Differential Privacy and the 2020 Census

Simson L. Garfinkel
U.S. Census Bureau

March 3, 2020
Presentation at Google

The views in this presentation are those of the author, and not those of the U.S. Census Bureau.
Abstract

The goal of the 2020 Census is to count everyone once, only once and in the right place. The decennial activity, mandated by the US Constitution, was first overseen by Thomas Jefferson in 1790 is the oldest continuously operating statistical program on the planet.

As part of the 2020 Census, each household in the United States will be asked to provide the number of residents, their ages, sex, race, ethnicity, as well as the inter-household relationship. This information will used to apportion the House of Representatives and to distribute more than $675 billion in federal aid to the US states.

The Census Bureau is legally prohibited from making publications in which the data contributed by a specific individual or establishment can be identified. Advances in computer performance, computer science, and the availability of “big data” makes that harder today than ever before.

In 2018-19, the Census Bureau conducted a “red-team” attack against the data that it published from the 2010 census and discovered that it could reconstruct microdata for all 308,745,538 residents, and that it could correctly re-identify data from 52 million.

Differential privacy was created in 2006 to precisely solve this problem. With differential privacy, it is possible to bound the privacy loss that results from a data publication, but doing so decreases the accuracy of the published data. It does this by introducing uncertainty, or error, into the published statistics. While the naïve application of differential privacy can result in substantial error for even modest privacy protection, it is possible to create sophisticated algorithms that do a better job balancing accuracy and privacy loss.
Acknowledgments

This presentation incorporates work by:
John Abowd (Chief Scientist)
Dan Kifer (Scientific Lead)
Simson Garfinkel (Senior Computer Scientist for Confidentiality and Data Access)
Rob Sienkiewicz (Chief, Center for Enterprise Dissemination)
Outline

Motivation
Technology change and the US Census Bureau
Privacy protection for the 2020 Census
Challenges deploying differential privacy
The public policy questions
Motivation

Article 1, Section 2

“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.”
“...in such Manner as they shall by Law direct.” Public Law 94-171


PL94-171 and SF1 Statistics per Census Block:
P1 – Total population by block x RACE (PL94-171)
P2 – Total population, (Hispanic & Not Hispanic) x RACE (PL94-171)
P3 – Race for Population 18 years and over (PL94-171)
P4 – (Hispanic & Not Hispanic) 18 years and over x RACE (PL94-171)
P12 and P12A-H – Sex By Age (23 age buckets) x RACE (SF1)
P14 – Sex By Age For Population Under 20 (20 age buckets) (SF1)
P22 – Household Type by Age of Householder (5 year buckets) (SF1)
P42 – Group Quarters population by GQ type (PL94-171)
H1 – Occupancy Status (Occupied & Vacant) (PL94-171)

SF1 Statistics per Census Tract:
PCT12 – Sex By Age (105 age buckets)
...
Uses of the Decennial Census Data

Apportioning the House of Representatives (U.S. Constitution)
50 numbers of total state population as of April 1

Enforcing Voting Rights Act of 1965 Section 2
Prohibits every state and local government from imposing any voting law that results in discrimination against racial or language minorities.

Distributing Federal Funds
$675 billion in FY2015
Privacy and the Decennial Census

Title 13 Section 9 of the US Code Prohibits the US Census Bureau from making any publication that reveals data provided by a person or an establishment.

Respondent data cannot be used for non-statistical purposes.

Census Bureau employees are sworn for life to protect respondent data.
Data that we collect:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Range</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>6,207,027 inhabited blocks</td>
<td>23</td>
</tr>
<tr>
<td>Sex</td>
<td>2 (Female/Male)</td>
<td>1</td>
</tr>
<tr>
<td>Age</td>
<td>103 (0-99 single age year categories, 100-104, 105-109, 110+)</td>
<td>7</td>
</tr>
<tr>
<td>Race</td>
<td>63 allowable race combinations</td>
<td>7</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>2 (Hispanic/Not)</td>
<td>1</td>
</tr>
<tr>
<td>Relationship</td>
<td>17 values</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>44</td>
</tr>
</tbody>
</table>

2010 values:
308,745,538 people x 6 variables = 1,852,473,228 measurements
308,745,538 people x 44 bits = 13,584,803,672 bits ≈ 1.7 GB
# 2010 Census: Summary of Publications (approximate counts)

<table>
<thead>
<tr>
<th>Publication</th>
<th>Released counts</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>Total Statistics in PL94-171 and Balance of SF1:</td>
<td>5,578,897,932</td>
</tr>
<tr>
<td>Published Statistics/person</td>
<td>18</td>
</tr>
<tr>
<td>Recall: Collected variables/person:</td>
<td>6</td>
</tr>
<tr>
<td>Published Statistics/collected variable</td>
<td>$18 \div 6 \approx 3$</td>
</tr>
</tbody>
</table>
Publishing too many queries on a confidential database with too much accuracy reveals the contents of the database. Today we call this the “fundamental law of information recovery.”

Dinur & Nissim proposed a generalized solution of adding noise.
Differential Privacy tells us how much noise to add!

Key features:
- Lower bound for the amount of noise that needs to be added
- Upper bound for privacy loss
- Mechanisms are composable
Technology and the Decennial Census
Punch Cards were invented for the 1890 Census

https://www.census.gov/history/www/innovations/technology/the_hollerith_tabulator.html
The Census Bureau bought UNIVAC 1, the world’s first commercial general-purpose electronic digital computer

https://www.census.gov/history/www/innovations/technology/univac_i.html

June 14, 1951
Differential Privacy was invented for the US Census

Cynthia Dwork at the Harvard Data Initiative Conference, October 25, 2019
The Census Bureau deployed DP “OnTheMap” in 2008!

https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf
Differential privacy was not ready for the 2010 census.

September 26, 2005 – Census Bureau awarded $500+ million contract to Lockheed Martin Corporation for the 2010 Census Decennial Response Integration System (DRIS)

March 30, 2009 – Census Bureau launches a massive operation to verify and update more than 145 million addresses as it prepares to mail out the 2010 census questionnaire.

March 1, 2010 – 2010 census questionnaires begin arriving in mailboxes throughout the United States and Island Areas

April 1, 2010 – Census Day.

December 21, 2010 – The Census Bureau announces the 2010 population counts and delivers the apportionment counts to the president.

https://www.census.gov/history/www/through_the_decades/overview/2010_overview_1.html
The 2010 Privacy Mechanism

Some statistics were published at the block level.
A single household on a block might be highly identifiable!
The 2010 privacy mechanism protected these households.

Hypothetical block 00010000001
The 2010 privacy mechanism swapped households with others the same size.

**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
- Privacy protection is not quantified
- Swap rate and details of swapping must remain secret
- Privacy guarantee based on the lack of external data
Household–level swapping was applied after editing, before tabulation.

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
We now know that the privacy techniques we used in the 2010 Census were flawed. These were the best available techniques at the time!

Assumed that disclosure avoidance modifications made for two products from the same confidential data are compatible

Released exact counts at the block, tract and county level.

Released exact counts for age in years, OMB race/ethnicity, sex, relationship to householder, in Summary File 2: detailed race data
Statistical agencies aggregate data from many households together into a single publication.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>7</td>
<td>30</td>
<td>38</td>
</tr>
<tr>
<td># female</td>
<td>4</td>
<td>30</td>
<td>33.5</td>
</tr>
<tr>
<td># male</td>
<td>3</td>
<td>30</td>
<td>44</td>
</tr>
<tr>
<td># black</td>
<td>4</td>
<td>51</td>
<td>48.5</td>
</tr>
<tr>
<td># white</td>
<td>3</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td># married</td>
<td>4</td>
<td>51</td>
<td>54</td>
</tr>
<tr>
<td># black F</td>
<td>3</td>
<td>36</td>
<td>36.7</td>
</tr>
</tbody>
</table>
We now know how to take many aggregate publications and “solve” for the original microdata.

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>7</td>
<td>30</td>
<td>38</td>
</tr>
<tr>
<td># female</td>
<td>4</td>
<td>30</td>
<td>33.5</td>
</tr>
<tr>
<td># male</td>
<td>3</td>
<td>30</td>
<td>44</td>
</tr>
<tr>
<td># black</td>
<td>4</td>
<td>51</td>
<td>48.5</td>
</tr>
<tr>
<td># white</td>
<td>3</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td># married</td>
<td>4</td>
<td>51</td>
<td>54</td>
</tr>
<tr>
<td># black F</td>
<td>3</td>
<td>36</td>
<td>36.7</td>
</tr>
</tbody>
</table>

This table can be expressed by 164 equations. Solving those equations takes 0.2 seconds on a 2013 MacBook Pro.
We performed a database reconstruct and re-identification attack for all 308,745,538 people in the 2010 Census.

Reconstructed all 308,745,538 microdata records
Used four commercial databases of the 2010 US population acquired 2009-2011 in support of the 2010 Census.

Link rate: 45%
Validated Re-identification Rate: 38% (17% of the US population)
Privacy protection for the 2020 Census
In 2017, the Census Bureau announced that it would use differential privacy for the 2020 Census.

DP is a tool for controlling privacy-loss/accuracy trade-off

DP lets us put the accuracy where it is needed.

DP privacy is “future-proof”

Records in the tabulation data will have no exact counterpart in the confidential data

Explicitly protected tabulations have provable, public accuracy levels

- PL 94-171
- Demographic and Housing Characteristics (DHC)
There was no off-the-shelf system for applying differential privacy to a national census

We had to create a new system that:
Produced higher-quality statistics at more densely populated geographies
Produced consistent tables

We created a new differential privacy algorithm and system that:
Produces statistics from the top-down
• e.g. National Level -> State Level -> County Level -> Tract Level -> Block Level
• Creates protected microdata that can be used for any tabulation without additional privacy loss
Fits into the decennial census production system
We planned to create a “Disclosure Avoidance System” that dropped into the Census production system.

**Features of the DAS:**
Operates on the edited Census records
Create microdata that would be “safe to tabulate.”
Capture all necessary statistics in the microdata.
The Disclosure Avoidance System allows the Census Bureau to enforce global confidentiality protections*


- Pre-specified tabular summaries: PL94-171, DHC, Other Scheduled Tabulations
- Special tabulations and post-census research

Privacy-Loss Budget, Accuracy Decisions

*Note: See later in this presentation for the design that includes “Group II” data products
Our DP mechanism protects histograms of person types.

Census “block”

**Census “block” histogram**

<table>
<thead>
<tr>
<th>Count</th>
<th>Age</th>
<th>Sex</th>
<th>Race</th>
<th>Ethnicity</th>
<th>REL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>F</td>
<td>B</td>
<td>-</td>
<td>Child</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
<td>M</td>
<td>W</td>
<td>H</td>
<td>Single</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
<td>F</td>
<td>W</td>
<td>H</td>
<td>Single</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>M</td>
<td>W</td>
<td>-</td>
<td>HH</td>
</tr>
<tr>
<td>1</td>
<td>36</td>
<td>F</td>
<td>B</td>
<td>-</td>
<td>Spouse</td>
</tr>
<tr>
<td>1</td>
<td>66</td>
<td>F</td>
<td>B</td>
<td>-</td>
<td>HH</td>
</tr>
<tr>
<td>1</td>
<td>84</td>
<td>M</td>
<td>B</td>
<td>-</td>
<td>Spouse</td>
</tr>
</tbody>
</table>
First system applied DP to every block. This was the “block-by-block” system.
There are roughly 8 million blocks
We released public test results using data from the 1940 Census
In 2018 we adopted the TopDown Algorithm (TDA)

Computes and protects a histogram for various geographical units at various geographical levels

Illustrated for the current specification of the Demographic and Housing Characteristics Person tables with proposed, approximate 2020 tabulation geography

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
<th>Histogram size</th>
</tr>
</thead>
<tbody>
<tr>
<td>National</td>
<td>1</td>
<td>1.2M</td>
</tr>
<tr>
<td>State</td>
<td>51</td>
<td>1.2M</td>
</tr>
<tr>
<td>County</td>
<td>3142</td>
<td>1.2M</td>
</tr>
<tr>
<td>Census Tract</td>
<td>75,000</td>
<td>1.2M</td>
</tr>
<tr>
<td>Block group</td>
<td>275,000</td>
<td>1.2M</td>
</tr>
<tr>
<td>Blocks</td>
<td>8M</td>
<td>1.2M</td>
</tr>
</tbody>
</table>
The TopDown algorithm*

330M records

State-level measurements

County-level measurements

Tract-level measurements

Block-level measurements

National Histogram

51 state histograms

3,142 county histograms

75,000 census tract histograms

8 million block histograms

*v1 geographies
TDA part 1: protecting the data

records → Histogram → $\varepsilon$ → Noisy Histogram
TDA part 2: post-processing

Find \( B + C + D = A \)
Minimize 1:
\[ |b - B| + |c - C| + |d - D| \]
Minimize 2:
\[ (b - B)^2 + (c - C)^2 + (d - D)^2 \]
TDA part 2: post-processing

- 330M records
- State-level measurements
- County-level measurements
- Tract-level measurements
- Block-level measurements

Optimization happens here.

- National Histogram
- 51 state histograms
- 3,142 county histograms
- 75,000 census tract histograms
- 8 million block histograms
TDA statistics for runs on the 1940 Census data.

TOP-DOWN DIFFERENTIAL PRIVACY ALGORITHMS (1940 CENSUS DATA)

- Enumeration district
- County
- State
- Nation
TDA produces better statistics at all geography levels!

**COMPARISON OF DISTRICT RESULTS BY ALGORITHM**
(1940 CENSUS DATA)

- Top-down
- District-by-district

**COMPARISON OF NATIONAL RESULTS BY ALGORITHM**
(1940 CENSUS DATA)

- Top-down
- District-by-district
Challenges deploying differential privacy at the US Census Bureau

Oh, no—not another learning experience!
Our experience with OnTheMap did not prepare the organization for the challenge.

OnTheMap was a new product, designed from the start to be DP on the residential side.

(Haney et al. (2017) extends to the employment side)

The decennial Census of Population and Housing, first performed under the direction of Thomas Jefferson in 1790, is the oldest and most expensive statistical undertaking of the U.S. government.

Transitioning existing data products has revealed:

The limits of today’s formal privacy mechanisms

The difficulty of retrofitting legacy statistical products to conform with modern privacy practice
Managing the Tradeoff

We thought that one of our primary problems would be managing the privacy loss-accuracy tradeoff.

The DP research community has not created any significant theory or tools in this area.
Basic Principles

Based on recent economics (2019, American Economic Review)

The marginal social benefit is the sum of all persons’ willingness-to-pay for data accuracy with increased privacy loss

The marginal rate of transformation is the slope of the privacy-loss v. accuracy graphs we have been examining

This is exactly the same problem being addressed by Google in RAPPOR or PROCHLO, Apple in iOS 11, and Microsoft in Windows 10 telemetry
Marginal Social Benefit Curve

Social Optimum: 
MSB = MSC 
(0.25, 0.64)

Production Technology
Scientific Issue: Setting Epsilon

Where computer scientists act like Marginal Social Cost = Marginal Social Benefit

Where social scientists act like MSC = MSB

Production Possibilities for Alternative Mechanisms

- Proposed 2020 Census differential privacy implementation with use-case based accuracy improvements
- Simple differential privacy implementation with no accuracy improvements
- Randomized response method pioneered by Google, Apple and Microsoft
In theory, practice and theory are the same. In practice, they aren’t.
Initial Operational Issues

Obtaining Qualified Personnel and Tools

Recasting high-sensitivity queries

Identifying Structural Zeros

Obtaining a Suitable Computing Environment

Accounting for All Uses of Confidential Data
Hierarchical Mechanisms
We needed a novel mechanism that:
Assured consistent statistics from US->States->Counties->Tracts
Provided lower error for larger geographies.

Invariants
C1: Total population (invariant at the county level for the 2018 E2E)
C2: Voting-age population (population age 18 and older) (eliminated for the 2018 E2E)
C3: Number of housing units (invariant at the block level)
C4: Number of occupied housing units (invariant at the block level)
C5: Number of group quarters facilities by group quarters type.(invariant at the block level)

https://github.com/uscensusbureau/census2020-das-e2e/blob/master/etl_e2e/ipums_1940_validator.py
Invariants in the 2020 Census

Invariants in the 2020 will be decided by the Census Bureau’s Data Stewardship Executive Policy Committee (DSEP). DSEP replaced the policy of holding voting age population invariant because of grave concerns about its effects on the Census Bureau’s ability to protect confidentiality, especially in block and block-group level tabulations. The policy was replaced with explicit management of invariants and privacy-loss budgets beginning with the decision memos for Apportionment, the 2018 End-to-End Census Test data products, and the 2010 Demonstration Data Products. Those memos are:


DSEP has made no final decisions regarding invariants or the privacy-loss budget for the 2020 Census data publications.
Scientific Issues: Quality Metrics

What is the measure of “quality” or “utility” in a complex data product?

Options we considered:
L1 error between “true” data set and “privatized” data set
Impact on an algorithm that uses the data (e.g. voting rights enforcement)

Options others considered:
Enumeration District-by-Enumeration District differences of specific demographic subgroups

David Van Riper & Tracy Kugler, IPUMS (APDU 2019)
Many in the data user community are apprehensive about DP.

Steven Ruggles
@HistDem

I am increasingly convinced that DP will degrade the quality of data available about the population, and will make scientifically useful public use microdata impossible. 3/

3:07 PM - 5 Jul 2019
9 Retweets 32 Likes

I also believe that the DP approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau. 4/

3:07 PM - 5 Jul 2019
2 Retweets 13 Likes
2010 Demonstration Data Products

October 29, 2019 Data and Software Release
2010 Census Edited File as processed by the 2020 DAS (as of October 2019)
Release of the 2020 DAS Source Code (as of October 2019)
https://github.com/uscensusbureau/census2020-das-2010ddp/

December 11-12 Workshop
National Academies of Science
https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518
“2010 Demonstration Data Products” were not well received.

https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html
Issues Faced by Data Users

Access to Micro-data
Many users expect access to microdata.

Difficulties Arising from Increased Transparency
Many users were not aware of prior disclosure avoidance practices.
The swap rate was never made public.

Misunderstandings about Randomness and Noise Infusion
By summer 2019, it was clear that we would not realize our microdata goal

MDF will not include Household-Person Joins and Detailed Race/Ethnicity/Tribe
# 2020 DAS: Data Products

<table>
<thead>
<tr>
<th>Group I</th>
<th>Group II</th>
<th>Group III</th>
</tr>
</thead>
<tbody>
<tr>
<td>• PL94-171</td>
<td>• AIAN</td>
<td>• PUMS</td>
</tr>
<tr>
<td>• Demographic Profiles</td>
<td>• Detailed Race</td>
<td>• Special Tabulations</td>
</tr>
<tr>
<td>• DHC-Persons</td>
<td>• Person &amp; Household joins</td>
<td></td>
</tr>
<tr>
<td>• DHC-Households</td>
<td>• Averages</td>
<td></td>
</tr>
<tr>
<td>• CVAP (special tabulation)*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Supported by MDF*  
*Generated directly by 3rd Party DB*  
*Under discussion*

*Citizen Voting-Age Population by Race and Ethnicity block-level special tabulation will be produced by March 31, 2021. See: [https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/cvap.html](https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/cvap.html), and technical documentation posted therein.*
Recommendations

Keep decision makers “in-the-loop.”
Adopt a controlled vocabulary.
Have an integrated communications strategy.
Minimize confusion from source code releases and make code runnable by outsiders.
Involves data users early and often.
practice

These attacks on statistical databases are no longer a theoretical danger.

BY SIMSON GARFINKEL, JOHN M. ABOWD, AND CHRISTIAN MARTINDALE

Understanding Database Reconstruction Attacks on Public Data

IN 2020, THE U.S. Census Bureau will conduct the Constitutionally mandated decennial Census of Population and Housing. Because a census involves collecting large amounts of private data under the promise of confidentiality, traditionally statistics are published only at high levels of aggregation. Published statistical tables are vulnerable to database reconstruction attacks (DRAs), in which the underlying microdata is recovered merely by finding a set of microdata that is consistent with the published statistical tabulations. A DRA can be performed by using the tables to create a set of mathematical constraints and then solving the resulting set of simultaneous equations. This article shows how such an attack can be addressed by adding noise to the published tabulations, so the reconstruction no longer results in the original data. This has implications for the 2020 census.

The goal of the census is to count every person once, and only once, and in the correct place. The results are used to fulfill the Constitutional requirement to apportion the seats in the U.S. House of Representatives among the states according to their respective numbers.

In addition to this primary purpose for the census, the U.S. Census Bureau has mandated many other uses for the data. For example, the U.S. Department of Justice uses block-by-block counts by race for enforcing the Voting Rights Act. More generally, the results of the decennial census, combined with other data, are used to help distribute more than $675 billion in federal funds to states and local organizations.

Beyond collecting and distributing data on U.S. citizens, the Census Bureau is also charged with protecting the privacy and confidentiality of survey respondents. The Census Bureau is responsible for upholding the confidentiality standard prescribed by Title 13, Section 9 of the U.S. Code, which states that Census Bureau publications are prohibited from identifying “the data furnished by any particular establishment or individual.” This section prohibits the Census Bureau from publishing respondents’ names, addresses, or any other information that might identify a specific person or establishment.

Upholding this confidentiality requirement frequently poses a challenge, because many statistics can inadvertently provide information in a way that can be attributed to a particular entity. For example, if a statistical agency accurately reports there are two persons living on a block and the average age of the block’s residents is 35, that would constitute an improper disclosure of personal information, because one of the residents could look up the data, subtract their contribution, and infer the age of the other.

These attacks on statistical databases are no longer a theoretical danger.

BY SIMSON GARFINKEL, JOHN M. ABOWD, AND CHRISTIAN MARTINDALE

Understanding Database Reconstruction Attacks on Public Data

IN 2020, THE U.S. Census Bureau will conduct the Constitutionally mandated decennial Census of Population and Housing. Because a census involves collecting large amounts of private data under the promise of confidentiality, traditionally statistics are published only at high levels of aggregation. Published statistical tables are vulnerable to database reconstruction attacks (DRAs), in which the underlying microdata is recovered merely by finding a set of microdata that is consistent with the published statistical tabulations. A DRA can be performed by using the tables to create a set of mathematical constraints and then solving the resulting set of simultaneous equations. This article shows how such an attack can be addressed by adding noise to the published tabulations, so the reconstruction no longer results in the original data. This has implications for the 2020 census.

The goal of the census is to count every person once, and only once, and in the correct place. The results are used to fulfill the Constitutional requirement to apportion the seats in the U.S. House of Representatives among the states according to their respective numbers.

In addition to this primary purpose for the census, the U.S. Census Bureau has mandated many other uses for the data. For example, the U.S. Department of Justice uses block-by-block counts by race for enforcing the Voting Rights Act. More generally, the results of the decennial census, combined with other data, are used to help distribute more than $675 billion in federal funds to states and local organizations.

Beyond collecting and distributing data on U.S. citizens, the Census Bureau is also charged with protecting the privacy and confidentiality of survey respondents. The Census Bureau is responsible for upholding the confidentiality standard prescribed by Title 13, Section 9 of the U.S. Code, which states that Census Bureau publications are prohibited from identifying “the data furnished by any particular establishment or individual.” This section prohibits the Census Bureau from publishing respondents’ names, addresses, or any other information that might identify a specific person or establishment.

Upholding this confidentiality requirement frequently poses a challenge, because many statistics can inadvertently provide information in a way that can be attributed to a particular entity. For example, if a statistical agency accurately reports there are two persons living on a block and the average age of the block’s residents is 35, that would constitute an improper disclosure of personal information, because one of the residents could look up the data, subtract their contribution, and infer the age of the other.

These attacks on statistical databases are no longer a theoretical danger.
The Most Technical 2020 Publications

Most recent public source code:
https://github.com/uscensusbureau/census2020-das-2010ddp

System Design Specification

Scientific paper describing mechanism:
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, 2018 CSAC (detailed discussion of algorithms and choices) https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf#


June 6, 2019 Blog explaining how to use the code base with the 1940 Census public data from IPUMS https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html

June 11, 2019 Keynote address “The U.S. Census Bureau Tries to Be a Good Data Steward for the 21st Century” ICML 2019 abstract, video

June 29-31, 2019 Joint Statistical Meetings Census Bureau electronic press kit (See talks by Abowd, Ashmead, Garfinkel, Leclerc, Sexton, and others)