International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems
Vol. 0, No. 0 (1993) 000—000

(Leave $1\frac{1}{2}$ inch blank space for Publisher)

TRENDS IN AGGREGATION AND SECURITY ASSESSMENT FOR INFERENCE CONTROL IN STATISTICAL DATABASES

VICENÇ TORRA
Institut d'Investigació en Intel·ligència Artificial
Campus de Bellaterra, E-08193 Bellaterra, Catalonia
vtorra@iiia.csic.es

JOSEP DOMINGO-FERRER
Universitat Rovira i Virgili, Dept. of Computer Eng. and Maths
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
jdomingo@etse.urv.es

Received (received date)
Revised (revised date)

As e-commerce and Internet-based data handling become pervasive, companies and statistical agencies have the need to exploit the data they accumulate without violating citizens' privacy. Inference control is a discipline whose goal is to prevent published/exchanged data from being linked with the individual respondents they originated from. This special issue illustrates that inference control largely draws on soft computing and artificial intelligence techniques.

Keywords: Disclosure risk, Inference control, Statistical disclosure control, Statistical databases, Aggregation, Uncertainty Models, Clustering.

1. Introduction

E-commerce companies and statistical agencies have the need to take advantage of the huge amount of information they collect. The former would like to exchange customer information and the latter are required to provide the society with accurate statistical information, both as summaries (tables) and individual data (microdata). However, exploiting information related to individuals or customers (respondents) should not result in a loss of privacy. For statistical agencies (and also for private companies in some countries), there is a legal obligation to protect respondent confidentiality. Beyond legal obligations, any data collecting entity should create trust among respondents; otherwise, these will become more and more reluctant to supply any information, let alone accurate information.

Inference control, also called statistical disclosure control, is a discipline whose goal is to allow dissemination/transfer of respondent data while preserving respondent privacy. To that end, inference control techniques transform an original dataset

into protected dataset such that: i) analyses on the original and protected datasets yield similar results (data utility); 2) information in the protected dataset is unlikely to be linkable to the particular respondent it originated from (data safety).

Computer science and, more specifically, artificial intelligence offer plenty of tools which are of great use for inference control (*e.g.* uncertainty models, clustering, information fusion, optimization, data mining methods). Articles in this issue focus on three main topics:

Aggregation Aggregating original data is one of the principles used by several inference control techniques to obtain protected data.

Sensitivity assessment For inference control methods to efficiently protect data, the original data set must be analyzed to establish how risky or sensitive (in terms of potential disclosure) is the publication of a particular piece of data.

Re-identification Re-identification happens when the respondent corresponding to published data is identified by an intruder. Data mining can be used on protected data to attempt re-identification and perform empirical disclosure risk assessment.

2. Aggregation

Microaggregation is an inference control method for microdata which relies on aggregating information. In microaggregation, records in the original dataset are partitioned into small groups of at least k records, in such a way that within group homogeneity is maximal. Records in the protected dataset are obtained as summaries of groups by using an aggregation operator on each group. Depending on the type or scale of data (numerical, ordinal, nominal), different operators can be used, *e.g.* average, median, OWA [9], etc. See [7,3] for details.

The first article "Exact and Approximate Methods for Data Directed Microaggregation", by Sande, proposes two approaches to univariate and multivariate microaggregation for inference control. The interesting point about this paper is that within group homogeneity is defined as the inverse of within group range, rather than as the inverse of within group variance. The author argues that, for skewed data such as business data, the group range is a better measure of group spread than the group variance.

The article "On the Security of Microaggregation with Individual Ranking: Analytical Attacks", by Domingo-Ferrer, Mateo-Sanz, Oganian and Torres, deals with the safety of individual ranking microaggregation, a special form of microaggregation very used by statistical offices. Individual ranking is attractive because it greatly preserves the analytical properties of the original data set; however, the paper analytically shows that a dataset protected in this way can provide intruders with very narrow interval estimates on the original data. This confirms recent

empirical results about the unsafety of individual ranking microaggregation [2].

3. Sensitivity assessment

Sensitivity assessment attempts to determine which information in the original dataset is more risky or sensitive, *i.e.* could be easily linked with the respondent who supplied it [4]. Sensitivity analysis is used to govern inference control techniques, which should provide a degree of protection proportional to the sensitivity of data.

"A Computational Algorithm for Handling the Special Uniques Problem", by Elliot, Manning and Ford, presents the new SUDA algorithm, which locates risky records in a categorical microdata set by first identifying all unique attribute sets within each record and then grading the risk of each record by considering the number and distribution of unique attribute sets within each record.

"Modelling User Uncertainty for Disclosure Risk and Data Utility", by Trottini and Fienberg, presents a general model for capturing user uncertainty when seeing protected statistical information. The authors distinguish the intruder's uncertainty and the scientist's uncertainty, and both uncertainties are subsequently used to define measures for data utility and disclosure risk (data safety).

Two classes of software systems for releasing tabular summaries of an underlying database are described in "Software Systems for Tabular Data Releases", by Dobra, Karr, Sanil and Fienberg. Both system classes rely on sensitivity assessment to decide which information can be released at a given moment. The first class are table servers, which respond to user queries for marginal tables of the full table that represents the entire database; sensitivity assessment in table servers is dynamically performed and depends on the past history of answered queries. The second class are optimal tabular releases, which are static releases of sets of sub-tables and try to maximize the amount of information released, subject to a constraint on disclosure risk.

In "A Critique of the Sensitivity Rules Usually Employed for Statistical Table Protection", by Domingo-Ferrer and Torra, general counterexamples are constructed which show that the most commonly used sensitivity rules for table protection (dominance and p% rules) fail to adequately capture disclosure risk when coalitions of cell contributors behave as intruders. A cell declared non-sensitive by those rules can actually imply higher disclosure risk than a cell declared sensitive. An alternative entropy-based sensitivity rule is proposed.

4. Re-identification

Re-identification happens when the individual respondent corresponding to some published protected data can be identified. Record linkage methods attempt re-identification by trying to link records in different datasets that correspond to the same individuals. Depending on the underlying assumptions about the data, record linkage methods in the literature can be classified as probabilistic [8], distance-based, fuzzy matching-based [6]... When applied to link original data and protected data,

record linkage methods can be used for empirical disclosure risk assessment. Note that the main difference with sensitivity assessment is that re-identification operates *a posteriori*, *i.e.* takes both the original and the protected datasets into account, whereas sensitivity assessment is an *a priori* approach, which is based only on the original dataset and provides an input to inference control methods.

"k-Anonymity", by Sweeney, reviews the difficulties of releasing sensitive data focusing on re-identification issues. To avoid re-identification, the article introduces the k-anonymity formal protection model. The idea is that released data are called k-anonymity compliant when the data of an individual cannot be distinguished from at least $k-1$ other individuals in the same database. Nevertheless, the computation of the optimal k-anonymization (with respect to a proposed distortion measure) turns out to be computationally intractable.

In "Preferred Minimal Generalization Algorithm", Sweeney introduces an algorithm that is k-anonymity compliant but computationally efficient. The price paid for becoming computationally efficient is that the resulting data are not optimally protected but overprotected. The article compares the results of the proposed system with those of earlier versions of μ-Argus [5], which were found to underprotect data.

The last article of the issue is "Re-Identifying Register Data by Survey Data Using Cluster Analysis: An Empirical Study", by Bacher, Brand and Bender. A novel approach for re-identification based on cluster analysis is introduced. The new method is applied to assessing the safety provided by inference control methods based on sampling (where the protected dataset is a sample of the records in the original dataset). Empirical results confirm that the number of re-identifiable records increases with the sampling fraction and decreases as the number of irrelevant variables increases (the latter are random variables added to the datasets to test the new method).

5. Concluding remarks

The purpose of this issue is to bridge the gap between the soft computing and AI community and the statistical disclosure control community. We hope that the articles in the issue will clearly show that the gap is very narrow because a lot of methodological connections exist. For example, data mining techniques are used for re-identification and safety assessment of inference control; aggregation and clustering are used for building inference control methods; uncertainty models are useful in sensitivity assessment. Further reference material on inference control can be found in [7,41].

Acknowledgments

We would like to thank the authors for their work and Prof. Bernadette Bouchon-Meunier for giving us the opportunity to offer this special issue.

References

1. D. E. Denning, *Cryptography and Data Security*. Reading MA: Addison-Wesley, 1982.
2. J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure and Data Access* (eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz). Amsterdam: North-Holland, pp. 111-133, 2001.
3. J. Domingo-Ferrer and V. Torra, "Aggregation techniques for statistical confidentiality", in *Aggregation Operators: New Trends and Applications* (eds. T. Calvo, G. Mayor and R. Mesiar). Berlin: Physica-Verlag, pp. 260-271, 2002.
4. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz (eds.), *Confidentiality, Disclosure and Data Access*. Amsterdam: North-Holland, 2001.
5. A. Hundepool, L. Willenborg, A. Wessels, L. van Gemerden, S. Tiourine and C. Hurkens, *μ-Argus 3.0 User's Manual*. Voorburg: Statistics Netherlands, 1998.
6. http://www.integrity.com
7. L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*. New York: Springer-Verlag, 2001.
8. W. E. Winkler, "Advanced methods for record linkage", American Statistical Association, *Proceedings of the Section on Survey Methods*, pp. 467-472, 1995.
9. R. R. Yager, "On ordered weighted averaging aggregation operators in multi-criteria decision making", *IEEE Trans. on Systems, Man and Cybernetics*, vol. 18, pp. 183-190, 1988.