Privacy & Confidentiality Technologies

Advisory Committee on Data for Evidence Building

19 March 2021

Len Burman (lburman@urban.org)
Urban Institute and Syracuse University

TPC TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

Opportunities and Challenges

- Protecting data is essential and difficult
- New technologies offer the prospect of making data protection easier, faster and safer for agencies, while making it easier for researchers to access and use the data
- Caveat: these technologies are still in an early stage of development; we have a lot to learn, but the prospects are exciting

Current Challenges

- Traditional statistical disclosure limitation methods are labor intensive and imperfect
- Some agencies do not have the resources to implement them; some datasets and statistics are released with minimal protection
- Methods such as adding noise, rank swapping, microaggregation, and dropping or combining sensitive variables to address perceived threats require a lot of staff time and extensive vetting to balance risk against utility

Opportunity

- New methods could allow more data to be released, more quickly, with strong privacy guarantee
- Privacy protection could be largely automated reducing the need for staff time to implement
- More users could access the data more easily and perform more statistical analyses

Three Promising Technologies

- Fully synthetic data
- Validation server
- Secure multi-party computation

Fully Synthetic Data

- If carefully designed, safe to release with no restrictions
  - No real data so impervious to linkage attacks
- Could be useful for a variety of analyses
- Over time, these data could become quite good
- Can be used as “training dataset” for validation server

Validation Server

- Researcher submits statistical program; receives statistically valid parameter estimates with noise added to protect privacy
- If carefully designed, safe way to access administrative data for research purposes
- Researchers never see the confidential data—just statistics derived from those data

Secure Multiparty Computation (SMC)

- Link records across datasets without ever explicitly merging data records
- Unlike the validation server, which adds privacy protection at the back end before statistics are released, SMC protects privacy at the front end as the data are being constructed for analysis
- In principle could be done in a way that doesn’t violate data sharing restrictions

SMC Could be a Game Changer

- Strong demand from researchers to merge datasets, but legal and ethical concerns are a hurdle
- Examples: match education data with income from tax returns
  - Education scorecard, Chetty, et al, research
  - Combine demographic information from Census (race, ethnicity, education, etc.) with high quality income and tax data from tax returns
- SMC could allow similar applications without ever combining the datasets

Advantages for Data Stewards

- Strong privacy protections—and accessible and transparent methods for implementing them—might encourage more agencies, private sector entities, to share sensitive data in NSDS
  - Might allow agencies that are releasing datasets with no statistical disclosure limitation (SDL) to adopt privacy measures
- Automating privacy protection processes could reduce staff time required to release databases and review statistical output and would be safer than ad hoc methods

Advantages for Researchers

- Systematic privacy protection can be designed to preserve statistical validity of estimates, albeit with somewhat larger standard errors
- Researchers would have unrestricted access to the synthetic data and may be able to access the validation server on their own PC or in a secure data room
  - If fully automated, no wait for manual review before release of statistics
- In contrast, standard ad hoc SDL measures can make it difficult or impossible to produce accurate confidence intervals, hypothesis tests

Challenges

- Producing high-quality synthetic data when underlying distribution is complex and highly nonlinear (without overfitting)
- Developing and implementing algorithms and methodologies with strong, measurable privacy guarantees
  - Pure DP works for some applications, but it assumes an overly expansive attack model, only developed for limited applications

Challenges (continued)

- Relaxed DP methods offer promise, but it is hard to measure privacy risk in some implementations and unknown how to measure the privacy budget
- Measuring and enforcing privacy budget further complicated when some agencies have unlimited ability to release unaltered statistics

Conclusions

- Validation server and SMC are two promising methods to safely expand research data access

- Both need further piloting and learning within the Federal system
  - Our project with the IRS will produce methods and evidence that could provide a template for other sensitive datasets—as well as an agenda for future research

- The Data Service could spur innovation by promoting R&D of new methods, improving and expanding existing methods, and helping agencies adopt best practices

Value Proposition is Key (Julia's Law)

- These technologies must work for agencies *and* data users

- Our big challenge is not only to work out the math, but to implement technology that produces timely datasets, can be managed by agencies with limited resources, and that researchers will want and be able to use to produce high-quality evidence

- The Data Service can play a key role in moving this agenda forward