Journal of Privacy and Confidentiality (2011) **3**, Number 2, 65–71

Toward a Reconceptualization of Confidentiality Protection in the Context of Linkages with Administrative Records

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*

Stephen E. Fienberg*

1 Introduction

Gates has contributed over an extended period of time to efforts to advance the use of administrative records for statistical purposes, especially at the U.S. Census Bureau, and thus it is a pleasure to be able to comment on this paper which reflects his experiences and his thinking on the topic. He focuses largely on the policy aspects of expanding the use of administrative records using ideas and approaches already in the Bureau's repertoire.

In these comments I have chosen to focus on what I believe are several important technical issues that the statistical community needs to consider as the U.S. and other countries move forward into an era of expanded data integration and, I hope, data access. The first issue is a semantic one: we need to use terms such as security, privacy, confidentiality, and disclosure in much more careful ways as we move forward with this new agenda. The semantics are linked to my second issue: we need to address the technical details of how we protect data and the promises we make about such protection. Third, I address some of the challenges of record linkage in the context of using administrative data for statistical purposes. Finally, I address the issue of the quality of sharable linked data released by statistical agencies, once we have adequately addressed the issue of confidentiality protection. The world of official statistics stands at a precipice, and we need new ways of thinking about statistical agency data and new technical approaches, cf. Fienberg and Prewitt (2010).

2 The Semantics of Data Protection: Privacy versus Confidentiality

All too often we as statisticians mix both the goals and requirements of privacy protection and confidentiality protection. As it so happens, the name of this publication, the *Journal of Privacy and Confidentiality*, may not help either, and while Gates makes

*Department of Statistics, Machine Learning Department, Heinz College, and Cylab, Carnegie Mellon University, Pittsburgh, PA mailto:fienberg@stat.cmu.edu

the distinction I wish to emphasize in a few places, he mixes the two notions in several other places.

I want to echo Prewitt (2011) in his comments that the role of statistical agencies is largely about the protection of confidentiality and not the protection of privacy. He makes the nice distinction that protecting *privacy* is akin to *don't ask*, whereas protecting *confidentiality* is about *don't tell.* Nissenbaum (2004) writes about privacy as contextual integrity, in the context of "three principles concerned with: (1) limiting surveillance of citizens and use of information about them by agents of government, (2) restricting access to sensitive, personal, or private information, and (3) curtailing intrusions into places deemed private or personal." Nissenbaum goes on to elaborate on this notion of privacy in terms of societal and legal norms. Prewitt's *don't ask* fits nicely with these principles and suggests that when statistical agencies deal with protecting privacy they should focus on it at the front-end of the operation—when they are deciding what and how to ask in their surveys and censuses.

Confidentiality protection is quite a different kettle of fish and deals with the contractual obligation of the statistical agency with its respondents. In the case of the Census Bureau, confidentiality is governed by language in Title 13 of the U.S. Code, which restricts the use of Bureau-collected data to "statistical purposes for which it is supplied" and states that the Bureau should not publish individual data in a form that can be identified. Title 13 does not mention the word privacy, although the Census Bureau's website in referring to it makes the claim that "Private Information is Never Published;" it then goes on to refer to names, addresses, telephone numbers, and social security numbers. This seems to refer to the publication of information in a form that is directly identifiable. The language does not extend the notion of confidentiality to any form of absolute protection for all information provided, and this opens the door to various forms of probabilistic notions of confidentiality protection and potentially looser standards regarding identification of some forms of statistical information.

The use of administrative records by statistical agencies does raise new questions about privacy since the administrative records may be viewed as a form of surveillance and laws surrounding their use are often viewed as protection of individual privacy. Thus when administrative records enter the statistical domain, agencies have special obligations they need to attend to in order to meet such legal requirements.

The differences between privacy and confidentiality that I have attempted to elucidate here are not simply a matter of semantics. They are crucial to the nature of statistical agencies and their mission, which includes the sharing of the statistical data they collect for the public good. And the differences imply that techniques for privacy protection may not and perhaps should not be the same as techniques for confidentiality protection, although they clearly are related. It is for this reason that in my work on this topic I use the term "disclosure limitation" to describe my methods, in lieu of the older and what I deem to be less appropriate language of "disclosure avoidance," used by Gates. The word "avoidance" suggests precisely the absoluteness I think is not achievable if we want to maintain the tradition of access to useful government statistical data.

3 Technical Aspects of Protecting Privacy and Protecting Confidentiality

Gates's references to the literature on privacy protection and confidentiality protection are substantially out-of-date as a quick perusal of prior issues of this journal and others would suggest.

From the perspective of privacy protection, the major advances of the past decade have come from the introduction of differential privacy by Dwork et al. (2006) and its elaboration and application to statistical problems, e.g., see Dwork and Smith (2009). Differential privacy provides strong privacy protection guarantees by assuring that the probability associated with any statistical quantity is "essentially unchanged" by the addition or removal of any individual from the database. This approach makes no assumptions regarding the external information of a potential intruder, and the protection mechanism is via the addition of noise to the released statistical quantities, typically drawn from the Laplace or double-exponential distribution. The strong guarantees come at a price, which involves reduced data utility—often substantial—and the methodology doesn't really help with the release of public use microdata (PUMS) files.

In contrast to the literature on privacy protection, that on confidentiality protection is more diverse in its methodology and in terms of the criteria it invokes. Techniques include but are not limited to: (1) sampling, (2) aggregation including variable coarsening and the use of marginal releases from contingency tables, (3) data swapping, and (4) synthetic data, e.g., in the form of multiple imputation. One of the ways statistical researchers have chosen to look at these methods is via something akin to the risk-utility tradeoff, with different aggregate criteria to assess disclosure risk and different measures of data utility, e.g., see Trottini and Fienberg (2002), Duncan and Stokes (2009), and Cox et al. (2011). Few of these approaches measure up to the strictness of the differential privacy approach, and when differential privacy is overlaid upon them utility tends to be undermined. For example, see Barak et al. (2007) on making marginal releases from contingency tables differentially private and Fienberg et al. (2010) on the impact of doing so on utility. Similarly, Charest (2010) describes how making multiply-imputed data differentially private affects their utility.

Sampling with top-coding and data swapping and some noise addition have, for example, formed the basis for release of the census long-form and now the American Community Survey PUMS by the Census Bureau. There are no simple criteria that are useful in assessing either the risk of disclosure from such files (although I know of no evidence that the PUMS are unsafe) nor their utility or inaccuracies, e.g., see Alexander et al. (2010). What does seem pretty clear is that current methods of Bureau data release would not meet privacy protection criteria such as differential privacy, nor do I think they should.

In his brief discussion of disclosure limitation methodology, and in the related discussion of the synthetic LED files in Section 6 of his paper, Gates suggests that the goal of such files is that they have the same specified statistical properties as the true microdata. This clearly is not correct. Adding noise to data or altering them in other ways of necessity adds uncertainty and possibly bias to anything we hope to estimate. Consider a regression situation where we add normal noise with zero mean to both the the outcome variable, y, and the predictors, $\mathbf{x}$. The noise added to y increases the error variance and thus the uncertainty associated with the regression coefficients. The noise added to $\mathbf{x}$ changes the standard regression problem into one involving "errors in the variables," and this produces more uncertainty as well as bias. Different methods are required to deal with estimation in this new setting as the statistical literature makes abundantly clear. The same is true of essentially all other methods mentioned here, e.g., see Raghunathan et al. (2003) on multiple imputation. Thus our goal in confidentiality limitation is to to be able to minimize the added uncertainty and to have in hand methods to remove any resulting bias associated with the transformed data.

4 Statistical Aspects of Record Linkage

While I agree with Gates that administrative records will play an increasingly important role for statistical agencies in the future, we must all recognize that record linkage across administrative databases is a highly non-trivial statistical activity. While the basic methodology of record linkage goes back to pioneering papers by Newcombe et al. (1959) and Fellegi and Sunter (1969), and the methodology has been refined over the years by many at the Census Bureau such as Winkler (2006), there remain major obstacles to accurate linkage and implications regarding the uncertainty of linkage for the subsequent analysis of linked files: (1) different units of analysis, e.g., households vs. taxpayers, (2) different frames, i.e., differential population coverage, (3) errors in databases, e.g., spelling and transcription errors,(4) timeliness of files and information in them, (4) statistical errors associated with the probabilistic record linkage algorithms, etc. For more details, see Herzog et al. (2007, 2010).

Gates refers to the Statistical Administrative Records System (StARS) that was built as a component of the 2000 Administrative Records Experiment from the 2000 decennial census. In StARS, the Bureau attempted to merge information from six different administrative lists. In essence what the Bureau did was match lists in pairs and then attempt to resolve discrepancies, e.g., when a record in list A matched one in list B and also one in list C, but where the two in lists B and C didn't match. To do this more systematically, one needs methods for multiple record linkage such as those described by Sadinle et al. (2011).

Record linkage errors propagate into the linked files. This has two implications. (1) protection of confidentiality, and (2) additional uncertainty in analyses of the resulting linked files. Although the fact is not widely recognized in the disclosure limitation literature, the uncertainties associated with the linkage and the probabilistic linkage model clearly offer some measure of protection, although we do not really know how much. This protection may be counterbalanced by added risk of disclosure because of the availability of more information on the individuals with data in the linked files. Thus I ask the question: What does it mean to "ensure that confidentiality is protected" for linked data? An equally important issue is the need to carry forward the uncertainties into the analyses of linked data, and not simply treat the linked files as if they had been gathered without error. Designing record linkage methods to optimize the analyses raises new and related research questions. See the related work of Scheuren and Winkler (1993). Both of these classes of problems cry out for additional research.

5 Towards Expanded Access to Linked Data

As the Census Bureau and other U.S. statistical agencies move forward to utilize administrative records and other information from public and private sources, we will face increasing efforts to control the linked files, especially through restricted data centers. This will be in part a consequence of legitimate confidentiality concerns and legal restrictions. As I indicate above, an expansion of the current research on confidentiality protection should help alleviate the concerns, but education of non-statisticians and perhaps changes in legal restrictions will still be necessary.

But I think that, as part and parcel of such developments and changes, we will also need new forms of restricted access to such files for legitimate researchers that do not force them to comply with the physical and technical restrictions associated with Census Bureau Data Research Centers and those of other statistical agencies. Researchers need to be able to use new methodology that will not necessarily be understood by those who control the linked databases and they will need to be able to carry out analyses remotely, but in a way that continues to protect in a reasonable fashion data released for use in publications. This poses yet another set of technical challenges.

As I noted at the outset of my comments, the world of official statistics stands at a precipice, and we need new ways of thinking about statistical agency data and new technical approaches. Demands for data will escalate and data collection methods will need to change. Methods for confidentiality protection and data access will also need to change.

Acknowledgments

The research reported here was supported in part by Army contract DAAD19-02-1-3-0389 to Cylab at Carnegie Mellon University and in part by NSF Grant BCS0941518 to the Department of Statistics, Carnegie Mellon University.

References

Alexander, J. T., Davern, M., and Stevenson, B. (2010). The Polls—Review. Inaccurate age and sex data in the census PUMS files: Evidence and implications. *Public Opinion Quarterly*, 74(3): 551–569.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. *PODS 2007*, 273–282.

Charest, A.-S. (2010). How can we analyze differentially-private synthetic datasets? *Journal of Privacy and Confidentiality*, 2(2): 21–33.

Cox, L. H., Karr, A. F., and Kinney, S. K. (2011). Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act (with discussion). *International Statistical Review*, 79(2): 160–199.

Duncan, G. and Stokes, L. (2009). Data masking for disclosure limitation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1: 83–92.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Third Theory of Cryptography Conference (TCC 2006)*. New York: Springer.

Dwork, C. and Smith, A. (2009). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2): 135–154.

Fellegi, I. P. and Sunter, A. B. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 40: 1183–1210.

Fienberg, S. E. and Prewitt, K. (2010). Save your census. *Nature*, 466(26): 1043.

Fienberg, S. E., Rinaldo, A., and Yang, X. (2010). Differential privacy and the risk-utility tradeoff for multidimensional contingency tables. In Domingo-Ferrer, J. and Magkos, E. (eds.), *Privacy in Statistical Databases 2010 (PSD 2010),* volume 6344 of *LNCS*, 187–199. Springer.

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

— (2010). Record linkage (Advanced Review). *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2): 83–92.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic linkage of vital records. *Science*, 130: 954–959.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, 79(1): 119–158.

Prewitt, K. (2011). Why it matters to distinguish between privacy & confidentialty. *Journal of Privacy and Confidentiality*, 3(2): 41–47.

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

Sadinle, M., Hall, R., and Fienberg, S. E. (2011). Approaches to multiple record linkage. In *Proceedings of ISI (2011)*. Dublin, Ireland.

Scheuren, F. J. and Winkler, W. E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19: 39–58.

Trottini, M. and Fienberg, S. E. (2002). Modelling user uncertainty for disclosure risk and data utility. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5): 511–528.

Winkler, W. E. (2006). Overview of record linkage and current research direction. Research Report Series (Statistics No. 2006-2), Statistical Research Division, U.S. Census Bureau, Washington, DC.