NATIONAL ACADEMIES Sciences Engineering Medicine

Consensus Study Report
Issue Brief

A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation

What Privacy Researchers Can Learn

In 2022–2023, the Committee on National Statistics within the National Academies of Sciences, Engineering, and Medicine convened a panel of experts to examine disclosure avoidance in the Survey of Income and Program Participation (SIPP). Following are some of the key implications of the report for privacy researchers.

SIPP PRESENTS DIFFERENT TYPES OF RISKS

The number and granularity of the variables in a database are commonly a concern for disclosure avoidance. As a longitudinal household survey, SIPP has the potential for risks that often do not appear in other databases.

- SIPP collects data on all members of a household. This means that not just the characteristics of an individual respondent are important, but also the characteristics of other members of the same household. For example, interracial marriages are relatively uncommon, as also are large age differences between spouses; thus, knowing that a black male is married to an Asian female who is 12 years younger helps to quickly narrow the number of possible matches.

- SIPP collects data on changes over time. Information on topics such as moving to a different state, changes in marital status or number of children, or loss of a job are all potentially disclosive.

MEASUREMENT OF DISCLOSURE RISK

The panel concluded that measuring both relative risk and absolute risk is important (Conclusion 3-1). Relative risk may be low while absolute risk is high; conversely, if absolute risk is low, then relative risk might be allowed to be high.

The panel offered advice on how the Census Bureau could improve its conduct of reidentification studies with respect to SIPP.

- Examine the importance of high imputation rates in SIPP on disclosure risk.

- Examine the full range of options available to data intruders, including the use of alternative thresholds, the measurement of change over time, and allowing for data across multiple members of a household.

- Extend the analysis beyond the use of IRS data to other databases that may be used, and allow for limitations in the IRS databases (such as that some people do not pay taxes).

ALLOWANCES FOR USER NEEDS

The privacy protections that are needed may vary depending on the planned uses of the data. In the particular case of SIPP, which is a highly complex database, data users tend to be highly sophisticated,

conducting complex analyses and making use of multiple modules within SIPP. These needs set constraints on what types of disclosure avoidance approaches might be workable. For example, reducing the public use file to a set of core variables would not be workable, and unless a table generator or remote analysis platform is highly sophisticated, it will not meet most needs of current users (though it might increase the number of SIPP data users).

The panel concluded that providing a highly comprehensive data file is needed for SIPP users, making the availability of a public-use file very important (Recommendation 9-3), along with readily accessible means of conducting those types of analysis that a privacy-protected public-use file cannot support (Recommendation 9-4).

METHODS FOR MAKING DATA AVAILABLE

A key strategy, consistent with recent legislation, is to make the data available through multiple tiers of access (Recommendation 4-3). Currently, SIPP is only available through two modes of access: a public-use file and access through a Federal Statistical Research Data Center (FSRDC). Future re-identification studies may find that additional limits will be needed on the public-use file, but relying on FRSDCs to be the sole alternative to a public-use file would be burdensome on researchers and create inequities of access (Conclusion 2-3). Thus, there is a need for an intermediate tier of access that will offer fuller access to the data without imposing all of the restrictions that are associated with FSRDCs. Secure online data access seems especially well suited for this need (Recommendation 5-1).

INVOLVEMENT OF OUTSIDE RESEARCHERS

Researchers outside the Census Bureau (or other agencies) can bring new perspectives and expertise to aid in identifying and handling disclosure risks (Recommendation 3-3).

THE CURRENT STATE OF TECHNOLOGY

Some powerful tools for disclosure avoidance do not yet have the capacity for handling databases with the size and complexity of SIPP. These limitations both constrain what options are available for protecting SIPP respondents and create a research agenda for further work.

- Synthetic data have been used in SIPP to allow the merging of administrative data with SIPP data while protecting privacy, initially with files in which selected data were synthetic, and ultimately with files in which all data are synthetic. However, these files were limited in size. Current technology would support replacing a limited number of variables with synthetic data, but not the creation of a fully synthetic and complete dataset (Recommendation 6-4).

- Differential privacy is another tool that might be useful in limited applications, such as within a table generator, but current technology is not well enough developed to support creating a complete SIPP data file with differential privacy.

- The technology for providing secure online data access exists and has been used successfully by multiple agencies and for multiple datasets (Recommendation 5-1). However, in order to limit the burden on Census or its contractors in supporting data validation and disclosure review, there would be value in creating automated systems to support such work.

THE IMPORTANCE OF TRANSPARENCY

The use of disclosure avoidance techniques affects which modes of access will be most appropriate for researchers, what analytic techniques may be applied, and how the data are interpreted. It is important for Census Bureau communications to support such needs (Recommendation 3-2).

FOR MORE INFORMATION

This Consensus Study Report Issue Brief was prepared by the Committee on National Statistics and based on the Consensus Study Report *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation* (2023). The study was sponsored by the U.S. Census Bureau. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project. The Consensus Study Report is available from the National Academies Press, (800) 624-6242 or https://www.nap.edu/catalog/27169.

Division of Behavioral and Social Sciences and Education

NATIONAL ACADEMIES *Sciences Engineering Medicine*