Advances in Inference Control in Statistical Databases: An Overview

Josep Domingo-Ferrer

Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain
jdomingo@etse.urv.es

Abstract. Inference control in statistical databases is a discipline with several other names, such as statistical disclosure control, statistical disclosure limitation, or statistical database protection. Regardless of the name used, current work in this very active field is rooted in the work that was started on statistical database protection in the 70s and 80s. Massive production of computerized statistics by government agencies combined with an increasing social importance of individual privacy has led to a renewed interest in this topic. This is an overview of the latest research advances described in this book.
Keywords: Inference control in statistical database, Statistical disclosure control, Statistical disclosure limitation, Statistical database protection, Data security, Respondents' privacy, Official statistics.

1 Introduction

The protection of confidential data is a constant issue of concern for data collectors and especially for national statistical agencies. There are legal and ethical obligations to maintain confidentiality of respondents whose answers are used for surveys or whose administrative data are used to produce statistics. But, beyond law and ethics, there are also practical reasons for data collectors to care about confidentiality: unless respondents are convinced that their privacy is being adequately protected, they are unlikely to co-operate and supply their data for statistics to be produced on them.

The rest of this book consists of seventeen articles clustered in three parts:

1. The protection of tabular data is covered by the first five articles;
2. The protection of microdata (*i.e.* invididual respondent data) is addressed by the next seven articles;
3. Software for inference control and user case studies are reported in the last five articles.

The material in this book focuses on the latest research developments of the mathematical and computational aspects of inference control and should be regarded as an update of [6]. For a systematic approach to the topic, we

J. Domingo-Ferrer (Ed.): Inference Control in Statistical Databases, LNCS 2316, pp. 1–7, 2002.

strongly recommend [13]; for a quicker overview, [12] may also be used. All references given so far in this paragraph concentrate only on the mathematical and computational side of the topic. If a broader scope is required, [7] is a work where legal, organizational, and practical issues are covered in addition to the purely computational ones.

This overview goes through the book articles and then gives an account of related literature and other sources of information.

2 Tabular Data Protection

The first article "Cell suppression: experience and theory", by Robertson and Ethier, emphasizes that some basic points of cell suppression for table protection are not sufficiently known. While the underlying theory is well developed, sensitivity rules in use are in some cases flawed and may lead to the release of sensitive information. Another issue raised by the paper is the lack of a sound information loss measure to assess the damage inflicted to a table in terms of data utility by the use of a particular suppression pattern. The adoption of information-theoretic measures is hinted as a possible improvement.

The article "Bounds on entries in 3-dimensional contingency tables subject to given marginal totals" by Cox deals with algorithms for determining integer bounds on suppressed entries of multi-dimensional contingency tables subject to fixed marginal totals. Some heuristic algorithms are compared, and it is demonstrated that they are not exact. Consequences for statistical database query systems are discussed.

"Extending cell suppression to protect tabular data against several attackers", by Salazar, points out that attackers to confidentiality need not be just external intruders; internal attackers, *i.e.* special respondents contributing to different cell values of the table, must also be taken into account. This article describes three mathematical models for the problem of finding a cell suppression pattern minimizing information loss while ensuring protection for different sensitive cells and different intruders.

When a set of sensitive cells are suppressed from a table (primary suppressions), a set of non-sensitive cells must be suppressed as well (complementary suppressions) to prevent primary suppressions from being computable from marginal constraints. Network flows heuristics have been proposed in the past for finding the minimal complementary cell suppression pattern in tabular data protection. However, the heuristics known so far are only appropriate for two-dimensional tables. In "Network flows heuristics for complementary cell suppression: an empirical evaluation and extensions", by Castro, it is shown that network flows heuristics (namely multicommodity network flows and network flows with side constraints) can also be used to model three-dimensional, hierarchical, and linked tables.

Also related to hierarchical tables is the last article on tabular data, authored by De Wolf and entitled "HiTaS: a heuristic approach to cell suppression in hierarchical tables". A heuristic top-down approach is presented to find suppression

patterns in hierarchical tables. When a table of high level is protected using cell suppression, its interior is regarded as the marginals of possibly several lower level tables, each of which is protected while keeping their marginals fixed.

3 Microdata Protection

The first three articles in this part describe methods for microdata protection:

- Article "Model based disclosure protection", by Polettini, Franconi, and Stander, argues that any microdata protection method is based on a formal reference model. Depending on the number of restrictions imposed, methods are classified as nonparametric, semiparametric or fully parametric. An imputation procedure for business microdata based on a regression model is applied to the Italian sample from the Community Innovation Survey. The utility of the released data and the protection achieved are also evaluated.
- Adding noise is a very used principle for microdata protection. In fact, results in the article by Yancey *et al.* (discussed below) show that noise addition methods can perform very well. Article "Microdata protection through noise addition", by Brand, contains an overview of noise addition algorithms. These range from simple white noise addition to complex methods which try to improve the tradeoff between data utility and data protection. Theoretical properties of the presented algorithms are discussed in Brand's article and an illustrative numerical example is given.
- Synthetic microdata generation is an attractive alternative to protection methods based on perturbing original microdata. The conceptual advantage is that, even if a record in the released data set can be linked to a record in the original data set, such a linkage is not actually a re-identification because the released record is a synthetic one and was not derived from any specific respondent. In "Sensitive microdata protection using Latin hypercube sampling technique", Dandekar, Cohen, and Kirkendall propose a method for synthetic microdata generation based on Latin hypercube sampling.

The last four articles in this part concentrate on assessing disclosure risk and information loss achieved by microdata protection methods:

- Article "Integrating file and record level disclosure risk assessment", by Elliot, deals with disclosure risk in non-perturbative microdata protection. Two methods for assessing disclosure risk at the record-level are described, one based on the special uniques method and the other on data intrusion simulation. Proposals to integrate both methods with file level risk measures are also presented.
- Article "Disclosure risk assessment in perturbative microdata protection", by Yancey, Winkler, and Creecy, presents empirical re-identification results that compare methods for microdata protection including rank swapping and additive noise. Enhanced re-identification methods based on probabilistic record linkage are used to empirically assess disclosure risk. Then the

performance of methods is measured in terms of information loss and disclosure risk. The reported results extend earlier work by Domingo-Ferrer *et al.* presented in [7].
- In "LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection", Dandekar, Domingo-Ferrer, and Sebé report on another comparison of methods for microdata protection. Specifically, hybrid microdata generation as a mixture of original data and synthetic microdata is compared with rank swapping and microaggregation, which had been identified as the best performers in earlier work. Like in the previous article, the comparison considers information loss and disclosure risk, and the latter is empirically assessed using record linkage.
- Based on the metrics previously proposed to compare microdata protection methods (also called masking methods) in terms of information loss and disclosure risk, article "Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets", by Sebé, Domingo-Ferrer, Mateo-Sanz, and Torra, demonstrates how to improve the performance of any microdata masking method. Post-masking optimization of the metrics can be used to have the released data set preserve as much as possible the moments of first and second order (and thus multivariate statistics) of the original data without increasing disclosure risk. The technique presented can also be used for synthetic microdata generation and can be extended to preserve all moments up to m-th order, for any m.

4 Software and User Case Studies

The first two articles in this part are related to software developments for the protection of statistical data:

- "The CASC project", by Hundepool, is an overview of the European project CASC (Computational Aspects of Statistical Confidentiality,[2]), funded by the EU 5th Framework Program. CASC can be regarded as a follow-up of the SDC project carried out under the EU 4th Framework Program. The central aim of the CASC project is to produce a new version of the Argus software for statistical disclosure control. In order to reach this practical goal, the project also includes methodological research both in tabular data and microdata protection; the research results obtained will constitute the core of the Argus improvement. Software testing by users is an important part of CASC as well.
- The first sections of the article "Tools and strategies to protect multiple tables with the GHQUAR cell suppression engine", by Gießing and Repsilber, are an introduction to the GHQUAR software for tabular data protection. The last sections of this article describe GHMITER, which is a software procedure allowing use of GHQUAR to protect sets of multiple linked tables. This software constitutes a very fast solution to protect complex sets of big tables and will be integrated in the new version of Argus developed under the CASC project.

This last part of the book concludes with three articles presenting user case studies in statistical inference control:

- "SDC in the 2000 U. S. Decennial Census", by Zayatz, describes statistical disclosure control techniques to be used for all products resulting from the 2000 U. S. Decennial Census. The discussion covers techniques for tabular data, public microdata files, and on-line query systems for tables. For tabular data, algorithms used are improvements of those used for the 1990 Decennial Census. Algorithms for public-use microdata are new in many cases and will result in less detail than was published in previous censuses. On-line table query is a new service, so the disclosure control algorithms used there are completely new ones.
- "Applications of statistical disclosure control at Statistics Netherlands", by Schulte Nordholt, reports on how Statistics Netherlands meets the requirements of statistical data protection and user service. Most users are satisfied with data protected using the Argus software: τ-Argus is used to produce safe tabular data, while μ-Argus yields publishable safe microdata. However, some researchers need more information than is released in the safe data sets output by Argus and are willing to sign the proper non-disclosure agreements. For such researchers, on-site access to unprotected data is offered by Statistics Netherlands in two secure centers.
- The last article "Empirical evidences on protecting population uniqueness at Idescat", by Urrutia and Ripoll, presents the process of disclosure control applied by Statistics Catalonia to microdata samples from census and surveys with some population uniques. Such process has been in use since 1995, and has been implemented with μ-Argus since it first became available.

5 Related Literature and Information Sources

In addition to the above referenced books [6,7,12,13], a number of other sources of information on current research in statistical inference control are available. In fact, since statistical database protection is a rapidly evolving field, the use of books should be directed to acquiring general insight on concepts and ideas, but conference proceedings, research surveys, and journal articles remain essential to gain up-to-date detailed knowledge on particular techniques and open issues.

This section contains a non-exhaustive list of research references, sorted from a historical point of view:

1970s and 1980s. The first broadly known papers and books on statistical database protection appear (*e.g.* [1,3,4,5,11]).

1990s. Eurostat produces a compendium for practitioners [10] and sponsors a number of conferences on the topic, namely the three *International Seminars on Statistical Confidentiality* (Dublin 1992 [9], Luxemburg 1994 [14], and Bled 1996 [8]) and the *Statistical Data Protection'98* conference (Lisbon 1998,[6]). While the first three events covered mathematical, legal, and organizational aspects, the Lisbon conference focused on the statistical, mathematical, and computational aspects of statistical disclosure control and data

protection. The goals of those conferences were to promote research and interaction between scientists and practitioners in order to consolidate statistical disclosure control as a high-quality research discipline encompassing statistics, operations research, and computer science. In the second half of the 90s, the research project SDC was carried out under the EU 4th Framework Program; its most visible result was the first version of the Argus software. In the late 90s, other European organizations start joining the European Commission in fostering research in this field. A first example is Statistisches Bundesamt which organized in 1997 a conference for the German-speaking community. A second example is the United Nations Economic Commission for Europe, which has jointly organized with Eurostat two *Work Sessions on Statistical Data Confidentiality* (Thessaloniki 1999 [15] and Skopje 2001). Outside Europe, the U.S. Bureau of the Census and Statistics Canada have devoted considerable attention to statistical disclosure control in their conferences and symposia. In fact, well-known general conferences such as *COMPSTAT*, *U.S. Bureau of the Census Annual Research Conferences*, Eurostat's *ETK-NTTS* conference series, *IEEE Symposium on Security and Privacy*, etc. have hosted sessions and papers on statistical disclosure control.

2000s. In addition to the biennial *Work Sessions on Statistical Data Confidentiality* organized by UNECE and Eurostat, other research activities are being promoted by the U.S. Census Bureau, which sponsored the book [7], by the European projects CASC [2], and AMRADS (a co-sponsor of the seminar which originated this book).

As far as journals are concerned, there is not yet a monographic journal on statistical database protection. However, at least the following journals occasionally contain papers on this topic: *Research in Official Statistics*, *Statistica Neerlandica*, *Journal of Official Statistics*, *Journal of the American Statistical Association*, *ACM Transactions on Database Systems*, *IEEE Transactions on Software Engineering*, *IEEE Transactions on Knowledge and Data Engineering*, *Computers & Mathematics with Applications*, *Statistical Journal of the UNECE*, *Qüestiió* and *Netherlands Official Statistics*.

Acknowledgments

Special thanks go to the authors of this book and to the discussants of the seminar "Statistical Disclosure Control: From Theory to Practice" (L. Cox, G. Ronning, P. M. Steel, and W. Winkler). Their ideas were invaluable to write this overview, but I bear full responsibility for any inaccuracy, omission, or mistake that may remain.

References

1. N. R. Adam and J. C. Wortmann, "Security-control methods for statistical databases: A comparative study", *ACM Computing Surveys*, vol. 21, no. 4, pp. 515-556, 1989.
2. The CASC Project, `http://neon.vb.cbs.nl/rsm/casc/menu.htm`
3. T. Dalenius, "The invasion of privacy problem and statistics production. An overview", *Statistik Tidskrift*, vol. 12, pp. 213-225, 1974.
4. D. E. Denning and J. Schlörer, "A fast procedure for finding a tracker in a statistical database", *ACM Transactions on Database Systems*, vol. 5, no. 1, pp. 88-102, 1980.
5. D. E. Denning, *Cryptography and Data Security*. Reading MA: Addison-Wesley, 1982.
6. J. Domingo-Ferrer (ed.), *Statistical Data Protection*. Luxemburg: Office for Official Publications of the European Communities, 1999.
7. P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), *Confidentiality, Disclosure and Data Access*. Amsterdam: North-Holland, 2001.
8. S. Dujić and I. Tršinar (eds.), *Proceedings of the 3rd International Seminar on Statistical Confidentiality (Bled, 1996)*. Ljubljiana: Statistics Slovenia-Eurostat, 1996.
9. D. Lievesley (ed.), *Proceedings of the International Seminar on Statistical Confidentiality (Dublin, 1992)*. Luxemburg: Eurostat, 1993.
10. D. Schackis, *Manual on Disclosure Control Methods*. Luxemburg: Eurostat, 1993.
11. J. Schlörer, "Identification and retrieval of personal records from a statistical data bank", *Methods Inform. Med.*, vol. 14, no.1, pp. 7-13, 1975.
12. L. Willenborg and T. de Waal, *Statistical Disclosure Control in Practice*. New York: Springer-Verlag, 1996.
13. L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*. New York: Springer-Verlag, 2001.
14. *Proceedings of the 2nd International Seminar on Statistical Confidentiality (Luxemburg, 1994)*. Luxemburg: Eurostat, 1995.
15. *Statistical Data Confidentiality: Proc. of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality (Thessaloniki, 1999)*. Luxemburg: Eurostat, 1999.