Considerations for Developing Disclosure Avoidance Systems for Longitudinal Survey Data

V. Joseph Hotz
University of Chicago
hotz@uchicago.edu

Trivellore Raghunathan
University of Michigan
teraghu@umich.edu

Abstract

This article summarizes the findings of a recent National Academy of Sciences, Engineering and Medicine Panel and its report, *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation*. While this report focused on the Survey of Income and Program Participation (SIPP), the issues discussed below apply to challenges that all, if not most, longitudinal studies face in balancing protection of the confidentiality of information provided by respondents to such study with the usability of the data released by researchers, policy makers and the genera public. The article provides a background on the structure and unique features of the SIPP, disclosure risks of releasing its data products, and strategies that have been used to protect confidentiality. It summarizes the findings of this Panel that are relevant for the dissemination of data not only from the SIPP but also from many other longitudinal studies.

1. Introduction

Longitudinal surveys, such as the Panel Study of Income Dynamics (PSID), the Health and Retirement Study (HRS) and Survey of Income and Program Participants (SIPP) are indispensable to study the changes in society, impacts of various factors on those changes and to gain deeper understanding of policy decisions needed for the betterment of society. Data from such studies should be widely distributed, analyzed from multiple perspectives for developing and implementing policy decisions. Releasing such data, though beneficial, also can lead to disclosure of private information shared by the respondents with the data gathering agencies. This breach of confiden-

tiality of the responses can harm the respondents in many ways and, therefore, a disclosure avoidance system needs to be in place that limits disclosure but at the same time allows the use of data for statistical purposes, i.e., to draw statistical inferences about the population aggregates.

In 2022-2023, the Committee on National Statistics (CNSTAT) within the National Academies of Sciences, Engineering, and Medicine (NASEM) convened a panel of experts to examine disclosure avoidance in SIPP. This Panel produced a report, *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation*, hereafter the *Roadmap* report, that was published by the Academies 2024 (NASEM, 2024).

In this short article we, who served as the chair (Raghunathan) and a member (Hotz) of this Panel, provide a brief summary of the SIPP, including its design and key features. We then provide a brief discussion of the nature of disclosure risks in the release of data from any survey, the obligations of data providers to protect the confidentiality of information they collect from individuals, firms or other entities, the features and strategies used in disclosure avoidance systems for such data and the fundamental challenge of balancing privacy protection with ensuring the usability of released data for alternative types of users. Finally, we summarize the key findings of and conclusions drawn by this Panel for addressing disclosure avoidance in the SIPP.

While this article is focused on the SIPP and the issues for addressing disclosure risks and data usability for this particular database, we argue below that most, if not all, of the issues confronted by this Panel apply to the data releases from other longitudinal surveys, be they disseminated by statistical agencies or by other research or non-governmental entities.

2. Background on the SIPP

The SIPP is designed to be a nationally representative sample of households in the civilian non-institutionalized population that are followed over time to provide longitudinal data on the

households and its members. Since its inception in 1983, the SIPP has been redesigned several times. We focus on its current design, which was last updated in 2014.

The current design of the SIPP consists of overlapping "panels" which are initiated periodically and run for a period of 4 years. Households in each panel are interviewed in annual "waves."

The number of households in each panel have varied over time, with the recent panels ranging from as large as 33,302 (2022 Panel) to as small as 13,293 (2021 Panel). During the Wave 1 interview, all members of the household who are aged 15 years and older. (Information about household members under 15 years of age is collected via a proxy interview.) In subsequent, waves, the SIPP attempts to interview all members interviewed in Wave 1 and attempts to find and continue to interview them, regardless of whether they move. In each wave's interview, most SIPP questions asked during an annual interview are about the preceding calendar year, using an Event History Calendar to aid in the collection of data on intra-year dynamics. (Some questions, such as about disability status and parent-child relationships, are asked as of the time of the interview.)

The SIPP collects information about the costs, impacts, and effectiveness of government social programs designed to help the poor, and on short-term financial and family dynamics. While largely focused on collecting financial data on income, assets, and expenditures, it also collects detailed information about the composition of households and demographic information about their members, including information on marriages, divorces, etc. A detailed description of the data collected in the SIPP is found in the *Roadmap* report (NASEM, 2024). In short, data from the SIPP provides the most comprehensive information available on how the nation's economic well-being changes over time.

SIPP is unique because it combines three features: (1) an extremely large number of variables collected on all major domains of importance to U.S. individuals and households; (2) its high level of detail on social program participation across a large number of programs; and (3) its ability, afforded by its longitudinal dimension, to follow individuals and households over time to chart the evolution of their status in the large number of domains on which it collects information. Furthermore, by creating an almost-complete record in these dimensions over the past 50 years, SIPP is a unique source for studying historical trends in the country. In the panel's view, preserving these unique characteristics of SIPP deserves high priority.

Two other characteristics also contribute greatly to the value of SIPP, though they might not necessarily be unique. SIPP collects multiple waves of data within a single panel year, thus making a greater number of cases available for analysis in any given year. Moreover, if needed and the questionnaire items have not changed, SIPP data can be combined across multiple years, providing a greater capacity to study small groups. Larger surveys, such as the American Community Survey (ACS), may not have the same need to expand the number of cases for analysis, but SIPP's design makes it very efficient in providing data even for rare groups. SIPP also is well designed for making linkages with administrative data and even provides SIPP synthetic data in which the links have already been made (and synthesized). Other surveys also allow linkages, but SIPP's collection of data on program participation makes the potential for linkages especially useful.

3. Disclosure Risks: Census Bureau's Obligations and Strategies for Mitigating these Risks

In this section, we briefly describe the risk of disclosure from the release of data to the public, the obligations the Census Bureau, as a governmental entity, has for minimizing these risks

when disseminating the data it collects, and briefly outline existing strategies for minimizing such risks.

Disclosure risk is a major concern in data privacy and confidentiality when it comes to sharing data from any longitudinal survey data for research or public-use purposes. If not managed properly, sensitive information, such as income, health insurance status, and fertility history, could be revealed, which could have significant consequences for the individual involved. As a result, the dissemination of data collected from individuals or firms is covered by confidentiality pledges to survey respondents that any data released to the public will not contain personal identifiers, which could directly reveal individuals' confidential information, and will be designed to minimize the risks that third parties can deduce individuals' identities from released data. Governmental agencies, such as the U.S. Census Bureau, are subject to laws, or titles, that establish its obligation to protect the confidentiality and privacy of the information they collect from individuals or firms. To comply with these laws, agencies develop strategies, or systems, for the data they distribute to minimize such disclosure risks.

As discussed in the *Roadmap* report, there is a sizeable body of research that has sought to characterize measures of disclosure risk associated with data releases (see chapter 4 of NASEM, 2024). To address these risks, data providers, including the U.S. Census Bureau, have used various disclosure avoidance systems throughout its history to control the disclosure risks of data it releases to the public. These include methods that alter, or mask, features of the data before releasing for public use in order to mitigate the ability of researchers, and "intruders," can identify respondents and their information. The methods include recoding variables to coarsen their content, suppressing more sensitive variables, swapping data elements across sample members, infusing random noise into variables, an approach used in disclosure avoidance systems that are based on the

privacy criteria of differential privacy, or, in a related approach, creating partial or full synthetic versions of data. In addition, agencies provide access to data with greater disclosure risks through restricted use enclaves to qualified users, such as Federal Statistical Research Data Center (FSRDC) program initiated by the U.S. Census Bureau and now used by other federal statistical agencies. (see Chapter 4 of NASEM, 2024). These different approaches are described in detail in the *Roadmap* report. One of the important challenges in the use of any of these approaches to disclosure avoidance is the balancing of protecting the privacy and confidentiality of information provided by respondents with the usability of such data the data related to users, including academic researchers, policy makers and the general public (Hotz et al. 2022). This challenge was an important issue addressed in the *Roadmap* report to which we discuss below.

4. Disclosure Avoidance and the SIPP

Mitigating disclosure risks in the context of the SIPP or in the release of longitudinal data more generally confronts a number of challenges. They arise from at least three unique features of such data:

(1) an extremely large number of variables collected on all major domains of importance to U.S. individuals and households;

(2) its high level of detail on social program participation across a large number of programs; and

(3) its ability, afforded by its longitudinal dimension, to follow individuals and households over time to chart the evolution of their status in the large number of domains on which it collects information.

Together, these features present important disclosure risks, even though the SIPP, as is the case

with other longitudinal databases, consists of samples drawn from the U.S. population of households as opposed to a census in Decennial Census data. Furthermore, devising a disclosure avoidance system in face of these features is both a unique challenge and one that has not been adequately addressed in the existing literature.

As such, the *Roadmap* Panel's task was more challenging as there was no existing disclosure avoidance system that can be implemented for the SIPP. Furthermore, the Panel noted that there was only limited evidence available that had attempted to quantify the disclosure risks associated with the SIPP that could guide the Census Bureau in its development of a system. Finally, as noted above, developing such a system was all the more challenging in terms of balancing the mitigation of disclosure risks while maintaining the usability of this database, given the diverse set of uses to which the SIPP has been used by its users, be they researchers, policy makers and the public. Below, we provide a summary of some of the conclusions that Panel drew for what it saw as the key elements of a disclosure avoidance system for the SIPP and the importance of this system being an on-going process.

5. Findings and Conclusions for a Disclosure Avoidance System for the SIPP

After a thorough assessment of the SIPP's unique features and challenges facing its disclosure avoidance risks, the *Roadmap* Panel discussed a number of features in the design of a disclosure avoidance system for the SIPP that the Census Bureau should consider. These included consideration of the modes of access that it should consider using to try to balance the tradeoff between privacy protection and the usability of its data for different sets of users. The Report provides a detailed discussion of these findings and conclusions (see both chapter 10 and the Executive Summary in NASEM, 2024). Here we try to provide a broad characterization of the Panel's findings

with a focus on not just the SIPP but also those that are relevant to the dissemination of data from longitudinal studies and surveys more generally. We focus on five key points from this report.

First, assessing the disclosure risk in a longitudinal setting is very different compared to cross-sectional surveys. As noted above, longitudinal surveys like SIPP typically collect a rich set of information from its respondents, both within and especially across its waves. And, because of longitudinal structure, such data captures changes in individuals' and households' demographic, economic and social circumstances. In addition, the information such longitudinal studies collect cumulates across the waves of data collection. Finally, and related to this latter feature, data from longitudinal studies are typically disseminated across waves, unlike data from one-time, cross-sectional surveys like the Current Population Survey. The combination of these features, all else equal, increases the risks of disclosing information from respondents that may comprise its confidentiality.

Second, given the combination of the above features of SIPP (or any longitudinal data), the panel concluded that there is no off-the-shelf disclosure avoidance system that is ideally suited for minimizing or controlling the disclosure risk associated with the release of data from the SIPP. For example, existing strategies, such as infusion of statistical noise in variables according to the differential privacy criteria or the construction of synthetic versions of the data when disseminating data products across waves of one of the SIPP panels will compromise the ability of users to measure changes of income, program participation or marriage stability across the subsequent waves. Measuring changes in such phenomena is a key use of the data from the SIPP or other longitudinal surveys. As a result, the NASEM panel did not make a recommendation on what disclosure avoidance system to use. Rather, it discussed elements of an overall strategy to which we now turn.

Third, in part because of the findings noted above, the Panel discussed the potential value

to users of having the Census Bureau make use of a broader array of modes of data release and access in disseminating SIPP products. This included the dissemination of tabular data from the SIPP using an on-line flexible table generator. For many users, especially students and the general public, this approach allows access to summary data from the SIPP and allows for the control of disclosure risk via the use of existing "static" methods for disclosure avoidance that have been employed in the release of tabular data.

However, the Panel noted that tabular data is not likely to suffice for many users, including those who need access to household or individual level data from the SIPP for use with analyses that employ regression or more sophisticated statistical methods that involve the combinations of variables both within and across waves. To address the needs of such users, the Panel investigated and discussed the use of one or more Secure Online Data Access (SODA) platforms to disseminate the micro data from the SIPP. A variant of this mode of data dissemination has been employed in what is known as the Synthetic SIPP Beta (SSB). The SSB provides access to users who apply to and who are approved by the Census Bureau. They access a synthetic data set produced from individual-level data the SIPP that is linked to tax and some program benefits data. In the past, such users have been able to analyze these synthetic data on the SSB platform and then have the final versions of their statistical programs re-run on a verification server that contains the actual linked data. As the Panel noted in its report, to Access this SSB is available to users who apply to and are approved by the Census Bureau.

While a useful step in providing access to users, the Panel noted that the SSB program has not been a stable one[1] and it has only received limited usage. The Panel discussed the benefits of

[1] The SSB server was initially hosted on a server at Cornell University but was suspended in 2023 and, as of this writing, a new host for the server has not been available, although the Census Bureau announced in July 2024 that a new host would be made available soon.

making the micro-level SIPP available on a more accessible platform that required either no or a limited approval process. Such a platform would allow access to users, while restricting users from downloading these data and, via the platform, would enable the Census Bureau or its designated agent the ability to monitor its usage.

Fourth, the Panel concluded that in developing its public data release strategies for the SIPP, it is essential for the Census Bureau establish a framework with which to evaluate the usability of data for different user communities and use purposes. The Panel viewed this as essential step in the Bureau's efforts both protect the confidentiality of SIPP data while not denigrating its usability for the wide range of uses which the SIPP was designed to use and which it has been employed by its users in the past. To achieve this balance, the Panel encouraged the Census Bureau to continue to assess the levels of disclosure risk associated with the individual- and household-level data it collects across its waves, which is the version of its data thought to be the most vulnerable to disclosure risk. We note that the Census Bureau did conduct what is known as a "re-identification attack" in which they linked the 2014 Panel of the SIPP to various administrative records (1040 tax forms, SSA Numident file, etc.) to uncover vulnerabilities to an external intruder attach (see Singer et al., 2023 for details). While the *Roadmap* Panel found this to be a useful assessment, it also found some important limitations to this study and recommended that the Bureau expand this assessment to include other sources of external data, including more commonly available commercial databases to which potential intruders be more likely to have access. In its Report, the Panel discussed the benefits of developing such an on-going program given the increasing availability of external data sources, especially commercial ones, which could be used by potential intruders could access.

Fifth, and finally, the Panel noted that a key, and sometimes overlooked, obligation of data

disseminators is communication. This includes explaining the importance of protecting privacy, providing clear and honest characterizations of the disclosure risks association with such data releases, and making sure there is a continual process of communication with data users and privacy advocates about the changing nature of disclosure risks threats and the changing nature of research and other uses of released data. But the Panel also identified the need for the Bureau to enhance its communications and engagement with SIPP users to ensure that the Bureau is aware of the users' needs in their uses of SIPP data. Such communication is essential for ensuring that the Bureau meets its obligations to not only protect the privacy and confidentiality of the data it releases but also ensure that these data continue to be useable for the range of researchers, policy makers and the general public that make use of the SIPP.

References

Singer, Phyllis; Aref Dajani and Steve Clark. 2023. “The Retrospective SIPP 2014 Panel PUF Re-Identification Research Study: Final Report.” U.S. Census Bureau, Center for Enterprise Dissemination – Disclosure Avoidance, Septembers 19.

Hotz, V. Joseph; Christopher R. Bollinger; Tatiana Komarova; Charles F. Manski; Robert A. Moffitt; Denis Nekipelov; Aaron Sojourner and Bruce D. Spencer. 2022. "Balancing Data Privacy and Usability in the Federal Statistical System." *Proceedings of the National Academy of Sciences*, 119(31), e2104906119.

National Academies of Sciences, Engineering and Medicine. 2024. *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation*. Washington, DC: The National Academies Press.

United States Census Bureau. 2024. "2023 Survey of Income and Program Participation Users' Guide," U.S. Census Bureau, July.