Differential Privacy and the 2020 Decennial Census

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

Baltimore Regional Transportation Board
Cooperative Forecasting Group
February 26, 2020

Shape your future
START HERE >

United States® Census 2020

Acknowledgements

This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, including: John Abowd, Tammy Adams, Robert Ashmead, Craig Corl, Ryan Cummings, Jason Devine, John Fattaleh, Simson Garfinkel, Nathan Goldschlag, Michael Hawes, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Kyle Irimata, Dan Kifer, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Claudia Molinar, Brett Moran, Ned Porter, Sarah Powazek, Vikram Rao, Chris Rivers, Anne Ross, Ian Schmutte, William Sexton, Rob Sienkiewicz, Matthew Spence, Tori Velkoff, Lars Vilhuber, Bei Wang, Tommy Wright, Bill Yates, and Pavel Zhurlev.

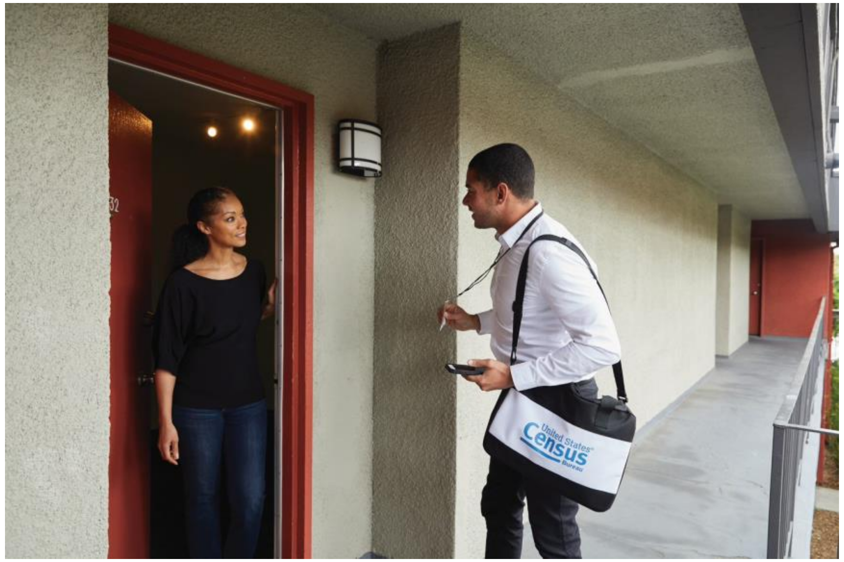
For more information and technical details relating to the issues discussed in these slides, please contact the author at michael.b.hawes@census.gov.

Any opinions and viewpoints expressed in this presentation are the author's own, and do not necessarily represent the opinions or viewpoints of the U.S. Census Bureau.

Our Commitment to Data Stewardship

Data stewardship is central to the Census Bureau’s mission to produce high-quality statistics about the people and economy of the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.



Shape your future START HERE >

United States® Census 2020

It's the Law

> *"To stimulate public cooperation necessary for an accurate census…Congress has provided assurances that information furnished by individuals is to be treated as confidential. Title 13 U.S.C. §§ 8(b) and 9(a) explicitly provide for nondisclosure of certain census data, and* ***no discretion is provided to the Census Bureau on whether or not to disclose such data****…"* (U.S. Supreme Court, Baldrige v. Shapiro, 1982)

Title 13, Section 9 of the United State Code prohibits the Census Bureau from releasing identifiable data "furnished by any particular establishment or individual."

Census Bureau employees are sworn for life to safeguard respondents' information.

Penalties for violating these protections can include fines of up to $250,000, and/or imprisonment for up to five years!

Keeping the Public's Trust

Safeguarding the public's data is about more than just complying with the law!

The quality and accuracy of our censuses and surveys depend on our ability to keep the public's trust.

In an era of declining trust in government, increasingly common corporate data breaches, and declining response rates to surveys, we must do everything we can to keep our promise to protect the confidentiality of our respondent's data.

Upholding our Promise: Today and Tomorrow

We cannot merely consider privacy threats that exist today.

We must ensure that our disclosure avoidance methods are also sufficient to protect against the threats of tomorrow!

The Privacy Challenge

Every time you release any statistic calculated from a confidential data source you "leak" a small amount of private information.

If you release too many statistics, too accurately, you will eventually reveal the entire underlying confidential data source.

Dinur, Irit and Kobbi Nissim (2003) "Revealing Information while Preserving Privacy" PODS, June 9-12, 2003, San Diego, CA

The Growing Privacy Threat

More Data and Faster Computers!

In today’s digital age, there has been a proliferation of databases that could potentially be used to attempt to undermine the privacy protections of our statistical data products.

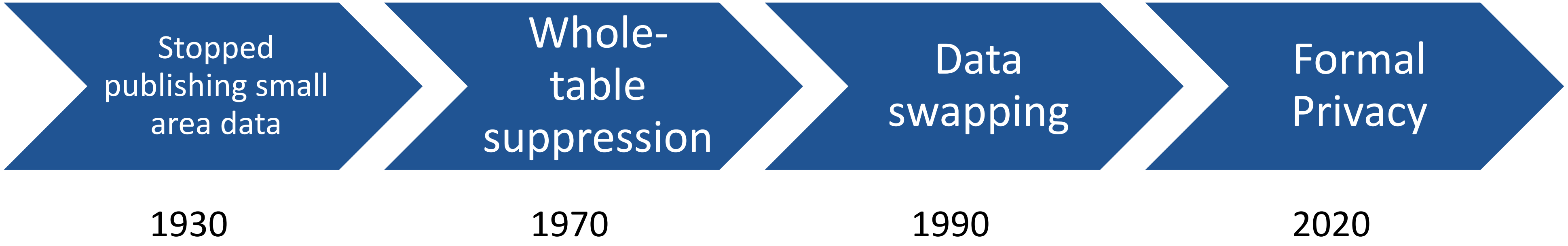
Similarly, today’s computers are able to perform complex, large-scale calculations with increasing ease.

These parallel trends represent new threats to our ability to safeguard respondents’ data.

The Census Bureau's Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.



Stopped publishing small area data

1930

Whole-table suppression

1970

Data swapping

1990

Formal Privacy

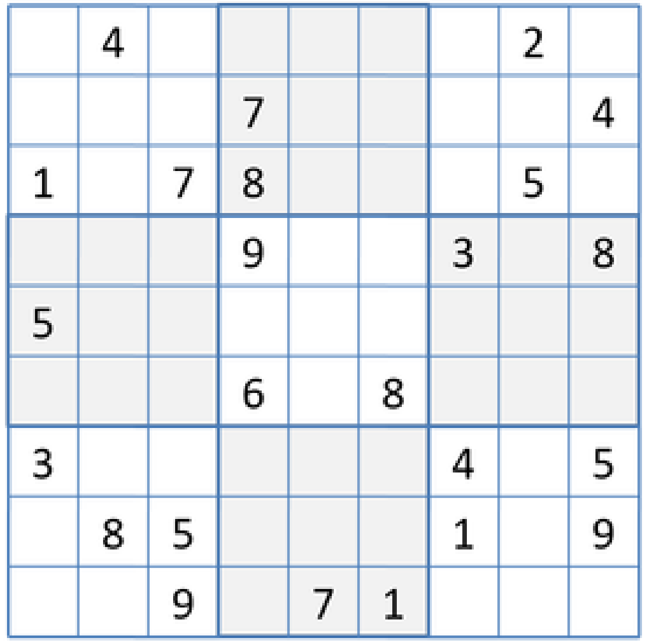
2020

Reconstruction

The recreation of individual-level data from tabular or aggregate data.

If you release enough tables or statistics, eventually there will be a unique solution for what the underlying individual-level data were.

Computer algorithms can do this very easily.



	4						2	
			7					4
1		7	8				5	
			9			3		8
5								
			6		8			
3						4		5
	8	5				1		9
		9		7	1			

Reconstruction: An Example

	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7

Reconstruction: An Example

	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7

Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married
36	Female	Black	Married
8	Female	Black	Single
18	Male	White	Single
24	Female	White	Single

This table can be expressed by 164 equations.
Solving those equations takes 0.2 seconds on a 2013 MacBook Pro.

Re-identification

Linking public data to external data sources to re-identify specific individuals within the data.

Name	Age	Sex
Jane Smith	66	Female
Joe Public	84	Male
John Citizen	30	Male

External Data

+

Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married

Confidential Data

In the News

Reconstruction and Re-identification are not just theoretical possibilities…they are happening!

- Massachusetts Governor's Medical Records (Sweeney, 1997)
- AOL Search Queries (Barbaro and Zeller, 2006)
- Netflix Prize (Narayanan and Shmatikov, 2008)
- Washington State Medical Records (Sweeney, 2015)
- and many more…

Reconstructing the 2010 Census

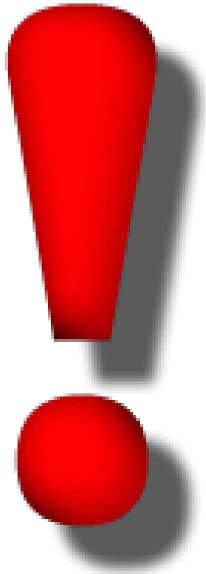
- The 2010 Census collected information on the age, sex, race, ethnicity, and relationship (to householder) status for ~309 Million individuals. (1.9 Billion confidential data points)
- The 2010 Census data products released over 150 billion statistics
- We conducted an internal experiment to see if we could reconstruct and re-identify the 2010 Census records.

Reconstructing the 2010 Census: What Did We Find?

1. On the 309 million reconstructed records, census block and voting age (18+) were correctly reconstructed for all records and for all 6,207,027 inhabited blocks.
2. Block, sex, age (in years), race (OMB 63 categories), and ethnicity were reconstructed:
    1. Exactly for 46% of the population (142 million individuals)
    2. Within +/- one year for 71% of the population (219 million individuals)
3. Block, sex, and age were then linked to commercial data, which provided putative re-identification of 45% of the population (138 million individuals).
4. Name, block, sex, age, race, ethnicity were then compared to the confidential data, which yielded confirmed re-identifications for 38% of the putative re-identifications (52 million individuals).
5. For the confirmed re-identifications, race and ethnicity are learned correctly, though the attacker may still have uncertainty.

The Census Bureau’s Decision

- Advances in computing power and the availability of external data sources make database reconstruction and re-identification increasingly likely.
- The Census Bureau recognized that its traditional disclosure avoidance methods are increasingly insufficient to counter these risks.
- To meet its continuing obligations to safeguard respondent information, the Census Bureau has committed to modernizing its approach to privacy protections.



Differential Privacy

aka “Formal Privacy”

-quantifies the precise amount of privacy risk…

-for all calculations/tables/data products produced…

-no matter what external data is available…

-now, or at any point in the future!

Assessing Privacy Risk

Traditional Disclosure Avoidance Considers Absolute Privacy Risk

Can an individual be re-identified in the data, and can some sensitive attribute about them be inferred?

Evaluates risk given a particular, defined mode of attack, asking: What is the likelihood, at this precise moment in time, of re-identification and inferential disclosure by a particular type of attacker with a defined set of available external information?

Formal Privacy is about Relative Privacy Risk

Does not directly measure re-identification risk (which requires specification of an attacker model).

Instead, it defines the maximum privacy “leakage” of each release of information compared to some counterfactual benchmark (e.g., compared to a world in which a respondent does not participate, or provides incorrect information).

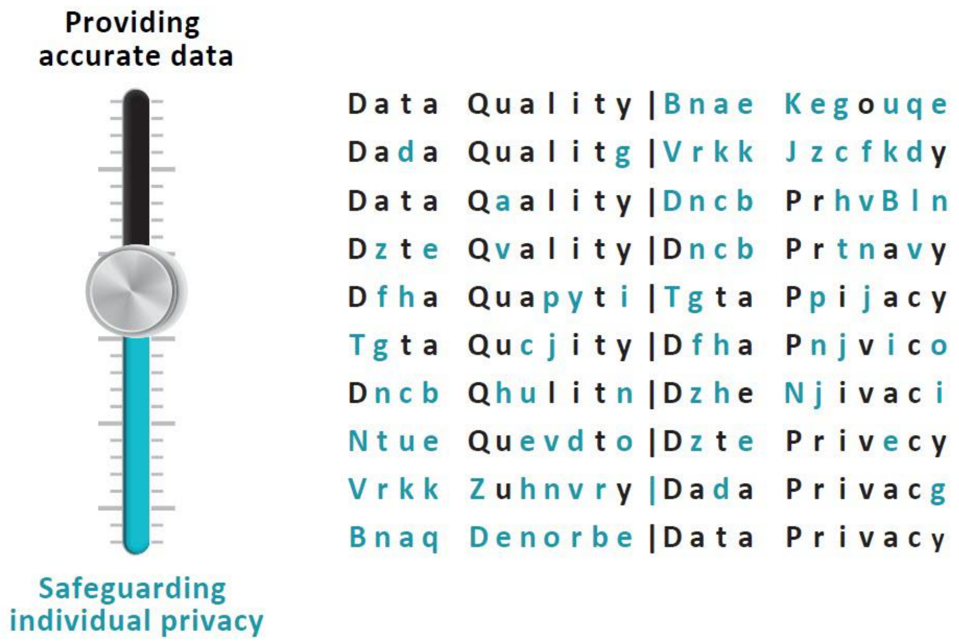
Precise amounts of noise

Differential privacy allows us to inject a precisely calibrated amount of noise into the data to control the privacy risk of any calculation or statistic.

Privacy vs. Accuracy

The only way to absolutely eliminate all risk of re-identification would be to never release any usable data.

Differential privacy allows you to quantify a precise level of “acceptable risk,” and to precisely calibrate where on the privacy/accuracy spectrum the resulting data will be.



Data Quality |Bnae Kegouqe
Dada Qualitg |Vrkk Jzcfkdy
Data Qaality |Dncb PrhvBln
Dzte Qvality |Dncb Prtnavy
Dfha Quapyti |Tgta Ppijacy
Tgta Qucjity |Dfha Pnjvico
Dncb Qhulitn |Dzhe Njivaci
Ntue Quevdto |Dzte Privecy
Vrkk Zuhnvry |Dada Privacg
Bnaq Denorbe |Data Privacy

Establishing a Privacy-loss Budget

This measure is called the “Privacy-loss Budget” (PLB) or “Epsilon.”

- ε=0 (perfect privacy) would result in completely useless data

- ε=∞ (perfect accuracy) would result in releasing the data in fully identifiable form

ε

Epsilon

The Formal Guarantee

Can Sara determine Joe’s exact age?

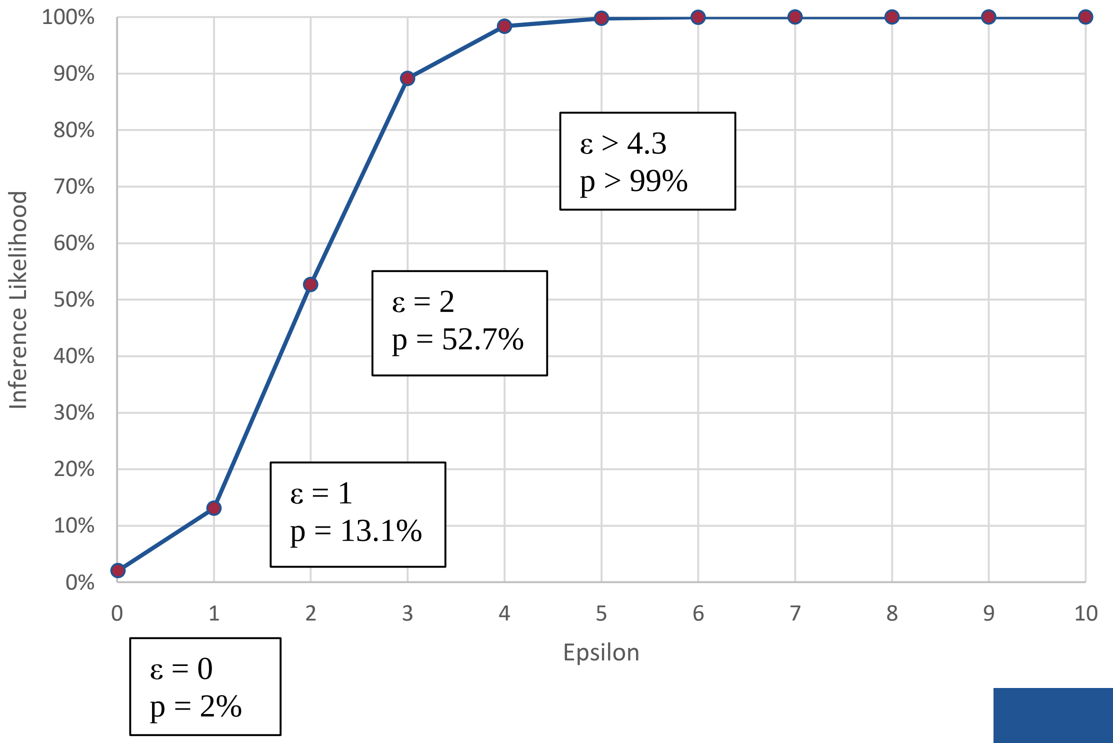
Suppose Joe submitted erroneous information for the Census, and the best Sara could otherwise do to determine Joe’s exact age (from other available information) is to predict that there is a 2% chance that Joe is 43 years old.

If Joe instead provides accurate information for the Census, then a small amount of information about him will “leak” through the publication of data products. This new information can improve Sara’s estimate.

The Privacy-loss Budget determines the amount of that leakage and the corresponding maximum possible improvement to Sara’s prediction.

Assumes that Sara has infinite computing resources, infinitely powerful algorithms, and allows her to have arbitrary side knowledge.

Inferred likelihood that Joe is actually 43 years old at varying levels of a privacy loss budget



ε > 4.3
p > 99%

ε = 2
p = 52.7%

ε = 1
p = 13.1%

ε = 0
p = 2%

Allocating the Privacy-loss Budget

Each calculation, query, or tabulation of the data consumes a fraction of the privacy-loss budget.

$(\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 \ldots + \varepsilon_n = \varepsilon_{Total})$

Calculations/tables for which high accuracy is critical can receive a larger share of the overall privacy-loss budget.

Keeping Accuracy High

When Differential Privacy is applied, the accuracy of the resulting data will be affected by:

- **The design of the algorithm**
- **The number of calculations being performed or tables being generated;**
- **The type of calculation being performed (e.g., count vs. mean);**
- **The size of the underlying populations for each calculation or table;**
- **The range of possible values;**
- **The overall privacy budget (epsilon); and**
- **The allocation of the privacy budget across calculations/tables.**

Comparing Methods

Data Accuracy

Differential Privacy is not inherently better or worse than traditional disclosure avoidance methods.

Both can have varying degrees of impact on data quality depending on the parameters selected and the methods’ implementation.

Privacy

Differential Privacy is substantially better than traditional methods for protecting privacy, insofar as it actually allows for measurement of the privacy risk.

Census TopDown Algorithm (TDA):
A Primer on Its Structure & Properties

Census TDA: Requirements and Properties I

TDA is the principal formally private 2020 Census disclosure limitation algorithm under development

Inputs:

- Post-edits-and-imputation microdata records (Census Edited File – CEF)
- Required structural zeros & data-dependent invariants

Processing:

- Convert CEF to an equivalent histogram
- Apply DP measurements & perform mathematical optimization
- Create noisy histogram; convert back to microdata

Output:

Return the Microdata Detail File (the MDF; microdata with same schema as CEF)

Example:

- Schema: Geography × Ethnicity × Race × Age × Sex × HHGQ
- This product yields a "histogram" (fully saturated contingency table)
- With shape: $\approx$ 10M × 2 × 63 × 116 × 2 × 43 = $\approx$ 10M × 1.25M

Census TDA: Requirements and Properties II

Data-dependent invariants:

Properties of true data that must hold exactly (*no noise*)

Current data-dependent invariants:

- State population totals
- Count of occupied GQ facilities by type by block (not population)
- Total count of housing units by block (not population)

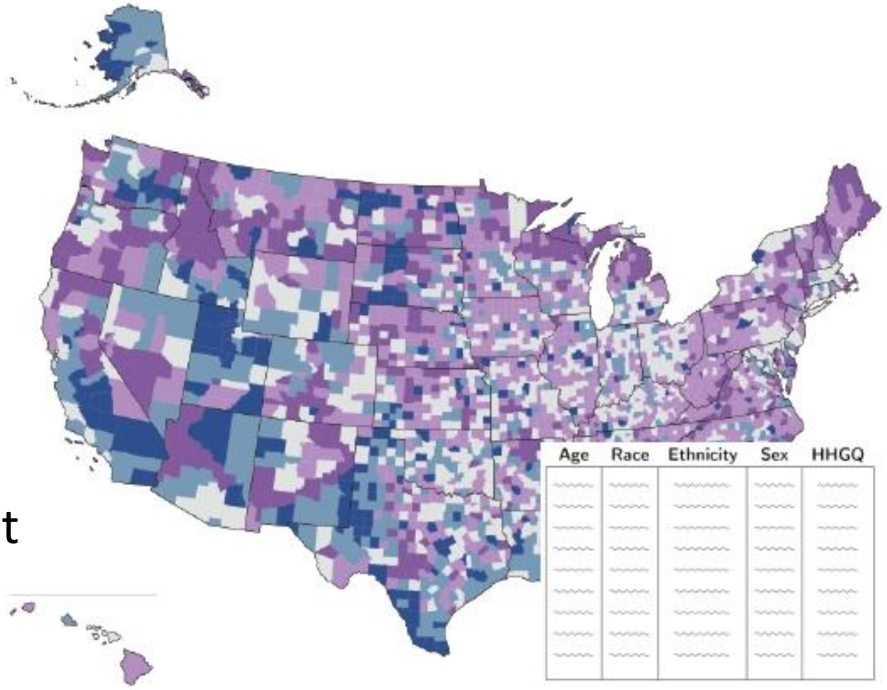
Utility/Accuracy for pre-specified tabulations

- Full privacy + full accuracy for arbitrary uses = impossible
- PL94-171: tabulations used for redistricting
- Demographic and Housing Characteristics File
  - Principal successor to 2010 Summary File 1
  - TDA creates separate Person and Housing Unit microdata sets

ϵ-consistency: error $\rightarrow 0$ as privacy loss $\epsilon \rightarrow \infty$

Transparency: source code and parameters made public

Basic Structure of TDA



1. Split privacy-loss budget ε into 6 pieces: $\varepsilon_{nat}, \varepsilon_{state}, \ldots$

2. Ignore geography, make national histogram $\tilde{H}^0$ using ε_{nat} budget

3. Using ε_{state} budget, make state histograms: $\tilde{H}^1_{AK}, \tilde{H}^1_{AL}, \ldots, \tilde{H}^1_{WY}$
   - Must be consistent
   - i.e., $\sum_{s \in states} \tilde{H}^1_s = \tilde{H}^0$

4. Recurse down the hierarchy

5. Invariants imposed as constraints in each optimization problem (with notable complications!)

Benefits of TDA

- Disclosure-limitation error does not increase with number of contained Census blocks
- A stark contrast with naïve alternatives (e.g., District-by-District)
- Yields increasing accuracy as number of observations increases
- “Borrows strength” from upper geographic levels to improve lower levels (for, e.g., sparsity)

Implications for the 2020 Decennial Census

The switch to Differential Privacy will not change the constitutional mandate to apportion the House of Representatives according to the actual enumeration.

As in 2000 and 2010, the Census Bureau will apply privacy protections to the PL94-171 redistricting data.

The switch to Differential Privacy requires us to re-evaluate the quantity of statistics and tabulations that we will release, because each additional statistic uses up a fraction of the privacy-loss budget (epsilon).

You Can Help Us to Help You!

Senior Census Bureau policymakers will be making important decisions – and they need your input!

The actual impact of Differential Privacy on the usability and accuracy of the 2020 Census data products will ultimately depend on the following factors:

- What will the overall privacy-loss budget (epsilon) be?
- What statistics will the Census Bureau release at which levels of geography?
- How will the overall privacy-loss budget be allocated across different geographies, tables, and products?

In order for the Census Bureau's senior leadership to make the most informed decisions on these questions, they need to know how you plan to use the 2020 Census data.

Shape your future START HERE >

United States Census 2020

2010 Demonstration Products

- Census Bureau has released a set of data products that demonstrate the computational capabilities of the DAS. The current version of the DAS was run on the 2010 internal data to produce two products:

  - PL 94-171

  - Demographic and Housing Characteristics File (selected tables)
- Allows data users to assess the impacts of the DAS implementation.
- Uses Privacy-Loss Budget of ε=6 (ε=4 for person records, ε=2 for household records)

Available at: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html

Shape your future START HERE >

United States® Census 2020

Impact on Privacy

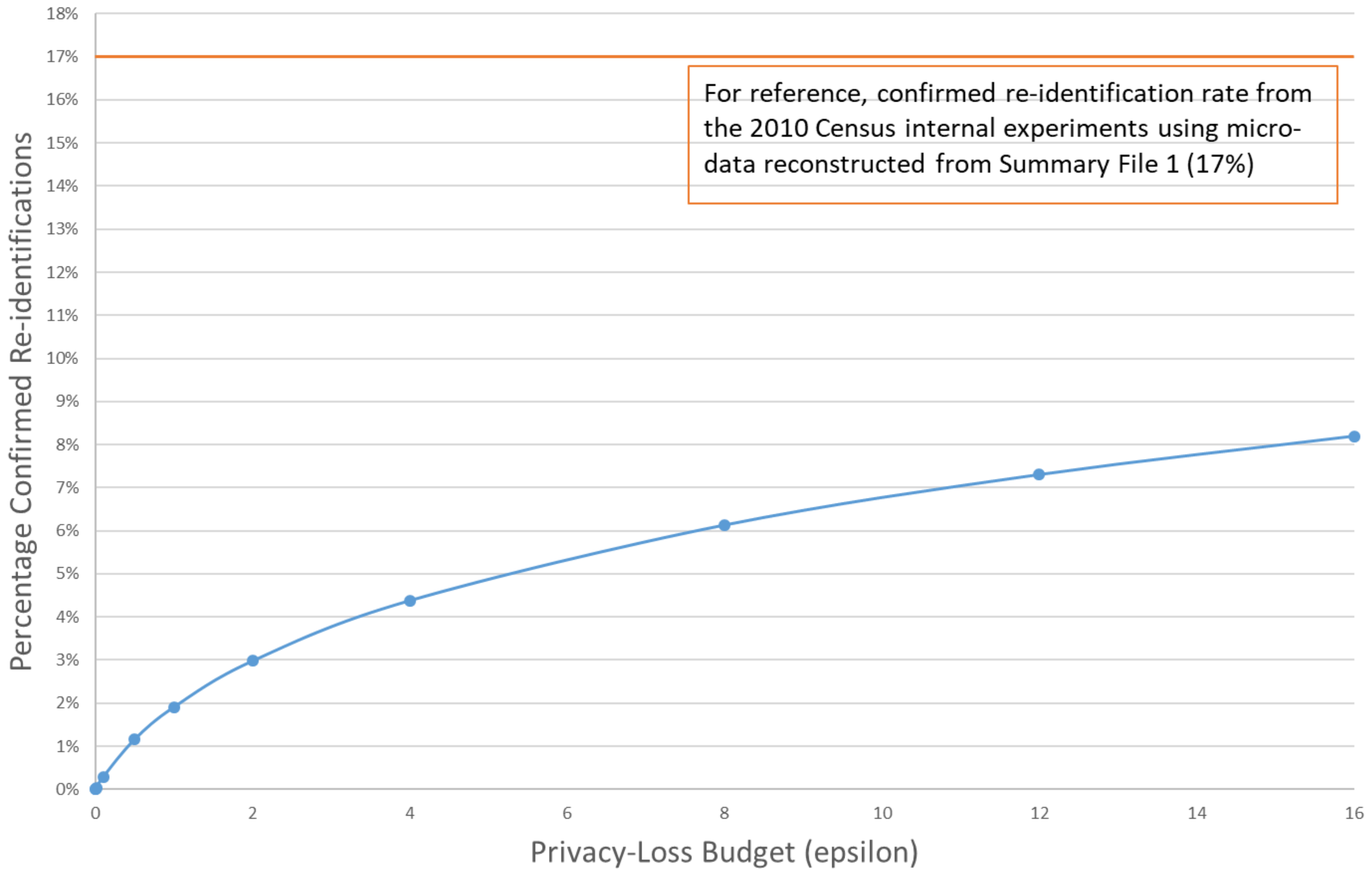
Using exactly the same re-identification strategy, we analyzed the differentially private microdata for persons at different privacy-loss budgets from ε=0 to ε=16.

We used ε=4 for the differentially private person-level microdata computed for the 2010 Demonstration Data Products.

Results varied from a confirmed re-identification rate of 0 at ε=0 to 8.2% at ε=16.

CBDRB-FY20-103

Confirmed Re-identifications as a Percentage of Total Population (2010 Census)



For reference, confirmed re-identification rate from the 2010 Census internal experiments using micro-data reconstructed from Summary File 1 (17%)

CBDRB-FY20-103

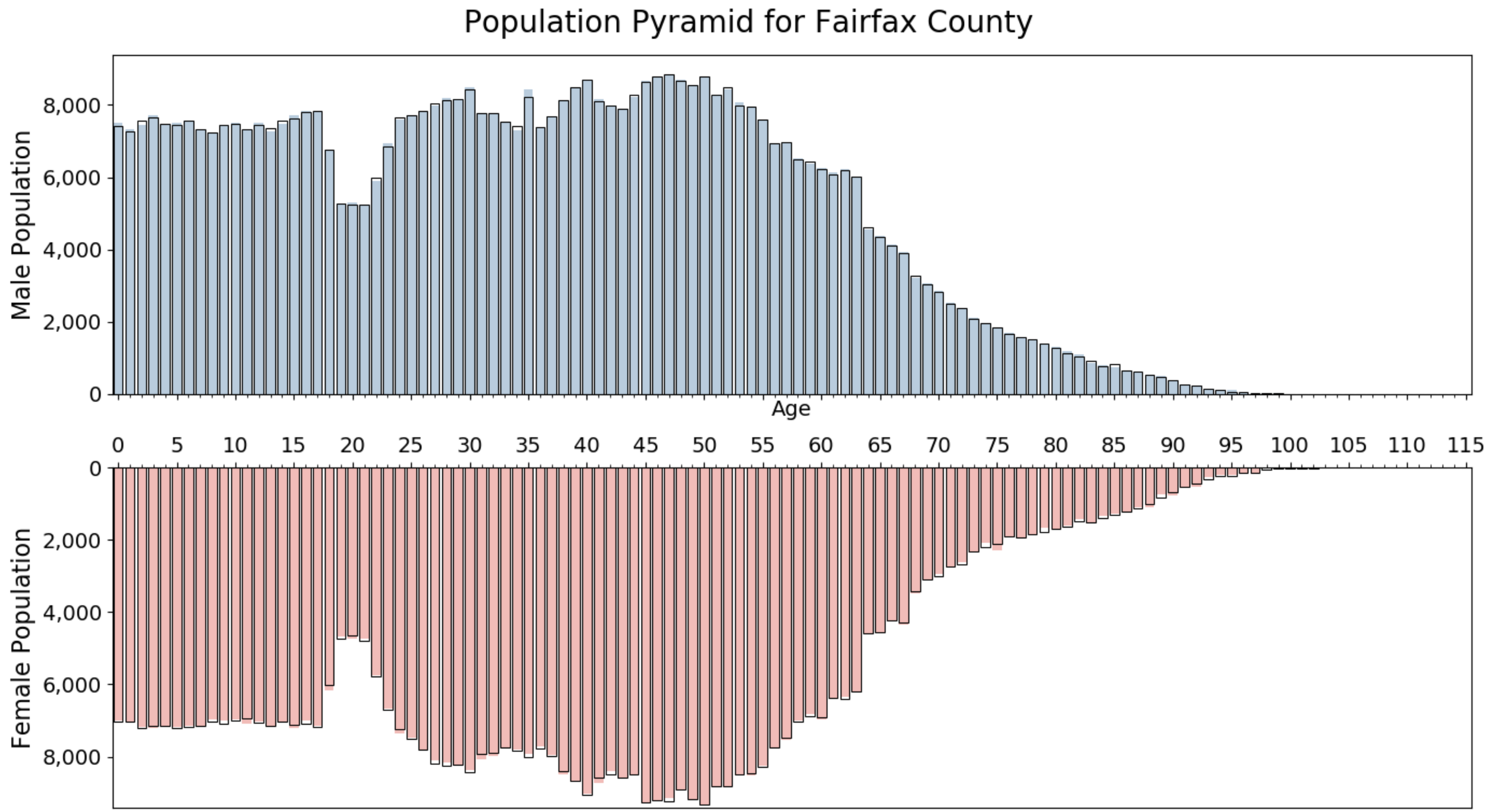
Impact on Accuracy

Comparison of TDA-generated MDF against the unswapped 2010 CEF

$\varepsilon = 4$

Fairfax County Population:

1,081,726



CBDRB-FY20-102

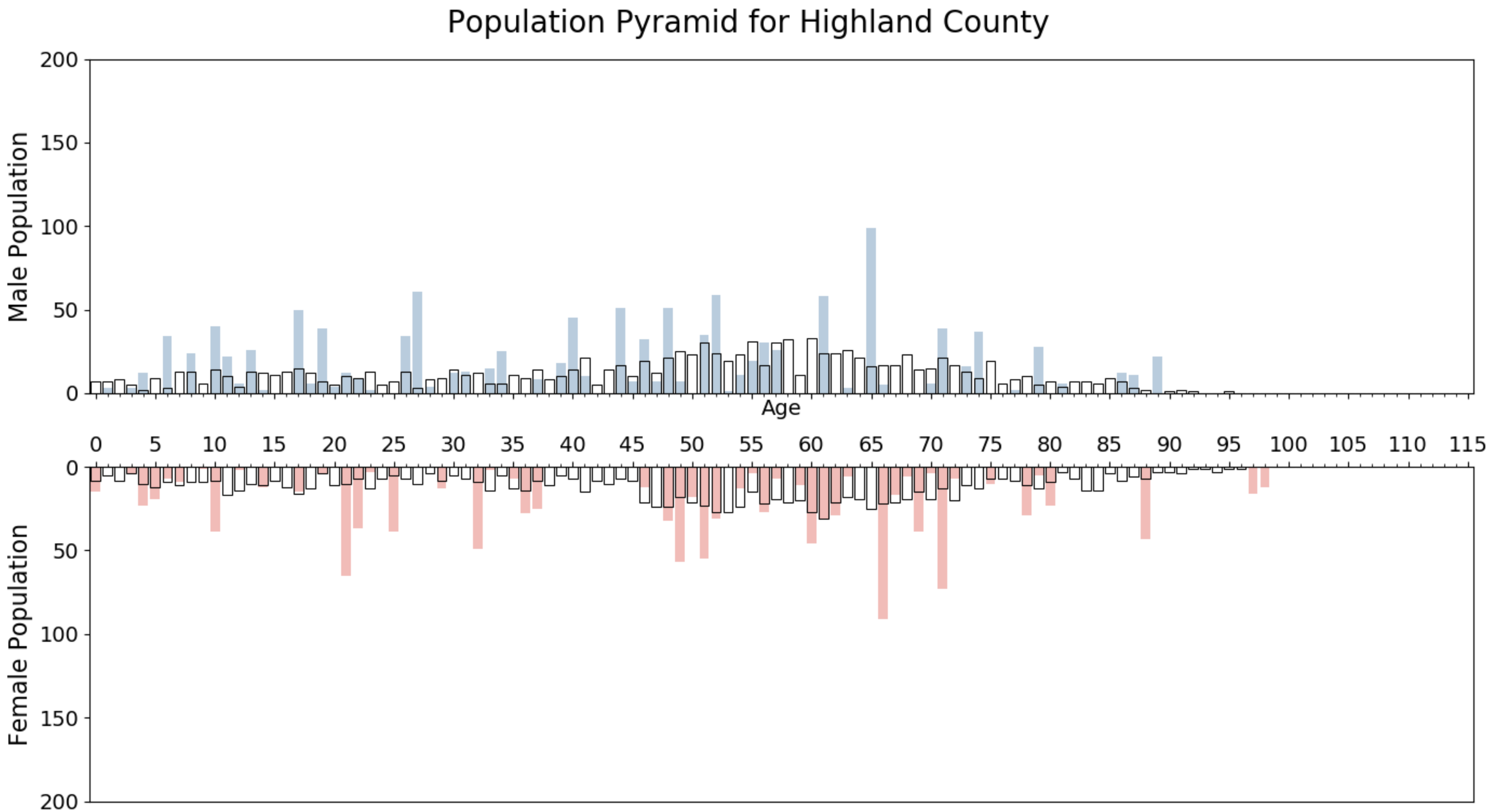
CBDRB-FY20-103

Impact on Accuracy

Comparison of TDA-generated MDF against the unswapped 2010 CEF

$\varepsilon = 4$

Highland County Population: 2,321



CBDRB-FY20-102

CBDRB-FY20-103

Impact on Accuracy

The National Academies of SCIENCES ENGINEERING MEDICINE

COMMITTEE ON NATIONAL STATISTICS
Division of Behavioral and Social Sciences and Education

Committee on National Statistics

[Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations]

December 11 - 12, 2019

CBDRB-FY20-103

Known Issues – Work is Ongoing!

- **There are two sources of error in the TopDown Algorithm (TDA):**
  - Measurement error due to differential privacy noise
  - Post-processing error due to statistical inference creating non-negative integer counts from the noisy measurements
- **Post-processing error tends to be much larger than differential privacy error**
- **Positive bias in small counts/negative bias in large counts is the result of**
  - Invariants
  - Post-processing error specifically introduced by our L2 optimization routine
- **Improving post-processing is not constrained by differential privacy**
- **Techniques to improve post-processing error may be drawn from demography, statistics, computer science, operations research, econometrics, etc. without increasing the privacy-loss budget**

Questions?

Disclosure Avoidance and the 2020 Census Website
https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

301-763-1960 (Office)
michael.b.hawes@census.gov