An Integrated System for Public/Private Access to Large-scale, Confidential Social Science Data

PI: Jerome Reiter
Duke University

Overall Goal of Project

Build pilot of an integrated system for accessing large-scale social science data that includes

- **synthetic data** intended for wide access, coupled with
- **secure remote access solutions** providing ways for approved researchers to access the confidential data via, glued together by
- **verification servers** that allow users to assess the quality of their analyses with the synthetic data so as to be more efficient with their use of remote access.

1. Motivation

- **Enormous benefits from wide access to social science data**
  - Facilitates policy making and research.
  - Enables students to learn the skills of data analysis.
  - Facilitates evaluation of new data science methodologies.
  - Allows citizens to understand their society.
- Data stewards obligated to **protect confidentiality** of data subjects.
- Removing direct identifiers not enough to protect confidentiality.
- Data stewards typically do not have expertise to deal with difficult data dissemination problems.

Our infrastructure is intended to help stewards share data with the public.

2. Testbed Data

- Data from the Office of Personnel Management (OPM)
  - Snapshot of every employee in U. S. government as of Sept. 30 stretching back to 1987. We exclude defense, CIA, etc.
  - Career trajectories, demographics, grades and steps, salaries....
  - Longitudinally linked.
  - About 3.5 million persons and 28 million person-year observations.
- Types of analysis questions
  - Do salaries differ by gender or race, holding all else constant?
  - What to typical career trajectories look like?
  - What happens to government after elections?

Data available to approved researchers via secure servers at Duke, not as public use files.

3. Synthetic Data

- Rubin (1993, *J. Offic. Statist*) proposed releasing **fully synthetic data.**
  - Build models for joint distribution of all variables using collected data, e.g., $f(y_1, y_2, y_3, \dots) = f(y_1)f(y_2 \mid y_1)f(y_3 \mid y_1, y_2)\dots$.
  - Release draws from the statistical models as public use files.
- Low disclosure risks, since matching to external databases is nonsensical.
- Can preserve main relationships in the data.
- Use machine learning algorithms to estimate conditional distributions.

4. Secure Remote Access

Protected research data network

- Data live on secure server.
- Approved users access data by **virtual machines** spun up specifically for their needs.
- Multi-factor authentication.
- Access protocols that allow approved researchers from InCommon institutions to log in.

5. Verification Servers

- No way for user to determine whether or not synthetic data offer high quality for specific analysis.
- Suggested in Reiter et al. (2009, *Comp. Statistic. Data Anal.*
  - Separate system with confidential and redacted data.
  - User submits query to system for verification of particular analysis.
  - Server reports back measure of similarity of analysis on confidential and redacted data.
- User can decide to publish if quality sufficient.
- But quality measures can leak information.

6. Verification Servers, Continued

- Provide different types of verification for different types of users
- Level 1 users
  - Verification only for agency-specified analyses
  - Broad measures, like similarity of signs and of significance in regression coefficients
  - Measures must satisfy (relaxed version of) **differential privacy**.
- Level 2 users
  - More freedom in types of analyses.
  - Overlap in confidence intervals for analyses based on confidential and synthetic data.
  - Lightly protected visual displays.

Synergies of System

- Use synthetic data to develop code, explore data, determine right questions to ask.
- User saves time and resources if synthetic data good enough for her purpose (and so does steward).
- If not, user can apply for special access to data.
- This user has not wasted time.
  - Exploration with synthetic data results in more efficient use of the real data.
  - Explorations done offline free resources (cycles and staff) for final analyses.

What Have We Done So Far?

- Fully synthetic OPM data, including synthetic careers.
- Verification measures for Level 1 users all satisfy differential privacy and include
  - Plots of residuals versus predicted values for linear regression.
  - ROC curves in logistic regression.
  - Sign of regression coefficient in any model.
  - Significance of regression coefficient in any model.
- Remote access tested by approved researchers from multiple universities.

Next Steps

- Get approval from the OPM to release the synthetic data, after additional quality improvements and disclosure risk evaluations.
- Develop and test software interface for users to run verification.
- Evaluate how to relax differential privacy for verification purposes (privacy budgets exhaust very quickly).
- Ensure sustainability – proposal for long-term continuation under consideration by OPM.

Acknowledgments: Supported by the National Science Foundation (ACI-14-43014).