Working Draft

Reconciling Access and Privacy: Building a Sustainable Model for the Future*

Katharine G. Abraham
University of Maryland

January 2019

*Without implicating them in any way, I thank John Abowd, Paul Ohm, Eric Slud, and Latanya Sweeney for past conversations about the reconciliation of data access and privacy protection.

In recent decades, social science researchers have benefitted enormously from access to survey and Census data collected and disseminated by the Federal statistical agencies. The research made possible by this access has generated invaluable insights. An important factor in the government's ability to collect data from individuals and businesses is the promise it gives to data subjects that their information will be kept private, and the statistical agencies are vigilant in their efforts to honor this promise. Given the explosion of data from numerous sources that increasingly is available in electronic form, however, the risk that information contained in data products released by federal agencies could compromise the privacy of data subjects has grown. I view it as an unavoidable conclusion that, in order to honor the promises of privacy made to data subjects, current modes for disseminating information based on survey and Census data will need to be rethought.

Tiered access seems certain to be a central feature of any new model for data access, with the needs of many data users met through tabulations or other data products that can safely be made public and behind-the-firewall access to more sensitive information provided to the smaller number of data users who truly require it. A similar mix of approaches can be used to increase access to administrative records for research purposes. While the broad outlines of what a new system will look like seem relatively clear, important practical questions about its implementation will need to be addressed.

Limitations of Existing Approaches to Statistical Disclosure Limitation

The problems associated with releasing individual-level microdata are by now well known. In a data file containing even a modest number of characteristics, a significant fraction of people are population uniques, meaning that it would be easy for an acquaintance or someone who could obtain information about the person from other sources to identify them in the data.

Sweeney (2000), for example, concluded that 87% of individuals in the 1990 Census were uniquely identifiable based on their gender, zip code and date of birth. Precisely because of the re-identification risk that would be created, statistical agency data microdata releases generally do not include information such as zip code or exact date of birth. The inclusion of characteristics that uniquely identify a much smaller share of the population nonetheless could imply a large number of identifiable individuals. It might sound like a small problem for, say, 0.5 percent of the population to be uniquely identifiable given a particular set of characteristics, but in a country the size of the United States, that would be more than 1½ million people. Re-identification in a federal data release in turn could facilitate the identification of information about a person in other record systems. Though generally considered to be less risky, detailed tabulations also may reveal sensitive information about individuals with unusual characteristics.

The challenges associated with protecting the confidentiality of business data are if anything more daunting. This is due primarily to the highly skewed distribution of businesses by size. Because even a small amount of information would make it easy to identify a large business's record, public use business microdata files seldom can be released, and detailed tabular releases also frequently are problematic.

Recognizing these issues, federal statistical agencies have acted to reduce the risk that specific individuals or businesses can be identified or information about them inferred based on microdata or tabular releases. Steps taken by agencies to reduce the risk of re-identification of subjects whose information is contained in public use microdata files have included suppressing detailed information about geography, date of birth and other characteristics; top-coding income and wealth variables; adding noise to variables contained on the file; and swapping records across households. Steps taken to control the exposure of confidential information in tabular

releases has included cell suppression (which can have the unfortunate side effect of creating swiss cheese tables that are missing information for many cells), data swapping or noise infusion prior to the creation of published tables, and rounding of cell values (see, e.g., Zayatz 2007; Lauger, Wisniewski, and McKenna 2014). Because providing too much information about the statistical disclosure methods applied to data could make it possible for them to be reverse engineered, exactly what has been done typically is not made public.

Unfortunately, absent a sound theoretical basis for their design and application, there is no guarantee that the disclosure limitation methods currently employed by the federal statistical agencies will be effective, especially against the emerging threats that are associated with the availability of ever-increasing amounts of information from external sources. Further, if data users are not provided with adequate information about the steps taken to reduce disclosure risks, inferences drawn from the data may be misleading (Abowd and Schmutte 2015). As John Abowd has argued eloquently (e.g., Abowd 2018), the existing state of affairs produces data products that are neither provably safe with regard to privacy nor as useful to analysts as they could be while still in fact protecting privacy.

It is true that I cannot point to harms to specific individuals that have come about as a result of information about them contained in a federal statistical data release. This does not mean, however, that the threat of future harms can be ignored. Many members of the public already have doubts about whether the government should be collecting information about them as well about the government's ability to protect the privacy of that information. A well-publicized incident in which a motivated hacker sought to identify and publicize individual-specific information discernible in public data releases could do irreparable damage to the agencies' ability to collect and disseminate data. As the availability of external information and

the computing power available to potential hackers grows, data products previously deemed to be safe will need to be reevaluated.

Outlines of a New Data Access Model

A starting point for the development of a new model for data access is the recognition that different data users have different needs, a fact that suggests thinking in terms of different tiers of access to data (Commission on Evidence-Based Policymaking 2017). At one end of the spectrum will be data users who need only data that can safely be made public. Published tabulations, perhaps modified by the addition of noise as required for disclosure control, are likely to meet the needs many data users. Other data users may be able to work with synthetic data files. Once an analysis using synthetic data has been completed, a data user could be given the option of checking how different the results would have been had the analysis been run against the original data (Reiter, Oganian and Karr 2009). This sort of a "verification server" could help a data user decide whether it was necessary to seek access to the original data.

Some number of data users are likely to require access to the original microdata for their research. If suitable public use microdata files cannot be created, these data users will need to work with data behind a firewall. The Federal Research Data Center (FRDCs) network is currently the primary vehicle for obtaining access to confidential federal microdata. As noted by the Task Force on Differential Privacy for Census Data (2018), the current process of gaining access to data housed in the FRDCs is cumbersome; there are limitations on the sorts of projects that can be approved; accessing the FRDCs can be inconvenient; the capacity of the FRDCs is limited; and there are potential data users who may have difficulty obtaining approval to access the FRDCs at all. If behind-the-fire-wall access is to become a viable alternative to the dissemination of public use microdata files, these limitations must be addressed.

One desirable change will be to make the process of applying for access to the FRDCs less cumbersome. The Foundations for Evidence-Based Policymaking Act (HR4174) recently passed by the Congress, which implements a number of the recommendations regarding data access and privacy made by the Commission on Evidence-Based Policymaking, calls for the Office of Management and Budget (OMB) to develop a common process for applications to access confidential Federal data. Work to establish this process already has begun.

Other provisions of HR4174 clarify that, absent an explicit legal prohibition to the contrary, data held by the federal government generally should be available for statistical evidence-building purposes. The main effect of these provisions may be to increase access to administrative data currently held by various federal program agencies, but their guidance also applies to survey and Census data. Under current law, users seeking access to confidential Census data must show that the project they have proposed will benefit the Census Bureau. This can be a limitation and the relevant statutes should be changed to recognize evidence-building that adds more generally to knowledge as an allowable purpose. Still, in practice, the range of analyses that have been approved under the existing legal structure is very broad and there is little reason to think this will change.

Although FRDC access currently is restricted to researchers who are able to travel to brick-and-mortar facilities, remote access options that would make it possible for larger numbers of researchers to use confidential microdata through the FRDCs may be on the horizon. Denmark and France, among other countries, offer models for how to do this while preserving essential privacy protections (Commission on Evidence-Based Policymaking 2017). An unanswered question is how many researchers in fact would require such access. The Task Force on Differential Privacy for Census Data (2018) notes that 60,000 or more individuals download

IPUMS data files each year, but this seems likely to be very much an upper bound on the number of data users who in fact would need access to restricted microdata.

Another issue noted by the Task Force is that certain groups of individuals, including graduate students and non-U.S. citizens, face barriers to accessing restricted data through the FRDCs. I would note that graduate students often are able to access such data under the umbrella of a broader project for which a faculty members has obtained approval. Non-U.S. citizens based outside of the United States are likely to need to partner with U.S.-based researchers.

There would be considerable advantages to having a centralized facility for coordinating researcher access to survey, Census and administrative data, especially access to linked data files. Staff of a centralized facility could develop expertise in both data linkage and the application of privacy-preserving technology to the preparation of data releases. Further, assessing the privacy risks associated with proposed data releases will require knowing what related releases already have occurred, something that would be facilitated by the release of data being managed through a centralized facility. While no detailed blueprint for establishing such a facility exists, the Commission on Evidence-Based Policymaking (2017) laid out one vision in the form of the National Secure Data Service (NSDS) proposed in its report.

Although it does not talk specifically about the NSDS, the Foundations for Evidence-Based Policymaking Act includes a provision that calls for the establishment of an Advisory Committee on Data for Evidence Building charged to "review, analyze and make recommendations on how to promote the use of Federal data for evidence building." The Advisory Committee is expected to provide recommendations to the OMB Director for "how to facilitate data sharing, enable data linkage and develop privacy enhancing techniques."

Implementation Challenges

The replacement of the existing model with a new model for data access will not happen overnight. What is to be done in the meantime? There is undoubted value to researchers having access to Census and survey microdata. In the short run, given their current configuration, relying on the FRDCs to provide this access is not a realistic alternative to the release of public use microdata files. The statistical disclosure methods currently applied by the statistical agencies, while less than fully satisfying for the reasons already cited, appear to have been largely successful in practice. Despite the risks of continuing with business as usual, the best course of action would seem to be to work within the current structure during some interim period as steps are taken to develop and implement a new data access protocol. Risks in other situations often are handled analogously. When engineers determine that a bridge may be at risk of failing, for example, the typical response is not to close the bridge immediately but rather to consider temporary repairs and then to accelerate efforts to address the problem in a more permanent fashion. I am suggesting a similar approach to the development of a new model for access to microdata, that is, tightening up the existing statistical disclosure limitation procedures as seems advisable but working towards the longer-term goal of a new model for data access.

As plans for a new data access model are developed, several difficult but important issues will need to be confronted. Here I highlight three—deciding on the right tradeoff between access and privacy; deciding on the best approach to allocating a limited "privacy budget" to different potential data users; and developing the capacity to operate the new model effectively.

With respect to the first of these issues, differential privacy offers an explicit characterization of the frontier representing the tradeoff between the amount of information that can be released from a given data set and the privacy protection afforded to the subjects of the

data. Where a statistical agency should locate along this frontier, however, is very much a policy decision rather than a technical decision. In essence, the choice of ε in the differential privacy framework is a choice about how much additional privacy risk data subjects potentially will incur as a result of whatever data releases are made. Choosing a value for ε requires weighing effects on the data products that will be permitted and the absolute disclosure risk created for different groups of data subjects. Effective means of communicating the implications of different choices are sorely needed.

Supposing that agreement can be reached about the appropriate level of ε and thus the aggregate "privacy budget" for a particular data set, there is then the knotty issue of how that privacy budget should be allocated. This again is a policy decision rather than a technical decision. The current model for access to the FRDCs is essentially first-come, first served, at least among the set of projects that satisfy the criteria for approval. It is possible, however, that a project that happens to get through the door first could lead to data releases absorbing much or even all of the entire agreed-upon privacy budget, but with a limited return in the form of additions to knowledge. Accumulating proposed projects over some period of time and then allocating the privacy budget by lottery might have the advantage of greater perceived fairness, but for a variety of reasons this mechanism also does not seem ideal. It seems unavoidable that explicit judgments will need to be made about whether the implied privacy budget expenditures associated with different proposed projects are merited.

At present, decisions about data access and dissemination largely are made by statistical agency staff. It would be preferable, however, for there to be broader input into decisions both about the most appropriate risk/information tradeoff and about the allocation of privacy budgets. With respect specifically to the allocation of any agreed privacy budget, mechanisms already

exist for allocating research dollars, another scarce resource, to potential projects. A peer review model similar to that used by the National Science Foundation or the National Institutes of Health to allocate research funding, with committees of scholars assessing the merits of competing proposals and their privacy budget costs, could perhaps be adopted for this purpose.

Developing the capacity to support the data access system I am envisioning undeniably will be a major undertaking. It will require expert staff the federal statistical agencies do not currently have, tools for the implementation of privacy-protecting approaches that do not currently exist, and a budget to support the necessary infrastructure. At present, a good deal of the work to make federal data accessible to researchers occurs outside of the federal statistical agencies and robust public-private partnerships are likely to be essential for the new model to succeed. One could imagine, to give just one example, existing IPUMS staff working as federal statistical system agents, as envisioned in the Foundations of Evidence-Based Policymaking Act, to develop curated data files and facilitate the use of restricted-access Census data. Funding for the necessary infrastructure could perhaps be provided by redirecting a small percentage of programmatic funding to support data access as an evidence-building tool, similar to what the Department of Labor has done to support the work of its Chief Evaluation Officer.

The transition to a new model for access to survey, Census and administrative data will not be either quick or easy, but changes to the existing model are needed. Ultimately it should be possible to strengthen the privacy protections afforded to data subjects while preserving the value of survey and Census data—and increasing the value of administrative data—for research purposes.

References

Abowd, John. 2018. "How Modern Disclosure Avoidance Methods Could Change the Way Statistical Agencies Operate," presentation to the Federal Economic Statistics Advisory Committee. December.

Abowd, John and Ian Schmutte. 2015. "Economic Analysis and Statistical Disclosure Limitation," *Brookings Papers on Economic Activity*, 2015, Spring, 221-23.

Commission on Evidence-Based Policymaking. 2017. *The Promise of Evidence-Based Policymaking.* Washington, DC.

Lauger, Amy, Billy Wisniewski and Laura McKenna. 2014. "Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research," Research Report Series Disclosure Avoidance 2014-02. September.

Reiter, Jerome P., Anna Oganian and Alan F. Karr. 2009. "Verification Servers: Enabling Analysts to Assess the Quality of Inferences from Public Use Data," *Computational Statistics and Data Analysis*, 53, 1475-1482

Sweeney, Latanya. 2000. "Simple Demographics Often Identify People Uniquely," Carnegie Mellon University, Data Privacy Working Paper 3.

Task Force on Differential Privacy for Census Data. 2018. "Implications of Differential Privacy for Census Bureau Data and Research," Minnesota Population Center Working Paper No. 2018-6. November.

Zayatz, Laura. "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update," *Journal of Official Statistics,* 23(2): 253-265.