The Survey Statistician, 2021, Vol. 83, 23-26.

Official Statistics at the Crossroads: Data Quality and Access in an Era of Heightened Privacy Risk[1]

John M. Abowd[2]

[2] Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau, john.maron.abowd@census.gov

Abstract

This paper discusses the challenging problem of balancing the competing interests of access to high-quality statistical data and privacy protection. The paper argues that an optimal choice must account for the preferences of data users and providers, and that technology developed by cryptographers, such as differential privacy, may help us find efficient algorithms for implementing such an optimal choice.

Keywords: confidentiality, privacy, data quality, data access

There are many stakeholders in the conversation about the appropriate response to heightened privacy risk. In the United States, the American Statistical Association, the Committee on National Statistics (a standing committee of the National Academy of Sciences), and the Chief Statistician of the U.S. (in the Office of Management and Budget), have all furthered the discussions of privacy protection. Statistical agency experts and directors that I have briefed over the course of the last four years, since I took this job at the Census Bureau, have also provided their frank and illuminating discussion of the issues that they face in attempting to modernize their disclosure avoidance methods.

Data users, including those from academia, research organizations, industry, and inside the Census Bureau, have also joined the discussion. Serious data users understand why it is worth multiple billions of dollars to conduct a population census. They know that conducting a quality census is more than just a statement by the statistical agency for public relations. They know that census data in the U.S. allocate at least six hundred and seventy-five billion dollars of federal funds every year with some estimates placing the annual value closer to one and a half trillion dollars. They know that the U.S. House of Representatives is apportioned based on data from the census, reminding us that it is important to conduct our censuses and surveys in a manner that maintains the integrity of those data and that allows the user community to have faith in their fitness for use. This is part of the U.S. Census Bureau's dual mandate.

The other part of that dual mandate is to protect the confidentiality of the respondents and the data that they provided. This is a challenging problem, and we all have to acknowledge that both sides of this discussion – really the entire continuum within this discussion – bring legitimate viewpoints to the table that must be respected and considered before making final decisions. It is important for me to acknowledge that, and I think most statisticians and researchers reading this article would not find anything controversial in that acknowledgement.

Traditional statistical disclosure limitation is broken. That does not mean it usually fails. It does not usually fail. It does have significant vulnerabilities exposed by the cryptographers who migrated from

[1] Based on a talk originally delivered at the Joint Statistical Meetings, Denver, CO, USA on July 30, 2019. The views expressed in this article are those of the author and not the U.S. Census Bureau.

Copyright © 2021 John M. Abowd. Published by International Association of Survey Statisticians. This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

computer science into safe data publication. We must now address those vulnerabilities. They are real. They are documented in thousands of carefully-prepared, well peer-reviewed scientific papers that explain what the vulnerabilities mean and why the traditional methods are so exposed—not in an extreme sense, but vulnerable in precisely the sense that computer scientists have properly defined.

There is a very steep learning curve for the official statistics community to climb, because these methods come from a different scientific tradition and involve very different methodologies. Extremely talented mathematical statisticians – well versed in the theory that underlies our statistical analyses – in particular, estimation based on complex multi-stage probability samples – were not exposed to the mathematical reasoning that underlies differential privacy and formal privacy systems. That is just a fact – not a statement of incompetence on anyone's part. My own mathematical background also did not include most of the tools necessary to understand the arguments that the cryptographers were making. But, we do need to face up to those vulnerabilities – we need to rethink how we approach confidentiality protection, and we need to do it so that our future disclosure limitation systems can deliver the same promise of quality and confidentiality protection that they delivered when they were originally conceptualized, primarily by mathematical statisticians in the 1970s. So now, let's dive into it.

Privacy protection is an economic problem. It is not a computer science problem; it is not a statistics problem; it is an *economic problem*. It is about the allocation of a scarce resource – the information in the confidential data that statistical agencies collected – between two competing uses: public data products and privacy protection. The confidential data are a scarce resource because they are finite. If finite, that must mean that the published data products can fully consume the information leaving no privacy protection, if we are not careful. That is precisely what the economic analysis of confidentiality protection teaches us: computer science informs the technology for transforming confidential data into useful information products, and making the publication algorithms produce accurate, fit-for-use information products that also protect confidentiality is a function of using good computer science.

Computer science has thus defined the production possibility frontier between privacy protection and accuracy of publications just the same way as the toy examples in your Introductory Microeconomics class defined the guns and butter production possibility frontier: if you consume more guns, you will have less butter. If we consume more accuracy, we will have less privacy – that is a mathematical fact. If we do it carelessly, then we will inefficiently spend privacy-loss, as I prefer to call it, and not get as much accuracy as we could. If we do it carefully, then we will use algorithms that are on that production possibility frontier – algorithms that are efficient. But if you claim that you can get more accuracy than an efficient privacy-enhancing data publishing technique, you are claiming something that is mathematically false. The comparable claim that traditional statistical disclosure limitation can be more accurate and just as privacy-preserving is also mathematically false. The traditional methods are Pareto dominated, in the economic sense, by the formally private methods, of which differential privacy is the leading example. That means the traditional methods are on the interior of the production possibility frontier – you can improve the accuracy, reduce the privacy loss or both at no cost by moving to the efficient frontier. The technology that cryptographers brought to data publication describes the efficient frontier. It is not constant. It changes every day with new research. This research can, and does, make the algorithms more efficient. It pushes the production possibility frontier outwards. That is the technology side.

What does it mean to have an optimal balance of accuracy relative to privacy protection? Answering this question requires more than technology. An optimal choice must account for the preferences of data users and providers. It must summarize the extraordinarily heterogeneous costs to the providers in terms of their privacy loss and the equally heterogeneous benefits to the users in terms of the accuracy of the publications for their intended uses. In short, we have to balance the social preferences of data users and providers to determine an optimal trade-off.

I want you to look yourself in the mirror now, because I have. How much weight do you personally put on data accuracy versus privacy protection? In your way of thinking about the world, whose job is it to protect the privacy interests of the data contributors? I think most official statisticians would say that is our job. Whose job is it to protect the fitness for use of the data products that we release? Again, I think most official statisticians would say that is also our job. Those are *competing* interests, and so it must be our job to balance those competing interests. We do not have a complete repertoire of tools with which to perform that job. We need to fix that, and that is a critical part of our research mission.

I want you to ask yourself this question. If Facebook said "*If you think you have re-identified someone in public data that we released for research purposes, you can't be sure that you are correct because we used disclosure limitation techniques for which we cannot give you the details,*" what would you say?

I also want you to ask yourself this question. If you were refereeing a scientific paper, and the author said "*My inferences may not be valid, because the agency that provided data access did not release details sufficient to correct for bias and variability due to statistical disclosure limitation,*" what would you say to the editor?

I think I can guess the answer to both of those questions. We have to find a way out of this situation. Statistical agencies should not behave in a manner that we would find unacceptable for Facebook or scientific journals. Official statisticians cannot continue to ask the users of our data to ignore the things that we have to do to protect confidentiality. We need to provide data analysis systems that are statistically valid – systems where the inferences are correct according the appropriate underlying mathematical theory, just as we did for design-based estimation from probability samples. We must be able to say that we protected the confidentiality of these data using these algorithms with these parameters. Statistical disclosure limitation should not remain a back-room operation exempt from the scrutiny of either the providers or the users. Users and providers must assess the quality of our confidentiality protections. National statistical agencies should correct, strengthen, or otherwise adjust their publication systems based on the reasoned analyses of both data and privacy advocates. Speaking only for myself, our research and methodology ought to step up to this challenge.

Selected references

The database reconstruction vulnerability:

Dinur, I. and Nissi, K. (2003) Revealing information while preserving privacy. In: *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03)*. ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.

Differential privacy:

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating noise to sensitivity in private data analysis." In: *Theory of Cryptography Conference* (eds. S. Halevi and T. Rabin), pp. 265-284. Springer Berlin Heidelberg. DOI: 10.1007/11681878_14.

Dwork, C. (2006) Differential Privacy, *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, Springer Verlag, 4052, 1-12, ISBN: 3-540-35907-9.

Dwork, C. and Roth, A. (2014) The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, **9**, 211–407, DOI: 10.1561/0400000042.

Economic analysis:

Abowd, J.M. and Schmutte, I.M. (2019) An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, **109**, 171-202, DOI:10.1257/aer.20170627. [AER, ArXiv preprint, Replication information]

Abowd, J.M., Schmutte, I.M., Sexton, W. and Vilhuber. L. (2019) Why the economics profession must actively participate in the privacy protection debate. *American Economic Association: Papers and Proceedings*, **109,** 397-402, DOI:10.1257/pandp.20191106. [download preprint].

Kifer, D. and Machanavajjhala, A. (2011) No free lunch in data privacy. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11)*. ACM, New York, NY, USA, 193-204. DOI:10.1145/1989323.1989345.

More information on the disclosure limitation system for the 2020 Census of Population and Housing in the United States:

Main U.S. Census Bureau web page on statistical disclosure limitation for the 2020 Census https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

International Conference on Machine Learning keynote address: Abowd, J.M. (2019) The Census Bureau tries to be a good data Steward in the 21st century. *International Conference on Machine Learning (ICML)*, keynote address. [video, start at minute 18:00]