Data Confidentiality, Data Quality and Data Integration for Federal Databases

Project Highlight

Alan F. Karr
National Institute of Statistical Sciences
PO Box 14006
Research Triangle Park, NC, 27709-4006
karr@niss.org

1. OBJECTIVES AND IMPACT

The high-level goal of the research is to develop abstractions, theory and methodology and software tools that allow federal statistical agencies to disseminate useful information derived from confidential data but protect the privacy of data subjects—individuals and establishments.

Specific scientific objectives include:

- Problem formulations and scalable tools that accommodate both *disclosure risk* and *data/information utility*;
- Understanding *consequences of data integration* for data confidentiality, data quality and statistical inference;
- Creation of fundamental quantifications, usable models, scalable methods for *data quality.*

Impacts of the research include:

- Protecting government-collected data on individuals and establishments from increasingly severe threats to confidentiality;
- Protecting privacy of individuals and establishments;
- Preventing federal data warehouses from becoming data cemeteries;
- Assisting agencies in preparing for a possible "world without releasable microdata."

2. SELECTED ACCOMPLISHMENTS

The project is leading to a paradigm shift in statistical disclosure limitation (SDL). New techniques are based on scalable risk-utility formulations that enable agencies to balance disclosure protection against the utility of released information.

Geographical Aggregation. Algorithms and software for achieving disclosability by aggregating adjacent geographical units such as counties were developed using data provided by the National Agricultural Statistics Service (NASS) [7, 8, 12]. These enable release of information below the state level, which was previously thought infeasible.

Tabular Data. NISS table servers are the first successful implementation of query systems containing principled, scalable methods for dealing with query interaction. The project has also developed:

- Scalable methods and software to compute bounds on cell entries from released marginals [1, 2];
- Scalable risk-utility formulations, which lead to optimal tabular releases maximizing data utility subject to a constraint on disclosure risk [1, 2];
- Initial methods for releasing conditional distributions (rather than or in addition to marginal distributions) [3].
- Prototype table server software [6].

Data Swapping. Principal products of the research include:

- A complete risk-utility formulation for data swapping as a decision problem, with multiple measures utility/distortion and disclosure risk [5].
- The NISS Data Swapping Toolkit (DSTK): operational software for performing swapping large-scale studies to select the swap attributes and swap rate, as well as for visualization of the results. The DSTK is available at `www.niss.org/software/dstk.html`.
- A Web service implementation of data swapping [14], for which software is available at `www.niss.org/WebServices/dg/WebSwap.html`.

Remote Access Analysis Servers. The project has created fundamental abstractions—query space, answer space, disclosure risk measure and data utility measure—for systems that disseminate the results of statistical analyses of confidential data. *Regression servers* that optimally protect a sensitive variable have been developed [4], and software is being built.

Secure Analysis of Distributed Data. Techniques for secure multi-party computation have been used to implement:

- Secure data integration, protecting data sources [10, 11].

dg.o 2005 Atlanta, GA USA

- Secure construction of (sparse representations of) contingency tables whose cells contain either counts or sums [11].
- Secure linear regression on multiple, horizontally partitioned databases [10, 11]. Least squares estimators, standard errors and the coefficient of determination R^2 can be calculated in a way that minimizes risk to individual data elements. Diagnostics can be performed via by means of securely integrated synthetic residuals.
- Secure linear regression on multiple, vertically partitioned data [9, 15].

Techniques for more complex partitioning, including missing data, are currently being developed [13].

Framework for Comparing Statistical Disclosure Limitation Methods. A framework is being developed for comparison and combination of SDL methods for microdata, including

- Rank swapping;
- Jittering;
- Resampling;
- Several forms of microaggregation.

The framework incorporates multiple measures of data utility—ranging from very general to analysis-specific—as well as multiple measures of disclosure risk. This line of research is expected to lead to an extensible NISS Microdata Toolkit (MDTK).

3. PROJECT STRUCTURE

The National Institute of Statistical Sciences (NISS) is lead institution for the project. University partners Carnegie Mellon University, Duke University, Iowa State University, Pennsylvania State University, Purdue University and Southern Methodist University. Federal statistical agency partners are the Bureau of Labor Statistics (BLS), Bureau of Transportation Statistics (BTS), Census Bureau (Census), NASS and National Center for Education Statistics (NCES). All have provided both data and support.

4. ACKNOWLEDGMENTS

This research was supported by NSF grant EIA–0131884 to NISS.

5. REFERENCES

[1] A. Dobra, S. E. Fienberg, A. F. Karr, and A. P. Sanil. Software systems for tabular data releases. *Int. J. Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):529–544, 2002.

[2] A. Dobra, A. F. Karr, and A. P. Sanil. Preserving confidentiality of high-dimensional tabular data: Statistical and computational issues. *Statist. and Computing*, 13(4):363–370, 2003.

[3] S. E. Fienberg and A. B. Slavkovic. Bounds for cell entries in two-way tables given conditional frequencies. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases '2004*. Springer–Verlag, New York, 2004.

[4] S. Gomatam, A. F. Karr, J. P. Reiter, and A. P. Sanil. Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.*, 2005. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

[5] S. Gomatam, A. F. Karr, and A. P. Sanil. Data swapping as a decision problem. *J. Official Statist.*, 2003. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

[6] A. F. Karr, A. Dobra, and A. P. Sanil. Table servers protect confidentiality in tabular data releases. *Comm. ACM*, 46(1):57–58, 2003.

[7] A. F. Karr, J. Lee, A. P. Sanil, J. Hernandez, S. Karimi, and K. Litwin. Disseminating information but protecting confidentiality. *IEEE Computer*, 34(2):36–37, 2001.

[8] A. F. Karr, J. Lee, A. P. Sanil, J. Hernandez, S. Karimi, and K. Litwin. Web-based systems that disseminate information but protect confidentiality. In W. M. McIver and A. K. Elmagarmid, editors, *Advances in Digital Government: Technology, Human Factors and Public Policy*, pages 181–196. Kluwer, Amsterdam, 2002.

[9] A. F. Karr, X. Lin, J. P. Reiter, and A. P. Sanil. Privacy preserving analysis of vertically partitioned data using secure matrix products. *J. Official Statist.*, 2004. Submitted for publication. Available on-line at www.niss.org/dgii/technicalreports.html.

[10] A. F. Karr, X. Lin, J. P. Reiter, and A. P. Sanil. Secure regression on distributed databases. *J. Computational and Graphical Statist.*, 2004. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

[11] A. F. Karr, X. Lin, J. P. Reiter, and A. P. Sanil. *Secure regression on distributed databases*, pages 000–000. ASA/SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, 2005. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

[12] J. Lee, C. Holloman, A. F. Karr, and A. P. Sanil. Analysis of aggregated data in survey sampling with application to fertilizer/pesticide usage surveys. *Res. Official Statist.*, 4:101–116, 2001.

[13] J. P. Reiter, A. F. Karr, C. N. Kohnen, X. Lin, and A. P. Sanil. Secure regression for vertically partitioned, partially overlapping data. *ASA Proc.*, 2004. To appear.

[14] A. P. Sanil, S. Gomatam, A. F. Karr, and C. Liu. NISSWebSwap: A Web Service for data swapping. *J. Statist. Software*, 8(7), 2003.

[15] A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter. Privacy preserving analysis of vertically partitioned data using secure matrix products. *J. Official Statist.*, 2004. Submitted for publication. Available on-line at www.niss.org/dgii/technicalreports.html.