Journal of Privacy and Confidentiality (2009) **1**, Number 2, pp. 125–134

Data Confidentiality: The Next Five Years Summary and Guide to Papers

Satkartar K. Kinney[*] and Alan F. Karr[†] and Joe Fred Gonzalez, Jr.[‡]

1 Introduction

On May 1–2, 2008, the National Center for Health Statistics/CDC (NCHS) and the National Institute of Statistical Sciences (NISS) co-sponsored a workshop "Data Confidentiality: The Next Five Years," which was held at NCHS' headquarters in Hyattsville, MD. The purpose of the workshop was to bring together the academic and federal data confidentiality research communities, including mathematicians, mathematical statisticians, statisticians, computer scientists, cryptographers, and federal agency "owners" of data confidentiality problems in order to:

- Identify important unsolved research problems associated with data confidentiality;
- Articulate a research agenda that both addresses those problems and responds to current and emerging needs among federal statistical agencies;
- Discuss and catalyze the kinds of collaborations among mathematicians, mathematical statisticians, statisticians, computer scientists, cryptographers, domain scientists, and data-owning agencies that are needed to pursue the research agenda.

The workshop was organized by Joe Fred Gonzalez, Jr. of NCHS and Alan Karr of NISS, with assistance from Lawrence Cox and Meena Khare of NCHS and others. More than 40 academic government and industry statisticians and computer scientists attended. The workshop was composed of five thematic sessions followed by a panel discussion on federal agency needs.[1] The topics for the sessions reflect the organizers' sense of where important unsolved issues lie:

> **Query Systems**: Cynthia Dwork (Microsoft Research) and Adam Smith (Pennsylvania State University) addressed differential privacy from a computer science perspective in the context of statistical theory in their presentations *Differential Privacy: What we Know and What we Want to Learn* and *Integrating Differential Privacy and Statistical Theory*.
>
> **Weighted Data**: Avinash Singh (NORC at the University of Chicago) presented *Maintaining Analytic Quality while Protecting Confidentiality of Survey Weighted*

[*] National Institute of Statistical Sciences, mailto:saki@niss.org

[†] National Institute of Statistical Sciences, mailto:karr@niss.org

[‡] National Center for Health Statistics/CDC, mailto:jgonzalez@cdc.gov

[1] The presentations themselves are available on the NISS website at http://www.niss.org/affiliates/dc200805/program.html.

© 2009 by the authors

Data. Stephen Fienberg (Carnegie Mellon University) presented *The Relevance or Irrelevance of Weights for Confidentiality and Statistical Analyses*.

Distributed Data: Alan Karr (NISS) addressed techniques for principled statistical analysis of distributed data in *Secure Statistical Analysis of Distributed Databases, Emphasizing What We Don't Know*, and Xiaodong Lin (University of Cincinnati) discussed maximum likelihood estimation in the same setting in *Privacy Preserving Distributed Maximum Likelihood Estimation*.

Synthetic Data: John Abowd (Cornell University) presented *Synthetic Data and Randomized Sanitizers*, and Jerome Reiter (Duke University) presented a research agenda for synthetic data in *Some Next Steps in Synthetic Data Research*.

Tabular Data: Research on SDL methods for tabular data was discussed by Lawrence Cox (NCHS) in *A Data Quality and Confidentiality Assessment of Complementary Cell Suppression*, and by Alexandra Slavkovic (Pennsylvania State University) in *Partial Information Releases for Confidential Contingency Table Entries*.

Federal Agency Needs: The panelists Jacob Bournazian (EIA), Larry Ernst (BLS), and Marilyn Seastrom (NCES) responded to the papers presentations, and described their agencies' practices and needs.

This special issue of the JPC includes eight papers based on nine presentations from this workshop—those by Dwork, Smith, Singh, Fienberg, Karr, Lin, Reiter, Cox and Slavkovic. In keeping with the theme of the next five years in data confidentiality research, these papers primarily review the current status of research and recommend new directions. To help link these papers, in Section 2 we lay out a conceptual structure and taxonomy for statistical disclosure limitation (SDL) that not only facilitates relating the papers to one another but also frames the discussion that appears in Section 3.

We emphasize that the world of privacy and confidentiality is broader than the official statistics-centric perspective of the workshop and this report. In particular, there are important legal, policy, and societal issues that underlie our perspective, but are not treated here.

2 Conceptual Structure and Taxonomy for SDL

Official statistics agencies have long faced conflicting missions. On the one hand, they must collect and assemble high-quality data about individuals and establishments in a manner that protects privacy of data objects and confidentiality of their databases. On the other hand, they must disseminate information for diverse purposes, the most notable of which are policy formulation/evaluation and research.

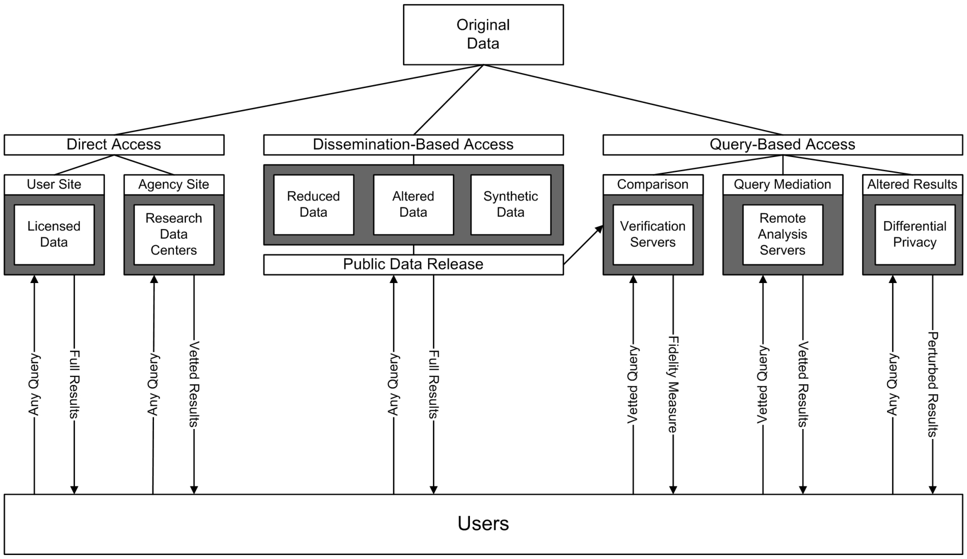
In the United States and elsewhere, statistical agencies have a long history of attention to statistical disclosure limitation (SDL)[2] (Doyle et al., 2001; Willenborg and de

[2] Otherwise known as statistical disclosure avoidance or statistical disclosure control

Waal, 2001). SDL strategies are meant to balance the inherently conflicting objectives of protecting confidentiality and disseminating information.

In a literal sense, SDL is what stands between users and the original confidential data. Figure 1 illustrates a variety of mechanisms for interaction between users and confidential data.[3] There are three major forms of interaction: direct access, dissemination-based access (public data releases), and query-based access. Direct access imposes the least interference between users and the confidential data. Dissemination-based access refers to the practice of releasing masked data in public files. In the query-based interaction mode, users cannot directly access individual data records, but are able to submit queries, either electronically or manually. A related interaction mode, not shown in Figure 1, is secure multi-party computation, where agencies holding related confidential data are able to perform a combined analysis without sharing the actual data.

Figure 1: Models for User-Data Interaction



The models for user-data interaction in Figure 1 include different levels of restriction on the queries posed and the results. Queries are either vetted or full; "Vetted Query" limits the questions the researcher may pose. Results may be full, vetted, or perturbed. "Vetted Results" typically involves forbidding users access to confidentiality-threatening items such as residuals from a fitted statistical model, while "Perturbed Results" involves distorting query results in the interest of disclosure protection.

[3] In this report, a database means a flat file in which rows represent data subjects and columns represent (numerical or categorical) attributes of those subjects.

Another high-level SDL classification is *restricted access* versus *restricting data*. The former refers to measures taken to limit the interaction between users and the confidential data, while the latter refers to masking or obscuring data for the purpose of confidentiality protection. SDL may involve both.

2.1 Direct Access

Under this mode, users interact with confidential data either at their own location or at the agency's location. Confidential data may be subject to access restriction but data restrictions tend to be low. Variables and units may be confined to those needed to conduct specific analyses and obvious identifiers are routinely excluded.

Licensed Access: Under this mode, users execute a licensing agreement with an agency, which then provides a complete, unaltered version of the data[4] that can be transferred to their own locations. Any query (ranging from a simple summary to a complex statistical analysis) can be applied to the data, using any software, and full results are obtained. Licenses may stipulate security measures and require researchers to submit publications for review. Among U.S. official statistics agencies, the National Center for Education Statistics is the most prolific issuer of licenses.[5]

Research Data Center (RDC): Users enter into an agreement with the agency to obtain access to unaltered confidential data; however, all work must be done at the agency location. For example, the U.S. Census Bureau provides access to confidential data, including datasets from other agencies under agreement, at several locations across the United States.[6] Access to an RDC requires project approval and security clearance, and outputs must be reviewed prior to removal from the RDC.

Overlapping these two types of direct access, agencies can license users to access data via secure internet connection to an agency computer. Remote desktop software can be used to allow researchers full access to data files while restricting the ability to transfer information to their local computers. The Cornell Virtual Research Data Center provides access to synthetic data in this manner.[7]

2.2 Dissemination-based Access

Dissemination-based access refers to the public release of data files that differ, perhaps dramatically, from the original database. As for direct access, any query can be posed, using any software, and full results are obtained, though these may not be identical to the results that would be obtained on the original confidential data. Public use files are often made available on the Internet; however, agencies may restrict access somewhat, for example by asking users to register and consent to terms of use. Typically one or

[4] Comparable to what the agency's internal analysts use, though data may be cleaned and small amounts of noise may be added.

[5] `http://www.nces.ed.gov/Pubsearch/licenses.asp`

[6] `http://www.ces.census.gov/index.php/ces/researchlocations`

[7] `http://www.vrdc.cornell.edu`

Figure 2: Data restriction methods and access modes

Tabular Data	Hidden Data	Microdata	
Model-based -CTA -Synthetic data	Query-Based -Differential privacy -Analysis server -Verification server -SMPC -Published Analyses	Model-based -Synthetic data -Data shuffling	**Synthethized data**
Distortion-based -Swapping -Noise addition		Distortion-based -Data swapping -Noise addition	**Altered data**
Suppression -Cell suppression Generalization -Coarsening -Marginals & conditionals		Suppression -Top coding Generalization -Microaggregation -Coarsening -Sampling	**Reduced data**
Complete tables		Direct Access -User license -RDC	**Little/no change**

more data restricting SDL methods are applied to construct the released dataset.

Figure 1 shows three qualitatively different approaches: data may be *reduced*-by dropping cases, dropping attributes or coarsening attributes, *altered*—for instance, by addition of noise to numerical attributes, or *synthesized*—by creating a surrogate for the original data that preserves important characteristics. By contrast, in query-based access the confidential data are left relatively intact but kept hidden from the user, while the query results may be reduced or altered. Figure 2 gives several examples of data restriction methods in these categories, and also relates them to direct access and query-based access. These are further sub-divided into methods for microdata (individual records) and methods for tabular summaries, reflecting the nature of SDL research.

2.3 Query-based Access

Under this mode, users interact with the data by posing queries (request for data summaries or statistical analyses), typically over a secure internet connection. Query-based access may involve a substantial amount of access restriction and data restriction.

Remote analysis servers that allow researchers to analyze confidential data without actually seeing the data are an emerging research area. Analyses may be complex linear regressions (Gomatam et al., 2005) or as simple as tabular summaries (Dobra et al., 2002, 2003). Analysis servers provide access to confidential data under the query mediation access mode. Both the queries that can be posed and the results that can be obtained may be subject to limitation.

Verification servers (Oganian et al., 2009) link dissemination-based access and query-based access. The principal shortcoming of public data releases is that analyses performed on them can differ from the same analyses performed on the original data. For instance, the magnitude or even the direction of an effect may be altered. Equally important, users of public data releases have no way of knowing whether this has happened. A user of a verification server, which is operated by the agency holding the original data, poses a query describing the analysis, and receives a measure of the fidelity of that analysis performed on the public data to the analysis performed on the original data. An example of a fidelity measure is whether there is overlap between the confidence interval computed on the confidential data and that computed on the public-use data.

Differential privacy is a research area stemming from work in the field of cryptography within computer science. In this framework the query results (analyses) are altered, often by adding noise, so that released information does not reveal any person's data with certainty. See the paper by Dwork and Smith.

2.4 Secure Multi-party Computation

Often agencies or other parties have related or overlapping confidential databases that cannot be shared but would provide better and more information if combined. Thus methods that allow unified analyses to be obtained without exchanging data are needed and of interest. In secure summation, for example, agencies combine additive sufficient statistics to obtain a combined analysis (Karr et al., 2007). Some challenges are that analyses must be agreed upon in advance, exploratory data analysis is not generally possible, and the parties involved must have some level of trust in each other. See the papers by Karr and Lin and Karr.

2.5 Data Restriction

As shown in Figure 2, methods for restricting microdata that alter data values can be divided into model-based methods that create surrogate datasets sharing statistical properties with the original data, and other distortion methods. Model-based methods include synthetic data, in which some or all of the data are replaced by multiple imputations (Raghunathan et al., 2003, Reiter paper), and data shuffling, in which values of continuous variables are swapped in a way that preserves correlations (Muralidhar and Sarathy, 2006). Other distortion methods include noise addition (Fuller, 1993) and data swapping (Dalenius and Rice, 1982), which is commonly used by U.S. agencies. Typically, data swapping entails selecting one or more variables to have a small and confidential percentage of their values swapped.

Data reduction methods include microaggregration, in which small groups of similar units have their values of confidential variables replaced by group centroids (Defays and Nanopoulos, 1992), top-coding, sampling, and coarsening or recoding of variables (Willenborg and de Waal, 2001). Common SDL methods for reducing tabular data include complementary cell suppression (Cox paper), partial contingency tables (Slavkovic paper), as well as rounding and collapsing categories. Controlled Tabular Adjustment (Cox, Orelian, and Shah, 2006) is an example of an SDL method that alters cell values.

The selection of method(s) for restricting data in given application depends on the acceptable risk, data type, user needs, and the resources available to generate and disseminate public use data. For some datasets, simple measures such as recoding variables or suppression of outliers may provide sufficient protection while still providing useful information to researchers. Other datasets may require substantial alteration of data values in order for the disclosure risk to be acceptably low. In such cases a synthetic data approach may be necessary in order to maintain acceptable data utility.

3 Research Agenda

This section summarizes research questions raised by workshop presenters, panelists, and participants. In no sense is this list comprehensive: it mirrors the views of the presenters and other participants at the workshop. Because both the academia-based "research" community and the agency-based "user" community were represented, many of the issues focus on research targeted at meeting agency needs.

Responding to a request from the workshop organizers, many of the presenters' papers list research issues specific to their contexts. These include the papers by Cox, Dwork and Smith, Karr, and Reiter. The presentations themselves also contain lists of research issues.[8] In this section, therefore, we focus on a set of cross-cutting issues.

Arguably, the most urgent need is not for development of more techniques for SDL, but for research that *provides agencies methods and tools for making sound decisions about SDL*. There is a broad consensus that, in principle, releasing data—either literally or through a set of allowable queries (and possibly incomplete responses)—is a decision problem in which each release is characterized by quantified measures of risk and utility from which a principled choice can be made in several ways.[9] But, inability to implement this paradigm remains nearly total.

In particular, construction of useful measures for utility and risk assessment is mired in a primitive state. Nearly all existing measures are either too broad and hence too blunt or too narrow (in particular, highly analysis-specific), and hence not generalizable (Karr et al., 2006, Gomatam, Karr, and Sanil, 2006). An array of new measures is clearly needed.

Actionable guidance for choosing among SDL methods (or, combining them, as in

[8] See Section 1 for the web site containing the presentations.

[9] For example, maximizing utility subject to an upper bound on acceptable risk, or choosing from candidate releases on the risk-utility frontier.

Oganian and Karr, 2006) or even choosing the parameters for a given method, do not exist. It is not known whether the choice of measures is a problem with theoretical or methodological structure or merely disconnected special cases amenable only to empirical analysis.

Possibly the risk-utility paradigm itself requires revision in order to incorporate multiple measures, or to address the problem posed sardonically but pointedly by one of us (Karr) that "One person's risk is another person's utility." More concretely, abstractions of "intruder" and "legitimate user" that unambiguously distinguish one from the other do not exist. An illustration of this is the current difficulty in dealing with transparency—the extent to which an agency cannot (risk perspective) or should (utility perspective) release information about the SDL that it has performed.

There are also risk and utility questions specific to SDL methods, for example, the relationship between risk and number of imputations in synthetic data (Reiter and Mitra, 2007) and the impact of query interaction in an analysis server (Gomatam et al., 2005). While the release of individual query results may not pose a threat, a series of queries could be used to re-identify units or attributes, and after some number of queries have been answered, there may turn out to be no more releasable queries. How, then, can an agency ensure that the "important" queries are among those released?

The structure and set of participants of the workshop responded to a growing sense that researchers in statistics and computer science have for some time been pursuing rather different approaches to data confidentiality largely in isolation from one another. That there is unrealized potential that would result from true collaboration seems apparent. For instance, statistical approaches to SDL almost exclusively entail altering data rather than query results, which facilitates (some level of) analysis of utility but often leaves risk nebulous. Differential privacy focuses on altering query results in ways that guarantee quantified levels of privacy, but there is little known about the effects on utility, or what happens for complex analyses. Nothing is known about how altering data could be traded off with altering results.

A second, and quite different point of intersection is that many current SDL strategies rely on the computational complexity of defeating them, but in ways that are poorly understood, if at all. The explosive growth of the World Wide Web created the last decade's most potent threat to data confidentiality—ready access to numerous, even if low quality, external databases that can be used to compromise confidentiality. Potentially, computing power is the next decade's big danger. To provide some concreteness, consider the following "universal" attack strategy for a public data release $\mathbf{M}$: identify the set $\mathbf{O}$ of all possible candidates $\mathbf{O}$ for the original data $\mathbf{O}^*$, and using standard Bayesian techniques, for each $\mathbf{O}$, calculate its posterior probability given all available knowledge, including $\mathbf{M}$, external databases, and what is released about the SDL. Then, choose as the estimate of $\mathbf{O}^*$ the $\mathbf{O}$ with maximum posterior probability. Now, this strategy is impossible; ten years from now, it might not be.[10]

[10] In Oganian, Reiter, and Karr (2009), a much simpler approach is shown to be able to defeat such common SDL strategies as top coding and data swapping when there is a verification server that both provides infinitely precise fidelity measures and allows arbitrary subsetting of the data.

As the definition of data grows to encompass such objects as images, audio, video, and (digitized or actual) biological samples, neither statistics nor computer science can "go at it alone."

While many workshop participants and others in the SDL research community agree on the need to bring together the approaches and terminology of SDL in the statistics and computer science research community, there still is no clear path to bridge the gap.

We conclude with several narrower issues:

- Dynamic databases, especially those associated with longitudinal studies, to which more attributes on the same subjects are added over time, present challenges that we do not know how to handle. SDL for existing data is hard enough, and there are few ideas about how to protect data whose values are not even known yet.

- Survey data often contain weights and, the papers and presentations by Singh and Fienberg notwithstanding, it is rarely clear what should be done with them in a SDL setting. Weights cannot be viewed as just another attribute. Their values may be highly confidential. Altering them may defeat a principal purpose—matching population-level estimates. Often there are multiple sets of weights, for example, initial weights representing the sample design and another set adjusted for nonresponse.

- In his paper, Singh discusses MASSC, an SDL approach that explicitly incorporates survey design, but often SDL considerations arise long after data have been collected. What benefits are there from a tighter integration of design and SDL?

Additional comments from the panel on federal agency needs include the need for more research focused on business data and more research on tabular data methods, which are the primary form of public data release for some agencies. Agencies appear willing to push for further public access of data if confidentiality can be protected. One suggestion for agencies to help the research community is to provide test datasets.

References

Cox, L.H., Orelien, J. G. and Shah, B. V. (2006) A method for preserving statistical distributions subject to controlled tabular adjustment. In J. Domingo-Ferrer and L. Franconi, eds., *Privacy in Statistical Databases*, volume 4302 of *Lecture Notes in Computer Science*. Springer, Berlin / Heidelberg, 1–11.

Defays, D. and Nanopolous, P. (1992). Panels of enterprises and confidentiality: The small aggregates method. In *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys* **92**:195–204.

Dalenius, T. and Reiss, S. P. (1982). Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6**:73–85.

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

Dobra, A., Karr, A. F. and Sanil, A. P. (2003). Preserving confidentiality of high-dimensional tabulated data: Statistical and computational issues. *Statistics and Computing* **13**:363–370.

Doyle, P., Lane, J. I., Theeuwes, J. J. M. and Zayatz, L. V. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*. Elsevier, Amsterdam.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**:383–406.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statistical Science*, **20**(2):163–177.

Gomatam, S., Karr, A. F., and Sanil, A. P. (2005). Data swapping as a decision problem. *Journal of Official Statistics*, **21**(4) 635–656.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60**:224–232.

Karr, A. F., Fulp, W. J., Lin, X., Reiter, J. P., Sanil, A. P., Vera, F. and Young, S. S. (2007) Secure, privacy-preserving analysis of distributed databases. *Technometrics* **49**(3):335–345.

Muralidhar, K. and Sarathy, R. (2006). Data shuffling—A new masking approach for numerical data. *Management Science* **52**:658–670.

Oganian, A. and Karr, A. F. (2006). Combinations of SDC methods for microdata protection. In J. Domingo-Ferrer and L. Franconi, eds., *Privacy in Statistical Databases*, volume 4302 of *Lecture Notes in Computer Science*. Springer, Berlin / Heidelberg, 102–113.

Oganian, A., Reiter, J. P. and Karr, A. F. (2009) Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis* **53**(4):1475–1482.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**:1–16

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* **1**(1):99–110.

Willenborg, L. C. R. J. and deWaal, T. (2001). *Elements of Statistical Disclosure Control*. Springer-Verlag, New York.