Full text of John M. Abowd's remarks "Official Statistics at the Crossroads: Data Quality and Access in an Era or Heightened Privacy Risk"

July 30, 2019

Joint Statistical Meetings, Denver, CO, Session 389 (CC-703)

Good afternoon. I want to start by thanking the Government Statistics Section, the Committee on National Statistics, and the ASA Committee on Professional Ethics for organizing this session. I also want to acknowledge the efforts that Nancy Potok, as the Chief Statistician, has made in furthering our discussions of privacy protection, and that all of the agency experts and heads that I've had to opportunity to talk to over the course of the last three years, since I took this job at the Census Bureau in discussing frankly and very openly the issues that they face in attempting to modernize their disclosure avoidance protections.

And I want to thank the data users, including those from at IPUMS, ICPSR, and inside the headquarters building at the Census Bureau. Serious data users understand why it is worth multiple billions of dollars to conduct a census, that it is more than just a statement by the Census Bureau for public relations, that six hundred and seventy-five billion dollars of federal funds are allocated in part every year based on data from the census, that the House of Representatives is apportioned based on data from the census, reminding all of us how important it is 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. It is part of our dual mandate.

And the other part of our dual mandate is to protect the confidentiality of the respondents and the data that they've provided in doing so. It is a challenging problem, and we all have to acknowledge that both sides of this discussion, or the entire continuum of this discussion, bring legitimate viewpoints to the table that have to be respected and considered before making final decisions. I think it is important for me to acknowledge that, I think most people in this room would not find anything controversial in that acknowledgement.

That said, traditional disclosure limitation is broken. That's not the same thing as saying that it usually fails. It doesn't usually fail. It has vulnerabilities that have been exposed by the group of cryptographers who migrated from computer science into safe publication of data, and those vulnerabilities now need to be addressed. Those vulnerabilities are real. They are documented in carefully-prepared, well peer-reviewed scientific papers that explain both what they mean and why the traditional methods are so vulnerable. Not in an extreme sense, but vulnerable in the sense that the computer scientists are precisely defining.

There is a very steep learning curve for the official statistics community to walk up 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, particularly multi-stage complex probability samples—haven’t been exposed through most of their career to the mathematics that underlies differential privacy and formal privacy systems. That’s just a fact; that’s not a statement of incompetence on anyone’s part. My own mathematical background did not include most of the tools necessary to understand the privacy arguments that the computer scientists 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 avoidance systems can deliver the same promise of quality and 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’s not a computer science problem; it’s not a statistics problem; it’s an economic problem. It is about the allocation of a scarce resource, namely the confidential data that we invested, in the case of a decennial census, fifteen billion dollars in assembling. That is a scarce resource because it is finite. And, if it is finite, that must mean that it can be used up if you’re not careful.

That’s precisely what the economic analysis of confidentiality protection teaches us: that the technology for transforming confidential data into useful information can be informed by computer science, indeed has been elaborately informed by computer science, and making the algorithms that transform confidential data into accurate fit-for-use public products is a function of using good computer science.

The computer science defines the production possibility frontier just the same way as those toy examples in your Intro Micro class between guns and butter define the production possibility frontier: if you consume more guns, you will have less butter. If you consume more accuracy, you will have less privacy—that is also a mathematical fact. If you do it carelessly, then you will inefficiently spend privacy, or ‘privacy loss’ as I prefer to say it, and not get as much accuracy as you could. If you do it carefully, then you will use algorithms that are on that production possibility frontier—algorithms that are efficient. But if you claim that you can get more accuracy out of an efficient privacy-enhancing data analysis technique, you’re 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 dominated 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 either improve the accuracy, or reduce the privacy loss, at no cost, by moving to the efficient frontier. The efficient frontier is described by the technology that the cryptographers brought in to data publication, but it’s not constant, it’s changing. Research can, and does, make the algorithms more efficient. It pushes the production possibility frontier outwards.

So, that’s the technology side. The other side of the problem is: what does it mean to say that you’ve made an optimal choice of accuracy relative to privacy protection? That has nothing to do with technology. That has to do with describing the preferences of the users and the contributors of the data in a manner that lets you summarize what the costs are to the providers of the data in terms of their privacy loss versus what the benefits are to the users in terms of the accuracy of the data. That is not described by the technology. That trade-off is described by preferences, and preferences are extraordinarily heterogeneous in this area.

I want you to look yourself in the mirror when you go home, because I have. How much weight do you 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 of us would say it’s our job. Whose job is it to protect the fitness for use of the data products that we release? Well, I think most of us would say it’s also our job. Those are *competing* interests, and so it’s our job to balance those competing interests.

I think it will be clear from today’s discussion that we don’t have a complete repertoire of tools with which to perform that job, and we need to fix that, and that’s part of our research mission.

I also want you to ask yourself: If Facebook said the following, *quote*, “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 use disclosure limitation techniques, for which we cannot give you the details” *unquote*. What would you say?

If you are refereeing a scientific paper, and the author said, *quote*, “My inferences may not be valid, because the agency that provided access did not release details sufficient to correct for bias and variability due to statistical disclosure limitation.” *unquote*. What would you say to the editor?

We have to find a way out of this situation. We can’t behave in a manner that we would not find acceptable for Facebook or journal editors, and we can’t continue to ask the users of our data to ignore the things that we have to do to protect confidentiality. We need to give them data analysis systems that are statistically valid, meaning the inferences that are made are correct according the mathematical theory they were constructed from, and we need to be able to say we protected the confidentiality in these data using these algorithms with these parameters, and you may assess the quality of that confidentiality protection as much as you wish and we are open to comments about whether it should be corrected because it was wrong, strengthened, or adjusted. I think that, speaking only for myself, that's a research and technology mission that we ought to step up to.

At the conclusion of my remarks, I read the July 30, 2019 blog post by Ron Jarmin, Deputy Director, U.S. Census Bureau (source: https://www.census.gov/newsroom/blogs/random-samplings/2019/07/boost-safeguards.html):

"The U.S. Census Bureau takes its responsibilities for data stewardship very seriously. As we have previously discussed, we are working diligently to honor our sworn oath to keep respondent data strictly confidential by implementing differential privacy for our 2020 Census data products.

The Census Bureau is rigorously testing this modernization of our disclosure avoidance methods for 2020 Census products. Because of priority of the decennial count, these data products will be the Bureau's differential privacy focus for the near future.

The data user community has raised many questions about the impact the adoption of differential privacy may have on the data products generated by the American Community Survey (ACS). ACS data products are critical to public, private, and not-for-profit sectors. Given the complexity of implementing differential privacy for a complex survey like the ACS, we anticipate that the earliest we would implement differential privacy for the ACS would be 2025.

The solutions will be thoroughly vetted within the scientific and user communities. We will continue to apply the vigorous traditional disclosure avoidance methods we have always applied to ACS. Those methods are reviewed and strengthened every year, and meet the high standards of the Census Bureau's Disclosure Review Board."

Thank you very much.