The Secure Query Service: A Novel Application for Providing Privacy-Preserving Aggregate Statistics on Individual Income Tax Data

Joshua Snoke

Georgetown University

Access to individual income tax data is vital for researchers, non-profits, and state or local governments seeking to inform public policy, but it has traditionally been limited due to privacy and confidentiality concerns. The Internal Revenue Service Statistics of Income (SOI) Division provides access to income data for statistical purposes through their Joint Statistical Research Program and public facing products such as published tables or the Public Use File (PUF) microdata. These products are either highly limited to a small pool of researchers, or they only publish highly aggregated statistics. For policymakers, such as those seeking to understand postsecondary outcomes, it is crucial to receive statistics on linked samples of individuals. Linked statistics enable targeted analyses, such as the evaluation of training programs or randomized controlled trials for new policy initiatives. They can also enable longitudinal studies, tracking a cohort over multiple tax years. Statistics such as these are currently unavailable, requiring a new tier of access.

The Secure Query Service (SQS) is a collaborative project between the Georgetown Massive Data Institute (MDI), Yale University, and IRS SOI that seeks to fill this gap. The overarching goal of the SQS is to provide clients with aggregate statistics via securely linking client data with IRS SOI data and applying privacy technologies to ensure confidentiality is preserved in the output statistics. MDI has developed methodology and code to conduct the matching, tabulation, and disclosure avoidance protocols, and MDI serves as an intermediary between IRS and the client by managing client data governance and data sharing agreements. IRS maintains complete control over the execution and the data, ensuring compliance with all regulations.

Development and deployment of the system was performed with the user community in mind. The statistics which can be requested were selected based on extensive client outreach, the policy decision regarding the trade-off between accuracy and privacy was informed by additional outreach understanding users' needs, and the system is automated to substantially reduce the administrative burden at SOI. Currently, projects at SOI require tailored one-off data use agreements for each project; customized delivery of input files that then require data cleaning and preparation by SOI employees; and manual disclosure review. Such a process is burdensome and cannot scale, and it would not be viable as a long-term solution.

The system's privacy protecting technologies and automated disclosure avoidance approaches ensure compliance with IRS statutory and regulatory requirements. SOI can use tax data under existing tax administration authority under Internal Revenue Code (IRC) 6103(h) and special statistical studies (IRC 6108(b)). This project considers outputs from a secure query system as compilations of tax information under IRC 6108(b). The SQS focuses on individual tax data derived from information returns (Forms W-2, 1099) as well as individual income tax returns (1040). In all cases, the outputs of the system must

meet the agency's disclosure guidelines and, notably, cannot reveal fact-of-filing or any specific federal taxpayer information.

There are several key aspects to ensuring security and privacy in the system. First, the system imposes a schema and controls on the input data to avoid security risks through input poisoning. Second, data are linked using accurate and privacy-preserving record linkage techniques which do not share any PII. Third, the system minimizes output statistics to only those requested and applies disclosure avoidance through a combination of suppression and noise addition based on differential privacy. All legal, security, and administrative steps are managed through agreements between MDI with SOI and end-users.

This talk will cover the background of the SQS, focusing on the policy framework that was developed through extensive client outreach and understanding SOI's regulatory requirements. The talk will also discuss the technical solutions, specifically for privacy-preserving record linkage and disclosure avoidance based on differential privacy in the release of aggregate statistics. Of particular emphasis will be the practical application of balancing between meeting client needs and applying rigorous privacy methodologies. Ultimately, this talk will show how the SQS provides a solution for a broad range of clients to securely link and receive aggregate statistics about individuals, that increases access to earnings and wage data while adhering to SOI's legal obligations to preserve privacy.