Reconciling Access and Privacy: Building a Sustainable Model for the Future

Katharine G. Abraham
AEAStat at the ASSA Meetings
January 5, 2019
Overview

• Current statistical disclosure methods neither guarantee protection of data subjects’ privacy nor optimize usefulness of information that statistical agencies report

• Addressing these issues will require that access to some data previously made public become more restricted
  – Tiered access will be a central feature of any new model

• Challenges to implementation of new model include
  – Deciding on appropriate tradeoff between privacy and information
  – Deciding on how available “privacy budget” will be allocated
  – Marshalling resources required for new model to operate effectively
Existing privacy protection methods flawed

• Federal statistical agencies pledge to protect data subject privacy
  – Agencies take pledge very seriously
  – Affects microdata and tabular data they release

• Various statistical disclosure control methods in use
  – For microdata, include coarsening categorical variables, top-coding continuous variables, noise infusion and data swapping
  – For tabular releases, include cell suppression (Swiss cheese tables), noise infusion and data swapping in underlying microdata, and cell value rounding
  – Agencies do not make public exactly what has been done to data

• Lack of information about current statistical disclosure methods creates risk of erroneous inference (Abowd and Schmutte 2015)

• While reducing utility of available information, methods are 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
What might a new model look like?

• Tiered access
  – Many users’ needs met with publicly-available tabulations
  – Some users work with synthetic data and a “verification server”
  – Small number of users given behind-the-firewall access to original microdata

• Institutional changes required for new model to be functional
  – Streamlined process to apply for microdata access
  – Changes in law to recognize evidence-building as an allowable purpose
  – Expanded access capacity, including remote access capabilities
  – Entity to evaluate privacy implications of proposed releases

• Building on recommendations of Commission for Evidence-Based Policymaking (2017), Foundations for Evidence-Based Policymaking Act (HR4174) begins to address needed changes
Transitioning to new model

• Drastic immediate action does not seem wise
  – Not at present a viable option for all data users who need access to microdata to obtain that access through existing FRDCs
  – Though not provably private, existing statistical disclosure control methods appear to have been successful to date

• Best course of action may be to continue with existing structure for some period of time while working hard to put new model in place
  – Analogy to how responsible officials might respond to receipt of an engineering report that a bridge is at risk of failing

• Will need to address implementation challenges already mentioned
  – Deciding on appropriate tradeoff between privacy and information
  – Deciding on how available “privacy budget” will be allocated
  – Marshalling resources required for new model to operate effectively
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 $\varepsilon$?
  – How should the available privacy budget be allocated across different competing uses?

• Need more effective means of communicating the implications of different values of $\varepsilon$

• Need mechanisms to involve broader constituencies in decisions about $\varepsilon$ 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?
Resources needed for new model to succeed

• Implementation of new model a daunting challenge, requiring
  – Staff the Federal agencies do not currently have
  – Tools for implementation of privacy-protecting approaches that do not currently exist
  – Money to support the necessary infrastructure

• Robust public-private partnerships will be essential

• Allocation of a small portion of federal program dollars to this and other evidence-building activities one possible funding option
  – Mechanism used to fund activities of Chief Evaluation Officer at the Department of Labor
Conclusion

• Transition to a new model for access to survey, Census and administrative data will not be either quick or easy, but it is necessary

• Ultimately should be possible to strengthen the 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