June 2023

THE AMERICAN ECONOMIC ASSOCIATION

Committee on Economic Statistics: www.aeaweb.org/about-aea/committees/economic-statistics

Implications of New Privacy Protection Methods for Economic Research

Summary of an AEAStat Working Session held January 6, 2023, at the
2023 Allied Social Science Association meetings in New Orleans, LA

Premise: Federal statistical agencies have long worked to preserve the privacy of respondents to statistical data collections, using such traditional statistical disclosure limitation (SDL) methods as top-coding, swapping, and omitting geographic or industry detail from public-use micro data and tabular files. But with detailed digital information on people and businesses rapidly accumulating and computing firepower increasing, the traditional methods have become inadequate for the task. This point has been vividly illustrated by simulated database reconstruction attacks conducted by the Census Bureau, which found alarming levels of reidentification risks in data prepared for public release using traditional disclosure methods.

To adapt to the changing environment, federal statistical agencies are exploring new SDL methods combined with new ways of providing access to micro data on U.S. households, workers, and businesses. Impacts may be most immediate and particularly large for rich micro or tabular data sets (such as the Decennial Census, the Survey of Income and Program Participation, or the Current Business Patterns). Published aggregates from those and other data sets, which are often used for timely economic policy analysis, will also be affected. These changes will substantially alter how researchers access micro data from federal data collections outside of channels for accessing confidential data for approved projects through the Federal Statistical Research Data Centers (FSRDCs) and other restricted data-access programs. While expected changes have provoked lively—and sometimes contentious—discussion among core users of the data, the broader economics profession has thus far had limited engagement with the issues around new privacy protection methods. For example, the long program of the 2023 ASSA meetings had one paper on the topic. Given that the agencies are obliged by law to respond to the growing need for new protections, the profession's involvement will need to change and expand as well.

Objective: The AEA's Committee on Economic Statistics (AEAStat) organized a working session at the 2023 ASSA meetings on implications of new privacy protection methods for economic research. The session brought together 30 experts—economists, data scientists, statisticians, and others—from academia, statistical agencies, other federal agencies, and research organizations for a facilitated discussion of four central questions, which are listed below. While some experts with deep expertise have been working on these issues for years, the meeting was intended to mark the beginning of the broader economics profession's engagement in an ongoing process that will be part of and shape a new normal. This document summarizes some of the key issues that arose during the discussion, organized

by groups of questions that framed the meeting. It should not be inferred that all participants agreed on points that were raised.

1. **Contributions from economists:** *What are the most important questions about implementing new SDL methods for economic data on which statistical agencies need input from researchers? What are the bests ways for contributions from economists to be solicited and delivered?*

    - Most members of the economics profession are not aware of upcoming changes in privacy-protection methods and data access. It is in economists' self-interest to inform ourselves and find effective and productive ways to contribute to discussions of how best to improve privacy protection while maintaining the value of federal data for economic research.

    - Our challenge is not just to communicate what economists like or do not like. We also have work to do. We must figure out if and how the specific research methods that are most important to economists can be made consistent with privacy protection methods. No one else will do this work for us.

    - Economists' usual means of providing inputs into agency decisions, via advisory committee meetings and replies to requests for information, are insufficient for exchanging ideas and promoting the work needed on privacy protection. The issues are too broad and complex to be worked out through these narrow channels. A series of online and in-person seminars, workshops, and conferences is needed to bring together people working on privacy protection and data access, to move theory and practice forward with the necessary speed and rigor. Such sustained discussions should include data scientists, engineers, statisticians, even legal professionals, as well as economists and other social scientists. Research in this area has large public benefits and should be prioritized in research funding.

    - Economists' input on privacy protection is most helpful when it is situated in the context of broader changes in data infrastructure underway at the statistical agencies. Survey response rates are falling; survey costs are rising; and demand for timely, granular economic data for policy and research is ever increasing. To adapt to changing realities, the agencies are increasingly using administrative data and real-time business data in combination with and/or possibly eventually replacing traditional survey data. The whole "production function" for economic research will need to be re-thought.

    - Privacy protection entails a social-welfare function that has not been sufficiently considered in the literature or emerging practice. Broad data access facilitates better policy and advances the knowledge frontier. Thus, raising barriers to data use in order to reduce reidentification risks can reduce social benefits of collecting survey data. Economists could help develop tools for conceptualizing and evaluating trade-offs between data access and privacy protection, recognizing the substantial public-goods character of representative survey data collected by the statistical agencies.

    - Although new privacy protection methods may seem to impose only costs on data users, researchers should also recognize important potential benefits. First, prevention of harm to respondents is beneficial and likely to help support response rates and the research that relies on those responses. A data breach for sensitive administrative data could potentially close that source of data for future economic research. Second, users should not ignore the accuracy lost when privacy is protected using traditional methods. The shortcomings of micro data subjected

to previous SDL protocols are opaque to the user and could affect inferences drawn from the data as much or more than new ones. There are many reasons to be optimistic that modern privacy-protection methods combined with new data-access platforms like validation servers could improve researchers' ability to address research questions based on relatively unprocessed micro data, including data that have traditionally been very difficult to access for research purposes such as tax data.

- Data curators face pressures from multiple constituencies, each concerned about how adopting modern privacy protection methods will affect their research programs. Modern systems treat privacy protection in a systematic way. Specifically, systems that offer differential privacy guarantees are designed to cater to the nature of the dataset they protect, the queries they anticipate answering, and the privacy-loss budget afforded to them, which is in turn expended in answering the queries. Privacy guarantees are only meaningful when the privacy loss is finite. Thus, provable privacy guarantees may depend on policy decisions that do not permit answering as many questions about a given dataset as accurately as might be desired. To obtain the benefits of modern privacy methods, data curators need to be able to manage privacy-protection systems *as* systems; data users' requests for changes in methodology often overlook this issue. Rather than asking the data curator to change the privacy-protection system to maintain the validity of an estimator researchers want to use to test a hypothesis, researchers should ask, "What privacy protection methods have been applied to my queries of the data or the released data itself, and what estimators to do I need to use to test my hypothesis?"

2. **Preparing for implementation:** *How should economic researchers prepare for analysis of privacy-protected data? How can this preparation be facilitated? What will be the role of enclaves such as the FSRDCs for gaining access to data when the public domain version has been subject to new disclosure-avoidance methods?*
   - Average academic economists who use federal survey data in their research are generally not aware of agency plans to modernize disclosure-avoidance systems, nor aware of the pressing rationales for doing so. Much work needs to be done to educate economic researchers as to when and how *they* will be affected by upcoming changes for the specific data products they use. Critical information to share with users includes:
     - When exactly will new methods start to be applied to given data collections? Will statistics from that data collection published by the statistical agency be affected by new methods, or only public-use micro data files?
     - Will data accessed through the FSRDCs and other restricted-access data programs also be subject to new privacy protections? (In general, under current practices the answer is no – although the results must be reviewed by the statistical agencies' disclosure review boards. Tabular output is discouraged and more likely to be subject to new privacy protections. Multivariate statistical analysis (e.g., OLS regressions) is currently subject to legacy disclosure-review protocols).
     - Are there plans for changing the criteria or procedures used for disclosure review?
     - How, and under what conditions, can noise-infused public-use data files be used for econometric analyses? What steps are needed to take into account the impact of new methods on model coefficients?

    - Has a system been established for running models on the agency's internal (unprotected) data and passing results back to the researcher, and if so, how does it work?

- Devising clear communications that help researchers get up-to-speed relatively quickly will be key to ensuring the continued value of survey data sets for economic research, as well as securing researchers' buy-in to the process of modernizing systems for accessing micro data. Online workshops organized around changes in privacy protection and data access for given data collections could be very helpful.

- From economic researchers' point of view, losing access to public-use files that have traditionally been used in standard ways to estimate econometric models is a significant concern.[1] Current systems used to approve access to confidential data and get results cleared through disclosure review – e.g., through the FSRDCs and similar portals – would be insufficient for handling increased demand for access to relatively unprocessed data. If systems are not substantially overhauled, with new means of accessing confidential data in a secure and accurate manner put in place, many users will substitute away from high-quality representative federal data, towards less reliable but more accessible private data; perhaps only a small community of academic researchers would continue to use federal data collections. This would constitute a substantial reduction in the utility of federal data. It would also likely degrade the quality of evidence available to guide key policy decisions. These are very large costs to incur to reduce disclosure risks.

- Modernizing access to confidential data would benefit substantially from modernizing how economists are trained to interact with data. To most empirical economists, using R to interact with data as an object remains foreign. But disclosure avoidance becomes substantially easier if data is accessed via interfaces that employ R- and Python-based advances, and all that is passed back to the user is a limited number of parameters. Building application programming interfaces (APIs) to access and analyze confidential data should be a priority. To equip data users to use modern systems, researchers need facilitated access to basic training in necessary tools, as well as nudges or incentives to shift to preferred ways of accessing and using data.

- Economists could also work further on developing econometric estimators that provide valid econometric inferences when used to analyze privacy-protected data. Valid estimators are not general; they depend on the notion of privacy that the system was built to protect. Co-developing estimators with the privacy-protection system can potentially provide a way to take the extra uncertainty from noise infusion into account.

- Applying standard econometric methods to privacy protected data products without accounting for the privacy mechanism creates a potential "double whammy" problem: Parameter estimates may no longer converge to their true values, yet the researcher is overconfident in the wrong estimates because standard errors can understate the estimation uncertainty associated with privacy protection. The statistical power associated with research findings may also be exaggerated.

[1] Some session participants underscored that analyses run on public-use micro data files that have been privacy-protected using traditional methods like swapping have also involved accuracy loss, but under traditional methods the loss is opaque to the user.

- The double whammy problem can be understood as a *marginal* effect of privacy protection on otherwise perfect data. But perfect data is clearly an incorrect imaginary, as data contain many other types of sampling and non-sampling error that principled analysis should take into account. Accounting for all sources of errors (including variance introduced as part of the privacy-protection process) is clearly the right thing to do, and it requires the development of practical methodologies suitable for a broad range of econometric analyses.

3. **Implications for policy:** *How should implementation of economic policy and regulation adjust when new SDL methods are applied to statistics used for funding or eligibility criteria?*

   - For policies that use sharp thresholds to determine benefit eligibility, use of privacy-protected data may cause misleading inferences. For example, noise-infused data may show average income in a county to be above a threshold, when the confidential data (itself subject to measurement error) show it to be below, or vice versa, so that some counties eligible for benefits do not receive them while others that are ineligible do.

   - Ensuring that government agencies have facilitated access to confidential data for policy purposes would address this issue. Although releasing precise statistics, such as total population, blows up the privacy-loss budget usually associated with differential privacy, there may be other ways to construct and calibrate other kinds of formal privacy guarantees. Facilitated access could also occur via collaborations with statistical agencies, special validation servers, and/or the new National Secure Data Service (the entity intended to implement data-access provisions of the Evidence Act).[2] The threshold-policy example can be used to educate policymakers on uncertainties in statistical data (including, but not limited to, uncertainties due to noise infusion). It also underlines the value of shifting policy-decision criteria away from sharp thresholds and towards gradual ramps instead.

   - Uncertainty due to privacy protection should be put into perspective, as it may be small relative to other uncertainties in statistical data. An important study by Steed, Acquisti, *et al*., examines the potential for misallocating Title 1 education funding due to uncertainties in Census results, had they been released after noise infusion using typical differential privacy methods. Of $11.7 billion in federal education grants in 2021, they find that $1.1 billion would be misallocated due to some quantifiable source of statistical uncertainty. In contrast, for plausible degrees of privacy protection, misallocation due to noise infused for the purpose of privacy protection would be far less, on the order of $1-50 million.[3] This finding underlines the importance of healthy support for federal data collections. Reducing statistical uncertainty by improving data quality increases the evidence base for effective economic policy.

   - In general, government agencies that use federal data as an input into economic policy decision-making should be encouraged to take uncertainties in statistical data <u>of all kinds</u> into account. When formulating policies going forward, for example, gradual cutoffs may make more sense than sharp ones. Economists and the statistical agencies could and should play helpful roles in improving reliable use of data for policy purposes.

[2] H.R.4174 - Foundations for Evidence-Based Policymaking Act of 2018, 115th Congress (2017-2018).
[3] Steed R, Liu T, Wu ZS, Acquisti A. "Policy impacts of statistical uncertainty and privacy," *Science*, August 2022, 377(6609):928-931.

4. **Unresolved questions:** *What are the most important unanswered questions about the implementation of new SDL methods to economic data? How can these questions be addressed? How can a constructive ongoing dialogue about these issues be implemented?*

   - Resources are required to modernize privacy protection and build new ways for researchers to continue to be able to draw valid inferences from analysis of data; whole teams of data scientists and engineers are needed. What can be done to ensure healthy funding for the statistical agencies and research funding from NSF and private foundations to support work in this area?

   - Frameworks for understanding the implications of alternative privacy-protection methods on disclosure risks are relatively well developed, but much less work has been done on conceptualizing and evaluating data utility. How are data sets being used in designing and implementing policies, and/or in scholarly research? Are the contributions resulting from current uses relatively large or relatively small? Without having characterizations of social benefits that are derived from available data, assessing privacy-utility trade-offs is relatively abstract.

   - Analyses of privacy-utility trade-offs need to account for expected changes in researcher behavior brought about by given privacy-protection systems. Data utility is usually understood in terms of the accuracy of queries made to the data. But if a given privacy-protection system makes it substantially more costly to use the data for research, its use will decline, and the social value of the now-better-protected data will fall. How can this substitution margin be taken into account?

   - How can the needs of the broader community of data users be addressed, so that shifts in data access and publicly available data files do not cause significant disruptions to ongoing applied research agendas? Who should take the lead on reaching out to data users to make relevant information on upcoming changes available? Who will develop and implement training resources that meet applied researchers' needs?

This document was prepared by the organizing committee for the AEAStat Working Session, which consisted of Erica Groshen (chair), Ruobin Gong (Rutgers University), Daniel Goroff (Sloan Foundation), John Haltiwanger (University of Maryland), and V. Joseph Hotz (Duke University), with inputs from Karen Dynan (Harvard University and AEAStat Chair) and Martha Starr (AEA). Many thanks to session participants who provided valuable feedback on an earlier draft.

AEA Contact: Martha Starr, PhD, Washington, DC, Representative of the AEA and Director of Government Relations. martha.starr@aeapubs.org. 703-283-0054.