Staring Down the Database Reconstruction Theorem

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau
American Association for the Advancement of Science
Annual Meeting Saturday, February 16, 2019 3:30-5:00

The views expressed in this talk are my own and not those of the U.S. Census Bureau.
The challenges of a census:

1. collect all of the data necessary to underpin our democracy;
2. protect the privacy of individual data to ensure trust and prevent abuse.
• Too many statistics

• Noise infusion is necessary

• Transparency about methods helps rather than harms
Fundamental Tradeoff between Accuracy and Privacy Loss

Accuracy

Privacy Loss

No privacy

No accuracy
Good science and privacy protection are partners
What we did

• Database reconstruction for all 308,745,538 people in 2010 Census
• Link reconstructed records to commercial databases: acquire PII
• Successful linkage to commercial data: putative re-identification
• Compare putative re-identifications to confidential data
• Successful linkage to confidential data: confirmed re-identification
• Harm: attacker can learn self-response race and ethnicity
What we found

• Census block correctly reconstructed in all 6,207,027 inhabited blocks
• Block, sex, age, race, ethnicity reconstructed
  • Exactly: 46% of population (142 million of 308,745,538)
  • Allowing age +/- one year: 71% of population (219 million of 308,745,538)
• Block, sex, age linked to commercial data to acquire PII
  • Putative re-identifications: 45% of population (138 million of 308,745,538)
• Name, block, sex, age, race, ethnicity compared to confidential data
  • Confirmed re-identifications: 38% of putative (52 million; 17% of population)
• For the confirmed re-identifications, race and ethnicity are learned exactly, not statistically
We fixed this for the 2020 Census by implementing differential privacy
Acknowledgments

• The Census Bureau’s 2020 Disclosure Avoidance System incorporates work by Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Scientist for Confidentiality and Data Access), Rob Sienkiewicz (ACC Disclosure Avoidance, Center for Enterprise Dissemination), Tamara Adams, Robert Ashmead, Michael Bentley, Stephen Clark, Craig Corl, Aref Dajani, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Edward Porter, Sarah Powazek, Anne Ross, Ian Schmutte, William Sexton, Lars Vilhuber, Cecil Washington, and Pavel Zhuralev
Thank you.

John.Maron.Abowd@census.gov
More Background on the 2020 Census Disclosure Avoidance System

- September 14, 2017 CSAC (overall design)
  https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf#

- August, 2018 KDD’18 (top-down v. block-by-block)
  https://digitalcommons.ilr.cornell.edu/ldi/49/

- October, 2018 WPES (implementation issues)
  https://arxiv.org/abs/1809.02201

- October, 2018 ACMQueue (understanding database reconstruction)
  https://digitalcommons.ilr.cornell.edu/ldi/50/ or
  https://queue.acm.org/detail.cfm?id=3295691

- December 6, 2010 CSAC (detailed discussion of algorithms and choices)
  https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf#
Selected References


