Data Infrastructure for Studying Mobility: Challenges and Opportunities Regarding Data Governance

Katharine G. Abraham

Workshop on Strengthening the Evidence Base to Improve Economic and Social Mobility in the United States

National Academies of Sciences, Engineering and Medicine

February 15, 2022

Motivation

- Comments will focus on management of survey, census and administrative data held by federal statistical agencies

- Competing goals for management of these data resources
  - Make data available to inform decisions
  - Protect privacy of data subjects

- Challenge: Develop models for data infrastructure and data access/dissemination that best serve both goals

Data infrastructure models

- Data warehouse: Data assembled from different sources permanently housed in a single location
- Data facility: Core data permanently housed in a single location; additional data needed for specific projects brought in as needed but not permanently retained
- Commission on Evidence-Based Policymaking recommended a data facility rather than a data warehouse
  - Recognized that not storing data permanently in a single location would complicate the ability to replicate and extend prior analyses
  - Concerned about (real and perceived) privacy risks of a data warehouse

Data access/dissemination modes

- Federal statistical agencies make data available through
  - Published tabulations
  - Public use microdata files
  - Confidential microdata files accessed behind the firewall
    - Federal Statistical Research Data Centers
    - Internal Revenue Service, Statistics of Income Division Joint Statistical Research Program
- Different access/dissemination modes serve needs of different data users

Privacy key consideration governing data dissemination

- Statistical agencies required by law to protect data subjects' privacy
- Common statistical disclosure control methods for agency releases
  - For microdata, coarsening categorical variables, top-coding continuous variables, noise infusion and data swapping
  - For tabular releases, cell suppression (Swiss cheese tables), noise infusion and data swapping in underlying microdata, and cell value rounding
  - Exact methods generally not made public
- Privacy considerations also govern release of research results generated behind the statistical agency firewall

Limitations of prevailing privacy protection methods

- Lack of public information about statistical disclosure methods creates risk of erroneous inference (Abowd and Schmutte, “Economic Analysis and Statistical Disclosure Limitation,” BPEA 2015)
- Typical statistical disclosure methods not provably private
  - A determined hacker might be able to glean confidential information about individuals or businesses from existing releases
  - Publicity surrounding a successful breach of promised privacy protections could have very negative consequences for federal statistical system
- Census Bureau has begun to adopt disclosure avoidance methods based on differential privacy that address these weaknesses
  - Use not yet widespread, but expect will spread

Rethinking the data access/dissemination model

- Tiered access
  - Fewer public use microdata files
  - Some users work with synthetic data and a “verification server”
  - Smaller number of users (but more than now) given behind-the-firewall access to original microdata
- What is needed to make this work?
  - Streamlined process to apply for microdata access
  - Expanded access capacity, including remote access capabilities
  - Capacity to evaluate privacy implications of proposed releases

Moving to a new data access/dissemination model

- Foundations for Evidence-Based Policymaking Act of 2018 took some first steps
  - Directed creation of standardized data access procedures
  - Established Advisory Committee on Data for Evidence Building that is considering implementation plan for a National Secure Data Service
- Difficult questions related to privacy need to be addressed
  - Deciding on appropriate tradeoff between privacy and information
  - Deciding on how available “privacy budget” will be allocated

Decisions about privacy and the privacy budget

- Differential privacy can be used to characterize the tradeoff between privacy and information released from a data set, but does not offer answers to important policy questions
  - What is the right value of ε ?
  - How should the available privacy budget be allocated across different competing uses?
- Need more effective means of communicating the implications of different values of ε
- Need mechanisms to involve broader constituencies in decisions about ε and the allocation of the agreed-upon privacy budget
  - Steering committee for centralized data access facility?
  - NSF- or NIH-style peer review committees to evaluate proposals for privacy budget expenditures?

Conclusion

- Optimistic that it will be possible to strengthen privacy protections afforded to data subjects while preserving the value of survey and Census data—and increasing the value of administrative data—for research purposes