ACI-14-43014: An Integrated System for Providing Access to Large-scale, Confidential Social Science Data

PI: Jerry Reiter, Duke University, December 5, 2016

In this DIBBs project, we are generating infrastructure to enable organizations to share confidential social science data for research, education, and public use. We are building a pilot of an integrated system for disseminating large-scale social science data that includes (i) capability to generate highly redacted, synthetic data intended for wide access, coupled with (ii) means for approved researchers to access the confidential data via secure remote access solutions, glued together by (iii) a verification server that allows users to assess the quality of their analyses with the redacted data so as to be more efficient with their use of remote data access. We are using data on the work histories of federal employees provided by the Office of Personnel Management (OPM) as the test case for the system. The data comprise about 3.5 million employees recorded over 25 years, including demographic, career, and salary information.

Challenges Met Thus Far. We have developed a synthetic dataset that, with some additional tweaking, should meet the first objective. The basic idea is to estimate statistical models that describe the joint distribution of the data, and simulate new versions of plausible datasets from the estimated models. We have developed means for researchers from InCommon institutions to access protected data networks without needing Duke credentials. We have developed a suite of verification measures that satisfy differential privacy, which is a privacy criterion that comes with strong guarantees on the amount of information leaked in multiple queries. We have intiated development of a software interface that allows users to get verification results. We have met multiple times with the OPM to describe the project and to discuss long term sustainability.

Challenges Remaining. We have a host of challenges remaining, mostly around seeing how the system will work in practice. For the synthetic data (which do not satisfy differential privacy, as this is incredibly hard to achieve for data of this dimensionality and complexity), we have yet to characterize the disclosure risks fully, particularly in combination with verification measures. The most significant challenges center around the verification component of the system. If we adhere to differential privacy formally, we should enforce an overall budget on the number of queries that can be answered from the data. Many of the analyses of the OPM data envisioned by our social science colleagues require numbers of verifications that would exhaust any sensible privacy budget (from a disclosure risk perspective) almost immediately. Hence, we either have to develop verification measures with lower privacy leakage—a challenging task for the types of analysis done by social scientists—or reconceive what it means to have privacy budgets. For example, perhaps we can enforce personal privacy budgets without capping the total budget for the dataset. Or, perhaps we can disregard budgets completely and just consider one query at a time. This would abandon differential privacy to some extent, but it would increase the usefulness of the data product enormously. A related challenge is getting feedback from users on how they interact with the system, so we can assess whether or not the system meets their needs.