Results from a Consolidated Database Reconstruction and Intruder Re-Identification Attack on the 2010 Decennial Census

Philip Leclerc, Mathematical Statistician
Center for Enterprise Dissemination - Disclosure Avoidance
United States Census Bureau
philip.leclerc@census.gov

Challenges and New Approaches for Protecting Privacy in Federal Statistical Programs

This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not those of the U.S. Census Bureau.

The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this data product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. (DAO Delegated Approval # CBDRB-FY19-CED002-B0013)

Acknowledgements

2020 Disclosure Avoidance System (DAS) Project Lead:
John Abowd; U.S. Census Bureau & Cornell University

2020 DAS Scientific Lead:
Daniel Kifer, Pennsylvania State University

Re-identification & Reconstruction Sub-Project Lead:
Lars Vilhuber; U.S. Census Bureau & Cornell University

Reconstruction & Re-identification Sub-Project Team:
Tamara S Adams, Robert Ashmead (former), Simson Garfinkel, Nathan Goldschlag, Edward Porter; U.S. Census Bureau

Outline

Motivation: to study the security of legacy disclosure avoidance (DA) methods

- Traditional DA methods have generally reasoned in an *ad hoc* manner about privacy guarantees, or provided rigorous guarantees against only very narrow classes of attackers
- As an example, traditional DA often imposes population thresholds to limit the ease with which information about person-records can be inferred from public data releases

Motivation: to study the security of legacy disclosure avoidance (DA) methods

- Tabular summaries have typically been treated as distinct from public-use microdata, with microdata often receiving much more stringent DA controls
- For the 2010 Decennial Census, multiple forms of traditional DA were used, including both randomized *swapping* of households and imposition of geographic/population limits

Reconstruction attacks convert tabular summaries into microdata

- *Microdata reconstruction* processes a set of tabular summaries, each on a small number of variables, and infers the microdata (featuring a larger number of variables per record) likely to have produced them
- Microdata reconstruction highlights the already blurry line between tabular summaries and microdata, and suggests that traditional tabular summaries implicitly, unintentionally release a large amount of detailed microdata

Consolidated database reconstruction and intruder re-identification is an emerging class of disclosure/privacy threats

- There are many microdata reconstruction techniques
  - For the 2010 Decennial Census, it is natural to construct a system of equations for which any solution corresponds to microdata consistent with published tabular summaries
  - For DA systems that do not preserve inter- or intra-table consistency, more general reconstructions based on generic mathematical optimization can be effective
- Once microdata are reconstructed, it can be easy or trivial to perform *re-identification* attacks: inference about personal information about an individual based on published releases

In 2010, the U.S. Census Bureau published a lot of data

- A 10% public-use microdata sample of the Decennial Census, with geographic areas limited by 100K population threshold, & 10K national population threshold per categorical variable
- And a large number of tabular summaries organized into 4 major products:
  - PL94: $\approx$ 3.6B tabulations
  - SF1: $\approx$ 22B Person, $\approx$ 4.5B HH/GQ tabulations
  - SF2: $\approx 50B$ tabulations
  - AIANSF: $\approx 75B$ tabulations
- Did these tabular releases implicitly release highly accurate microdata at resolution greater than the Decennial public-use microdata sample?

Outline

To lower bound the 2010 Decennial Census privacy risk, we reconstructed microdata from these published tables:

- P001 (Total Population by Block)
- P006 (Total Races Tallied by Block)
- P007 (Hispanic or Latino Origin by Race by Block)
- P009 (Hispanic or Latino, and Not Hispanic or Latino by Race by Block)
- P011 (Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years And Over by Block)
- P012 (Sex by Age by Block)
- P012A-I (Sex by Age by Block, iterated by Race)
- P014 (Sex by Age for the Population under 20 Years by Block)
- PCT012A-N (Sex by Age by Tract, iterated by Major Race Alone)

Tract-level tables provided high-resolution age information:

P12. SEX BY AGE [49]

Universe: Total population

Total:
- Male:
  - Under 5 years
  - 5 to 9 years
  - 10 to 14 years
  - 15 to 17 years
  - 18 and 19 years
  - 20 years
  - 21 years
  - 22 to 24 years
  - 25 to 29 years
  - 30 to 34 years
  - 35 to 39 years
  - 40 to 44 years
  - 45 to 49 years
  - 50 to 54 years
  - 55 to 59 years

PCT12. SEX BY AGE [209]

Universe: Total populatio

Total:
- Male:
  - Under 1 year
  - 1 year
  - 2 years
  - 3 years
  - 4 years
  - 5 years
  - 6 years
  - 7 years
  - 8 years
  - 9 years
  - 10 years
  - 11 years
  - 12 years
  - 13 years

The reconstruction system of equations is simple

- We used a modern, commercial mixed-integer linear programming solver, *gurobi*, to solve for microdata consistent with the published constraints
- Solvers like *gurobi* support integer-valued equations
- For example, letting $T_{\text{Tract,Sex,Age}}$ be the count of persons in tract *Tract* with given *Sex*/*Age*, we formed equations like:

$$T_{t,M,27} = \sum_{p \in \text{personNumber}} \sum_{r \in \text{Race}} \sum_{b \in \text{Blocks}_t} \mathcal{B}_{p,M,27,r,b}$$

$$\mathcal{B}_{i,j,k,l,m} \in \{0,1\} \ \forall \ i,j,k,l,m$$

Outline

At the tract level, *gurobi* solves were rapid and stable

- Tract-level systems of binary equations were reliably solved quickly, and with few to no unexpected *gurobi* exceptions
- These solves are fast enough to infer complete nation-wide microdata in all 70,000 Census tracts and 11M Census blocks in a matter of weeks at modest expense (e.g. with virtual machines rented on "the cloud")

At the tract level, solves rapidly reconstructed a complete set of US microdata

- Solving the binary equation system in each tract nation-wide yielded a set of $308,745,538$ reconstructed microdata records at the block level
- The reconstructed microdata included variables:
  - Geocode (at Census block level)
  - 126 combinations (Hispanic ethnicity, binary; race, 63 OMB categories)
  - Sex (*sex*, binary sex flag)
  - Age (single-year-of-age, 111 levels)
- Reconstructed microdata features this information:
  - with no population threshold limits, unlike PUMS
  - for block-level single-year-of-age with all race-ethnicity attributes, unlike tabular inputs

Given the reconstructed microdata, we performed a simple re-identification

- To lower bound re-identification risk, we used commercial microdata acquired in support of the 2010 Census between 2009 and 2011 that contained name, address, sex, and birthdate information for all known members of the included households
- These data had limited race and ethnicity information, and could not have access to the self-reported values on the 2010 Census

Given the reconstructed microdata, we performed a simple re-identification

- To simulate an attacker trying to infer the 2010 Census self-reported race-ethnicity values, we *joined* the reconstructed and commercial data sets on the *Age*, *Sex*, & *Block* variables
- Reconstructed-commercial record pairs joined in this fashion were dubbed *putative* matches: these represent an attacker's guess that the commercial and reconstructed records describe the same person

We first compared the reconstructed microdata to the commercial data & our internal, sensitive data

- 46%, & 71% of reconstructed records matched correctly to the internal data on *Age*, *Sex*, *Race* (all 63 OMB categories), Hispanic ethnicity, & Block, using exact-age and ± 1 "fuzzy" *Age* matching, respectively
- 45% of reconstructed records were putatively mapped to a corresponding commercial database record, using combined exact-"fuzzy" *Age* matching (1)

We then checked the accuracy of the attacker's inferences about race-Hispanic ethnicity

- Of the 45% of reconstructed records that yielded putative matches, 38% of those matches were confirmed to match exactly, including race-Hispanic ethnicity: intuitively, these are *"The guesses the attacker got right"*
- For comparison, in the last re-identification published by U.S. Census Bureau researchers (on the *American Community Survey*), the putative re-identification rate was 0.017%, and the percent of those confirmed correct, 22% (2)

We have some additional computations to complete

- The aggregate putative & confirmed re-identification rates are large, & mark a sea change in how we think about privacy risk
- But not all inference is equal: it is much easier—and less privacy-eroding—to infer someone's race in populous blocks where race is homogeneous
- Because of this fact, we are performing follow-up investigations to characterize in greater detail the privacy risks suggested by the headline figures I have shared today
- This follow-up analysis will appear in a paper currently under preparation for submission

Importantly, the 2010 Decennial Census reconstruction is just the tip of a very large iceberg

- We used a modest set of variables, and did not use the Decennial public-use microdata sample (PUMS). With additional tables and the PUMS, the attack could be considerably expanded and sharpened
- Staff working on this project do not typically perform large-scale combinatorial optimization. With this expertise, some limiting factors may disappear (e.g., may be able to use County-level tables)
- We relied on *gurobi*'s branch-and-cut mixed-integer linear programming solver, but many algorithms can solve equivalent problems. Any NP-Hard problem-solver (of which there are many) is easily leveraged to perform reconstructions

What have we learned?

- The distinction between tabular summaries and microdata has always been somewhat blurry, with tabular summaries at coarse, populous geographic levels treated as "safe"
- But in the era of cheap, large-scale reconstruction, the microdata-tabular summary distinction is essentially superficial
- Reconstructed microdata can support simple, large-scale statistical inference about private information
- Together, these facts force DA practitioners to reason not about the particular form in which information is published, but about the total information implicit in publications

Contact Information

Philip Leclerc, *Mathematical Statistician*
Center for Enterprise Dissemination - Disclosure Avoidance
United States Census Bureau
Email: philip.leclerc@census.gov

References I

[1] J. Abowd and V. Velkoff, *Managing the privacy-loss budget for the 2020 census*. https://www2.census.gov/cac/sac/meetings/2019-03/managing-privacy-loss-budget-2020-census.pdf, 2019. U.S. Census Bureau Scientific Advisory Committee Report.

[2] A. Ramachandran, L. Singh, E. Porter, and F. Nagle, *Exploring re-identification risks in public domains*, Research Report Series, *Statistics #2012-13* (2012). U.S. Census Bureau.