, such a probability can be dangerously high; in that case, an intruder who locates a unique value in the released sample can be almost sure that there is a single individual in the population with that value. This could lead to identification of that individual.

The uniqueness property as stated above is no longer relevant for perturbative methods, since in this case the whole microdata set is published, but with some distortion. There is not much literature on disclosure risk that can be used for a broad class of perturbative methods; disclosure risk measures tend to be method-specific (measures described in [1] are still up-to-date). Empirical methods, like record linkage techniques, provide a more unified approach to disclosure risk assessment for perturbative methods. We briefly describe below two methods for record linkage which yield empirical disclosure risk measures. An analytical measure based on interval disclosure is also described.

2.1. Distance-based record linkage

Distance-based record linkage was first described in [12] for the specific case of numerical variables and using the Euclidean distance. We next discuss how to generalize it for any perturbative method provided that a distance between the original and the masked variables can be defined (note that a distance can be defined not only between numerical variables, but also between some types of categorical variables, such as ordinal variables).

Let the original and masked data sets consist both of d variables (it is assumed that both data sets contain the same variables). We define that a record in the masked data set corresponds to the nearest record in the original data set, where "nearest" means at shortest d-dimensional distance. Assume further that the intruder can only access i key variables of the original data set (such variables may be available in external data sets accessible to the intruder) and tries to link original and masked records based on these i variables.

Linkage then proceeds by computing i-dimensional distances between records in the original and the masked data sets (distances are computed using only the i key variables). If they are numerical, variables used are standardized to avoid scaling problems. A record in the masked data set is labeled as "correctly linked" when the nearest record using i-dimensional distance is the corresponding one (i.e., the nearest record using d-dimensional distance). The percentage of "correctly linked" records is a measure of disclosure risk.

Distance-based record linkage was originally designed for numerical variables. In order to extend it for dealing with categorical variables, we need to define a distance for this type of variables, which can be done as follows:

Definition 1.

(1) For a nominal variable V, the only permitted operation is comparison for equality. This leads to the following distance definition:

$$\mathrm{d}_V(c,c') = \begin{cases} 0 & \text{if } c = c', \\ 1 & \text{if } c \neq c', \end{cases}$$

where c and c' correspond to categories for variable V. An alternative definition applicable in some cases is to use a string matching algorithm to compute the distance between two nominal variables; this is especially appropriate for nominal variables such as names and addresses, which can contain typos or be written in several formats (see [15]).

(2) For an ordinal variable V, let $\leqslant_V$ be the total order operator over the range of V. Then, the distance between categories c and c' is defined as the number of categories between the minimum and the maximum of c and c' divided by the cardinality of the range (denoted by $D(V)$):

$$\mathrm{d}_V(c,c') = \frac{|c'' \colon \min(c,c') \leqslant_V c'' \leqslant_V \max(c,c')|}{|D(V)|}.$$

Note that Definition 1 specifies a distance for a single variable. To obtain a distance for pairs of records, the distances corresponding to the variables in the records should be aggregated.

2.2. Probabilistic record linkage

In [10], a probabilistic record linkage method was described and illustrated on the 1985 Census of Tampa, Florida. The matching algorithm uses the linear sum assignment model to "pair" records in the two files to be matched (the original file and the masked file in our case). The definition of "correctly linked" records is the same as in distance-based record linkage. The percentage of correctly paired records is a measure of disclosure risk.

Although slower and more complex than the distance-based method described in the previous section, this approach is sometimes attractive because it only requires the user to provide two probabilities as input: one is an upper bound of the probability of a false match, and the other an upper bound of the probability of false nonmatch. Unlike distance-based record linkage, probabilistic record linkage does not require rescaling variables nor makes any assumption on their relative weight (by default, distance-based record linkage assumes that all variables have the same weight).

2.3. Interval disclosure

The intruder may not be satisfied with pairing masked records and original records. For numerical or ordinal variables, she may wish to go further and find an interval around each variable value in a masked record which contains the corresponding value in the corresponding original record. We next describe the interval disclosure measure we used in our comparative study [4].

For a record in the masked data set, compute rank intervals as follows: (1) rank each variable independently; (2) define a rank interval around the value the variable takes for record r as the interval centered on the rank of the value of record r and comprising the surrounding ranks differing among them less than p% of the total number of records; (3) convert rank intervals into value intervals by mapping ranks to values. Then the proportion of original values which fall into the interval centered around their corresponding masked value is a measure of disclosure risk. A 100% proportion means that an attacker is completely sure that the original value lies in the interval around the masked value (interval disclosure).

2.4. Example

Let us consider the files **A** and **B** in Table 1. Both files contain 8 records and 3 variables (Name, Surname and Age). For the sake of understandability, the files are defined so that records in the same row correspond to matched pairs and records in different rows correspond to unmatched pairs. The goal of record linkage in this example is to classify all possible pairs so that pairs with both records in the same row are classified as linked pairs and all the other pairs are classified as nonlinked pairs.

Next, we illustrate how distance-based and probabilistic record linkage would be carried out in this example:

- In distance-based record linkage, a distance is defined for variable pairs (Name_A, Name_B), (Surname_A, Surname_B) and (Age_A, Age_B). One possibility to aggregate those distances and obtain a record-level distance is to use the sum of squares as an aggregation function.
- In probabilistic record linkage, we consider all record pairs $(a, b) \in \mathbf{A} \times \mathbf{B}$. For each record pair, a coincidence vector is computed as a binary string consisting of as many bits as variables are

Table 1
Records in the files **A** and **B** to be linked

Name_A	Surname_A	Age_A	Name_B	Surname_B	Age_B
Joan	Casanovas	19	Joan	Casanovas	19
Pere	Joan	17	Pere	Joan	17
J.M.	Casanovas	35	J.Manel	Casanovas	35
Juan	Garcia	53	Juan	Garcia	53
Ricardo	Garcia	14	Ricard	Garcia	14
Pere	Garcia	18	Pere	Garcia	82
Juan	Garcia	18	Juan	Garcia	18
Ricard	Tanaka	14	Ricard	Tanaka	18

in the records; the ith bit of the coincidence vector is 1 if the values of the ith variable in both record are the same and is 0 otherwise. For the files in Table 1, there are 8 possible different coincidence vectors. In general, the number of different coincidence vectors is much less than the number of record pairs in $\mathbf{A} \times \mathbf{B}$ (64 in the case of Table 1). Yet, in probabilistic record linkage, the classification of any pair (a, b) as linked or nonlinked is solely based on its coincidence vector.

3. Disclosure risk assessment in tabular data protection

Tabular data constitute the most traditional output released by statistical agencies. Being aggregate data, one might infer that tables cannot leak information about specific respondents. As argued in [8], it turns out that table cells often do contain information on a single or very few respondents, which implies a disclosure risk for the data of those respondents. In these cases, disclosure control methods must be applied to the tables prior to their release.

Disclosure assessment for tables is usually performed a priori, that is, before tables are protected. The standard approach is to use a *sensitivity rule* to decide whether a table cell is sensitive and should be protected.

3.1. A priori risk assessment through sensitivity rules

For magnitude tables (normally related to economic data), there are two widely accepted rules to decide whether a cell is sensitive:

(n, k)-*Dominance*: In this rule, n and k are two parameters with values to be specified. A cell is called sensitive if the sum of the contributions of n or fewer respondents represents a fraction k or more of the total cell value. Usually k is a fraction higher than 0.6.

pq-Rule: The prior–posterior rule is another rule gaining increasing acceptance. It also has two parameters p and q. It is assumed that, prior to table publication, each respondent can estimate the contribution of each other respondent to within less than q percent. A cell is considered sensitive if, posterior to the publication of the table, someone can estimate the contribution of an individual respondent to within less than p percent. A special case is the p%-rule: in this case, no knowledge prior to table publication is assumed, i.e. the pq-rule is used with $q = 100$.

3.2. A critique to the dominance rule

According to [7,9,11], the (n,k)-dominance rule is the most popular one for magnitude tables, followed by the p%-rule and the pq-rule. Yet, it is significant to note that the US Census Bureau switched in 1992 from the dominance rule to the p%-rule, and the German Statistisches Bundesamt did the same in 2001.

The dominance rule has received critiques for failing to adequately reflect the risk of disclosure, but these have been limited to numerical counterexamples for particular choices of n and k. The following is a counterexample from [13] for the particular case $n = 1$:

Example 1 (Robertson and Ethier, 2002). In the dominance rule, let $n = 1$ and $k = 0.6$ (60%). Then a cell with value 100 and contributions 59, 40, 1 is declared not sensitive, while a cell with value 100 and contributions 61, 20, 19 would be declared sensitive. Assume now that the second largest respondent of both cells knows the total 100 and is interested in estimating the contribution of the largest respondent. Then, for the $(59, 40, 1)$ cell, she removes her contribution and gets an upper bound $100 - 40 = 60$ for the largest contribution. For $(61, 20, 19)$ the upper bound she gets is $100 - 20 = 80$, much farther from the real largest contribution. So the cell declared nonsensitive by the rule allows better inferences than the cell declared sensitive!

We generalize below the critique in the above example for any values of n and k. Assume a cell X in a table takes a value x which is formed by N respondent contributions $x_1, \ldots, x_N$. Equivalently,

$$x = x_1 + x_2 + \cdots + x_N.$$

The dominance rule declares X to be sensitive if a few contributions (n or less) add up to a substantial fraction of x (k or more).

In order to construct a nonsensitive cell, we need the following result.

Lemma 1. *For any integer n and $k \in (0, 1]$, there exists an integer N and $r \in [0, 1)$ such that*

$$f(r) = \frac{r^n - 1}{r^N - 1} = k. \tag{1}$$

A nonsensitive cell is now constructed as follows:

Construction 1 (*Non-sensitive cell X_{ns}*).

(1) *Take $r \in [0, 1)$ and N as defined in Lemma 1. Then it holds that*

$$k = \frac{r^n - 1}{r^N - 1} = \sum_{i=1}^{n} \frac{r^i(r-1)}{r^{N+1} - r}. \tag{2}$$

(2) *Let*

$$R_i := \frac{r^i(r-1)}{r^{N+1} - r}. \tag{3}$$

(3) *Consider a cell* X_{ns} *whose* N *relative contributions are*

$$x_i/x = \begin{cases} R_i & \text{for } i = 1,\ldots,n-1 \text{ and } i = n+2,\ldots,N, \\ R_n - \varepsilon & \text{for } i = n, \\ R_{n+1} + \varepsilon & \text{for } i = n+1, \end{cases}$$

where $\varepsilon := (R_n - R_{n+1})/3$. *With this choice of* ε, *one still has* $x_n/x > x_{n+1}/x$.

(4) *According to Expression* (2), *the* n *largest relative contributions* $x_1/x,\ldots,x_n/x$ *add to* $k - \varepsilon$. *Therefore, there is no subset of* n *contributions adding to* k, *so* X_{ns} *is clearly not sensitive according to the dominance rule.*

A sensitive cell is constructed as follows.

Construction 2 (*Sensitive cell* X_{s}).

(1) *Take the same values* N, n *and* k *used in Construction* 1.
(2) *Consider a cell* X_{s} *whose relative contributions are*

$$x_i/x = \begin{cases} R_i & \text{for } i = 1,\ldots,n, \\ (1-k)/(N-n) & \text{for } i = n+1,\ldots,N. \end{cases}$$

(3) *By construction, the sum of the* n *relative contributions* $x_1/x,\ldots,x_n/x$ *is* k. *Thus,* X_{s} *is declared sensitive by the* (n,k)*-dominance rule.*

We next show that the cell declared nonsensitive by the dominance rule can yield a closer upper bound for the largest contribution than the cell declared sensitive.

Theorem 1. *Let* n *and* k *be the parameters of the dominance rule. Assume a coalition of the* n *second largest contributors want to upper bound the largest contribution. For any* n *and* k, *if* N *is taken large enough, then the coalition gets a proportionally closer upper bound for the case of* X_{ns} *than for the case of* X_{s}.

The following example illustrates that N does not need to be very large for Theorem 1 to hold.

Example 2. For $n = 1$, $k = 0.369$ and $N = 3$, we have $r = 0.9$. The largest relative contribution is $x_1/x = 0.369$ for both X_{s} and X_{ns}. The tail of the $N - n - 1 = 1$ smallest relative contribution is 0.298893 for X_{ns} and 0.315498 for X_{s}.

Theorem 1 highlights a major flaw in the dominance rule. Note that n can be as small as 1 and, in that case, *a single cell contributor* (the second largest) can, without any help, get more precise estimates on the largest contribution for a cell declared nonsensitive than for a cell declared sensitive. This gives some theoretical justification to the decision of leading statistical agencies to abandon the dominance rule in favor of sensitivity rules with more general definitions of sensitivity, like the pq-rule or the p%-rule.

4. Conclusion

For microdata protection, we have proposed empirical disclosure risk measures based on record linkage which should be preferred to conventional measures based on uniqueness because, unlike the latter, the former measures apply to both perturbative and nonperturbative disclosure control methods. In particular, a general distance-based record linkage method has been presented.

For tabular data protection, it has been shown that the most widely used sensitivity rule for a priori risk assessment, the dominance rule, is flawed. This justifies the current trend to abandon this rule in favor of other rules.

Acknowledgements

Thanks go to Sarah Giessing for useful comments on some parts of this paper.

Appendix A.

Proof of Lemma 1. For fixed n and N, with $n < N$, the function

$$f(r) = \frac{r^n - 1}{r^N - 1}$$

bijectively maps the interval $[0, 1)$ onto the interval $(n/N, 1]$. The lower bound of the image interval is determined as

$$\lim_{r \to 1} f(r) = \frac{n}{N}.$$

Thus, the lemma holds if we take N large enough so that $n/N < k$. □

Proof of Theorem 1. For both X_s and X_{ns}, the n second largest contributors know that the largest contribution is upper-bounded by the total x minus their own contributions, that is

$$x_1 \leqslant x - (x_2 + x_3 + \cdots + x_{n+1}). \tag{A.1}$$

The distance between x_1 and the upper bound (4) is exactly the sum of the $N - n - 1$ smallest contributions. We next show that, for large enough N, this sum is smaller for X_{ns} than for X_s (since x_1 is the same for both cells, this is equivalent to showing that the upper bound on the largest contribution is proportionally closer for X_{ns}). Now, both X_{ns} and X_s total to x, so we can use in what follows relative contributions rather than absolute contributions for both cells. For X_{ns}, the $N - n$ smallest relative contributions add to $1 - k + \varepsilon$ by construction; therefore, the $N - n - 1$ smallest relative contributions add to $1 - k + \varepsilon$ minus the $(n + 1)$th largest relative contribution

$$(1 - k) + \varepsilon - (R_{n+1} + \varepsilon) = (1 - k) - \frac{kr^n(r - 1)}{r^n - 1}. \tag{A.2}$$

To obtain the last term of expression (5), we have used that $k = (r^n - 1)/(r^N - 1)$. On the other hand, for X_s, the sum of the $N - n - 1$ smallest relative contributions is

$$\frac{(N-n-1)(1-k)}{N-n} = (1-k) - \frac{1-k}{N-n}. \tag{A.3}$$

If $N \to \infty$, expression (6) approaches $1 - k$. On the other hand, if we let $N \to \infty$, then r is such that $k = 1 - r^n$ (according to Lemma 1). In this case, expression (5) becomes

$$(1-k) - \frac{k(1-k)(r-1)}{1-k-1} = r(1-k) < (1-k).$$

Thus, N can be taken large enough so that expression (6) is larger than expression (5), which causes the theorem to hold. □

References

[1] N.R. Adam, J.C. Wortmann, Security-control methods for statistical databases: a comparative study, ACM Comput. Surv. 21 (1989) 515–556.

[2] J. Domingo-Ferrer (Ed.), Inference Control for Statistical Databases, Lecture Notes in Computer Science, Vol. 2316, Springer, Berlin, 2002.

[3] J. Domingo-Ferrer, V. Torra, Disclosure control methods and information loss for microdata, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access, North-Holland, Amsterdam, 2001, pp. 91–110.

[4] J. Domingo-Ferrer, V. Torra, A quantitative comparison of disclosure control methods for microdata, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access, North-Holland, Amsterdam, 2001, pp. 111–133.

[5] G.T. Duncan, S.E. Fienberg, R. Krishnan, R. Padman, S.F. Roehrig, Disclosure limitation methods and information loss for tabular data, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access, North-Holland, Amsterdam, 2001, pp. 135–166.

[6] M.J. Elliot, C.J. Skinner, A. Dale, Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk, Res. Official Statist. 1 (2) (1999) 53–67.

[7] F. Felsö, J. Theeuwes, G.G. Wagner, Disclosure limitation methods in use: results of a survey, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access, North-Holland, Amsterdam, 2001, pp. 17–42.

[8] S. Giessing, Nonperturbative disclosure control methods for tabular data, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access, North-Holland, Amsterdam, 2001, pp. 185–213.

[9] J. Holvast, Statistical dissemination, confidentiality and disclosure, in: Proceedings of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality, Eurostat, Luxembourg, 1999, pp. 191–207.

[10] M.A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, J. Amer. Statist. Assoc. 84 (1989) 414–420.

[11] T. Luige, J. Meliskova, Confidentiality practices in the transition countries, in: Proceedings of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality, Eurostat, Luxembourg, 1999, pp. 287–319.

[12] D. Pagliuca, G. Seri, Some results of individual ranking method on the system of enterprise accounts annual survey, Esprit SDC Project, Deliverable MI-3/D2, 1999.

[13] D. Robertson, R. Ethier, Cell suppression: theory and experience, in: J. Domingo-Ferrer (Ed.), Inference Control in Statistical Databases, Lecture Notes in Computer Science, Vol. 2316, Springer, Berlin, 2002, pp. 9–21.

[14] C. Skinner, C. Marsh, S. Openshaw, C. Wymer, Disclosure control for census microdata, J. Official Statist. 10 (1994) 31–51.

[15] V. Torra, J. Domingo-Ferrer, Record linkage methods for multidatabase data mining, in: V. Torra (Ed.), Information Fusion in Data Mining, Springer, Berlin, 2003, pp. 99–130.

[16] L. Willenborg, T. de Waal, Elements of Statistical Disclosure Control, Springer, New York, 2001.