Introduction to the special issue on Privacy in Statistical Databases

Privacy in statistical databases is more than a need nowadays. While the demand for detailed statistical information grows steadily, so does the legal and ethical obligation to protect the right of individuals about keeping their data away from the general public access. It makes sense that respondents are willing to participate in surveys only if there is a guarantee that their individual responses will not be disclosed and they may be worried about the access to data, which is easier than ever.

Both collectors, who spend time and effort to gather individual data, and respondents, who want to preserve their private records, have concerns about the need to keep to minimal confidentiality standards. However dealing with the protection and disclosure of statistical information poses a bunch of relevant and interesting methodological problems that have open up a fruitful area of research, which has also benefited from contributions in the field of computer science.

The current issue of SORT-Statistics and Operations Research Transactions is composed by a selection of papers by authors that participated to the PSD'2010-Privacy in Statistical Databases international conference held in Corfu (Greece) in 2010. PSD'2010 was a conference sponsored and organized by the UNESCO Chair in Data Privacy and the CONSOLIDER ARES project, with proceedings published by Springer-Verlag in Lecture Notes in Computer Science. Its purpose was to attract world-wide, high-level research in statistical database privacy. The conference was a successor to PSD 2008 (Istanbul, Sep. 24-26, 2008), PSD 2006 (Rome, Dec. 13-15, 2006) and PSD 2004 (Barcelona, June 9-11, 2004), all with proceedings published by Springer in LNCS 5262, LNCS 4302 and LNCS 3050, respectively. Those four PSD conferences follow a tradition of high-quality technical conferences on SDC which started with "Statistical Data Protection-SDP'98", held in Lisbon in 1998 and with proceedings published by OPOCE, and continued with the AMRADS project SDC Workshop, held in Luxemburg in 2001 and with proceedings published in Springer LNCS 2316.

The editorial committee of SORT-Statistics and Operations research Transactions has had the pleasure to invite Josep Domingo-Ferrer and Vicenç Torra to become guest editors for this special issue of this journal.

The selected articles are an extended version of those presented to the PSD'2010 conference and differ in more than 25% to the original content published in the proceedings. All featured articles have undergone the usual blind referee process and were handled by the invited editors of this special ssue.

The first article is by Philipp Bleninger, Jörg Drechsler and Gerd Ronning. It is a very interesting piece of research entitled "Remote Data Access and the Risk of Disclosure from Linear Regression". Here the authors point out to the risk that an intruder who makes educated queries to a database can disclose sensitive individual information using a simple linear regression. The results are of interest to agencies who are about to implement security barriers and it can also be useful to determine which type of queries should be allowed.

"Coprivacy: An Introduction to the Theory and Applications of Co-operative Privacy" is the topic addressed by Josep Domingo-Ferrer in his article. He presents the concept of coprivacy or co-operative privacy to make privacy preservation attractive. After a brilliant discussion of the new theory, concepts are illustrated in P2P anonymous keyword search, in content privacy in social networks, in vehicular network communications and in controlled content distribution and digital oblivion enforcement.

Arnau Erola, Jordi Castellà-Roca, Guillermo Navarro-Arribas and Vicenç Torra present an article about "Using the Open Directory Project to protect query logs with semantic microaggregation". In their work they focus on the anonymization of web search logs and indicate that existing classical methods can pose a problem of loss of utility of those logs. Their notable contribution is based on methods that are typical for statistical disclosure control, which improve data usefulness when compared to other alternatives.

The next contribution is made by Sarah Giessing and Jörg Höhne on "Eliminating Small Cells From Census Counts Tables: Empirical vs. Design Transition Probabilities". These authors present a splendid analysis the software SAFE, that has been used in the State Statistical Institute Berlin-Brandenburg already for several years. The authors compare empirically observed transition probabilities that arise once the protection algorithm is implemented to transition matrices in the context of variants of micro-data key based post-tabular random perturbation methods that have been proposed in the literature.

Jason Lucero, Michael Freiman, Lisa Singh, Jiashen You, Michael DePersio and Laura Zayatz present an article on "The Microdata Analysis System

at the U.S. Census Bureau". They show the features of a system that is under development, which will allow users to receive certain statistical analysis of Census Bureau data, such as cross-tabulations and regressions, without ever having access to the data themselves. Such analyses must satisfy several statistical confidentiality rules (including the requirement to remove some observations before the analysis is performed) and those that fail these rules will not be output to the user. Approaches to creating a system of this sort, evaluation of its effectiveness and some directions for future research are discussed in this exceptional contribution.

Finally, Anna Oganian also makes a remarkable contribution on "Multiplicative Noise for Masking Numerical Microdata with Constraints". In her paper, she presents several multiplicative noise masking schemes that are applied by statistical agencies under the form of Statistical Disclosure Limitation (SDL) methods, which are applicable to microdata (i.e. collection of individual records) and are often called masking methods. The new schemes that are proposed by Anna Oganian are designed to preserve positivity and inequality constraints in the data together with means and covariance matrix.

I wish you enjoy the reading. Besides the topics represented in this selection of excellent articles, there is much more to be done. The methodological challenges and the increasing concern in our society about the need to protect privacy are obvious, and there is no doubt that our audience of practitioners and academics is waiting for deeper insights.

Let me finish with a sincere acknowledgment to the invited editors Josep Domingo-Ferrer and Vicenç Torra for producing this remarkable special issue of SORT-Statistics and Operations Research Transactions, and for giving our readers the opportunity to plunge into the knowledge of privacy in statistical databases.

Montserrat Guillén
Chief Editor
Barcelona, March 24th, 2011.