9-6-2016

How Will Statistical Agencies Operate When All Data Are Private?

John M. Abowd
Cornell University, John.Abowd@cornell.edu

How Will Statistical Agencies Operate When All Data Are Private?

Abstract

The dual problems of respecting citizen privacy and protecting the confidentiality of their data have become hopelessly conflated in the "Big Data" era. There are orders of magnitude more data outside an agency's firewall than inside it—compromising the integrity of traditional statistical disclosure limitation methods. And increasingly the information processed by the agency was "asked" in a context wholly outside the agency's operations—blurring the distinction between what was asked and what is published. Already, private businesses like Microsoft, Google and Apple recognize that cybersecurity (safeguarding the integrity and access controls for internal data) and privacy protection (ensuring that what is published does not reveal too much about any person or business) are two sides of the same coin. This is a paradigm-shifting moment for statistical agencies.

Comments

Published in Journal of Privacy and Confidentiality Volume 7 (2015-2017) Iss. 3 (2017):1-15. http://repository.cmu.edu/jpc/vol7/iss3/1/ .

Abowd acknowledges support through NSF Grants 1131848 (NCRN) and 1012593 (TC:Large), and through the Labor Dynamics Institute.

Washington Statistical Society Julius Shiskin Memorial Award Seminar

How Will Statistical Agencies Operate When All Data Are Private?

John M. Abowd[1]

September 6, 2016

Abstract

The dual problems of respecting citizen privacy and protecting the confidentiality of their data have become hopelessly conflated in the "Big Data" era. There are orders of magnitude more data outside an agency's firewall than inside it—compromising the integrity of traditional statistical disclosure limitation methods. And increasingly the information processed by the agency was "asked" in a context wholly outside the agency's operations—blurring the distinction between what was asked and what is published. Already, private businesses like Microsoft, Google and Apple recognize that cybersecurity (safeguarding the integrity and access controls for internal data) and privacy protection (ensuring that what is published does not reveal too much about any person or business) are two sides of the same coin. This is a paradigm-shifting moment for statistical agencies.

Preliminaries

I would like to thank the Washington Statistical Society, the National Association for Business Economics, and the Business and Economics Statistics Section of the American Statistical Association for honoring me with the 2016 Julius Shiskin Memorial Award. The award explicitly cites my contributions to Statistical Disclosure Limitation (SDL) and increasing access to valuable individual, business, and job microdata. I want to take a moment at the beginning of this talk to acknowledge several colleagues, past and present, without whom much of this work would never have occurred.

The first of these is Nancy Gordon, who was the Associate Director for Demographic Programs when I first joined forces with the Census Bureau in 1998 and who remains a trusted friend and colleague to this day. I consider her my mentor in the federal statistical system. Nancy was part of a team of Associate Directors that included current Census Bureau Director John Thompson and Deputy Director Nancy Potok. Those senior Census executives understood that the Bureau needed to modernize and that modernization meant looking beyond traditional business models. But Nancy understood two other things that she articulated as well as anyone in the federal statistical system: (1) new methods of collecting microdata, whether it be the continuous measurement of the American Community Survey or the administrative records-based Longitudinal Employer-Household Dynamics Program—two programs she particularly championed—would require creative new methods for accessing and publishing data and (2) the research activities associated with modernizing access and publication methods—specifically confidentiality protection through multiple access modes and creative new SDL methods—had to be

[1]Associate Director for Research and Methodology and Chief Scientist, U.S. Census Bureau and Edmund Ezra Day Professor of Economics, Statistics and Information Science, Cornell University. The opinions expressed in this talk are my own, not those of the Census Bureau. Research support from the National Science Foundation (SES 1131848, TC 1012593), the Alfred P. Sloan Foundation, and the U.S. Census Bureau (prior to my appointment) is gratefully acknowledged. I have benefitted from numerous conversations with Cynthia Dwork, Steve Fienberg, Matthew Graham, Sam Haney, Dan Kifer, Mark Kutzbach, Ashwin Machanavajjhala, Kobbi Nissim, Jerry Reiter, Ian Schmutte, and Lars Vilhuber.

done by those with a subject-matter interest in the data. It took me 16 years to prove the theorem that embodies those insights, and yes I am going to show it to you today. Thank you, Nancy, for dragging me, sometimes quite reluctantly, into this arena. My understanding of how the responsibilities of an official statistical agency should be discharged owes a huge debt to you.

The second person I want to acknowledge is my long-time Cornell colleague Lars Vilhuber. Lars is my intellectual grandchild—his thesis supervisors at the Université de Montréal were two of my former doctoral students. He came to the Census Bureau in 1999 as part of the first team of economists hired by the LEHD Program.

It was clear from the very beginning that Lars was cut from different cloth than most of the other economists or survey statisticians at the Census Bureau. He understood the intimate connection between doing research on confidential microdata and documenting all the sources and methods to standards that would permit a researcher who was completely unfamiliar with any part of the work to reproduce it beginning with the original inputs as delivered from their original sources. This duty is an imperative for any organization that grants research access to its confidential microdata—if the scientific community, which cannot directly peer-review this research, loses confidence in its reproducibility, it will be very difficult to restore that confidence.

In 1999, Lars, in collaboration with my colleagues at the Cornell Institute for Social and Economic Research, designed and implemented the first Virtual Research Data Center, called the Cornell Restricted-Access Data Center. After he had left the Census Bureau and was working as a Cornell employee under contract to the Bureau, he and two other Census contractor colleagues designed the metadata database that still drives the LEHD data acquisition and production systems—a marvel that is able to reproduce any of the released statistics from the Quarterly Workforce Indicators back to the beginning of time (2003) from the actual inputs and codebase as they existed when that release was produced. These are two seminal early contributions to applied information science that deserve explicit commendation from the social science community. Thank you, Lars, for having the intellectual fortitude to insist that the research infrastructure was as critical to the mission of a statistical agency as its data publications.

I also want to explicitly acknowledge two former students, Simon Woodcock, now at Simon Fraser University, and Ian Schmutte, now at the University of Georgia.

Simon was the one who realized that synthetic data had “legs.” Our early work, published in 2001, laid the groundwork for the synthetic data projects undertaken with the Census Bureau in the mid-2000s. These systems paved the way for the formally private ones we are now developing for flagship Census Bureau publications.

Ian and I are currently working on the social science underlying disclosure limitation. I’m going to be talking about that work extensively today.

Finally, I would like to acknowledge the role of Stephen Fienberg of Carnegie-Mellon University. I’m sure almost everyone in this auditorium can cite a path-breaking contribution of Steve’s that had a major impact on statistics and the federal statistical system. I want to highlight the foresight that he had in gathering researchers from the SDL community and the emerging computer science data-privacy community in Bertinoro, Italy, in 2005. This is where I first met Cynthia Dwork and the team of young cryptographers who were shattering the received wisdom in SDL with methods that Steve recognized as revolutionary. I’m also going to spend much of this lecture on those methods. The last time Steve and I talked about this, at this year’s JSM, he confided to me that our big mistake was that “we did not grow the community fast enough.” I hope this lecture helps solve that problem too.

The Social Science of Data Privacy

To understand why statistical agencies are facing a paradigm-shifting challenge we need to examine the origins of privacy research in economics and confidentiality protection in statistics and computer science.

The intellectual giant of information economics, George Stigler, produced the first rigorous analysis of the economics of privacy in 1980 using, as Richard Posner noted in 1981, a legal analysis of privacy based on an individual's incentive to conceal personal details from trading partners, like banks, and law enforcement agencies—governments. In 1980, more than a decade before the Internet, he said "[g]overnments (at all levels) are now collecting information of a quantity and in a personal detail unknown in history" (p. 623). While acknowledging that governments played a valuable role when they published statistics about their populations, Stigler correctly predicted that the problem would be how to constrain the use of this private information rather than how to defend against its acquisition in the course of law enforcement.

It is not just governments that face a paradigm shift. In the private sector, the rise of Internet-intermediated commerce, especially through market giants like Amazon, Apple, Facebook, Google, Microsoft, Twitter, and the late Yahoo bestowed upon these intermediaries an information advantage over the buyers that Alessandro Acquisti and Hal Varian described, in 2005, as privileging the seller's informational position in a manner that allows price discrimination on a massive scale. That price discrimination happens every time one of these giants offers you a "special deal." A recent survey by Acquisti and co-authors (2016) focuses almost exclusively on the private value of the information that Internet giants have acquired through voluntary exchanges with consumers who shared the information as part of a mutually-beneficial commercial transaction. They spend just a few paragraphs on the scientific value of these commercial databases and similar ones that government agencies now possess. They conclude that "[h]ow to balance researchers' and society's needs to access granular data with the need to protect individuals' records is a question that simultaneously involves economists and scholars in other disciplines, such as statisticians and computer scientists" (p. 43).

Indeed it does.

Data collected by statistical agencies must be published in some form. The enabling legislation for a statistical agency makes that obligation clear. A statistical agency cannot use the data that it acquires to enforce laws. This prohibition lies at the heart of their second statutory obligation: to protect the confidentiality of the identity and attributes of individual and business data sources.

Publications are the public good that justifies the expense of taxpayer revenue on the agency. And the quality of those publications, measured in terms of their usefulness to the society that financed them, is the social benefit from that public good.

Confidentiality protection is also a public good. Or, more pointedly, Stigler's privacy protection—protection from harmful use of private data—is the public good that is also produced when a statistical agency discharges its statutory mission by publishing summaries that have been altered using statistical disclosure limitation methods.

It is easy to see why the quality of an agency's statistical data is a public good. Better quality data allow all users to make better decisions. And when one of these users does so, that person doesn't use up any of the data quality. It is all still there for the next user. There is no rivalry in the consumption of quality statistical data. There is also no plausible excludability. Any person in the world can download data from the American Community Survey, the Economic Census, the National Income Accounts, the Consumer Price Index, and a host of other statistics produced by our agencies.

It is harder to see why privacy protection is a public good, and indeed, it need not be. In this country, and in most western democracies, the ideal of equal protection under the law is an important safeguard built into the justice system. When a statistical agency is charged with protecting the confidentiality of the individuals and businesses from whom it collects data, the equal protection provision matters. The agency cannot, without explicit authorization in law, disadvantage one person by publishing his private data while simultaneously safeguarding the privacy of another. The requirement that every person obligated to supply data is entitled to the same confidentiality protection when those data are published translates into non-rivalry in consumption for privacy protection because the technology is implemented using worst-case analysis. Whatever mechanism an agency designs for protecting the confidentiality of its data safeguards that confidentiality precisely when one person's protection doesn't come at the expense of another person's (or business's). That is non-rivalry in consumption—I can't use up your confidentiality protection by invoking mine. If I could, we wouldn't be equally protected under the legal mandate of Title 13 or CIPSEA.

In a minute, I will begin my discussion of the technologies that limit the feasible production of quality statistics and privacy protection. Before I do, I want to stress that there are always two sides to an economic analysis—supply and demand. The technology determines the supply of quality statistics and privacy protection. What determines the demand?

The Census Bureau produces the American Community Survey, for example, as part of the constitutionally mandated decennial census—the part specifically mandated by the instruction to conduct the enumeration "in such a manner as they [Congress] shall by law direct." What part of American society benefits from the data published as part of the ACS? The answer is clearly those subpopulations, places, towns, counties, cities, states and regions that use the data or the funds allocated in part based on those data to make better decisions and to enforce laws protecting the rights of residents of those places.

The demand for the ACS is a derived demand for a public good—quality data on many subpopulations. The demand for the Economic Censuses is a derived demand for quality data on economic activity as reflected in the National Income and Product Accounts and other measures. The demand for the Current Population Survey is a derived demand for quality data on the state of the labor market at high frequency.

We are used to thinking of the cost of these public goods as the taxpayers' willingness to spend money on data collection as manifested by their representatives' willingness to fund the various data collection activities. But there is a much subtler cost that matters just as much. Holding constant the spending on the ACS, for example, the most accurate way to release the data would be to publish the exact microdata as collected. Such a publication would have no privacy protection, since the ACS is a mandatory survey, and would clearly be a violation of the Census Bureau's current statutory confidentiality protection mandate. But the same Congress that directed the collection of the ACS could also direct its publication in full detail. Congress could legislate that the social interest in data quality outweighs all countervailing interests in privacy protection. It is a social choice about demand not about technology: "Does Congress want to direct statistical agencies to sacrifice some data quality in order to provide privacy protection?" not "Can statistical agencies do this if so directed?"

Congress has mandated both data publication and privacy protection. It has, however, given little guidance as to how a statistical agency might reflect a social choice for better data quality at the expense of some privacy loss. Navigating our way through this social choice is what the second half of this lecture is about. I hope it might influence policymakers into thinking more proactively about how to provide legislative guidance on privacy-loss choices.

What Is an Inferential Disclosure?

Just as I began the section on the social science of data privacy with its intellectual underpinnings in the economics of privacy, I want to begin the analysis of the technology of producing quality data with privacy protections by noting the roots of this problem in statistics and computer science. And unsurprisingly, that takes us right to Ivan Fellegi, in many ways one of the most influential official statisticians of the 20th Century. In 1972, Fellegi showed that publishing too many related summary tables could, if the agency did not take care in their preparation, lead to what he called a "residual disclosure"—what we would now call a differencing (or subtraction) attack in computer science or a complementary disclosure in SDL: a user could exactly reconstruct at least one confidential data record from the published data.

By proving that a combination of primary and complementary suppressions in the published tables could produce publications that were free from residual disclosure, Fellegi provided the first tool in the field of statistical disclosure limitation. And he opened the door for statistical agencies around the world to structure their published tables in a manner that satisfied the prevailing standards of confidentiality protection.

Fellegi also provided the first practical example of what is now known in computer science as the Fundamental Law of Information Reconstruction or, sometimes, the Database Reconstruction Theorem. First proven by Irit Dinur and Kobbi Nissim in 2003, this result states that it is possible to reconstruct any finite confidential database to within an arbitrary level of accuracy using a finite series of queries.

Consider a series of tables protected by primary and complementary suppression. Each time one of these suppressions is relaxed, a record in the underlying database can be, at least partially, reconstructed. In 2010, Scott Holan and his co-authors provided a direct example of how suppression could be almost completely undone by database reconstruction as applied to the Quarterly Census of Employment and Wages. Ian Schmutte and I (2015a) showed similar results for the Quarterly Workforce Indicators and County Business Patterns. None of these reconstructions resulted in an exact disclosure—direct identification of an attribute of a particular respondent because the SDL had been properly implemented. But this line of research taken in combination with the Fundamental Law of Information Reconstruction implies that statistical agencies need stronger theoretical underpinnings for their confidentiality protection technologies. And I'm sure that most of you are aware of famous database reconstructions like the one Arvind Narayanan and Vitaly Shmatikov accomplished in 2008 using the Netflix Prize data, which did result in many exact re-identifications.

Enter the concept of inferential disclosure—the modern method for quantifying the incremental information contained in a data release. First introduced by Tore Dalenius in 1977, an inferential disclosure occurs when the user of a statistic can make too improved an inference about a respondent's identity or attributes once the data have been released as compared to the best possible inference before the statistic was published. In 1982, Shafi Goldwasser and Silvio Micali defined a cryptogram as semantically secure if the information about the actual message (cleartext) that could be extracted from the encrypted text could also be extracted without using the encrypted text.

Perfect inferential disclosure limitation and perfect semantic security are intimately related; indeed, they are really the same thing. And they are both impossible to achieve in any statistical publication. What I like to call the Impossibility Theorem was first proven by Cynthia Dwork and Moni Naor in 2008, but a simpler version of the argument lies behind the original papers in differential privacy: Cynthia's work with Frank McSherry, Kobbi Nissim and Adam Smith in 2006, and her paper entitled "Differential Privacy" in 2006.

The impossibility theorem begins by quantifying the amount of information about any respondent or attribute released in any published statistic using the Bayes factor associated with that statistic when it is computed with and without the contribution of that respondent. The theorem then states that any published statistic for which this Bayes factor is unity for all possible respondents in all possible configurations of the data has provably zero inferential disclosure or, equivalently, full semantic security. That is, any statistic that is perfectly safe—zero inferential disclosure or perfect semantic security—is also perfectly useless: it is a full encryption of the confidential data. It is impossible to fully control inferential disclosure.

It is simple to relate inferential disclosure to the more common notions of identity and attribute disclosure. Inferential disclosure can be quantified by the Bayes factor associated with the hypothesis that a particular respondent's values are or are not in the database used to compute a particular statistic. An exact identity disclosure occurs when the Bayes factor associated with this hypothesis is infinite. A similar definition implies that an exact attribute disclosure has occurred when its associated Bayes factor is infinite.

The result that differential privacy works by bounding a particular Bayes factor has been so poorly understood, and sometimes maligned, in the practice of SDL, that I want to take some time to clarify its meaning and explain why it is the most important result in disclosure limitation since Fellegi's initial contribution.

Consider a confidential database, *D*, where rows are respondents and columns are variables. Consider another database *D'* where the data on respondent *i* has been deleted. The difference between databases *D* and *D'* is, therefore, the data on a single respondent whose row has been deleted from *D'*.

Now consider a pre-defined set of tabulations *F.* We will do this in its simplest form. *F* contains the formulas for computing one statistic from *D*. For simplicity, imagine that *F* allows only subpopulation totals—counting the selected rows of *D* where some combination of the data in that row conforms to the conditions of the query. For example, our counting query could be all men in the database, all women living in Texas with a college education, etc.

Finally, let *M* be a disclosure limitation system that uses output noise infusion. *M* takes as inputs *D* and *F* and outputs the statistic *S* = *a* + *u*. That is, *M* takes *D*, makes the computations according to *F* to produce the exact answer *a*, then adds noise *u*.

Suppose the databases *D* and *D'* both produce the same statistic *S*. The Bayes factor for the hypothesis that respondent *i* was used to compute statistic *S* is just

$$BF_i = \frac{\frac{Pr[D|S,M,F]}{Pr[D'|S,M,F]}}{\frac{Pr[D|M,F]}{Pr[D'|M,F]}}$$

That is, the Bayes factor is equal to the odds for *D* v. *D' a posteriori* divided by the odds *a priori*. The incrementa2l event is the publication of *S*. By convention, we exchange *D* and *D'* so that the Bayes factor is always at least one.

Our publication system *M* might be structured so that the Bayes factor for every respondent *i*, BF_i, is bounded by some constant, say e^{ε}. Our publication system would then have the property, given *D*, that for all *D'*

$$\sup_i BF_i \leq e^{\varepsilon}$$

That is, the Bayes factor associated with the hypothesis that respondent *i* was used to compute the statistic *S* from *D* never exceeds e^{ε}.

We might think that bounding the Bayes factor for all possible respondents in a given database is sufficient to make a quantitative statement about confidentiality protection. But, as in the case of survey design, we really want to know how the system *M* works for any database *D* we might collect—just as we try to understand the theoretical error from a particular survey design over all possible realizations of the sample data. So, let's modify the condition on the Bayes factor so that we can say

$$\sup_{\forall D,D'}\{\sup_i BF_i\} \leq e^{\varepsilon}$$

The databases are defined exactly as above but now the bound says that for any pair of databases *D* and *D'* that differ by the deletion of a single respondent and for any choice of that respondent, the Bayes factor is always bounded by e^{ε}.

A disclosure limitation system *M* that satisfies this Bayes factor bound for all input databases *D* that can be produced by the data collection system under study is called ε-differentially private. It is the leading example of a privacy-preserving data publication system. Some useful generalizations can be found in the Pufferfish system of Daniel Kifer and Ashwin Machanavajjhala (2014).

The bound in ε-differential privacy is a worst-case bound. Why should we use such a worst case as the standard for a disclosure limitation system? The answer lies in a remarkable paper published by Arpita Ghosh and Aaron Roth in 2011. They asked the sensible question: Suppose a scientist wants to compute a simple statistic from a database already held by a trusted custodian. The custodian has said that the scientist may offer monetary compensation to the respondents already in the database to induce them to opt-in to the calculation. The custodian must compute the statistic using ε-differential privacy.

Ghosh and Roth set the problem up so that the scientist determines the minimum expenditure to produce a statistic with given accuracy defined over the whole population. Each respondent has a unique distaste for privacy loss. Each respondent is offered a pair of outcomes $\{\varepsilon_i, p_i\}$—a privacy guarantee and a payment. The solution to the minimum expenditure problem is a Vickrey-Clarke-Groves (VCG) auction in which every willing participant receives the same payment and gets the same level of privacy protection. The payment and protection are determined by the marginal participant in the calculation of the statistic. That marginal participant has the largest distaste for privacy loss of all willing participants, and the number of willing participants is just large enough to achieve the desired accuracy relative to the population value of the statistic.

The result in Ghosh and Roth implies that the privacy loss in ε-differential privacy is non-rival. Every member of the population gets at least the protection offered to the least-willing participant. In order to buy more accuracy, more participants have to opt-into the calculation. To get them to do so, the privacy loss offered to the marginal participant must be reduced (smaller ε). Once that happens, everyone in the population gets more privacy protection because in order to insure that the marginal participant gets the promised level of privacy protection the statistic must be computed using the smallest value of ε offered to any participant. This is the value offered to the marginal participant; hence, that participant is determining the privacy loss for all members of the population, not just himself.

Optimal Data Quality and Privacy Protection

Armed with the knowledge that privacy protection is a non-rival public good, we can now ask the question: What is the optimal level of data quality and privacy protection? This is a different question from the one that Ghosh and Roth asked. They allowed the scientist buying the statistic to set

the level of data quality to meet her desired accuracy. The scientist is, in their setup, a private data supplier. She is going to publish the statistic once for the benefit that it affords her career. Her willingness to pay for that statistic is determined entirely by her private motivation to pay for it, just as Google sells advertising space surrounding search results based on the highest bidder's willingness to pay to advertise to the specific person whose search results are about to be published.

We already saw that data quality, especially the quality of data published by statistical agencies, is a public good. The right amount of data quality to publish might very well be more than any one person would be willing to buy. Ian Schmutte and I (2015b) proved that this intuition is correct. Here is the theorem that Nancy Gordon presaged in 1998.

The technology for jointly producing data quality and privacy protection defines a production possibilities frontier (PPF) along which increased data quality requires increased privacy loss—larger ε. The slope of this PPF measures the marginal social cost of increased data accuracy in terms of foregone privacy protection. Statisticians have another name for this curve—the Risk-Utility map (Duncan and Fienberg 1999). I'm an economist, so I am not going to use the term "utility" that way. They mean "data quality" or "suitability for use," not utility the way economists define the term. It turns out that statisticians also have another name for this curve—the Receiver Operating Characteristics curve. I'll make this analogy clear when I get to my example.

The preferences for data quality and privacy protection come from a social welfare function that aggregates every person's preferences for data accuracy and distaste for privacy loss. The slope of the isovalue curves from this social welfare function defines the marginal willingness to pay for incremental data quality with incremental privacy loss in the whole population.

Optimal data quality and privacy protection occur when the marginal social cost of incremental data quality, measured in terms of incremental privacy loss, equals the marginal willingness to pay for incremental data quality, also measured in terms of incremental privacy loss.

Using this setup, Ian and I proved that the VCG auction from Ghosh and Roth described the correct technology for jointly producing data quality and privacy protection. This means that if a statistical agency uses the PPF associated with the best-known differentially private data publication mechanism, it is provably producing the best quality statistics possible using a technology with formal bounds on the population loss of privacy. If it uses an ad hoc SDL method, it is provably sacrificing either data quality or privacy protection or both.

We also proved that a statistical agency should produce higher quality statistics than a private, profit-maximizing data publisher would produce. The agency should also impose the required increase in privacy loss to achieve the improved quality. A private supplier, like the scientist in the original Ghosh and Roth work, only buys enough data quality to meet her private goals. But the general population can reuse her statistic many times. For example, if the statistic is the conditional cure rate for a particular cancer therapy, one might imagine that many physicians and patients would reuse that number. Similarly, many property developers might reuse data on the proportion of households with income in excess of $200,000/year in a particular place. The sum of the willingness to pay for all of the uses of a statistic determines the demand side of the socially optimal choice of data quality and privacy loss.

Can You Do This with Real Data?

Yes. In fact, the Census Bureau was the very first organization in the world to implement a production data publishing system that was formally private (Machanavajjhala et al. 2008).

Figure 1 (based on results in my 2015b paper with Ian) shows the complete solution for publishing the distribution of income in 1,000 bins using the differentially private algorithm known as

private multiplicative weights. The *x*-axis measures the privacy loss, parameterized as the maximum log Bayes factor for any adult in the United States—the ε in differential privacy. Hence, the *x*-axis measures a public "bad"—privacy loss. The *y*-axis measures data quality stated as the absolute error when the statistic is produced with the differentially private method divided by the absolute error when the statistic is produced with an infinite privacy loss.

The point (0,1) represents the unattainable best outcome: perfect accuracy and no privacy loss. The point (1,0) represents the worst outcome shown: no data accuracy at all with a privacy loss of e^1, a bound on the Bayes factor of 2.72.

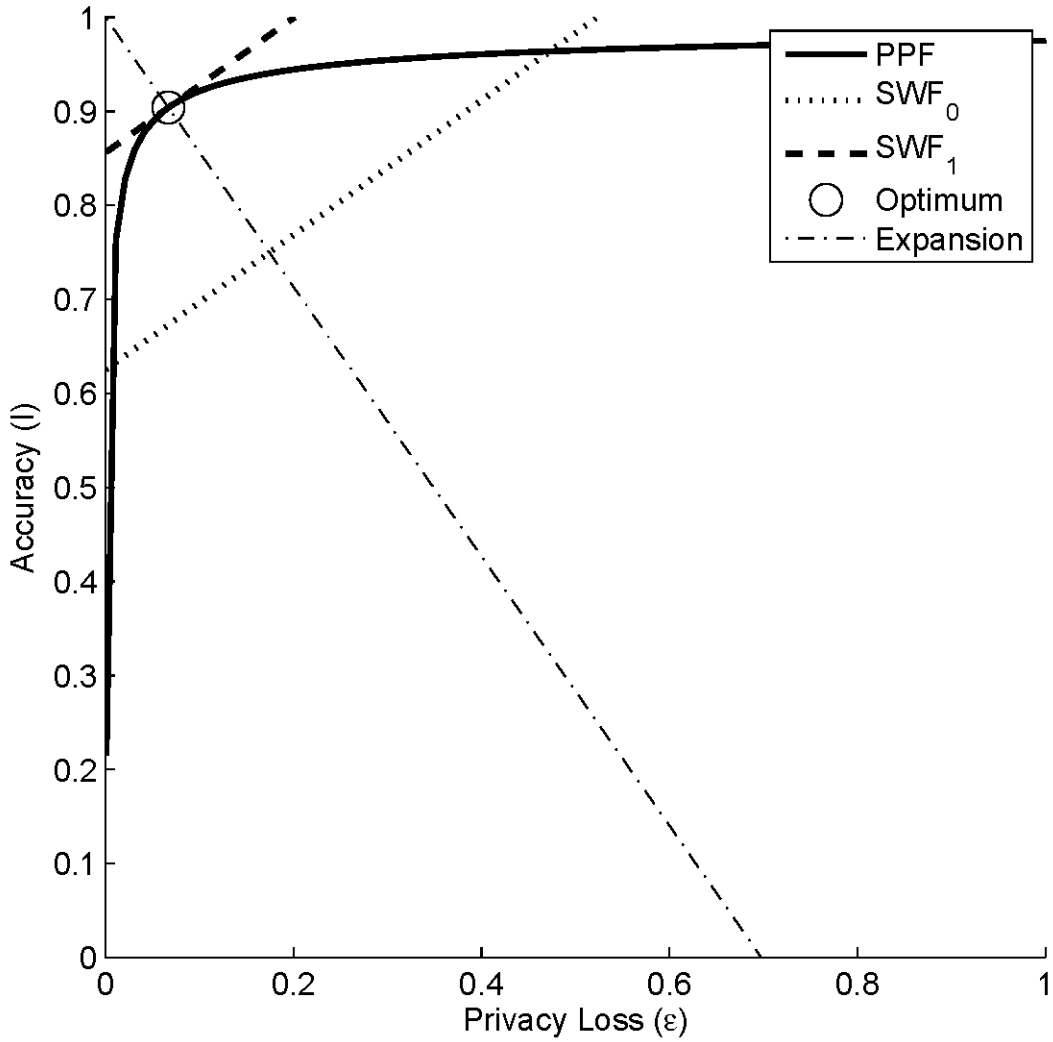

Figure 1
Publishing Income Data with Differential Privacy

The PPF shows the feasible pairs of data quality and privacy loss. It should now be clear why I used the ROC analogy above. We would like to attain the point (0,1), which is also the best outcome in the ROC case—false positive rate of zero and true positive rate of unity.

The PPF in Figure 1 uses accurate parameters for the 2010 population of the United States ages 18-64 varying the ε parameter in the differentially private data publication mechanism. It passes through (0,0), as it must. It asymptotically approaches data quality of one as privacy loss goes to infinity.

Infinite privacy loss means the published data attain the same quality as statistics computed from the confidential data directly.

Two isovalue curves for the social welfare function are shown. The one labeled SWF_0 is just tangent to the PPF. It represents the socially optimal data quality-privacy loss pair. The one labeled SWF_1 is the inferior social welfare level that would be achieved by a private publisher who overprotected privacy by using a differential privacy value that was too low while computing the same statistic.

We estimated the social welfare function using two different sources of data: the 2006 General Social Survey (Smith et al., 2011) and the Federal Statistical System Trust Survey (Childs et al., 2012 and 2015). The pictured results are for the GSS. Marginal distaste for privacy loss was measured by the answer to the following question:

> "The federal government has a lot of different pieces of information about people which computers can bring together very quickly. Is this a very serious threat to individual privacy, a fairly serious threat, not a serious threat, or not a threat at all to individual privacy?"

Marginal preferences for data accuracy were measured by answers to this question:

> "How important [is] the following in making something scientific? The conclusions are based on solid evidence."

The results imply that the adult population of the U.S. prefers data that are 90 percent as accurate as those that would be produced with infinite privacy loss. At the same time, the population prefers privacy loss of $\varepsilon = 0.067$, a Bayes factor of 1.07. The point (0.90, 0.067) is just feasible, and therefore can be attained using the posited differentially private publication mechanism. Results using the FSS Trust Survey, which has somewhat different questions and is much more current, imply that the adult population would prefer slightly less accurate statistics and slightly less privacy loss.

A Research Program for Modernizing Disclosure Limitation

Almost all current disclosure limitation methods used by statistical agencies around the world are based on ad hoc criteria for measuring their effectiveness. They fail the criterion of equal protection under the law because their effectiveness is measured in terms of an agency's best efforts to insure that the ensemble of publications does not violate the confidentiality of any respondents. Those best efforts, while diligently and competently delivered, were predicated on the assumption that most of the information that could be used to compromise the disclosure limitation procedure was inside the agency's firewall.

Such an assumption is simply no longer tenable. It must be replaced by assumptions that allow the agency to release the statistical summaries without fear of future attacks. Formally private disclosure limitation procedures meet this condition. And they are really the only players left standing.

Formally private systems always respect the Fundamental Law of Information Reconstruction. This means that they state a privacy-loss budget for all publications based on a particular confidential database. Then, they provably respect that budget. There are no exact identity or attribute disclosures by construction. All publications cumulatively imply a stated privacy loss. And that privacy loss is guaranteed against all future external data and attacking algorithms.

We need to guard against the violations of the Fundamental Law of Information Reconstruction that arise from doing an incomplete analysis of the data publication problem. I'm thinking, in particular, of the analysis many researchers and agencies do when considering whether it is safe to publish summary results from an additional study using a confidential data source. We usually hear this incomplete analysis as some form of the question "How can my six regression coefficients possibly

compromise the confidentiality of these data?" It is a reasonable reaction, but the argument is flawed because it is a partial equilibrium analysis. Each publication, compared to no publication at all, involves a small but measurable privacy loss. The cumulative effect of this partial equilibrium reasoning is either a strong limitation on the ensemble of publications or a violation of the Fundamental Law of Information Reconstruction. If we can't measure how much privacy loss has already been incurred, can we reasonably hope to make a rational decision about whether the next publication is still safe? And either way, how do we decide which analyses to publish and which to limit in order to respect the privacy-loss budget?

Formally private publication systems automatically allow correct inferences about the published data and automatically share all new results with the public. There is nothing secret about the disclosure limitation process except the seed for the random number generator used to create the noise in the publication tables and microdata. The complete statistical process can be published and combined with other sources of error to produce real measures of total variability. At present, key parameters of the SDL, like swap rates and algorithms, are closely guarded secrets. The published data have deliberate disclosure limitation errors, but we don't talk about them or release any information that might be used to correct inferences even though these systems may have infinite privacy loss in a formal setting. Ad hoc SDL methods like swapping are technically dominated once an agency agrees to limit its publications to those that are within a stated privacy-loss budget, even a very large one like 10 (Bayes factor bound = 22,000).

Formally private publication systems provide statistical agencies with the tools needed to deliver both data quality and privacy protection in a manner consistent with their statutory missions. They do not help the agencies decide which data quality, privacy-loss combinations to use. Those decisions require a model of the demand-side of data publication. The computer scientists are silent on this subject. They are waiting for the social scientists to provide models and data suitable for making this choice. I have just shown you some examples. Let's not keep the computer scientists, or the statistical agencies, waiting any longer for better ones. Thank you.

References

Abowd, John and Ian Schmutte (2015a) "Economic Analysis and Statistical Disclosure Limitation," Brookings Papers on Economic Activity (Spring): 221-267.

Abowd, John and Ian Schmutte (2015b) "Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods," Cornell University Labor Dynamics Institute (curated preprint) http://digitalcommons.ilr.cornell.edu/ldi/22/.

Abowd, John M. and Simon Woodcock (2001) "Disclosure Limitation in Longitudinal Linked Data," in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), 215-277.

Acquisti, Alessandro, Curtis Taylor and Liad Wagman (2016) The Economics of Privacy, Journal of Economic Literature 54(2): 442–92.

Acquisti, Alessandro and Hal R. Varian (2005) Conditioning Prices on Purchase History, Marketing Science 24(3): 367–381.

Childs, Jennifer H., Stephanie Willson, Shelly W. Martinez, Laura Rasmussen and Monica Wroblewski (2012) "Development of the Federal Statistical System Public Opinion Survey," JSM Proceedings Survey Research Methods Section, https://www.amstat.org/sections/srms/proceedings/y2012/files/400242_500695.pdf

Childs, Jennifer H., Ryan King, and Aleia C. Fobia (2015) "Confidence in U.S. federal statistical agencies," Survey Practice 8(5).

Dalenius, Tore (1977) "Towards a Methodology for Statistical Disclosure Control," Statistik Tidskrift 15: 429–444.

Dinur, Irit and Kobbi Nissim (2003) "Revealing Information while Preserving Privacy," Proceedings of the 22nd Symposium on Principles of Database Construction (SIGMOD-SIGACT-SIGART), pp. 202-210, DOI:10.1145/773153.773173.

Duncan, George T. and Stephen E. Fienberg (1999) "Obtaining information while preserving privacy: A Markov perturbation method for tabular data," Statistical Data Protection (SDP 1998), Eurostat, pp. 351–362.

Dwork, Cynthia (2006) "Differential Privacy," Proceedings of the International Colloquium on Automoat, Languages and Programming (ICALP), pp. 1-12.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim and Adam Smith (2006), "Calibrating Noise to Sensitivity in Private Data Analysis," Theory of Cryptography (Lecture Notes in Computer Science 3876), pp. 265-284, DOI: 10.1007/11681878_14.

Dwork, Cynthia and Moni Naor (2008) "On the Difficulties of Disclosure Prevention in Statistical Databases or the Case for Differential Privacy," Journal of Privacy and Confidentiality, pp. 93-107.

Fellegi, Ivan P. (1972) "On the Question of Statistical Confidentiality," Journal of the American Statistical Association 67(337): pp. 7–18.

Ghosh, Arpita and Aaron Roth (2011). Selling privacy at auction, Proceedings of the 12th ACM Conference on Electronic Commerce, EC '11, ACM, New York, NY, USA, pp. 199–208.

Goldwasser, Shafi and Silvio Micali (1982) “Probabilistic Encryption & How to Play Mental Poker Keeping Secret All Partial Information, STOC ’82 Proceedings of the 14th annual ACM Symposium on Theory of Computing, pp. 365–377. URL: http://dl.acm.org/citation.cfm?id=802212

Holan, Scott H., Daniell Toth, Marco A. R. Ferreira and Alan F. Karr (2010) “Bayesian Multiscale Multiple Imputation with Implications for Data Confidentiality,” Journal of the American Statistical Association, Vol. 105, No. 490 (June): 564-577.

Kifer, Daniel and Ashwin Machanavajjhala (2014) “Pufferfish: A framework for mathematical privacy definitions,” ACM Transactions on Database Systems (TODS), Vol. 39, No. 1 (January), DOI: 10.1145/2514689.

Machanavajjhala Ashwin, Daniel Kifer, , John M. Abowd, Johannes Gehrke and Lars Vilhuber “Privacy: Theory Meets Practice on the Map,” International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.

Posner, R. A. (1981) “The Economics of Privacy,” American Economic Review, Vol. 71, No. 2 (May): 405–409.

Smith, T.W., Marsden, P., Hout, M. and Kim, J. (2011) “General Social Surveys, 1972-2010: cumulative codebook,” National Data Program for the Social Sciences Series, no. 21, Chicago: National Opinion Research Center.

Stigler, George J. (1980) “An Introduction to Privacy in Economics and Politics,” Journal of Legal Studies 9(4): 623–644.