Journal of Privacy and Confidentiality (2011) **3**, Number 2, 73–76

Comment on Article by Gates

A comment on Gerald W. Gates's *How Uncertainty about Privacy and Confidentiality Is Hampering Efforts to More Effectively Use Administrative Records in Producing U.S. National Statistics*

Jerry Reiter[*]

I congratulate Jerry Gates on a lucid and important discussion of the role of privacy and confidentiality in sharing government data. He makes a strong case that increased and easier data sharing would improve the collection, estimation, and dissemination of federal data—which in turn would lead to better research and public policy—and that confidentiality concerns raise barriers to realizing those benefits. His suggestion for a multi-faceted approach based on legal reform, inter-agency cooperation, and methodological research is a solid path for finding ways to lower the barriers.

As someone who works on statistical disclosure limitation techniques, and especially the synthetic data approaches that Gates mentions, I appreciate and second his call for continued research in data sharing and dissemination. In this commentary, I would like to suggest some additional topics for methodological research. These are focused on dissemination of data to the public; sharing data across government agencies seems to me foremost a matter of policy and law, which are not my areas of expertise. I should add a disclaimer: I have not done a proper literature search to place these additional topics in context. For all I know, they already have been investigated and discarded! I apologize in advance if others have suggested similar ideas and do not receive credit here.

Expressing confidentiality preferences in data collection.

Gates calls for system-wide, ongoing surveys of data subjects' attitudes on privacy and confidentiality. While such surveys can inform strategies for assuring data subjects about the confidentiality of their data, they are insufficient alone as guides for data dissemination. To illustrate, suppose surveys indicate that 90% of individuals in a population do not care about the confidentiality of a particular set of variables and are willing to share these data without redaction. The agency is still obligated to protect the confidentiality of the 10% who care, and the agency is unlikely to know who those people are in a particular dataset without other information.

To get this information, perhaps agencies could ask data subjects about their confidentiality preferences during the data collection. Agencies could release data "as is" for subjects who do not mind having them do so, and redact confidential data of subjects who request greater protection. For example, the agency could use partially synthetic data, i.e., simulate new values of only the confidential variables, using models based on the original data (Little, 1993; Reiter, 2003). When the synthetic data models capture

[*] Department of Statistical Science, Duke University, Durham, NC, mailto:jerry@stat.duke.edu.

the distribution of the confidential values, the released data are not subject to selection biases, even with selection among the sampled subjects who agree to the "release-as-is" option. In this way, higher fractions of genuine data might be shared with the public while data subjects request their own confidentiality comfort level.

For this idea to work operationally, the agency would have to balance the increase in survey length (and monotony) from repeatedly querying subjects about confidentiality preferences with ensuring sufficient opportunities to express preferences. As a speculative possibility, the agency could group questions into categories (e.g., demographics, housing, income, health), and ask respondents for their confidentiality preferences about the group, e.g., something along the lines of "Would you be willing to share your exact answers to the demographic questions with others?" Data subjects could be given opportunities later in the survey to change their minds about any preference. Agencies might be able to develop algorithms that learn what data subjects deem sensitive in real-time, so as to avoid asking unnecessary preference questions.

There are a host of survey methodology questions associated with this approach to data collection. Does it lead to more, or perhaps less, measurement error than the combination of current data collection plus statistical disclosure limitation? What is the impact on response rates? How do we ensure people truly understand the implications of stating confidentiality preferences? What is the optimal implementation when considering respondent and interviewer burden, costs, errors, and confidentiality? How and when should the questions about confidentiality preferences be asked? Clearly, there are many challenges to getting this to work well in practice. But, in an age when people are accustomed to expressing their privacy preferences for social networking and other websites, this approach to data collection has the potential for decreasing barriers to data sharing.

Incentivizing participation in sensitive surveys.

Gates reviews some excellent research on why individuals do not participate in surveys, including the role of confidentiality concerns. This research suggests a variety of strategies for increasing incentives for participation. I would like to suggest research on another type of incentivization scheme, which I motivate with a brief personal story. Several years ago, I designed some applets for an introductory statistics text book. A key part of my contract with the book publisher was a re-use fee; that is, I was paid every time one of my applets was adopted by another text book. This increased my incentive to provide high quality designs, since I presumed that a better product would enable me to make more money on re-use fees.

I propose that it would be beneficial to adapt the re-use fee model to survey data collection and dissemination. For example, the federal agency would pay a modest amount of money to a respondent each time his or her data were downloaded from the agency's website (or obtained by any other access medium). Agencies could pay differentially by question, for example, paying more for answers to sensitive questions than for answers to routine ones. Building on the idea of asking respondents for their confidentiality preferences, the agency could pay more for data that the respondent allows to be released without redaction and less for data that the respondent requires to be redacted. Similarly, the agency could pay more for data judged to be high quality, e.g., by independent audits, and less for data deemed of dubious quality. Taking the re-use model even further, perhaps agencies could set up a market whereby potential respondents could negotiate their prices for data dissemination (but probably not data collection) with the agency.

The re-use fee approach has many unknowns. Would individuals provide accurate answers and allow for easier data dissemination if compensated with re-use fees? What prices would increase data quality yet still be affordable? How do those costs compare to the costs of nonresponse follow-up plus data redaction? Would selection bias from who accepts the fees undermine the utility of the data? While clearly there are many challenges to successful implementation, a re-use fee may incentivize respondents to share more high quality data, perhaps at a cost that is comparable or cheaper to the costs of existing approaches to dissemination.

Placing more trust in selected users.

Gates calls for amending confidentiality legislation to facilitate greater sharing and dissemination. I think it is crucial that data stakeholders be at the table when those amendments are discussed. With intense concerns about privacy among the public and legislators, it could be easy for amended or new legislation to err on the side of over-protection at the expense of data access and quality. Data stakeholders can provide a voice to help ensure confidentiality laws do not become more restrictive than they are now.

I would like to add a plea to Gates's call for revising confidentiality legislation: give trusted researchers greater access to confidential data. The overwhelming majority of researchers do not use government data for malicious purposes. Their careers depend on adhering to pledges of confidentiality protection. I believe that, with reasonable vetting, agencies can trust researchers more than at present, and hence provide them with greater and easier access to confidential data than at present. Agencies can impose stiff penalties on approved researchers (e.g., the five years in prison and $250,000 fine in CIPSEA) and their institutions (no use of government data for some time period or significant fines) for those who abuse that trust.

The most serious threat posed by researchers is one of accident or negligence. There are many stories of researchers who share licensed data with others without permission, or who do not properly store media containing sensitive data. Many of these problems could be solved by moving toward virtual data enclaves, such as the one developed by the National Opinion Research Center. With virtual data enclaves, the data live on a computer server at the agency, and researchers access that data remotely via secure connections. Researchers can view and analyze the genuine data, but the system prevents functions like local saving and printing. Outputs from analyses of the confidential data are screened by the agency before release. Virtual data enclaves reduce the temptation of unauthorized sharing and avoid disclosures from lost media or laptops. These features, when combined with serious penalties for misuse and education about the importance of protecting confidentiality, might reduce risks sufficiently to allow agencies to give more access to trusted researchers.

Concluding remarks.

One reading of Gates's article reveals that current policies and methodologies must change if agencies are to engage in more and better data sharing and dissemination. I suggest that we also re-examine the process of data collection itself, for example with confidentiality preferences and data re-use fees. If the ultimate goal is data dissemination (which admittedly is not for some government data collections), then the costs of providing access to high-quality data should be incorporated at the survey design and data collection stages.

References

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9: 407–426.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29: 181–189.