Discussion on New Data Release Techniques

Josep Domingo-Ferrer

Dept. of Computer Engineering and Maths (ETSE), Universitat Rovira i Virgili
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
e-mail jdomingo@etse.urv.es, http://www.etse.urv.es/~jdomingo

Abstract. An overview of papers submitted to the *3rd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality* on Topic 2 (New data release techniques) is given. A list of key issues that were at the core of the discussion for those papers is also given.

Keywords: Statistical disclosure control (SDC), Remote access systems, SDC for remote access, SDC threats resulting from mass e-access.

1 Introduction

There were four invited papers and two contributed papers presented on Topic 2 (New data release techniques). Authors of those papers came from seven different countries. According to the presented contents, three groups of papers can be distinguished in this topic:

- *Technology for remote access systems*. This includes the paper by T. Desai, the paper by J. Coder and M. Cigrang, and the paper by O. Andersen.
- *SDC for remote access* Papers in this group are by N. Shlomo and by L. Franconi and G. Merola.
- *SDC threats resulting from mass e-access* The paper by V. Torra, J. Domingo-Ferrer and À. Torres is the only one in this topic.

Section 2 below summarizes the contents of papers on technology for remote access systems and lists the related key issues for discussion. Section 3 is a tour to the contents of papers on SDC for remote access and lists related discussion issues. Section 4 does the same for the paper on SDC threats resulting from mass e-access. Some conclusions are listed in Section 5.

2 Technology for remote access systems

Desai in her paper identifies the criteria for choosing a remote access system (speed, familiarity, flexibility, graphics, cost, security). The paper also lists the issues related to supporting a remote access system (human resources, networking and partnerships).

Coder and Cigrang describe the technology of the LISSY Remote Access System, originally developed for the Luxembourg Income Study project. LISSY provides easy remote access to datasets by accepting requests submitted as e-mail messages. SAS, SPSS and STATA code can be used.

Andersen gives an account of the (r)evolution of the Danish system for access to microdata, from on-site to remote data access. The paper gives some detail on the technological aspects of both the on-site and the remote access systems in use.

2.1 Key questions on technology for remote access systems

Desai's paper contains a very interesting overview of the evolution of remote access systems. Flexibility is indeed a difficult requirement to meet. The key issue is how to prevent confidential analyses, because the assumption that academics have neither the inclination nor the time to identify individuals may indeed be too optimistic. In addition, there are users which are pseudo- or non-academic, and these may have interests other than science. Another issue that appeared during the discussion of that paper was related to the internal operation of the technique for preventing confidential analyses based on "blocking at source".

Coder and Cigrang's paper describes the LISSY system for remote access from the technological standpoint. A question that was raised during the discussion was whether manual prevention of confidential analyses (such as the one offered by LISSY) was sufficient, practical and safe.

Similar remarks to those reported for the LISSY paper were made in connection with the Danish system described in Andersen's paper.

3 SDC for remote access systems

The paper by Shlomo reports on work done at CBS-Israel on SDC for remote access to microdata. R-U maps are used to compare SDC methods. Different methods are proposed to measure the disclosure risk and the information loss. Rather than using record linkage, analytical measures are used to estimate the expected number of correct matches.

Franconi and Merola give in their paper a thorough account of SDC issues related to the release of tabular data through the Web. Problems peculiar to SDC in Web-based Systems for Data Dissemination (WSDD) are identified. WSDD are

understood as sites allowing users to query tables at their choice. The approaches examined are:

- Source data perturbation
- Output perturbation
- Query set restriction

Each approach above can be applied before the query is submitted (PRE) or after that (POST).

3.1 Key questions on SDC for remote access

In Shlomo's paper, an estimate of global risk is computed for microdata protected using non-perturbative methods (sampling, collapsing, etc.). An interesting line of work would be to investigate analytical risk measures for perturbative masking in order to avoid to the extent possible the burden of empirical record linkage. Some work has already been accomplished for particular perturbative methods (*e.g.* the MASSC method for categorical data) but a generalization is not straightforward.

Franconi and Merola offer in their paper a very detailed and comprehensive analysis of SDC for remote access. Some issues about it that emerged during the discussion were:

- The authors conclude that the choice of the release policy depends on the data to be released. A general recommendation (rule of thumb) about when to choose a PRE or a POST approach was felt to be very useful.
- A necessary assumption for POST SDC to be safe seems to be that users/intruders do not co-operate. The question arises on when is such an assumption reasonable. The very fact that POST SDC requires that kind of assumption would seem to suggest that PRE SDC is to be preferred (?).

4 SDC threats resulting from mass e-access

Torra, Domingo-Ferrer and Torres analyze in their paper the SDC issues associated with the mass release of data in electronic format. Indeed, multiple database data mining is within reach of an increasing number of intruders thanks to electronic dissemination of both administrative and statistical datasets:

- The steps of data mining are discussed, with a focus on data pre-processing and model estimation.
- Data mining in SDC is discussed, with a focus on record linkage across databases. Record linkage aims at re-identification and can be carried out even if the databases do not share any variables.

4.1 Key questions on SDC threats resulting from mass e-access

Mass e-access results in users being able to link several data sources. This raises the following related issues for discussion:

- Given the enormous amount of information an intruder can access (on her own via Web or through co-operation with other intruders), a guidance is definitely needed for deciding what are the disclosure scenarios that should be considered when empirically computing disclosure risk via record linkage.

- Another hot topic is how to include disclosure risk computation in a general SDC package such as μ-Argus. In particular, identification is needed of the input on disclosure scenarios to be considered that can be reasonably requested from a standard user.

5 Conclusions

Remote access systems should evolve towards automating the prevention of confidential analyses (SDC for remote access). Manual prevention is hard and qualified labor for it is scarce.

At the same time, best practice recommendations on SDC for remote access would be very useful if available.

Regarding disclosure risk for microdata, record linkage is still the only general tool for computing such a risk for a broad range of protection techniques. However, record linkage is costly in terms of computation and requires skilled users that can specify realistic disclosure scenarios. Therefore, in a similar way as analytical disclosure risk estimation has been developed for non-perturbative methods, it would be most interesting to be able to assess disclosure risk for perturbative methods without resorting to record linkage. Holy Grail?

References

O. Andersen, "From on-site to remote data access-The revolution of the Danish system for access to microdata", in *3rd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxemburg, April 2003. See this volume.

J. Coder and M. Cigrang, "LISSY remote access system", ibidem.

T. Desai, "Providing remote access to data: the academic perspective", ibidem.

L. Franconi and G. Merola, "Implementing statistical disclosure control for aggregated data released via remote access", ibidem.

N. Shlomo, "Accessing microdata via the Internet", ibidem.

A. C. Singh, F. Yu and G. H. Dunteman, "MASSC: A new data mask for limiting statistical information loss and disclosure", ibidem.

V. Torra, J. Domingo-Ferrer and À. Torres, "Data mining methods for linking data coming from several sources", ibidem.