MSRI
Mathematical Sciences Research Institute

MULTISTAGE MASKING METHODS FOR MICRODATA PROTECTION

ANNA OGANIAN* AND ALAN F. KARR

Publication of statistical information by national agencies in the form of microdata (i.e., individual records) raises the problem of preventing disclosure of confidential information about particular respondents without significantly damaging the utility of the data being protected.

Often, statistical agencies disseminate information only in the form of tables. But, microdata—records which contain information about individuals or establishments—offer far greater flexibility for statistical research, especially of an exploratory nature, than tables. As a result, there has been an increasing demand from users for such data, and agencies would like to be able to comply this demand, provided that confidentiality is not compromised. In particular, there is a well-recognized need to prevent both identity and attribute disclosure.

Before releasing microdata, a statistical office deletes from the data direct identifiers, such as names and addresses; however, the risk of identification still exists, for example, by means of linkage of existing databases, available to the users, to the released data. So, in addition, released microdata are typically perturbed, in order to make disclosure more difficult.

For such a purpose, a number of Statistical Disclosure Control (SDC) techniques have been developed. Following are the most widely used SDC methods for microdata:

- *Rankswapping:* First, values of variable V_i are ranked in ascending order. Then, each ranked value of V_i is swapped with another ranked value randomly chosen within a restricted range; e.g., the rank of two swapped values cannot differ by more then p percent of the total number of records.
- *Microaggregation:* Records are clustered into small aggregates or groups of size at least k. Rather than publishing an original variable V_i for a given record, the average of the values of V_i over the group to which the record belongs is published. Classical microaggregation,

Key words and phrases. Statistical Disclosure Control (SDC), microdata, disclosure control methods, data utility, disclosure risk, combinations of SDC methods.

according to [4], required that all groups, except perhaps one, be of size k.

Variants considered include microaggregation on projected data using z-scores projection and principal components projection (multivariate data are ranked by projecting them onto a single axis, using the first principal component or the sum of z-scores, and then all the groups, except perhaps one, are taken of the same size [1], [4], [3]), and microaggregation on unprojected multivariate data.

- *Additive noise:* Random noise, usually normal noise with the same correlation structure as the original unmasked data, is added [10], [14], [11], [15], [2].

All the aforementioned methods have very different properties and affect the data in a different way. The common feature for SDC methods is that all of them are designed with two goals in a mind. First is to minimize disclosure risk, that is, the risk to respondent confidentiality that the data releaser would experience as a consequence of releasing the data. And the second is to maximize data utility, that is, the value of the released data to a legitimate data user. Note that these are usually conflicting goals, in the sense that the methods with high utility usually have high disclosure risk and the methods with low disclosure risk very often have low utility.

Different proposals can be found in the recent literature regarding utility measures; see, for example, [6], [16], and [12].

In this paper, we adopt a broad utility measure, called *propensity score utility*, recently proposed in [9]. This measure is both suitable for any distribution of the data and not tied to a particular data analysis.

Propensity score utility measures the distance between distributions of original and masked data by the means of classification of the pooled data into two groups: one corresponding to the original and the other corresponding to the masked data.

Two varieties of disclosure risk are usually considered. *Identification disclosure* occurs when an intruder[1] can associate a released record with the individual or establishment to which it pertains. Typically, identification disclosure is effected by record linkage (see, e.g., [7], [13], [12]) to an external database containing identifiers. One measure of disclosure risk, then, is the percentage of masked records that are linked “correctly” to their parent records in the original data.

Attribute disclosure occurs when the intruder’s target is an original value of a particular attribute, for example, the salary of a particular individual. Attribute disclosure risk can be measured by the tightness of bounds for attribute values in the original data given the masked data, as in [5] in the context of tabular data.

In this paper, we show that for some masking methods, a considerable gain of information about particular attributes is possible if only a single

[1] The generic term for an illegitimate use of the data.

method is applied. This happens if the information about the parameters of this method (such as the maximal range of ranks of the records which could be swapped in the case of rankswapping or the number of records per cluster for microaggregation) was reported. Almost always, however, only a single method of perturbation, chosen in advance, is applied. A risk-utility framework for selecting the method (as well as associated parameters) is described in [8].[2] As shown in [8], the methods differ rather dramatically in terms of their effects on disclosure risk and data utility. Some are good for one but not the other, some seem not especially good for either risk or utility, and none is uniformly good for both.

These results raise the intriguing possibility of *combining SDC methods* in such a way that the combined method is superior in terms of both disclosure risk and data utility to either method alone. Focusing for simplicity on combining two methods, the intuition is as follows. The first method should be one that is

- good from the perspective of disclosure risk,
- not necessarily good in terms of data utility, but whose utility consequences can be characterized.

Then, the second method should be chosen to "reverse" the utility consequences of the first without harming disclosure risk.

In this paper, we show how these methods should be combined in order to increase the utility of the released data set. We report the results of a simulation study in which the first method is a particular form of microaggregation and various second methods are employed. The results of the simulation study showed that the combined methods significantly outperform a single method in terms of the utility. Regarding identification disclosure risk, the disclosure risk of the combinations was larger than that of a single method; however, the gain in utility is in general much more significant. For most of the combinations, the risk remains still very low: less than one percent of records are correctly identified. Clearly, even with combined methods, there is no "free lunch." As we noted earlier, high utility and low disclosure risk are conflicting goals, in the sense that methods leading to substantially increased data utility also carry increased disclosure risk. What combining methods accomplishes is dramatic increases in utility accompanied by only modest increases in risk.

References

[1] N. Anwar, *Micro-aggregation - the small aggregates method.* Internal Research Report. Luxembourg: Eurostat, 1993.

[2] R. Brand, *Microdata protection through noise*, in Inference Control in Statistical Databases. Ed. J. Domingo-Ferrer. Lecture Notes in Computer Science, 2316. Berlin: Springer, 2002. 97–116.

[2] Utility is measured differently than in is this paper, but the concepts are the same.

[3] D. Defays and N. Anwar, *Micro-aggregation: A generic method*, in Proceedings of the 2nd International Symposium on Statistical Confidentiality. Luxembourg: Office for Official Publications of the European Communities, 1995. 69–78.

[4] D. Defays and P. Nanopoulos, *Panels of enterprises and confidentiality: the small aggregates method*, in Proceedings of the '92 Symposium on Design and Analysis of Longitudinal Surveys. Ottawa: Statistics Canada, 1993. 195–204.

[5] A. Dobra, A. F. Karr, A. P Sanil, and S. E. Fienberg, *Software systems for tabular data releases*, International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, **10** (2002), no. 5, 529–544.

[6] J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra, *Comparing SDC methods for microdata on the basis of information loss and disclosure risk of disclosure control methods*, in Proceedings of the ETK-NTTS 2001 Conference. Ed. P. Nanopoulos and D. Wilkinson. Rome: European Communities, 2001. 807–825.

[7] M. A. Jaro, *Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida.* Journal of the American Statistical Association **84** (1989), 414–420.

[8] A. F. Karr, C. N. Kohnen, A. Oganian, J. R. Reiter, and A. P. Sanil, *A framework for evaluating the utility of data altered to protect confidentiality.* To appear in The American Statistician.

[9] A. F. Karr, A. Oganian, J. P. Reiter, and M.-J. Woo, *New Measures of Data Utility.* Preprint.

[10] J. Kim, *A method for limiting disclosure in microdata based on random noise and transformation*, in Proceedings of the Survey Research Methods Section, ASA. Alexandria, VA: American Statistical Association, 1986. 370–375.

[11] R. J. A. Little, *Statistical analysis of masked data*, Journal of Official Statistics **9** (1993), no. 2, 407–426.

[12] A. Oganian *Security and Information Loss in Statistical Databases Protection.* PhD thesis. Universitat Politecnica de Catalunya, 2003.

[13] D. Pagliuca and G. Seri, *Some results of individual ranking method on the system of enterprise accounts annual survey*, Deliverable MI-3/D2, Statistical Disclosure Control Project, Esprit, 1998.

[14] G. Sullivan and W. A. Fuller, *The use of measurement error to avoid disclosure*, in Proceedings of the ASA Survey Research Methods Section. Alexandria, VA: American Statistical Association, 1989. 802–807.

[15] P. Tendick and N. Matloff, *A modified random perturbation method for database security*, ACM Transactions on Database Systems **19** (1994), no. 1, 47–63.

[16] W. E. Yancey, W. E. Winkler, and R. H. Creecy, *Disclosure risk assessment in perturbative microdata protection*, in Inference Control in Statistical Databases. Ed. J. Domingo-Ferrer. Berlin: Springer-Verlag, 2002. 135–152.

(Oganian) National Institute of Statistical Sciences; Research Triangle Park, NC 27709-4006 USA
E-mail address: aoganian@niss.org

(Karr) National Institute of Statistical Sciences; Research Triangle Park, NC 27709-4006 USA
E-mail address: karr@niss.org