Modernizing the Disclosure Avoidance System for the 2020 Census

Simson L. Garfinkel
Senior Scientist, Confidentiality and Data Access
U.S. Census Bureau

CS Colloquium, Georgetown University
11:00AM
Friday, February 16, 2018

This presentation is made with the hope that their content may be of interest to the general statistical community. The views in this presentation are those of the author(s), and do not necessarily represent those of the U.S. Census Bureau.
Acknowledgments

This presentation incorporates work by:

- Dan Kifer (Scientific Lead)
- John Abowd (Chief Scientist)
Outline

Motivation
The flow of census response data
Disclosure Avoidance for the 2010 census
Disclosure Avoidance for the 2018 census End-to-End test
Disclosure Avoidance for the 2020 census
Conclusion
Motivation
Article 1, Section 2

The House of Representatives shall be composed of Members chosen every second Year by the People of the several States, and the Electors in each State shall have the Qualifications requisite for Electors of the most numerous Branch of the State Legislature.

No Person shall be a Representative who shall not have attained to the Age of twenty five Years, and been seven Years a Citizen of the United States, and who shall not, when elected, be an Inhabitant of that State in which he shall be chosen.

Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.

The Number of Representatives shall not exceed one for every thirty Thousand, but each State shall have at Least one Representative; and until such enumeration shall be made, the State of New Hampshire shall be entitled to chuse three, Massachusetts eight, Rhode-Island and Providence Plantations one, Connecticut five, New-York six, New Jersey four, Pennsylvania eight, Delaware one, Maryland six, Virginia ten, North Carolina five, South Carolina five, and Georgia three.

When vacancies happen in the Representation from any State, the Executive Authority thereof shall issue Writs of Election to fill such Vacancies.

The House of Representatives shall chuse their Speaker and other Officers; and shall have the sole Power of Impeachment.
“in such Manner as they shall by Law direct.”

Public Law 94-171
Dec. 31, 2018

We will report (per block):

- **P1. RACE/ETHNICITY**
  
  *Universe: Total population*

  *Group by: BLOCK*

- **P2. RACE/ETHNICITY**
  
  *Universe: Total population age 18 and over*

- **H1. OCCUPANCY STATUS**

- **P42. GROUP QUARTERS POPULATION**
  
  *Universe: Population in Group Quarters*

---

**DEPARTMENT OF COMMERCE**

Bureau of the Census

[Docket Number 170824606–7606–01]

**Proposed Content for the Prototype 2020 Census Redistricting Data File**

**AGENCY:** Bureau of the Census, Department of Commerce.

**ACTION:** Notice and request for comment.

**SUMMARY:** The 2020 Census Redistricting Data Program provides states the opportunity to specify the small geographic areas for which they wish to receive 2020 decennial population totals for the purpose of reapportionment and redistricting. This notice pertains to Phase 3, the Data Delivery phase of the program, as the U.S. Census Bureau is providing notification and requesting comment on the content of the prototype 2020 Census Redistricting Data File that will be produced from the 2018 End-to-End Census Test. The Census Bureau anticipates publishing the content for the prototype 2020 Census Redistricting Data File from the 2018 End-to-End Census Test in the second quarter of fiscal year 2018 in a final notice. In that final notice, the Census Bureau also will respond to the comments received on this notice.
But, we need to protect privacy!

13 U.S. Code § 9 - Information as confidential; exception

(a) Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison may, except as provided in section 8 or 16 or chapter 10 of this title or section 210 of the Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act, 1998.

(1) Use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or

(2) Make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or

(3) Permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports. No department, bureau, agency, officer, or employee of the Government, except the Secretary in carrying out the purposes of this title, shall require, for any reason, copies of census reports which have been retained by any such establishment or individual. Copies of census reports, which have been so retained, shall be immune from legal process, and shall not, without the consent of the individual or establishment concerned, be admitted as evidence or used for any purpose in any action, suit, or other judicial or administrative proceeding.

(b) The provisions of subsection (a) of this section relating to the confidential treatment of data for particular individuals and establishments, shall not apply to the censuses of governments provided for by subchapter III of chapter 5 of this title, nor to interim current data provided for by subchapter IV of chapter 5 of this title as to the subjects covered by censuses of governments, with respect to any information obtained therefore that is compiled from, or customarily provided in, public records.
“This is the official form for all the people at this address.”

“It is quick and easy, and your answers are protected by law.”
Example: 2010 Census of Population

Basic results from the 2010 Census

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population</td>
<td>308,745,538</td>
</tr>
<tr>
<td>Household population</td>
<td>300,758,215</td>
</tr>
<tr>
<td>Group quarters population</td>
<td>7,987,323</td>
</tr>
<tr>
<td>Households</td>
<td>116,716,292</td>
</tr>
</tbody>
</table>
Example: 2010 Census II

High-level database schema

<table>
<thead>
<tr>
<th>Variables</th>
<th>Distinct values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habitable blocks</td>
<td>10,620,683</td>
</tr>
<tr>
<td>Habitable tracts</td>
<td>73,768</td>
</tr>
<tr>
<td>Sex</td>
<td>2</td>
</tr>
<tr>
<td>Age</td>
<td>115</td>
</tr>
<tr>
<td>Race/Ethnicity (OMB Categories)</td>
<td>126</td>
</tr>
<tr>
<td>Race/Ethnicity (SF2 Categories)</td>
<td>600</td>
</tr>
<tr>
<td>Relationship to person 1</td>
<td>17</td>
</tr>
<tr>
<td>National histogram cells (OMB Categories)</td>
<td>492,660</td>
</tr>
</tbody>
</table>
### Example: 2010 Census III

**Summary of the publications (counts are approximate)**

<table>
<thead>
<tr>
<th>Publication</th>
<th>Released counts (including zeros)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL94-171 Redistricting</td>
<td>2,771,998,263</td>
</tr>
<tr>
<td>Balance of Summary File 1</td>
<td>2,806,899,669</td>
</tr>
<tr>
<td>Summary File 2</td>
<td>2,093,683,376</td>
</tr>
<tr>
<td>Public-use micro sample</td>
<td>30,874,554</td>
</tr>
<tr>
<td>Lower bound on published statistics</td>
<td>7,703,455,862</td>
</tr>
<tr>
<td>Statistics/person</td>
<td>25</td>
</tr>
</tbody>
</table>
ABSTRACT

We examine the tradeoff between privacy and usability of statistical databases. We model a statistical database by an n-bit string \(d_1, ..., d_n\), with a query being a subset \(q \subseteq [n]\) to be answered by \(\sum_{i \in q} d_i\). Our main result is a polynomial reconstruction algorithm of data from noisy (perturbed) subset sums. Applying this reconstruction algorithm to statistical databases we show that in order to achieve privacy one has to add perturbation of magnitude \(\Omega(\sqrt{n})\). That is, smaller perturbation always results in a strong violation of privacy. We show that this result is tight by exemplifying access algorithms for statistical databases that preserve privacy while adding perturbation of magnitude \(O(\sqrt{n})\).

For time-\(T\) bounded adversaries we demonstrate a privacy-preserving access algorithm whose perturbation magnitude is \(\approx \sqrt{T}\).
2006: Differential Privacy
The 2000 and 2010 Disclosure Avoidance System operated as a filter, on the Census Edited File:

- Raw data from respondents: Decennial Response File
- Selection & unduplication: Census Unedited File
- Edits, imputations: Census Edited File
- Confidentiality edits (household swapping), tabulation recodes: Hundred-percent Detail File
- Pre-specified tabular summaries: PL94-171, SF1, SF2 (SF3, SF4, ... in 2000)
- Special tabulations and post-census research
The protection system used in 2000 and 2010 relied on swapping households:

Advantages of swapping:
- Easy to understand
- Does not affect state counts if swaps are within a state
- Can be run state-by-state
- Operation is “invisible” to rest of Census processing

Disadvantages:
- Does not consider or protect against database reconstruction attacks
- Does not provide formal privacy guarantees
- Swap rate and details of swapping must remain secret.
- Privacy guarantee based on the lack of external data
The US Census Bureau embraces formal privacy.
Motivation:
To protect the privacy of individual survey responses

2010 Census:
- 7.7 billion independent tabular summaries published
- 25 records per person

Database reconstruction (Dinur and Nissim 2003) is a serious disclosure threat that all statistical tabulation systems from confidential data must acknowledge.

The confidentiality edits applied to the 2010 Census were not designed to defend against this kind of attack.
Our plan is to create a “Disclosure Avoidance System” that drops into the Census production system.

Features of the DAS:

- Operates on the edited Census records
- Designed to make records that are “safe to tabulate.”

![Diagram showing the process from Census Edited File through Disclosure Avoidance System to Microdata Detail File (2020).]
The Disclosure Avoidance System allows the Census Bureau to enforce global confidentiality protections.
The Census disclosure avoidance system will use differential privacy to defend against a reconstruction attack,

Differential privacy provides:

- Provable bounds on the accuracy of the best possible database reconstruction given the released tabulations.

- Algorithms that allow policy makers to decide the trade-off between accuracy and privacy.

Final privacy-loss budget determined by Data Stewardship Executive Policy Committee (DSEP) with recommendation from Disclosure Review Board (DRB)
The Disclosure Avoidance System relies on infusing formally private noise.

Advantages of noise infusion with formal privacy:
- Easy to understand
- Provable and *tunable* privacy guarantees
- Privacy guarantees do not depend on external data
- Protects against database reconstruction attacks
- Privacy operations are *composable*

Disadvantages:
- Entire country must be processed at once for best accuracy
- Every use of private data must be tallied in the *privacy loss budget*
Differentially Private Disclosure Avoidance System: Requirements

DAS must be able to read the Census Edited File (CEF):
- CEF must be exactly specified and contain all information necessary for all tabulation recodes
- CEF must be kept confidential after DAS runs (as it was for historical censuses)

DAS must generate the Microdata Detail File (MDF):
- Must contain all information that appears in any publicly released table (e.g. PL94-171, SF1, SF2)
- Should not contain any information that does not appear in a publicly released table
- May be publicly released (in whole or in part)

Non-functional requirements:
- The disclosure avoidance system must provably move information from the CEF to PL94/SF1/SF2 with an adjustable total privacy-loss budget
- The source code and parameters for the DAS will be made publicly available
Why generate a differentially private MDF?

- Familiar to internal and external stakeholders
- Operates with legacy tabulation systems to produce PL-94 and SF-1 tabulations
- Guarantees population totals (voting age, non-voting age, householder) exact at all levels of geography
- Consistency among query answers
Challenges in creating a differentially private MDF

Changes required to Census business processes:
- All desired queries on MDF must be known in advance.
- All uses of confidential data need to be tracked and accounted.
- Data quality checks on tables cannot be done by looking at raw data.

Communications challenges:
- Differential privacy is not widely known or understood.
- Many data users want highly accurate data reports on small areas.
- Users in 2000 and 2010 didn’t know the error introduced by swapping.
Differential privacy meets Article 2, Section 1 and PL-94

“Invariants”
Specific PL-94 queries must be exact:
- Block population
- Block voting age population
- Block householders & vacancies

“Privacy protected”
Other PL-94 and SF-1 queries will not be exact:
- Age distribution under 18
- Age distribution 18 and over
- Race and ethnicity distribution
- Household relationship distribution
- Household ownership distribution
2018 End-to-End Test
Providence County, R.I.
2018 “Block-by-Block” System: High-level Overview:

Person-level histogram for each block → Differential privacy geometric mechanism → Optimizer applies invariants → PL-94 Published Results

Invariants:
# total population
# age >= 18

Advantages:
Easy to implement
Provable privacy protection
Provable application of invariants

Published Results: Census.gov
2020 Census of Population and Households
How the 2020 System Works: High-level Overview

- Every record in the population may be modified
  
  *But modifications are bounded by the global privacy budget.*

- Records in the tabulation data have no exact counterpart in the confidential data
  
  *There is no one-to-one mapping between CEF and MDF records.*
  
  *But there are the same number of records for every block.*

- Explicitly protected tabulations (PL-94 and SF-1) have provable, public accuracy levels
  
  *2020 will publish the algorithms, the parameters and the accuracy of the tabulations.*
**Proposed “Top-Down” Algorithm**

- **National table of US population**
  - $2 \times 255 \times 17 \times 115$
- **Spend $\epsilon_1$ privacy-loss budget**
- **National table with all 500,000 cells filled, structural zeros imposed with accuracy allowed by $\epsilon_1$**
  - $2 \times 255 \times 17 \times 115$
- **Reconstruct individual micro-data without geography**
  - 325,000,000 records

- **Sex**: Male / Female
- **Race + Hispanic**: 255 possible values
- **Relationship to Householder**: 17
- **Age**: 0-114

325,000,000 records
State-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1.

Target state-level tables required for best accuracy for PL-94 and SF-1. Exact state voting-age, non-voting age, and householder counts as enumerated.

Construct best-fitting individual micro-data with state geography.

325,000,000 records now including state identifiers.

Spend $\varepsilon_2$ privacy-loss budget.
County-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Target county-level tables required for best accuracy for PL-94 and SF-1

Exact county voting-age, non-voting age, and householder counts as enumerated.

Construct best-fitting individual micro-data with state and county geography

325,000,000 records now including state and county identifiers

Spend $\epsilon_3$ privacy-loss budget
Tract-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend $\epsilon_4$ privacy-loss budget

Target tract-level tables required for best accuracy for PL-94 and SF-1

Exact tract voting-age, non-voting age, and householder counts as enumerated.

Construct best-fitting individual micro-data with state, county, and tract geography

325,000,000 records now including state, county, and tract identifiers
Block-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend $\varepsilon_5$ privacy-loss budget

Block tract-level tables required for best accuracy for PL-94 and SF-1
Exact block voting-age, non-voting age, and householder counts as enumerated.

Construct best-fitting individual micro-data with state, county, tract and block geography

325,000,000 records now including state, county, tract identifiers
MDF for tabulating

Construct best-fitting individual micro-data with state, county, tract and block geography

325,000,000 records now including state, county, tract, and block identifiers

MDF used for tabulating PL-94, SF-1
MDF for tabulating

How accurate is the MDF?
- Disclosure Avoidance Certificate
  - Certifies that the DAS passed tests
  - Reports the accuracy of the MDF
  - Requires $\varepsilon_A$

Construct best-fitting individual micro-data with state, county, tract and block geography

325,000,000 records now including state, county, tract, and block identifiers

MDF used for tabulating PL-94, SF-1
Operational Decisions

Set total privacy loss budget: $\varepsilon$

- Ensure that $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 + \varepsilon_A = \varepsilon$

Within each stage, allocate privacy-loss budget between:

- PL-94
- Parts of SF-1 not in PL-94

These are policy levers provided by the system.
Levers are set by the Data Stewardship Executive Policy Committee
Inputs Used by the Development Team

Lists of matrices in technical documentation express core queries in the workload


Over 1,000 pages of edit specifications for 2010 CEF

Uncurated tabulation recode programs
We are creating

A framework for Disclosure Avoidance Systems:
- Development & Test Mode
- Production Mode

Testing Systems:
- DAS0 — 100% accuracy, no privacy
  (No disclosure avoidance)
- DAS1 — 100% privacy, no accuracy
- DAS2 — “bottom-up” engine

Operational System:
- DAS3 — “top-down” engine
Plans for the 2018 End-to-End Test

The 2018 End-to-End test will incorporate differential privacy
- Likely DAS2 — Bottom-up algorithm

Only the prototype PL94-171 files will be produced

No decisions yet regarding the privacy-loss budget or accuracy level

Questions?