Journal of Privacy and Confidentiality (2011) **3**, Number 2, 77–82

Trust but Pre-Verify?

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*

Fritz Scheuren*

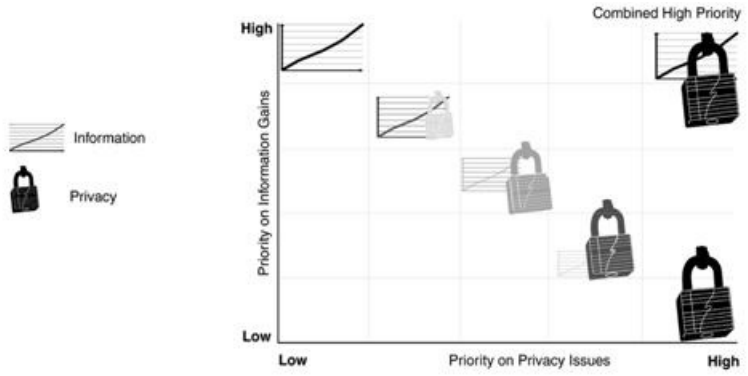
1 Introduction

There is much to discuss concerning Jerry Gates's paper. This is a wise man that has spelled things out very well. However, as you will see, my discussion is not just an "Amen" but also a sketch of a few "Next Steps." Much too early to stop matters where they are today.

2 Goals

Governments should be uncompromising in jointly seeking both information and privacy goals for the data that is in their stewardship. The GAO's 2001 report on Record Linkage and Privacy places a combined high priority on information gains and privacy issues–and presents this position in contrast to an alternative "zero-sum approach" that conceptualizes privacy and information mainly, or only in terms of, trade-offs. The 2001 GAO report further indicates that the way to achieve a combined high priority (for both information gains and privacy issues) is through an improved "privacy-protection tool box" to insure linkages that preserve privacy:

The graphic below, borrowed from GAO (2001), visualizes the possible trade-off matrix as follows.



*Senior Fellow and Vice President for Statistics, National Opinion Research Center, University of Chicago, Chicago, IL mailto:scheuren-fritz@norc.org.

© 2011 by the authors

Decades ago, when Jerry and I were just getting started in government statistics, the upper right corner was seldom, if ever, attainable. Something less was the standard most of the time. No more! Breakthroughs that Jerry mentions by Rubin, Little, and more recently Reiter (and others) have increased the amount of time in which both goals can be achieved. My take-away here is **"No more tradeoffs on goals!"**

To quote the GAO report directly, "Those who prioritize both information gains and privacy issues may be more likely to champion techniques designed to build in personal privacy, confidentiality, or security while still allowing information gains and [to] work to foster improved stewardship or decision-making processes that better balance or, where possible, maximize both personal privacy and information gains."

The Gates paper, in part, reflects this dual-priority perspective, but is still ambivalent in places on the need to achieve both goals. And here I disagree. No more compromise. But there is a lot to like too in what Gates's paper recommends. Notably, Gates (1) advocates extending data stewardship programs to more statistical agencies (such as the one he himself helped establish at the Census Bureau two decades or more ago); (2) calls for increasing both research on privacy issues and public engagement efforts–to better clarify "what conditions would make data sharing for statistical purposes workable or unworkable;" and (3) champions the notion that statistical agencies' jointly undertake, with administrative agency counterparts, research on the limitations and potential of synthetic data (such as proposed by Rubin, Abowd, and Reiter, among others).

3 A Process based on Trust

Gates rightly talks about the need for trust—"Trust but Verify" as we have come to say. Or, as I would augment, "Trust but Pre-verify." Now, those who live in this world of data sharing know there have been statistical systems failures by the statistical agencies, places where data collected for an administrative purpose was reused **inappropriately** by statistical agency staff.

Gates tells the gist of one such story—an incident between the US Census Bureau and the US Department of the Treasury. In that incident, as he tells it, employees and researchers at the Census Bureau used Internal Revenue Service (IRS) data in an unauthorized manner. Although seemingly no great harm was done, that was a real blow to trust and reputations.

What is our takeaway concerning this incident and what it illustrates? That it should not have happened! Clearly! That it cannot happen again. Not so clearly!

The Census Bureau introduced much tighter management controls after the incident, and there was also more monitoring by the IRS on how its data were used. So all to the good!

But was enough done? We are not sure and would propose that still more be considered. Specifically, that there be added statistical safeguards. That more use of

the "privacy-protection tool box" be employed. How might this work?

To start with, administrative agencies might want to study, adapt, and possibly adopt tools the Census Bureau uses to safeguard its own statistical products. To illustrate we will take one of the safeguards Gates discusses (now applied by the Census Bureau in another context) and see if or how it applies to the administrative agency providing data to the Census Bureau for the Census Bureau's statistical purposes.

4 Improving Practice?

But before going further, let me remind you what the salient elements are of Census Bureau practice, repeating in part what you just read but casting my comments in the context of—

If the Census does it for themselves, why can't Administrative agencies do likewise?

What are Census practices? Well, naturally, they vary depending on circumstances. For survey public use data, like the Current Population Survey (CPS), there has long been a tradition of public use files that were only lightly modified to protect against re-identification. Top coding, the suppression of some geographic detail and more recently data substitution were the extent. This was possible because the risk was thought to be low.

But other files, where administrative data were linked to Census survey data have had "tougher sledding." Take the linked file developed by the Census Bureau where the Census Bureau links Social Security Administration (SSA) and IRS data that it obtains to its Survey of Income and Program Participation (SIPP).

Now before the SSA-IRS data linkage was made, the SIPP data were already public. Obviously, adding SSA-IRS data to the SIPP file created an additional re-identification risk. The concerns were not really the added risk that the general public might pose. These were judged to be very small. Rather the risk that the Census Bureau was worried about was the re-identification risk that the agencies whose data were being linked might pose to the SIPP respondents. After all, the potential exists to "relink" using the data themselves as indirect identifiers. The administrative data on the statistical file could be used to link back to the basic administrative file, which has identifiers.

That an administrative agency, the SSA or IRS in this context, might be able theoretically to re-identify Census respondent data was possible. And no amount of reassurance by the administrative agencies to the Census Bureau that administrative agency managerial controls would be sufficient to protect against a breach was viewed as acceptable. Notice the lack of parallel here with the response to the Census Bureau breach mentioned earlier.

5 Possible Approach

What to do? Reiter (2005) developed an algorithm for a synthetic data file alternative which makes for a safe and, many could argue, "satisfying" re-identification resolution. Its "genesis" is found in papers by Rubin (1993) and Little (1993). Gates mentioned all this work in his paper. Anyway, census staff under the guidance of Professor Abowd successfully used this technology to produce the SSA-SIPP synthetic file. There were problems with the algorithms speed, but Bill Winkler at the Census Bureau has resolved these now satisfactorily (Winkler, 2011).

The creation of synthetic files by Census Bureau staff was precedent setting and all involved should be applauded. In an earlier age the two agencies agreed upon administrative procedures that had the same effect (e.g., Kilss, Herriot and Scheuren 1980). But that was before the loss of confidence by the Census Bureau in any agreement that an administrative agency might make regarding re-identification. Of course, recent events alluded to above have led the IRS to also have a lack of confidence in the Census Bureau too. Ironically, in neither the SSA nor the IRS case were there any breaches cited, just fears. But there had been breaches by Census staff, as Gates documents.

So what should the administrative agencies do in response to the lack of a level playing field? One response is to live with this change in Census Bureau practice and continue things as they are. This is what has happened so far. Another alternative is for the administrative agencies that provide data to the census to look again at their own data sharing practices and see if synthetic files **created by them** could be provided to the Census Bureau in lieu of some of the original files provided now. This certainly seems viable. The IRS and SSA have talented in-house statistical staff such that, supported by the technological breakthroughs at the Census Bureau, they could develop synthetic files themselves before turning over (now synthetic) data to the Census Bureau.

6 Implications

Now the assertion is made, as Gates attests, by its Census Bureau implementors and by the original developers of synthetic data that there is no real information loss. So what is the harm to the intent of Census Bureau purposes, if synthetic data are provided in place of the original files? Certainly there is a gain in reducing the possibility of unauthorized disclosures at the Census Bureau of administrative data provided by another agency.

Certainly there is a gain in reducing the possibility of an unauthorized disclosure by the Census Bureau of administrative data provided by another agency. Ironically, when Census staff developed the first synthetic SIPP-SSA-IRS file they did it very well but, because they did not understand the SSA system fully, there were several mistakes that no one from SSA would even have made.

What are some of the downsides to creating synthetic files in lieu of providing original data? Of course, there are many, mostly related to time and cost issues. However, this is not a crazy idea. The Census Bureau and others have shown that the information quality of a synthetic data source can be separated from any re-identification risk and

kept fit for use.

Will the synthetic approach always work? Probably not in all cases (e.g., Mulcahy and Scheuren 2011)! But it will work in many cases and the due diligence steps taken when it does not, could stand everyone in good stead—especially if another breach at the Census Bureau were to occur, this time perhaps with more adverse publicity.

7 Trust again

Let me end with the idea of trust again. The need for trust by all of us in our public institutions is essential. What we do must be open and transparent. There are lots of ways to accomplish this. The way Gates describes has limitations for the administrative agencies involved. Only one alternative, of many perhaps, has been covered here. It could be termed "Trust but Pre-verify." Can a still better way be found? Let's work together for that end, keeping to our principles and adhering to our missions.

Acknowledgments

This is a nice moment to show my respect for a good friend and long time colleague, Jerry Gates. Of course, thanks to Steve Fienberg I get to join with many others. So thanks Jerry and Steve.

References

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*, vol. 201. 1st edition. Lecture Notes in Statistics, Springer.

Herzog, T. N., Scheuren, F., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Kliss, B. and Scheuren, F. (1980). The 1973 CPS-IRS-SSA Exact Match Study. Studies from Interagency Data Linkages: Measuring the Impact on Family and Personal Income Statistics of Reporting Differences Between the Current Population Survey and Administrative Sources. Report No. 11.

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

Mulcahy, T. M. and Scheuren, F. (2011). 21st century data dissemination: Practice & innovations. In *Proceedings of the Conference on New Techniques & Technologies for Statistics, Eurostat*. Brussels, Belgium.

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

Reiter, J. P. (2005). Significance tests for multicomponent estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131: 365–377.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9: 462–268.

United States General Accounting Office (2001). Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information. Report No. GAO-01-126SP.

Winkler, W. E. (2011). References. `http://www/hcp.med.harvard.edu/statistics/survey-soft/docs/WinklerRecli\%nkRef.pdf`.