Using statistics to protect privacy

Jerry Reiter
Duke University

Alan Karr
National Institute of Statistical Sciences

Releasing Record-level Data

- Enormously beneficial for society
  - Facilitates research and policy-making (as argued in other chapters)
  - Student training, methods development
- But risky for data subjects and stewards
  - Data often from administrative sources, hence available to others
  - Large number of variables means everyone is a population unique
  - Examples: Netflix challenge, Finding accidents in news

Might typical disclosure control methods provide an answer?

- Many data stewards alter data before releasing them
  - Aggregate data, swap records, add noise…
  - Usually minor perturbations for quality reasons
- Typical methods not likely to be effective
  - Low intensity perturbations not protective
  - High intensity perturbations destroy quality

A Potential Path Forward

- An integrated system including
  - unrestricted access to highly redacted data, most likely some version of synthetic data, followed with
  - means for approved researchers to access the confidential data via remote access solutions, glued together by
  - verification servers that allow users to assess the quality of their inferences with the redacted data so as to be more efficient with their use (if necessary) of the remote access to the confidential data.

We Have the Building Blocks

- Synthetic data
  - Synthetic Longitudinal Business Database, Synthetic Survey of Income and Program Participation
  - Automated methods based on machine learning
- Remote access solutions
  - NORC virtual data enclave
  - Virtual machines and protected data networks
- Verification servers
  - Not been built yet, but we have ideas for quality measures

Comments and questions?

- Jerry Reiter
- jerry@stat.duke.edu
- www.stat.duke.edu/~jerry/