Confidentiality, Disclosure, and Data Access: Excerpts from a Summary of a Conference

Compiled by Pat Doyle, U.S. Census Bureau

SUMMARY

The explosion of data and the increased ease with which large data bases can be mined is causing government agencies concern over the release of information collected for the public good. The U.S. government has the dual (and potentially conflicting) responsibilities of releasing data—to inform public debate on the economy and the general population—while protecting the confidentiality of respondents. Hence, some method has to be employed to release statistical data without revealing confidential information. To date, data disclosure methods have allowed release of data without risk of identifying respondents. But the information explosion and enhanced data access techniques lead the government to be concerned about future releases of statistical information. To address this topic, the Census Bureau and other agencies and organizations sponsored the development of a book and a conference entitled *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*.[1] This paper summarizes discussions that took place during three sessions of that conference and highlights plans for future research and development.

The conference had three parts: a pre-conference workshop to provide a primer for attendees who were not well-versed in the topic; a pair of sessions focused on policy issues (under the direction of Katherine Wallman, a roundtable of statistical agency representatives); and two days of technical sessions organized around the chapters in the book. The sections below summarize the two roundtable discussions and the concluding session that reflected comments from users. It also presents a research agenda suggested by the conference participants.

FEDERAL AGENCIES DISCUSS CONFIDENTIALITY POLICY

Katherine Wallman, Chief Statistician at the Office of Management and Budget, opened the conference by noting that this conference addresses the thorniest—and potentially most consequential—challenge facing the national statistical system. The challenge she described arises from the growing tension between the need to protect the confidentiality of individual respondents, and the desire to provide the broadest possible access to information. Government statistical offices derive their mandate for data collection and dissemination from a citizenry that demands both quality information (to drive public policy) and protection of individual respondents from privacy invasion and administrative harm.

Wallman went on to say that many researchers find access to government data increasingly desirable because the newer data bases are more comprehensive, of better quality, and—with improved data base management techniques—better structured. At the same time, she noted, the individuals and institutions that provide the data residing on government data bases (as well as the agencies that sponsor the collection of such information) are well aware that the same technologies that extend analytical capabilities also furnish the tools that threaten the confidentiality of data records. This awareness has the potential to erode (or at least to undermine) respondents' confidence that their privacy will be protected. Striking the proper balance between permitting access to accomplish compelling and legitimate research, and incurring the risk, however remote, of inadvertent revelation of individual information is a fundamental concern and challenge—or, as Ken Prewitt has termed it, the "train wreck" on the horizon.

Safe Data Issues

Wallman posed three questions to a panel of representatives from federal statistical agencies and asked each agency to offer their thoughts. The panel members were:

Rich Allen, National Agricultural Statistics Service (NASS)
William Barron, U.S. Census Bureau
Susan Grad, Social Security Administration (SSA)
Lawrence Greenfeld, Bureau of Justice Statistics (BJS)
Nancy Kirkendall, Energy Information Administration (EIA)
Thomas Petska, Statistics of Income, Internal Revenue Service (IRS).

A summary of the panel discussion, organized by question, follows.

How do we achieve the appropriate balance between data protection and data release, given that laws allow little flexibility? Should users share the risk and be liable for their actions if they attempt to breech confidentiality?

Allen reminded the audience that "one size does not fit all." Each agency's views on confidentiality have been, and will continue to be, shaped by the types of data it collects, by its confidentiality laws and regulations, and, importantly, by its data users and providers. NASS, for example, mainly publishes standard aggregated data designed to protect respondents' identities. Special tabulations are prepared and released but are subject to disclosure protection (and limited to nonproprietary data).

Barron believed that the ultimate responsibility for appropriate use of the statistical information federal agencies collect lies with the federal agencies. However, he also believed there should be laws penalizing bad behavior by users. Grad believed statistical agencies are ultimately responsible for protecting confidentiality. However, it still

[1] Doyle, Pat, Julia I. Lane, Jules J.M. Theeuwes, and Laura V. Zayatz (2001). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: North Holland.

makes sense to impose large fines for misuse of data (where appropriate), as this sends a strong message about the seriousness of the issue. SSA is not sure how to assess fines for misuse of public-use files, but the penalties can be imposed for data provided under a "memorandum of understanding" (discussed further below).

Grad noted that SSA supports continued enhancement of the ability to release data for statistical purposes—while protecting the identity of the data subjects, sharing techniques among federal agencies, and adapting a variety of release strategies optimized to provide maximum analytic utility of the data. She also expressed the opinion that the SSA penalty for misuse of data is fairly low (up to $10,000) and that perhaps this should be raised to be in line with other statistical agencies.

What are the real threats to confidentiality from published data? Are all data equally at risk? Does the Internet create an unacceptable risk (either real or perceived)?

Kirkendall noted that EIA does not have separate laws protecting the confidentiality of businesses but the agency takes steps to protect the data anyway. The agency delays release of time-sensitive information as a way to protect data on firms since, in the realm of statistics EIA disseminates, the data becomes less useful the older they are. EIA also suppresses information on certain firms if release of that information would do competitive harm.

Many agreed that linked administrative and survey data are more at risk than other types of data. Grad described the SSA's collaborative efforts to develop public-use files of various types of benefit and earnings data and of Survey of Income and Program Participation data linked with SSA administrative data. To address the higher risk of releasing linked survey and administrative data the agency is working with relevant Federal agencies and users to adapt the latest disclosure methods. The objective of that effort is to see if it is possible to adequately mask the data so that its public release will not permit identification of individuals but will adequately preserve the properties of the underlying data so that it will serve analytical needs.

Petska discussed IRS concerns that are shared by other agencies about nonresponse that might arise if the public did not trust the agency's pledge of confidentiality and IRS recognition that data are a public good. Indeed, Allen noted that NASS response rates were affected recently when an environmental activist organization acquired all government farm payments data for the past five years through Freedom of Information Act provisions and posted those totals by name and county to the Internet. While most agricultural producers understand that data they have provided to NASS will never be released, the backlash from that Web site did negatively affect response rates on recent major NASS surveys. In spite of this, he does not think the Internet creates a real risk for NASS since NASS data policies ensure that all published data products meet confidentiality guidelines or have waivers (see next paragraph).

According to Allen, NASS has a policy of publishing some otherwise confidential data if the company whose data might be identifiable will sign a waiver. These waivers are renewed annually to ensure the provisions are understood and acceptable, especially when there is staff turnover or policy change within the company.

Greenfeld pointed out that BJS faces a once unique problem: assuring respondents that their information will not be shared with litigating branches within the Department of Justice. Marilyn Seastrom from the National Center for Education Statistics (NCES) noted a similar problem as a result of the USA Patriot Act. This act includes a provision allowing the U. S. Attorney General to appeal to the courts for copies of records in the possession of NCES. The Attorney General must show that the Justice Department will keep the data confidential and will only use it for antiterrorist activities. NCES will be modifying their confidentiality messages to clarify to respondents that their data will remain confidential unless sufficient justification exists to use it for antiterrorist activities.

Should we seek a legal status for a new category of data that is not public and not confidential for certain research uses. (For example, data that do not meet strict confidentiality disclosure rules but are not easily identifiable could be provided under data use agreements that carry legal penalties.)

While not speaking directly to this issue, Grad noted the use of other mechanisms to provide nonpublic information to researchers for analysis. Where data are needed that are not publicly available, SSA tries to provide data in unidentifiable form in tables or to provide limited dissemination governed by a "memorandum of understanding" that lays out responsibilities such as the need to safeguard the identity of individuals.

Since NASS does have a strong specific confidentiality statute and good operating standards, Allen "selfishly" opposed legislation for a new category of data that is not public and not confidential for certain research uses. He would be leery of any new legislation that might open doors to more scrutiny of existing confidentiality laws and that might build in special provisions that would be harmful to current operations. He thought such legislation, by changing important restrictions on access to NASS data, might get in the way of confidentiality understandings that NASS presently has with the broad agricultural communities.

Greenfeld raised a special problem related to the law and maintaining the confidentiality of respondents: What should an agency representative do upon learning of an (as yet) unreported crime or victimization during the conduct of an interview? He recommended that each agency needs to consider a set of procedures that may need to be tailored to the individual program, but that follows a general set of principles with respect to both respondent concerns and disclosure concerns. He believes methods for addressing the respondent concerns are far less developed than the methods for protecting against disclosure of data. In BJS surveys,

providing assistance or intervention to a vulnerable respondent—who is experiencing emotional upset as a consequence of survey questions—could affect what BJS is trying to measure over time. But the single most important challenge, both statistically and morally, is that once an agency knows the respondent is experiencing a continuing exposure to victimization, what should it do about it?

Safe Setting Issues

Wallman then posed two questions to a different set of representatives of selected federal statistical agencies for their comments.[2] The Panel members were:

Lynda Carlson, Science Resources Statistics, National Science Foundation (NSF)
Lois Orr, Bureau of Labor Statistics (BLS)
Marilyn Seastrom, National Center for Education Statistics (NCES)
Edward Sondik, National Center for Health Statistics (NCHS)
John Thompson, U.S. Census Bureau.

A summary of the panel discussion, organized by question, follows.

What are acceptable options for providing safe settings for access to confidential data (licensing, data centers, remote access)? What are the legal issues and the pros and cons of each method?

Thompson pointed out that, like all of the statistical agencies represented at the conference, the Census Bureau has strong obligations to protect the confidentiality of respondents. The Census Bureau views providing safe settings as an important means of permitting critical research—where safe data are not useful and cannot be made such (e.g., linked data, business data, decennial data). Carlson mentioned that the viable safe settings for data analysis vary, depending on the size of the agency, the type of agency employees, the type and format of data provided by the agency, and the characteristics of users.

The Census Bureau uses the secure site model (in the form of Research Data Centers), limiting allowable projects to those that are well-developed and that will benefit Census Bureau programs. Carlson pointed out the impracticality of the secure site model for small statistical agencies like NSF, noting that they are simply unable to provide the support needed to establish or maintain the operation or oversee the use of the centers and approve data releases. She recommended that the government establish cross-agency data centers that are shared, recognizing there are some legal and other constraints that need to be overcome. Seastom responded to this proposal by noting that the biggest impediment to sharing software is the difference in the laws across agencies that dictate how confidentiality is to be protected.

Carlson thought that remote access might be a viable alternative for small agencies but only if the effort to develop software for remote access that can be shared across agencies. Thompson reported that the Census Bureau is exploring "remote access" via the American Fact Finder, Tier 3, an access system that will provide a means to allow users to get safe data using the underlying confidential data. It has been in development for several years and reflects compromises that address disclosure risks in the context of the Census Bureau legal and policy constraints.

A program of licensing is well established at NCES. Seastrom described NCES as having a distinct set of users, uses, and data so the licensing process works well. The agency recognizes its responsibility to protect the information and takes this seriously. They are committed to maximizing access safely. Carlson would like to adopt a licensing option at NSF but is faced with the ruling that licensing is not allowed under Title 13 (which governs some of their data collection programs). In the spirit of agency sharing suggested by Carlson, Seastrom thought some aspects of the licensing process, such as the inspections, might be more amenable to cross-agency sharing than software products that need to rely heavily on the details of each agency's disclosure requirements.

Orr characterized BLS as being behind the times on this topic as it has no licensing, no research data centers and no remote access. However, she did note that BLS does have programs that allow researchers to come on-site as fellows or temporary BLS employees and use their data.

How do we deal with the perception that confidentiality is not protected in these safe setting arrangements? What are the threats (real and perceived) and how do we evaluate and mitigate them?

Seastrom emphasized the importance of following through on all dimensions of the licensing process. It is critically important to have an enforceable licence, that the agency make adequate investment in security checks, and that the agency follow through with inspections to ensure compliance. In response to a question about volume, she indicated that NCES has 400 to 500 active licenses, all of which are subject to inspection.

Thompson noted that the Census Bureau views its "culture of confidentiality" as integral to maintaining the trust of respondents. Safe settings are designed to maintain this culture of confidentiality by having a Census Bureau employee onsite and requiring researchers to be trained on rights and responsibilities. The Census Bureau recently established the Data Stewardship Executive Policy Committee to focus attention on issues related to confidentiality, privacy, and security. This committee is composed of members of the senior executive staff of the Census Bureau and is charged

[2] Actually a third question was also posed (How do we make it clear that safe setting arrangements are different from data sharing arrangements that must be legislated?) but no one spoke to it.

with establishing policies on confidentiality and access. Of course, the Census Bureau continues to have an active and effective Disclosure Review Board (DRB) through which all data to be released to the public must pass for approval.

Sondik addressed six points that he believes are important as statistical agencies consider how to build, maintain, and provide access to safe settings for the release and use of the information entrusted to us by respondents.

- *Process.* There is no substitute for a careful review process prior to release.
- *Quantifying Risk.* Agencies should explicitly characterize the trade off between analytic utility and disclosure risk.
- *Research*. Statistical agencies should join forces to sponsor both intramural and extramural research to better understand disclosure strategies and risk.
- *Legal Issues*: Sondik expressed concern that current sanctions against those who willfully try to invade statistical systems are too weak.
- *Be Prepared*. Even though agencies continually protect against the possibility of a violation of confidentiality, they need to undertake the detailed planning and thinking required to address the issue if it ever arises.
- *Peer Review.* Because of the likelihood that problems will occur, it is important that the statistics community and the public review each statistical agency's plans. Sondik supported the notion of DRBs and Institutional Review Boards (IRBs).

Seastrom disagreed with Sondik on the issue of IRBs by noting that data collection carried out for social science research that does not involve invasive data collection need not be subjected to the extra scrutiny of an IRB if there are strong laws, strongly enforced governing the protection of confidential information. Greenfeld acknowledged that the BJS experience with its IRB was not completely positive citing the IRB's lack of appreciation for the need for repeated measures, the IRB's concern over administrative issus, the board's lack of knowledge of survey methods and the existing protections already in place.

TECHNICAL SESSIONS

In the interest of fitting this document within the page constraints of the ASA proceedings, the section has been deleted. Full details of the technical presentations are found in the book on which the conference was based. Discussion of these papers will be summarized in a future version of this paper (contact the author for information..)

REACTIONS FROM USERS AND ADVOCATES

The conference concluded with a panel of three individuals representing three constituencies beyond the federal statistical agencies. Stephen Fienberg, from Carnegie Mellon University, described himself as a researcher specializing in disclosure methods, a user of federal data subject to disclosure, and a respondent in federally sponsored household surveys. Stephen Tordella, from Decision Demographics, is a demographer with over 25 years of experience in secondary data analysis using federally produced microdata products. Robert Gellman is a privacy consultant who spent much of his career as a Congressional staffer focusing on privacy legislation and whose views differ dramatically from the users and producers of statistical data.

Fienberg believed the statistical system needs changes in laws to put the onus on users, offer even greater protection against misuse, and protect against retrospective exceptions (such as the USA Patriotic Act). He thinks the new laws need to be different, recognizing that confidentiality is not absolute and that disclosure limitation is probabilistic. He believes in maximizing data sharing and unrestricted access, and he opposes restricted access centers. Fienberg reminded the audience that data are measured with error, a powerful tool in disclosure avoidance that is not well used at the current time.

Tordella seconded the notion that the current approach needs to be changed. He noted that restrictive access policies on the government side will hobble the advance of research and understanding in many important demographic and economic areas and will stifle creativity in reaching solutions to society's problems. He also noted, in particular, the extra restrictions imposed on for-profit firms' access to data, which assume that research carried out in such firms will be worse along some dimension than research in nonprofit firms or universities. Tordella questions that assumption. In addition to his objection to restrictions on for-profit firms, Tordella also pointed out that research data centers are impractical for many users (because of the high cost in terms of time, money, and mission restrictions).

Tordella went on to explain that, even within the context of the current laws, the outcome would be improved if statistical agencies could find a way to work with users in developing the disclosure methods—so that the outcome was optimized for research and analysis while still protecting the identity of the respondents. He perceives that the agencies are trying to protect themselves from a rare event (a hacker trying to re-identify a survey respondent) instead of planning for the majority of uses. His perception is also that there is a tendency in access control to restrict the range of users, thereby hoarding the data and keeping it from everyone except "politically correct" users. He illustrates his point by recalling Eden's three goals for producers presented at the pre-conference workshop: protect yourselves, protect your data, and protect the people from whom you collect the data. He asked, "Where are users in this equation?"

Gellman made it clear that he does not perceive the likelihood of a violation of disclosure protection to be as rare an event as Tordella suggested. Gellman acknowledged that adjusting, modifying, or masking microdata is good; but he still has a concern about those in the private sector he calls voracious data collectors. These are the firms who will take anything, and it does not have to be all that accurate or current—making the point that there are institutions for whom accuracy is not important and who are motivated to amass large amounts of data from any source that is available. In

light of this, Gellman thought it was interesting to hear in the pre-conference workshop that no one has a list of external databases. He questioned why the statistical community has not beaten the public and commercial bushes to find all the databases, reflecting on the fact that you cannot assess risks if you do not know what the risks are.

Fienberg recommended some changes in the way statistical agencies do business and suggested avenues for further research. Agencies should explain issues and disclosure protections to the public and policy makers, revise and strengthen confidentiality laws, link disclosure limitation to editing as part of a broader strategy on data quality, rethink agency practices from sample design through disclosure limitation and data release, and expand data access.

On the topic of perceptions, Fienberg believes the public perceives the Census Bureau as not having the legal authority to protect confidentiality or as not being willing to exercise that authority. On the other hand, more people cooperate with federal surveys than would be expected, given their skepticism.

A contrary opinion on the public's perception came from Gellman. Although he was encouraged by the amount of effort, energy, and expertise put into this issue for this conference, he still had four areas of concerns. First, the definition of privacy in terms of personal control over data is meaningless and antediluvian. Second, statistical agency licensing of data, provision of data to temporary employees (special sworn status or deemed employees), and other sharing arrangement are something of a subterfuge (doing indirectly what you are not supposed to do directly.) Third, there is an apparent conflict of interest in the Census Bureaus's limitation on use of restricted data to projects that benefit the Census Bureau. Fourth, he believes (without exception) that all disclosures breach confidentiality. In addition, Gellman perceives that the Patriot Act may affect the ability of the statistical agencies to convince respondents of the seriousness of their confidentiality pledges through the citation of confidentiality laws and associated penalties for disclosure.

However, Gellman can see that providing access to data is justified at times, if there is adequate *independent* review for privacy, transparency in process and notice, and better enforcement. He believes that subjecting people to criminal penalties rarely invoked is not much of a deterrent. Professional embarrassment might be better, but there is a question as to whether one agency would know if a researcher violated rules at another agency.

Fienberg reminded the audience that the collection methods will continue to become more sophisticated, thus requiring new and more effective disclosure methods. His example was a survey administered on the Wright-Patterson Air Force Base, which included 50 megabytes of data from each respondent (generated by a 3-dimensional full-body laser surface scan).

SUGGESTIONS FOR FUTURE RESEARCH AND CONFERENCE

Sondik recommended that statistical agencies join forces to sponsor both intramural and extramural research to better understand disclosure strategies and risk. The final aspect of the conference was to gather participants thoughts on what this research should cover. Over one third of the conference participants offered suggestions for future research and encouraged a repeat of the conference, offering some suggestions for additional topics to address. Those suggestions are summarized below, by topic.

Research on Respondent Behavior and Perceptions

- Continue research into the public's perceptions (and into how to address misperceptions) about statistical uses of these data and about associated data protections. Consider measuring the public's faith in current deterrents and assessing what deterrents might work better (e.g., sanctions on receipt of federal funds for willful violation of the confidentiality protection).
- Investigate disseminators' ethical responsibilities to respondents how the agencies convey those in a productive and meaningful way.
- Continue to try to assess impacts of confidentiality concerns on willingness to respond and find ways to increase willingness to respond by increasing confidence in the data protection techniques employed.
- Continue to study the relationship between respondents' characteristics and their response to requests for information to be used for statistical purposes.
- Expand the research on business perceptions of confidentiality protections.

Instrument Development and Disclosure

- Investigate whether it is beneficial to get precise measures (through sophisticated questioning) if the increased precision results in the higher likelihood of a possible intrusion which, in turn, results in the application of disclosure methods that blur the data and make it less precise.
- Determine the impact of alternative means of data collection on disclosure and data availability (e.g., global positioning systems, satellite images, body scans, etc.)
- Consider administrative records as an alternative to direct data collection. Can the increased risk of disclosure be addressed in data protection strategies? Is the reduction in cost of collection offset by the increased cost of disclosure—or possibly the reduced accessibility of the data?

Research on User Needs and Analytic Utility

- Expand the research to assess impacts of disclosure methods and adequacy of access methods to suit users needs, and to consider perspectives and needs of data users in designing disclosure methods. In particular, try to distinguish among techniques that work for broad sets of

applications (and are, therefore, suitable for disclosure proofing general purpose public use files) and those that work for specific analytic tasks.
- Expand studies of the analytic utility of disclosure-proofed data to a broader set of applications, particularly those that require the preservation of the covariance structure among huge numbers of items in the surveys (such as microsimulation models). Continue evaluation of methods—by disclosure method, by type of data, by access methods (including Geographic Information Systems), and by type of use.

Application of Disclosure Techniques
- Expand to study use of disclosure techniques in the private and congressional sectors and among nonstatistical agencies or organizations. How does this impact small organizations? Need less technical, more practical, decision-making guidelines for units with fewer (or less sophisticated) resources.
- Develop a methodology that can quantify the risk of disclosure—without revealing information that would increase the risk of disclosure—so that statistical agencies can increase effectiveness of communication with stakeholders and solicit their input.
- Try to identify types of would-be data intruders and develop a cost-benefit analysis to assess likelihood of an intrusion.
- Generate information to make decisions about the cost-effectiveness of different strategies (where cost includes the cost of making decisions without adequate data).
- Evaluate the cost of disclosure by type, in terms of potential for error in government policy decisions (e.g., mistargeted programs, duplicate data collection). Also assess whether disclosure methods are making the data useless for the purposes for which they were collected.

Methods of Data Protection
- Consider all legal, regulatory, and policy options available to provide access to data for statistical purposes. Consider the increasing availability of data and the sophistication of technology, necessitating changes in disclosure methods—to the point where high quality research and analysis is inhibited. Also consider public perceptions and the influence of privacy advocates on public perceptions.
- Undertake the detailed planning and thinking required to address the issue of disclosure violation, if it ever arises.
- Continue to develop new and improved methods of disclosure for tabular data and microdata, covering economic and demographic censuses and surveys.
- Develop a better science for the impact of sampling fractions as a disclosure protection tool among rare populations.
- Develop new data analysis techniques, data access methods, or dissemination practices that will help to minimize disclosure risk without adversely affecting the analytic utility of the data.
- Consider expanding the role of editing and imputation as a disclosure method.

Suggestions for a Future Conference
- Repeat this conference and publish a new book updating all topics of the current conference (including the policy topics), reducing the repetition on access methods, keeping the balance between economic and demographic data, maintaining international perspective, and retaining the closing panel.
- Add new topics to the next conference: examples of successful break-ins; impact of the Freedom of Information Act, the Health Insurance Portability Act, and the Patriot Act on data release and disclosure protection; an expanded workshop for novices and a new workshop on applying techniques; training and practice in developing and maintaining a culture of confidentiality; function of IRB's versus DRB's; and an interagency panel of DRB members discussing what they do, their perspectives, communication with stakeholders, models for release decisions (objective versus probability measures), tradeoff between variables, benefits versus sensitivity of data, sampling as disclosure protection, and confidence that today's methods will not be undone by tomorrow's access methods.

ACKNOWLEDGMENTS AND DISCLAIMER

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. I would like to thank all participants in the January conference as this paper constitutes no more than a compilation of their excellent commentary and suggestions made while in attendance. All errors in interpretation of their remarks are my responsibility and I apologize if any exist. I would also like to thank Michael Morgan for his outstanding editorial assistance.