STATISTICS AND PRIVACY IN THE NEW MILLENNIUM: CONTINUING THE DIALOGUE FOR INCREASED ACCESS TO RESEARCH DATA

Stephen E. Fienberg, Carnegie Mellon University
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

Key Words: Confidentiality, Data sharing, Disclosure limitation methods, Perturbation methods.

1 Introduction

The title of this session emphasizes the word privacy and the abstract, for the panel which I prepared these remarks, stressed concerns in the domain of privacy protection. These concerns have been triggered by the rapidly expanding commercial uses of personal information for credit, as well as the privacy threats associated with Internet commerce and access to large commercial databases. There is a sense in which the impact of such commercial databases should impinge little upon the collection, analysis, and reporting of statistical databases. After all, most statistical databases have been assembled with consent and cooperation of the respondents. Nonetheless, in privacy as in other matters, things that should be separate become intertwined at least in terms of perception. And much of the response to privacy concerns the matter of perceptions.

Panel members were invited to participate in a millennium effort at forecasting on this topic this morning, and as I indicated in a somewhat different context last night, forecasting is an inherently statistical problem. And to forecast well we typically need data, at least if we are going to assess the quality of our forecast in the end. The area of privacy and confidentiality is remarkable for the extent of anecdotes on not necessarily data. So, in some sense, all of us must resort to stories. My remarks are based on a limited set of experiences and some deeply rooted prejudices about the privacy, confidentiality and data access. The three are, of course, related but I'll address them in what follows under separate headings.

2 Privacy

Among the topics we were asked to address during this session was the European Union's Data Protection Directive, implemented during the past year. Not knowing precisely what that was, I went to the World Wide Web and did a search which produced not only a copy of an actual directive, but also extensive commentary on it, largely about Articles 25 and 26. These put restrictions on the transfer of personal data to non-EU countries, such as the United States. I confess I was struck by the extent to which the entire directive was focused on commercial databases. Statistical data showed up in Article 6, which allows that, data collected for some specific purpose should not be considered as incompatible with statistical purposes and uses, with appropriate safeguards. That language allows a nice entry point for disclosure limitation tools. Statistical data are also the partial focus of Article 8, which identifies some specific variables such as racial and ethnic origin or political opinion as being especially sensitive, although it goes on to make exceptions for certain public health and medical uses.

My second observation was that the EU Directive reads as though it were written by lawyers, which of course it was, and that statisticians had at best a limited hand in the crafting of the language. I was reminded of my experiences at international statistical conferences on privacy and confidentiality, where there are usually three groups of people, the people from statistical agencies, the methodologists in universities, and the lawyers. There is usually some overlap of interest between the methodologists and the agency statisticians, and a lesser amount between the agency people and the lawyers, but there is no overlap of interest or language between the statistical methodologists and the lawyers. Thus, as I reviewed the provisions of the EU directive, I concluded that the fear that such laws may curtail the access to international statistical data is largely ungrounded. I believe that most privacy concerns can be addressed if there is a clear distinction made between administrative and commercial data on the

one hand, and statistical data on the other. Someone must educate the lawyers and the politicians, however, about the notion of statistical confidentiality and how various disclosure limitation techniques help to preserve it.

3 Confidentiality

Much of my research over the past six years has focused on issues of confidentiality. One of the breakthroughs in this field occurred over two decades ago when Tore Dalenius (1978) convinced most of those in this field that confidentiality was a statistical topic and that any release of statistical data produced a disclosure in that it increased the probability identification of some individual in the relevant population. As a consequence, promises of confidentiality cannot be absolute, and we must focus on the limitation of disclosure risk rather than its elimination.

The past decade has seen a remarkable growth of research literature on the topic of disclosure limitation, and new techniques have been created or are on the research horizon, which can radically change our approaches to data access in many statistical agencies. (The two special issues of the *Journal of Official Statistics* on this topic, in 1993 and 1998, are good sources for learning more about the topic.) Here are a few observations and thoughts about the impact of these statistical research developments and their implementation in the statistical agencies:

1. In general, the statistical agencies have been far too conservative and astatistical in their approaches to disclosure limitation, and thus they have restricted access to their data in ways that are unnecessary. Statistical data are a public good, and they need to be shared more broadly with the very public that funds their collection.

2. Some tools in common use such as cell suppression distort the inferences that are possible in released data and thus impede the use of released data for actual statistical analyses by others (e.g., see the discussion in Fienberg, Makov, and Steele, 1998, and the contrast with perturbation methods).

3. Agencies need to devote more research effort to the modern statistical approaches now appearing in the literature, since the scaling up of approaches and ideas to large-scale survey and census releases involves new research.

4. Many of the fears raised by reports of the accuracy of record linkage methods for breaking the confidentiality of statistical databases are misplaced. They typically require specification knowledge that only comes from other confidential sources or with heroic and untestable assumptions. What happens when they are untrue, is that an individual doing linkage presumes the accuracy of a match is typically far higher than is in fact the case.

4 Data Access

I have long been an outspoken advocate for access to research data, and I have always thought of data collected by statistical agencies as fitting this description. In 1985, the Committee on National Statistics issued a report on *Sharing Research Data* (Fienberg, Martin, and Straf, 1985) which focused largely on access to data collected by researchers outside the government. At long last, during the past year the specter of access to government data has come to the fore largely as a result of the Shelby Amendment to last year's appropriations bill (Public Law 105-277). Under this amendment, individuals and groups can use the Freedom of Information Act to obtain "all of the data produced" by a published study paid for with public funds. The amendment was aimed at gaining access by industry to data underlying reports and regulations issued by the Environmental Protection Agency and other government regulatory bodies, but it effects everyone in government and lots of us outside.

Early in 1999, the Office of Management and Budget (OMB) released preliminary guidelines for the implementation of the amendment in the form of a revision to OMB Circular A-110, "Uniform Administrative Requirements for Grants and Agreements with Institutions of Higher Education, Hospitals, and Non-Profit Organizations." The response to the proposed guidelines included articles in the *New York Times* and other newspapers, and petitions from scientists, including groups from the National Academy of Science, who fear that they will be required to surrender lab notebooks and complete data files to their scientific competitors. OMB received over 9,000 comments on these draft changes. My sense at the time was that the concerns expressed were by and large misplaced, especially since those scientists with NIH and NSF contracts are already bound to provide access to their research data within some specified period of time. The problem is of course that the provision has been honored more often in the break than in the observance.

When OMB released clarifying changes to the guidelines in August, immediately following the

Joint Statistical Meetings, it recognized the need to protect confidentiality of data sources at the same time as it mandated data access, in much the same language that we in the statistical community have advanced for years. The new language for Circular A-110 makes clear that investigators cannot hide behind the need to preserve confidentiality as a way to block permanent access to their data. The final text of the revision to OMB Circular A-110, which takes effect on November 6, 1999, is available at `http://whitehouse.gov/OMB`.

The intent of the Shelby Amendment to provide expanded access appears to run in the opposite direction from that of the EU Privacy Protection Directive to ensure privacy and expanded protection from data. Therefore, it would appear that the time has come to bring the various communities with interests at stake in the discussions of confidentiality and privacy together. Statisticians, both those in the statistical agencies who are collectors of data, and the methodologists who can explain both how the statistical approaches to disclosure limitation work, would have a special role to play in such discussions. After all, it is the statisticians who can explain how such approaches can facilitate the demand for access to basic research or statistical data.

5 The Future

So what then are my forecasts for the future in this area? Here are some thoughts on where attention should be focused and what the impact might be:

1. We need more and better research on confidentiality. And statistical agencies need to be more creative in the ways in which they draw on the expertise of those in the academic community. Disclosure limitation research and record linkage research need to go on outside of the statistical agencies but with suitable access to agency data.

2. We have a big education job ahead of us if we are to explain to researchers in other fields and the public at large what confidentiality of statistical data is really all about, and why it is not absolute.

3. Statistical agencies and researchers need to make more aggressive use of modern disclosure limitation research. Over time this should lead to greater access to statistical and research databases.

4. Statistical agencies and university researchers must be prepared to respond to requests for access to data in a responsible fashion.

5. When the matter of access to data becomes entwined with matters in legal dispute, only strict legal provisions supporting the confidentiality of databases are likely to survive judicial remedies. Thus we must work to strengthen the legal protections of confidentiality for statistical data in a variety of settings.

Clearly, public interest in privacy and confidentiality will remain high as we move into the new millennium, and statistical approaches for their preservation will remain on the research agenda.

6 References

Dalenius, T. (1978). The Swedish Data Act and Statistical Data. *Statistical Review*, **16**, 37-45.

Fienberg, S.E., Martin, M., and Straf, M.L. (1985). *Sharing Research Data.* National Academy Press, Washington, DC.

Fienberg, S.E., Makov, U.E., and Steele, R. (1998). Disclosure Limitation Using Perturbation and Related Methods for Categorical Data. *Journal of Official Statistics,* **14**, 485-502.

Fienberg, S.E., and Willenborg, L.C.R.J. (eds.) (1998). Special Issue on Disclosure Limitation for Protecting the Confidentiality of Statistical Data. *Journal of Official Statistics,* **14**, no. 4.

Lyberg, L. and Duncan, G. (eds.) (1993). Special Issues on Confidentiality and Data Access. *Journal of Official Statistics,* **9**, no. 2.