Disclosing Economists' Privacy Perspectives: A Survey of American Economic Association Members on Differential Privacy and Data Fitness for Use Standards

Aaron R. Williams[1], Joshua Snoke[2,*], Claire McKay Bowen[1], and Andrés F. Barrientos[3]

[1] *Urban Institute*
[2] *RAND Corporation*
[3] *Florida State University*
[*] *Corresponding author: jsnoke@rand.org*

Keywords— administrative data, data confidentiality, data privacy, differential privacy, social science, tax policy analysis

Proposed Article Description: Researchers and practitioners at universities, government agencies, not-for-profits, and industry widely use official statistics and government data. These data often contain sensitive information and undergo statistical disclosure control to protect the confidentiality of individuals or entities in the data. With the introduction of Differential Privacy (DP) (Dwork et al., 2006), the field of statistical disclosure control is experiencing a paradigm shift (Snoke and Bowen, 2020). Most notably, the U.S. Census Bureau updated their 2020 Disclosure Avoidance System for the 2020 Decennial Census with formal privacy (United States Census Bureau, 2021), and recently the Internal Revenue Service has started considering DP to protect tax data (Barrientos et al., 2021a,b). The shift to increase privacy protections through DP offers the exciting possibility of releasing new sources of administrative data from the government. On the other hand, these changes will affect results derived from official statistics, impact the availability of data, and demand new skills and knowledge for data analysis.

Despite these major changes and opportunities, little is known about data users' knowledge and perceptions of DP. Economists are a large group of empirical researchers who regularly use Census Bureau and Internal Revenue Service data. In an effort to understand economists' knowledge of DP and attitudes towards it, we conduct a convenience sample survey of members of the American Economic Association, receiving over 1,000 responses and a response rate in excess of 10%.[1] The survey includes demographics on the covariates shown in Table 1.

Beyond demographics, the survey contains three primary sections: (1) baseline knowledge and attitudes towards DP, (2) types of statistical methods that are most useful to economists for analyzing cross-sectional administrative data, and (3) tolerance levels for additional injection of

[1] For a preview of the survey, see https://urban.co1.qualtrics.com/jfe/preview/SV_2o4xajAcmanOcIK?Q_CHL=preview&Q_SurveyVersionID=current.

Table 1: Demographic Covariates Collected in the Survey of Economists

Demographic Covariate	**Response Type**	**Population Estimate**
Reside in U.S.	Binary	No
Student	Binary	No
Highest Degree Completed	5 Categories	Yes[2]
Year of Completion	Continuous	Yes[3]
Primary Employer	5 Categories	No
Peer-Reviewer in the Past Five Years?	Binary	No
Journal of Economic Literature Codes (Select up to three)	20 Categories	Yes[4]

noise under DP. We analyze the results primarily as a convenience sample that does not necessarily represent the population of economists. We conduct some sensitivity analyses for our findings by re-weighting the data to match known population estimates for a small number of the covariates. Table 1 shows which covariates have available population estimates that we use for the sensitivity analyses.

We use the results from the survey in the proposed article to inform the privacy conversation on two topics. First, regarding the economists general opinions and positions towards DP, our results show that: (i) most economists are unfamiliar with the concepts of DP and formal privacy and (ii) economists who are familiar with DP are more skeptical of how much noise it induces than the need for stronger protections. We further separate these results by location (i.e., reside in U.S.), student status, high degree completed, year of completion, primary employer, has peer-reviewed in the past five years, and research sub-field via journal of economic literature codes to discover further insights. Because broad opinions concerning DP are unknown, this survey represents a unique opportunity to communicate a better understanding of data users and their position towards innovations in data privacy research.

In addition to the general views on DP and formal privacy, we demonstrate how results from the survey can be used to inform implementations on differentially private algorithms. We do this through connecting the results to ongoing work developing an Internal Revenue Service (IRS) tax data validation server (Barrientos et al., 2021b) and providing a road-map for how future surveys

[2]See https://nces.ed.gov/programs/digest/d21/tables/dt21_325.92.asp?current=yes for the National Center for Education Statistics and https://ncses.nsf.gov/pubs/nsf22300/data-tables#group3 for the NSF National Center for Science and Engineering Statistics on degrees awarded in the United States.

[3]See https://nces.ed.gov/programs/digest/d21/tables/dt21_325.92.asp?current=yes for the National Center for Education Statistics on degrees awarded in the United States.

[4]The proportion of published papers in each topic is available, so we can weight individuals using these estimates if we assume per-person publishing rates are roughly equal in different sub-fields.

could be used to influence formally private implementations. As it relates to development of the IRS validation server, the survey results provide two substantial findings. First, we find that the use of simpler analyses, such as multiple linear regression and prediction, are more common among economists than more complex methods, such as Tobit or regression discontinuity design. This result means that a validation server that does not allow for more complex queries will still be useful to the economics research community. Second, we measure a quantifiable level of error tolerance among economists on four metrics. We show how we can use this to determine required ϵ levels that meet economists' desires and help determine the overall privacy budget that would be required to run a validation server.

In the proposed article, we discuss these results and more in further detail. The article will contribute to the data privacy literature through a thorough assessment of economists' perspectives towards DP. It also provide an example of how surveys might be used to determine the privacy budgets required to meet users' accuracy needs, which will help policy makers determine the feasibility of implementing new privacy algorithms. We provide suggestions for how the adoption of differentially private methods and tools can align with users' expectations, possible next steps to create better privacy-preserving tools for the economics community, and recommendations to improve communication and research relationships.

References

Barrientos, A. F., A. R. Williams, J. Snoke, and C. M. Bowen (2021a). Differentially private methods for validation servers.

Barrientos, A. F., A. R. Williams, J. Snoke, and C. M. Bowen (2021b). A feasibility study of differentially private summary statistics and regression analyses for administrative tax data. *arXiv preprint arXiv:2110.12055. Under Review*.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pp. 265–84. Springer.

Snoke, J. and C. M. Bowen (2020). How statisticians should grapple with privacy in a changing data landscape. *Chance 33*(4), 6–13.

United States Census Bureau (2021). Disclosure avoidance for the 2020 census: An introduction. Technical report.