MICROAGGREGATION FOR PROTECTING INDIVIDUAL DATA PRIVACY

Josep Domingo-Ferrer
CRISES Group, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona
e-mail josep.domingo@urv.net, http://vneumann.etse.urv.es

Abstract

Microaggregation is a technique for protecting the privacy of respondents in individual data (microdata) releases. This papers starts with a survey of the general definitions and concepts related to microdata protection and then reviews the state of the art of microaggregation, to which our group has substantially contributed.

Keywords: Privacy, Microdata, Statistical disclosure control, Microaggregation.

1 INTRODUCTION

Privacy in statistical databases, also known as Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL), seeks to protect statistical data in such a way that they can be publicly released without giving away confidential information that can be linked to specific individuals or entities. There are several areas of application of SDC techniques, which include but are not limited to the following:

- *Official statistics.* Most countries have legislation which compels national statistical agencies to guarantee statistical confidentiality when they release data collected from citizens or companies. This justifies the research on SDC undertaken by several countries, among them the European Union (*e.g.* the CASC project[2]) and the United States.
- *Health information.* This is one of the most sensitive areas regarding privacy. For example, in the U. S., the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA,[17]) requires the strict regulation of protected health information for use in medical research. In most western countries, the situation is similar.
- *E-commerce.* Electronic commerce results in the automated collection of large amounts of consumer data. This wealth of information is very useful to companies, which are often interested in sharing it with their subsidiaries or partners. Such consumer information transfer should not result in public profiling of individuals and is subject to strict regulation; see [11] for regulations in the European Union and [28] for regulations in the U.S.

The protection provided by SDC techniques normally entails some degree of data modification strictly between the following two extreme situations:

Data encryption If the released data are encrypted, then they are extremely protected, but they are useless if the data user cannot decrypt them.

No modification If the original data are released without modification, then their usefulness/accuracy is maximal, but no protection whatsoever against disclosure is offered.

The challenge for SDC is to modify data in such a way that sufficient protection is provided while keeping at a minimum the information loss, *i.e.* the loss of the accuracy sought by database users. In the years that have elapsed since the excellent survey by [1], the state of the art in SDC has evolved so that now at least three subdisciplines are clearly differentiated:

Tabular data protection This is the oldest and best established part of SDC, because tabular data have been the traditional output of national statistical offices. The goal here is to publish *static* aggregate information, *i.e.* tables, in such a way that no confidential information on specific individuals among those to which the table refers can be inferred. See [29] for a conceptual survey and [13] for a hands-on survey.

Dynamic databases The scenario here is a database to which the user can submit statistical queries (sums, averages, etc.). The aggregate information obtained by a user as a result of successive queries should not allow him to infer information on specific individuals. Since the 80s, this has been known to be a difficult problem, subject to the tracker attack [25]. One possible strategy is to perturb the answers to queries; solutions based on perturbation can be found in [10], [21] and [27]. If perturbation is not acceptable and exact answers are needed, it may become necessary to refuse answers to certain queries; solutions based on query restriction can be found in [3] and [15]. Finally, a third strategy is to provide correct (unperturbed) interval answers, as done in [14] and [12].

Microdata protection This subdiscipline is about protecting static individual data, also called microdata. It is only recently that data collectors (statistical agencies and the like) have been persuaded to publish microdata. Therefore, microdata protection is the youngest subdiscipline and is experiencing continuous evolution in the last years.

This paper starts with an introduction to microdata protection. Then, some background on microaggregation is given. In the final sections, several types of microaggregation algorithms we have contributed are described.

2 A CLASSIFICATION OF MICRODATA PROTECTION METHODS

A microdata set $\mathbf{V}$ can be viewed as a file with n records, where each record contains m attributes on an individual respondent. The attributes can be classified in four categories which are not necessarily disjoint:

- *Identifiers.* These are attributes that *unambiguously* identify the respondent. Examples are the passport number, social security number, etc.
- *Quasi-identifiers or key attributes.* These are attributes which identify the respondent with some degree of ambiguity. (Nonetheless, a combination of quasi-identifiers may provide unambiguous identification.) Examples are name, address, gender, age, telephone number, etc.
- *Confidential outcome attributes.* These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.
- *Non-confidential outcome attributes.* Those attributes which do not fall in any of the categories above.

Since the purpose of SDC is to prevent confidential information from being linked to specific respondents, we will assume in what follows that original microdata sets to be protected have been pre-processed to remove from them identifiers and quasi-identifiers with low ambiguity (such as name).

The purpose of microdata SDC mentioned in the previous section can be stated more formally by saying that, given an original microdata set $\mathbf{V}$, the goal is to release a protected microdata set $\mathbf{V}'$ in such a way that:

1. Disclosure risk (*i.e.* the risk that a user or an intruder can use $\mathbf{V}'$ to determine confidential attributes on a specific individual among those in $\mathbf{V}$) is low.
2. User analyses (regressions, means, etc.) on $\mathbf{V}'$ and on $\mathbf{V}$ yield the same or at least similar results.

Microdata protection methods can generate the protected microdata set $\mathbf{V}'$

- either by *masking original data*, *i.e.* generating $\mathbf{V}'$ a modified version of the original microdata set $\mathbf{V}$;
- or by *generating synthetic data* $\mathbf{V}'$ that preserve some statistical properties of the original data $\mathbf{V}$.

Masking methods can in turn be divided in two categories depending on their effect on the original data [29]:

- *Perturbative.* The microdata set is distorted before publication. In this way, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset; such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset.
- *Non-perturbative.* Non-perturbative methods do not alter data; rather, they produce partial suppressions or reductions of detail in the original dataset. Global recoding, local suppression and

sampling are examples of non-perturbative masking.

At a first glance, synthetic data seem to have the philosophical advantage of circumventing the re-identification problem: since published records are invented and do not derive from any original record, some authors claim that no individual having supplied original data can complain from having been re-identified. At a closer look, some authors (*e.g.*, [31] and [23]) claim that even synthetic data might contain some records that allow for re-identification of confidential information. In short, synthetic data overfitted to original data might lead to disclosure just as original data would. On the other hand, a clear problem of synthetic data is data utility: only the statistical properties explicitly selected by the data protector are preserved, which leads to the question whether the data protector should not directly publish the statistics he wants preserved rather than a synthetic microdata set.

So far in this section, we have classified microdata protection methods by their operating principle. If we consider the type of data on which they can be used, a different dichotomic classification applies:

- *Continuous.* An attribute is considered continuous if it is numerical and arithmetic operations can be performed with it. Examples are income and age. Note that a numerical attribute does not necessarily have an infinite range, as is the case for age. When designing methods to protect continuous data, one has the advantage that arithmetic operations are possible, and the drawback that every combination of numerical values in the original dataset is likely to be unique, which leads to disclosure if no action is taken.

- *Categorical.* An attribute is considered categorical when it takes values over a finite set and standard arithmetic operations do not make sense. Ordinal and nominal scales can be distinguished among categorical attributes. In ordinal scales the order between values is relevant, whereas in nominal scales it is not. In the former case, max and min operations are meaningful while in the latter case only pairwise comparison is possible. The instruction level is an example of ordinal attribute, whereas eye color is an example of nominal attribute. In fact, all quasi-identifiers in a microdata set are normally categorical nominal. When designing methods to protect categorical data, the inability to perform arithmetic operations is certainly inconvenient, but the finiteness of the value range is one property that can be successfully exploited.

3 BACKGROUND ON MICROAGGREGATION

Microaggregation is a family of perturbative SDC techniques for continous microdata. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if records correspond to groups of k or more individuals, where no individual dominates (*i.e.* contributes too much to) the group and k is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation.

To obtain microaggregates in a microdata set with n records, these are combined to form g groups of size at least k. For each attribute, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting (modified) records can be published.

Consider a microdata set with p continuous attributes and n records (*i.e.* the result of recording p attributes on n individuals). A particular record can be viewed as an instance of $\mathbf{X}' = (X_1, \cdots, X_p)$, where the X_i are the attributes. With these individuals, g groups are formed with n_i individuals in the i-th group ($n_i \geq k$ and $n = \sum_{i=1}^{g} n_i$). Denote by x_{ij} the j-th record in the i-th group; denote by $\bar{x}_i$ the average record over the i-th group, and by $\bar{x}$ the average record over the whole set of n individuals.

The optimal k-partition (from the information loss point of view) is defined to be the one that maximizes within-group homogeneity; the higher the within-group homogeneity, the lower the information loss, since microaggregation replaces values in a group by the group centroid. The sum of squares criterion is common to measure homogeneity in clustering. The within-groups sum of squares SSE is defined as

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i)$$

The lower SSE, the higher the within group homogeneity. The total sum of squares is

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})'(x_{ij} - \bar{x})$$

In terms of sums of squares, the optimal k-partition is the one that minimizes SSE.

For a microdata set consisting of p attributes, these can be microaggregated together or partitioned into

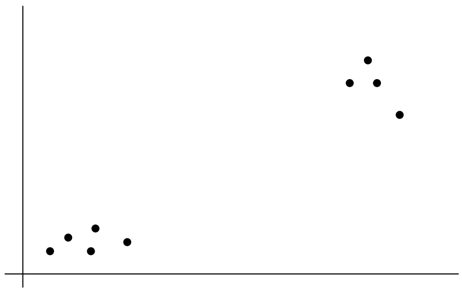

Figure 1: Variable-sized groups versus fixed-sized groups

several groups of attributes. Also the way to form groups may vary. We next review the main proposals in the literature.

4 FIXED VS VARIABLE GROUP SIZE

Classical microaggregation algorithms [4] required that all groups except perhaps one be of size k; allowing groups to be of size $\geq k$ depending on the structure of data was termed *data-oriented microaggregation* [19, 5]. Figure 1 illustrates the advantages of variable-sized groups. If classical fixed-size microaggregation with $k = 3$ is used, we obtain a partition of the data into three groups, which looks rather unnatural for the data distribution given. On the other hand, if variable-sized groups are allowed then the five data on the left can be kept in a single group and the four data on the right in another group; such a variable-size grouping yields more homogeneous groups, which implies lower information loss.

However, except for specific cases such as the one depicted in Figure 1, the small gain in within-group homogeneity obtained with variable-sized groups hardly justifies the higher computational overhead of this option with respect to fixed-sized groups. This is particularly evident for multivariate data, as noted by [24].

5 EXACT OPTIMAL VS HEURISTIC MICROAGGREGATION

For $p = 1$, *i.e.* a univariate dataset or a multivariate dataset where attributes are microaggregated one at a time, an exact polynomial shortest-path algorithm exists to find the k-partition that optimally solves the microaggregation problem [16].

For $p > 1$, finding an exact optimal solution to the microaggregation problem, *i.e.* finding a grouping where groups have maximal homogeneity and size at least k, has been shown to be NP-hard [22].

Unfortunately, the univariate optimal algorithm by [16] is not very useful in practice and this for two reasons: i) microdata sets are normally multivariate and using univariate microaggregation to microaggregate them one attribute at a time is not good in terms of disclosure risk (see [6]); ii) although polynomial-time, the optimal algorithm is quite slow when the number of records is large.

Thus, practical methods in the literature are heuristic:

- Univariate methods deal with multivariate datasets by microaggregating one attribute at a time, *i.e.* attributes are sequentially and independently microaggregated. These heuristics are known as individual ranking [4]. While they are fast and cause little information loss, these univariate heuristics have the same problem of high disclosure risk as univariate optimal microaggregation.
- Multivariate methods either rank multivariate data by projecting them onto a single axis (*e.g.* using the first principal component or the sum of z-scores, [4]) or directly deal with unprojected data [19, 5]. When working on unprojected data, we can microaggregate all attributes of the dataset at a time, or independently microaggregate groups of two attributes at a time, three attributes at a time, etc. In any case, it is preferable that attributes within a group which is microaggregated at a time be correlated [30], in order to keep as much as possible the analytic properties of the file.

To illustrate, we next give an algorithm called MDAV (Maximum Distance to Average Vector) for multivariate fixed group size microaggregation on unprojected data. We designed and implemented MDAV for the μ-Argus package [18].

Algorithm 1 (MDAV)

1. *Compute the average record $\bar{x}$ of all records in the dataset. Consider the most distant record x_r to the average record $\bar{x}$ (using the squared Euclidean distance).*
2. *Find the most distant record x_s from the record x_r considered in the previous step.*
3. *Form two groups around x_r and x_s, respectively. One group contains x_r and the $k-1$ records closest*

to x_r. The other group contains x_s and the $k-1$ records closest to x_s.

4. *If there are at least $3k$ records which do not belong to any of the two groups formed in Step 3, go to Step 1 taking as new dataset the previous dataset minus the groups formed in the last instance of Step 3.*

5. *If there are between $3k-1$ and $2k$ records which do not belong to any of the two groups formed in Step 3: a) compute the average record $\bar{x}$ of the remaining records; b) find the most distant record x_r from $\bar{x}$; c) form a group containing x_r and the $k-1$ records closest to x_r; d) form another group containing the rest of records. Exit the Algorithm.*

6. *If there are less than $2k$ records which do not belong to the groups formed in Step 3, form a new group with those records and exit the Algorithm.*

The above algorithm can be applied independently to each group of attributes resulting from partitioning the set of attributes in the dataset.

6 CATEGORICAL MICROAGGREGATION

Recently [26], microaggregation has been extended to categorical data. Such an extension is based on existing definitions for aggregation and clustering, the two basic operations required in microaggregation. Specifically, the median is used for aggregating ordinal data and the plurality rule (voting) for aggregating nominal data. Clustering of categorical data is based on the k-modes algorithm, which is a partitive clustering method similar to c-means. A version of categorical microaggregation similar based on the MDAV algorithm above is currently being developed [9].

7 CONCLUSION

Statistical disclosure control for microdata has been introduced and motivated. Microaggregation is a perturbative SDC masking approach to whose advancement our group has substantially contributed. Our enhancements and, in particular, the MDAV method, have turned microaggregation into one of the best options for the protection of continuous microdata (*e.g.* financial data). This statement is supported by empirical performance comparisons [8, 32] between microaggregation and other methods.

Acknowledgments

This work and our contributions cited herein were partly supported by the Spanish Ministry of Science and Technology and the FEDER Fund under projects no. TIC-2001-0633-C03-01 "STREAMOBILE" and TIC-2002-11942-E "REDEMAP"; by the Government of Catalonia under project 2002 SGR 00170; by the European Commission under projects IST-2000-25069 "CASC" and IST-2000-26125); and by the U. S. Census Bureau under project OBLIG-2000-29158-0-0.

References

[1] N. R. Adam and J. C. Wortmann. Security-control for statistical databases: a comparative study. *ACM Computing Surveys*, 21(4):515–556, 1989.

[2] CASC. Computational aspects of statistical confidentiality, 2004. European project IST-2000-25069 CASC, 5th FP, 2001-2004, http://neon.vb.cbs.nl/casc.

[3] F. Y. Chin and G. Ozsoyoglu. Auditing and inference control in statistical databases. *IEEE Transactions on Software Engineering*, SE-8:574–582, 1982.

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

[5] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14:1:189–201, 2002.

[6] J. Domingo-Ferrer, Josep M. Mateo-Sanz, A. Oganian, and À. Torres. On the security of microaggregation with individual ranking: analytical attacks. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10: 5:477–492, 2002.

[7] J. Domingo-Ferrer and V. Torra. Disclosure protection methods and information loss for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 91–110, Amsterdam, 2001. North-Holland.

[8] J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory*

and Practical Applications for Statistical Agencies, pages 111–134, Amsterdam, 2001. North-Holland.

[9] J. Domingo-Ferrer and V. Torra, Ordinal, continuous and heterogeneous k-anonymity through microaggregation, 2004 (manuscript).

[10] G. T. Duncan and S. Mukherjee. Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. *Journal of the American Statistical Association*, 95:720–729, 2000.

[11] E.U.Privacy. European privacy regulations, 2004. http://europa.eu.int/comm/internal_market/-privacy/law_en.htm.

[12] R. Garfinkel, R. Gopal, and D. Rice. New approaches to disclosure limitation while answering queries to a database: protecting numerical confidential data against insider threat based on data and algorithms, 2004. Manuscript. Available at http://www-eio.upc.es/seminar/04/garfinkel.pdf.

[13] S. Giessing. Survey on methods for tabular data protection in argus. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *LNCS*, pages 1–13, Berlin Heidelberg, 2004. Springer.

[14] R. Gopal, R. Garfinkel, and P. Goes. Confidentiality via camouflage: the cvc approach to disclosure limitation when answering queries to databases. *Operations Research*, 50:501–516, 2002.

[15] R. Gopal, P. Goes, and R. Garfinkel. Interval protection of confidential information in a database. *INFORMS Journal on Computing*, 10:309–322, 1998.

[16] S. L. Hansen and S. Mukherjee. A polynomial algorithm for optimal univariate microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 15:4:1043–1044, 2003.

[17] HIPAA. Health insurance portability and accountability act, 2004. http://www.hhs.gov/ocr/hipaa/.

[18] A. Hundepool, A. Van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing. *μ-ARGUS version 3.2 Software and User's Manual.* Statistics Netherlands, Voorburg NL, feb 2003. http://neon.vb.cbs.nl/casc.

[19] J. M. Mateo-Sanz and J. Domingo-Ferrer. A method for data-oriented multivariate microaggregation. In J. Domingo-Ferrer, editor, *Statistical Data Protection*, pages 89–99, Luxemburg, 1999. Office for Official Publications of the European Communities.

[20] R. Moore. Controlled data swapping techniques for masking public use microdata sets, 1996. U. S. Bureau of the Census, Washington, DC, (unpublished manuscript).

[21] K. Muralidhar, D. Batra, and P. J. Kirs. Accessibility, security and accuracy in statistical databases: the case for the multiplicative fixed data perturbation approach. *Management Science*, 41:1549–1564, 1995.

[22] A. Oganian and J. Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Comission for Europe*, 18:4:345–354, 2001.

[23] J. P. Reiter. Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, page forthcoming, 2004.

[24] G. Sande. Exact and approximate methods for data directed microaggregation in one or more dimensions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10:5:459–476, 2002.

[25] J. Schlörer. Disclosure from statistical databases: quantitative aspects of trackers. *ACM Transactions on Database Systems*, 5:467–492, 1980.

[26] V. Torra. Microaggregation for categorical variables: a median based approach. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *LNCS*, pages 162–174, Berlin Heidelberg, 2004. Springer.

[27] J. F. Traub, Y. Yemini, and H. Wozniakowski. The statistical security of a statistical database. *ACM Transactions on Database Systems*, 9:672–679, 1984.

[28] U.S.Privacy. U. s. privacy regulations, 2004. http://www.media-awareness.ca/english/issues-/privacy/us_legislation_privacy.cfm.

[29] L. Willenborg and T. DeWaal. *Elements of Statistical Disclosure Control.* Springer-Verlag, New York, 2001.

[30] W. E. Winkler. Masking and re-identification methods for public-use microdata: overview and research problems. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *LNCS*, pages 231–246, Berlin Heidelberg, 2004. Springer.

[31] W. E. Winkler. Re-identification methods for masked microdata. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *LNCS*, pages 216–230, Berlin Heidelberg, 2004. Springer.

[32] W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *LNCS*, pages 135–152, Berlin Heidelberg, 2002. Springer.