Sharing Confidential Data in an Era with No Privacy

Jerry Reiter
Department of Statistical Science
Duke University
Acknowledgements

• Research ideas in this talk supported by
  • National Science Foundation
    • SES-11-31897, CNS-10-12141
  • US Bureau of the Census

• Primary collaborators
  • Alan Karr, National Institute of Statistical Sciences
  • PhD students: Christine Kohnen, Saki Kinney, Robin Mitra, David McClure, Monika Hu
Public use social science data

- Record-level data are enormously beneficial for society
  - Facilitates research and policy-making
  - Trains students at skills of data analysis
  - Enables development of new analysis methods
  - Helps citizens understand their communities

- U.S. statistical agencies leaders at providing public use social science data

- Even in a world where analysis is brought to the data
But....

- Data stewards are ethically and often legally obligated to protect confidentiality
- Releasing record level data can be risky for data subjects (and for data stewards!)
  - Stripping direct identifiers necessary but not sufficient
  - Potential to link variables to external sources
- Examples
  - Netflix challenge
  - Identifying accident victims in newspaper stories
What about large-scale data?

- Arguably big data even riskier to release
  - Data often from administrative sources or social media, hence available to others
  - Large number of variables for matching
  - With many variables everyone is a population unique

- Threat potential increases as quality of data improves
What can be done?

- Many data stewards alter data before releasing them
  - Aggregate data (coarsen geography, top-code, collapse categories)
  - Suppress data
  - Swap variables across records
  - Add random noise

- Usually minor perturbations to preserve data quality
Good for large-scale data?

Existing methods not likely to be effective

- Low intensity perturbations not protective
- But 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

- Fully synthetic data (Rubin 1993)
  - Fit statistical models to the data, and simulate new records for public release
    - Low risk, since matching is not possible
    - Can preserve associations, keep tails, enable small area estimation….  
    - Can release simple random samples
- Synthetic Longitudinal Business Database, Synthetic Survey of Income and Program Participation
- Automated synthesizers based on machine learning
We have the building blocks

- Remote access solutions
  - Data stored on steward’s compute server
  - Approved user’s access server remotely
  - All computation done on remote server
  - No data downloaded to users’ machines
- NORC virtual data enclave,
  Duke Protected Data Network
- Servers cost stewards and hence users
We have the building blocks

- Verification servers (Reiter et al. 2009)
  - Separate system with real and redacted data
  - User submits query to system for verification of particular analysis
  - Server reports back measure of similarity of analysis on real and redacted data
- User can decide to publish if quality sufficient
- But quality measures can leak information
- Not been built yet, but we have plans
Synergies of integrated 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
Concluding remarks

- Long term goal to build this integrated system, including software for generating synthetic data and verification of results, and tools for stewards to spin up remote access solutions.

More information

- Duke/NISS NCRN node: sites.duke.edu/tcrn/
- The NCRN network: ncrn.info