Issues Encountered Deploying Differential Privacy

Simson L. Garfinkel
Senior Scientist, Confidentiality and Data Access
U.S. Census Bureau
Simson.l.garfinkel@census.gov

University of Massachusetts

Wednesday, September 26, 2018

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Abstract

When differential privacy was created more than a decade ago, the motivating example was statistics published by an official statistics agency. In attempting to transition differential privacy from the academy to practice, an in particular for the 2020 Census of Population and Housing, the U.S. Census Bureau has encountered many challenges unanticipated by differential privacy's creators. These challenges include obtaining qualified personnel and a suitable computing environment, the difficulty accounting for all uses of the confidential data, the lack of release mechanisms that align with the needs of data users, the expectation on the part of data users that they will have access to micro-data, and the difficulty in setting the value of the privacy-loss parameter, ε (epsilon), and the lack of tools and trained individuals to verify the correctness of differential privacy implementations.

Acknowledgments

This presentation incorporates work by:

- Dan Kifer (Scientific Lead)
- John Abowd (Chief Scientist)
- Tammy Adams, Robert Ashmead, Aref Dajani, Jason Devine, Michael Hay, Cynthia Hollingsworth, Meriton Ibrahimi, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, **Gerome Miklau**, Brett Moran, Ned Porter, Anne Ross and William Sexton

Outline

Motivation

The flow of census response data

Disclosure Avoidance for the 2010 census

Disclosure Avoidance for the 2020 census

Conclusion

Motivation


We the People
of the
insure domestic Tranquility, provide for the common defence,
and our Posterity, do ordain and establish this Constitution
Article. I.

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.

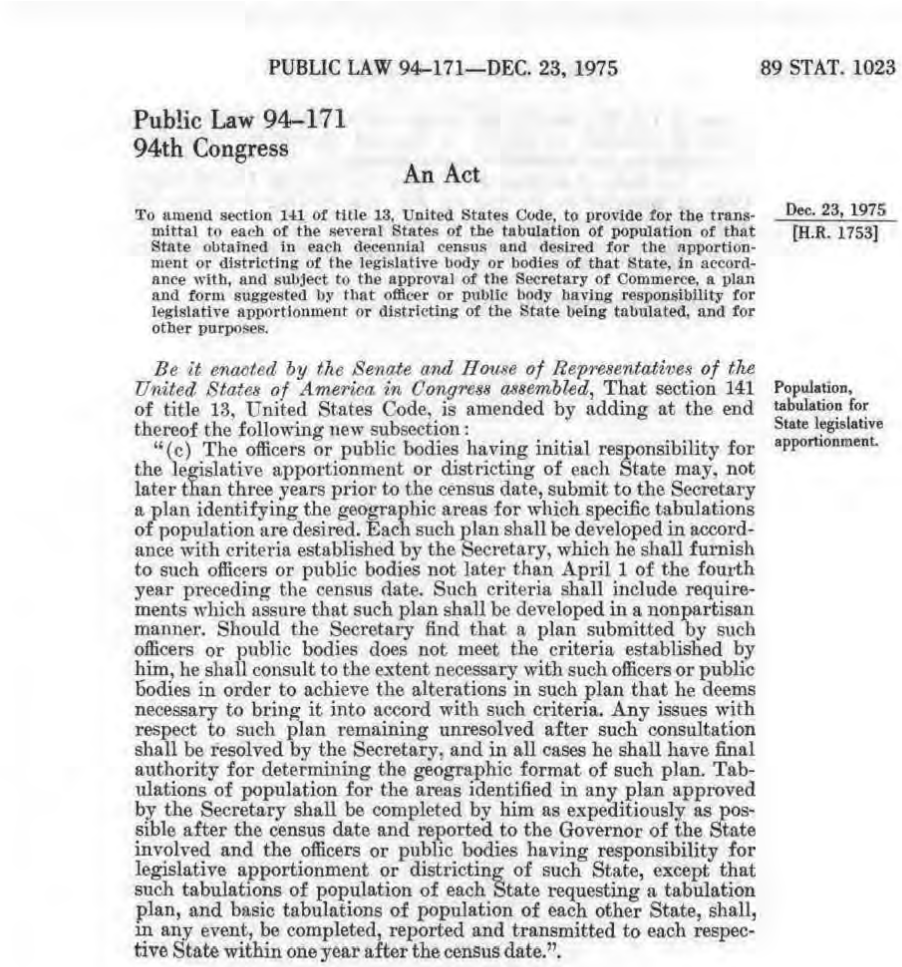
Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians 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



PUBLIC LAW 94–171—DEC. 23, 1975 89 STAT. 1023

Public Law 94–171
94th Congress

An Act

To amend section 141 of title 13, United States Code, to provide for the transmittal to each of the several States of the tabulation of population of that State obtained in each decennial census and desired for the apportionment or districting of the legislative body or bodies of that State, in accordance with, and subject to the approval of the Secretary of Commerce, a plan and form suggested by that officer or public body having responsibility for legislative apportionment or districting of the State being tabulated, and for other purposes.

Dec. 23, 1975
[H.R. 1753]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, That section 141 of title 13, United States Code, is amended by adding at the end thereof the following new subsection:

Population, tabulation for State legislative apportionment.

"(c) The officers or public bodies having initial responsibility for the legislative apportionment or districting of each State may, not later than three years prior to the census date, submit to the Secretary a plan identifying the geographic areas for which specific tabulations of population are desired. Each such plan shall be developed in accordance with criteria established by the Secretary, which he shall furnish to such officers or public bodies not later than April 1 of the fourth year preceding the census date. Such criteria shall include requirements which assure that such plan shall be developed in a nonpartisan manner. Should the Secretary find that a plan submitted by such officers or public bodies does not meet the criteria established by him, he shall consult to the extent necessary with such officers or public bodies in order to achieve the alterations in such plan that he deems necessary to bring it into accord with such criteria. Any issues with respect to such plan remaining unresolved after such consultation shall be resolved by the Secretary, and in all cases he shall have final authority for determining the geographic format of such plan. Tabulations of population for the areas identified in any plan approved by the Secretary shall be completed by him as expeditiously as possible after the census date and reported to the Governor of the State involved and the officers or public bodies having responsibility for legislative apportionment or districting of such State, except that such tabulations of population of each State requesting a tabulation plan, and basic tabulations of population of each other State, shall, in any event, be completed, reported and transmitted to each respective State within one year after the census date.".

89 STAT. 1024 PUBLIC LAW 94–171—DEC. 23, 1975

SEC. 2. (a) The heading for section 141 of title 13, United States Code, is amended by adding at the end thereof the following: "**; tabulation for legislative apportionment**".

(b) The table of sections for chapter 5 of title 13, United States Code, is amended by striking out the item relating to section 141 and inserting in lieu thereof the following:

"141. Population, unemployment, and housing; tabulation for legislative apportionment.".

Approved December 23, 1975.

LEGISLATIVE HISTORY:

HOUSE REPORT No. 94–456 (Comm. on Post Office and Civil Service).
SENATE REPORT No. 94–539 (Comm. on Post Office and Civil Service).
CONGRESSIONAL RECORD, Vol. 121 (1975):
Nov. 7, considered and passed House.
Dec. 15, considered and passed Senate.

United States Census Bureau

Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

http://uscode.house.gov/statutes/pl/94/171.pdf

Federal Register / Vol. 82, No. 215 / Nov 8, 2017 / Notices

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 170824806–7806–01]

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.

Disclosure Avoidance for the 2010 Census

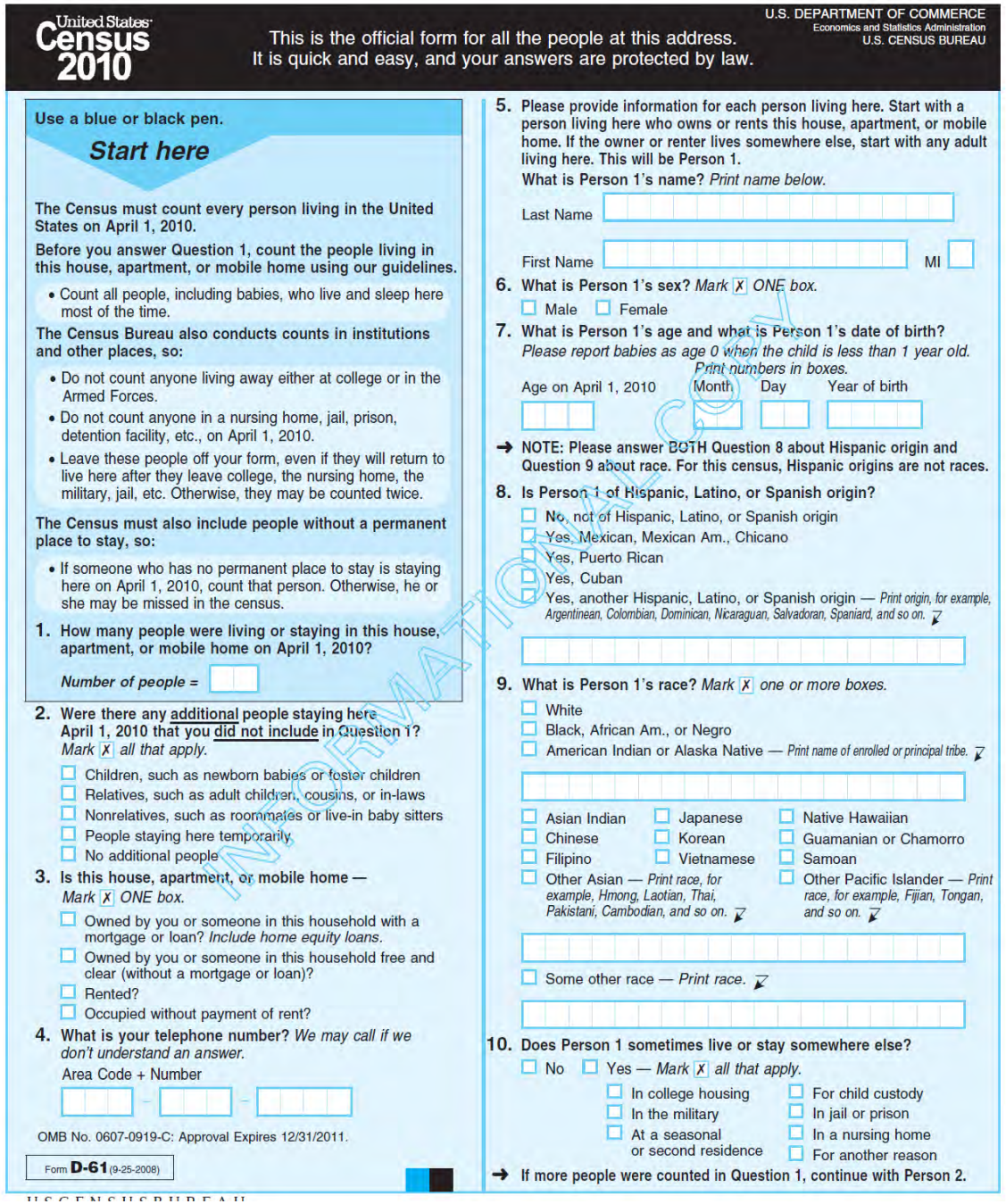
IT'S IN **OUR HANDS**

United States®
Census
2010

United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

"This is the official form for all the people at this address."



United States Census 2010

This is the official form for all the people at this address.
It is quick and easy, and your answers are protected by law.

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

Use a blue or black pen.

Start here

The Census must count every person living in the United States on April 1, 2010.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2010.
- Leave these people off your form, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

The Census must also include people without a permanent place to stay, so:

- If someone who has no permanent place to stay is staying here on April 1, 2010, count that person. Otherwise, he or she may be missed in the census.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2010?

   Number of people =

2. Were there any additional people staying here April 1, 2010 that you did not include in Question 1? *Mark ✗ all that apply.*
   - Children, such as newborn babies or foster children
   - Relatives, such as adult children, cousins, or in-laws
   - Nonrelatives, such as roommates or live-in baby sitters
   - People staying here temporarily
   - No additional people

3. Is this house, apartment, or mobile home — *Mark ✗ ONE box.*
   - Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
   - Owned by you or someone in this household free and clear (without a mortgage or loan)?
   - Rented?
   - Occupied without payment of rent?

4. What is your telephone number? *We may call if we don't understand an answer.*

   Area Code + Number

OMB No. 0607-0919-C: Approval Expires 12/31/2011.

Form D-61 (9-25-2008)

5. Please provide information for each person living here. Start with a person living here who owns or rents this house, apartment, or mobile home. If the owner or renter lives somewhere else, start with any adult living here. This will be Person 1.

   What is Person 1's name? *Print name below.*

   Last Name

   First Name MI

6. What is Person 1's sex? *Mark ✗ ONE box.*
   - Male
   - Female

7. What is Person 1's age and what is Person 1's date of birth?
   Please report babies as age 0 when the child is less than 1 year old.
   Print numbers in boxes.

   Age on April 1, 2010 | Month | Day | Year of birth

→ NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.

8. Is Person 1 of Hispanic, Latino, or Spanish origin?
   - No, not of Hispanic, Latino, or Spanish origin
   - Yes, Mexican, Mexican Am., Chicano
   - Yes, Puerto Rican
   - Yes, Cuban
   - Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

9. What is Person 1's race? *Mark ✗ one or more boxes.*
   - White
   - Black, African Am., or Negro
   - American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
   - Asian Indian
   - Japanese
   - Native Hawaiian
   - Chinese
   - Korean
   - Guamanian or Chamorro
   - Filipino
   - Vietnamese
   - Samoan
   - Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
   - Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*
   - Some other race — *Print race.*

10. Does Person 1 sometimes live or stay somewhere else?
    - No
    - Yes — *Mark ✗ all that apply.*
      - In college housing
      - In the military
      - At a seasonal or second residence
      - For child custody
      - In jail or prison
      - In a nursing home
      - For another reason

→ If more people were counted in Question 1, continue with Person 2.

United States Census Bureau

U.S. Department of Comm
Economics and Statistics Ad
U.S. CENSUS BUREAU
census.gov

"It is quick and easy, and your answers are protected by law."

Example: 2010 Census of Population

Basic results from the 2010 Census

Total population	308,745,538
Household population	300,758,215
Group quarters population	7,987,323
Households	116,716,292

Example: 2010 Census II

High-level database schema

Variables	Distinct values
Habitable blocks	10,620,683
Habitable tracts	73,768
Sex	2
Age	115
Race/Ethnicity (OMB Categories)	126
Race/Ethnicity (SF2 Categories)	600
Relationship to person 1	17
National histogram cells (OMB Categories)	492,660

Example: 2010 Census III

Summary of the publications (counts are approximate)

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

2003:
Database Reconstruction

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 $\tilde{O}(\sqrt{n})$.

For time-$\mathcal{T}$ bounded adversaries we demonstrate a privacy-preserving access algorithm whose perturbation magnitude is $\approx \sqrt{\mathcal{T}}$.

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Revealing Information while Preserving Privacy

Irit Dinur Kobbi Nissim*
NEC Research Institute
4 Independence Way
Princeton, NJ 08540
{iritd,kobbi}@research.nj.nec.com

ABSTRACT

We examine the tradeoff between privacy and usability of

research which is based (among other things) on statistics of the information in the database. On the other hand, the hospital is obliged to keep the privacy of its patients, i.e. leak no medical information that could be related to a specific patient. The hospital needs an access mechanism to the database that allows certain 'statistical' queries to be answered, as long as they do not violate the privacy of any single patient.

*Work partly done when the author was at DIMACS, Rutgers University, and while visiting Microsoft Research Silicon Valley Lab.

PODS 2003, June 9-12, 2003, San Diego, CA.

One simple tempting solution is to remove from the database all 'identifying' attributes such as the patients' names and

n their comparative survey of privacy methods for statistical databases, Adam and Wortmann [2] classified the approaches taken into three main categories: (i) query restriction, (ii) data perturbation, and (iii) output perturbation. We give a brief review of these approaches below, and refer the reader to [2] for a detailed survey of the methods and their weaknesses.

Query Restriction. In the query restriction approach, queries are required to obey a special structure, supposedly to prevent the querying adversary from gaining too much information about specific database entries. The limit of this approach is that it allows for a relatively small number of queries.

A related idea is of query auditing [7], i.e. a log of the queries is kept, and every new query is checked for possible compromise, allowing/disallowing the query accordingly.

[1]A patient's gender, approximate age, approximate weight, ethnicity, and marital status – may already suffice for a complete identification of most patients in a database of a thousand patients. The situation is much worse if a relatively 'rare' attribute of some patient is known. For example, a patient having Cystic Fibrosis (frequency $\approx 1/3000$) may be uniquely identified within about a million patients.

2006:
Differential Privacy

Calibrating Noise to Sensitivity in Private Data Analysis

Cynthia Dwork[1], Frank McSherry[1], Kobbi Nissim[2], and Adam Smith[3]*

[1] Microsoft Research, Silicon Valley. {dwork,mcsherry}@microsoft.com
[2] Ben-Gurion University. kobbi@cs.bgu.ac.il
[3] Weizmann Institute of Science. adam.smith@weizmann.ac.il

Abstract. We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function f mapping databases to reals, the so-called *true answer* is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Previous work focused on the case of noisy sums, in which $f = \sum_i g(x_i)$, where x_i denotes the ith row of the database and g maps database rows to $[0, 1]$. We extend the study to general functions f, proving that privacy can be preserved by calibrating the standard deviation of the noise according to the *sensitivity* of the function f. Roughly

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

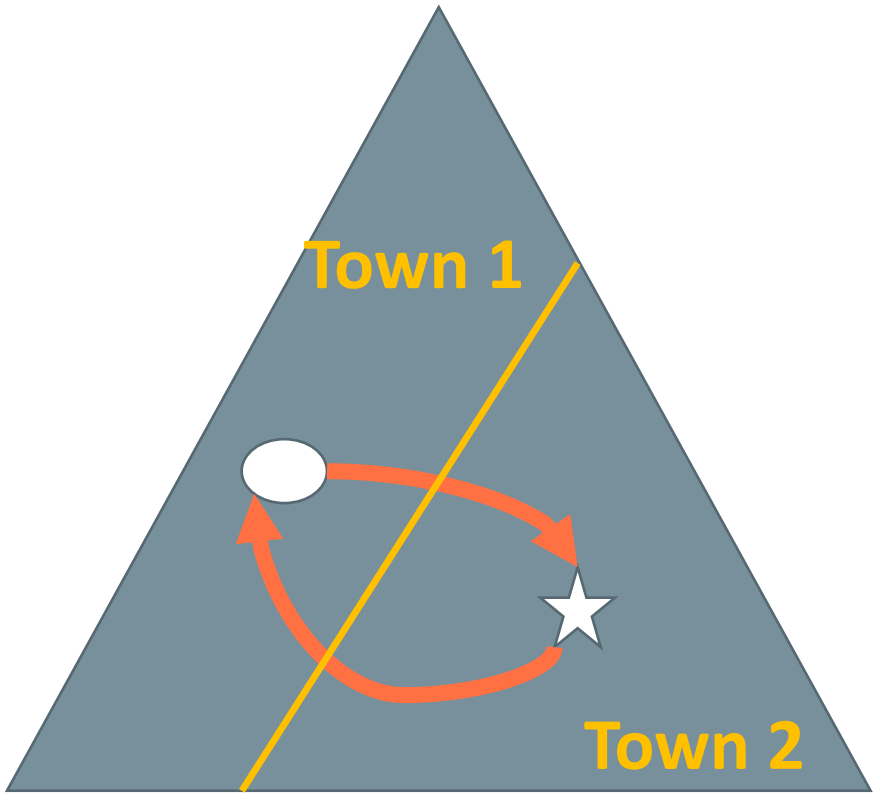
The protection system used in 2000 and 2010 relied on swapping households:

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

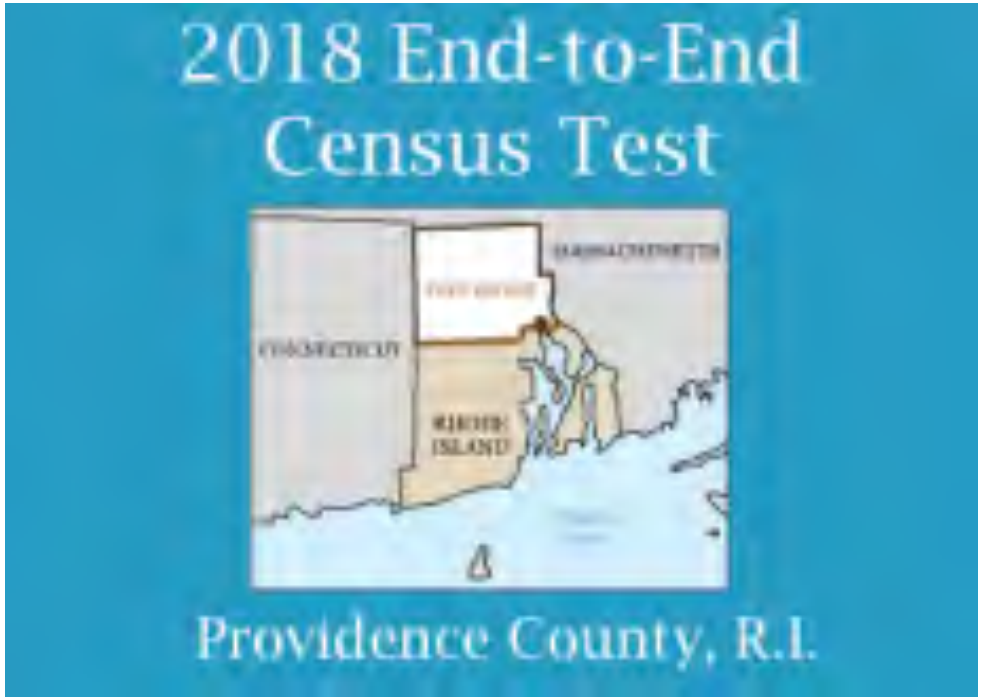
Disadvantages:

- 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



State “X”

The US Census Bureau embraces formal privacy.



United States
Census
2020

United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Motivation:
To protect the privacy of individual survey responses

2010 Census:

- 7.7 billion independent tabular summaries published
- 25 records per person

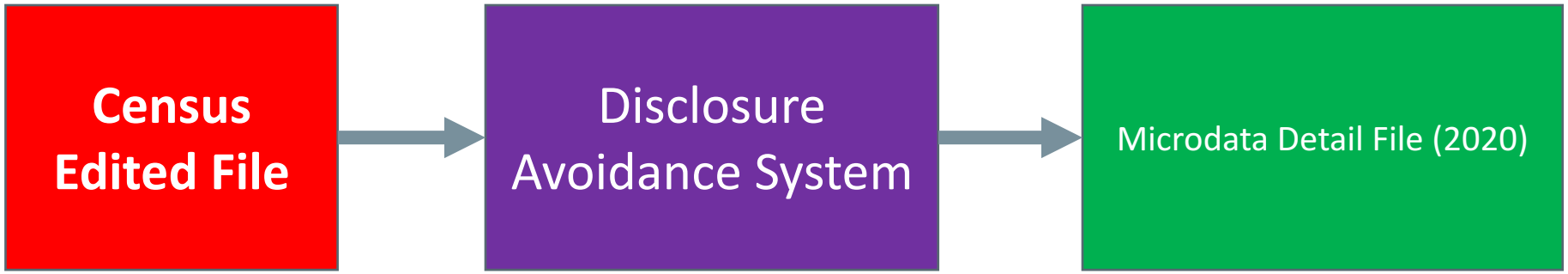
Database reconstruction (Dinur and Nissim 2003) is a serious disclosure threat that all statistical tabulation systems from confidential data must acknowledge.

The confidentiality edits applied to the 2010 Census were not designed to defend against this kind of attack.

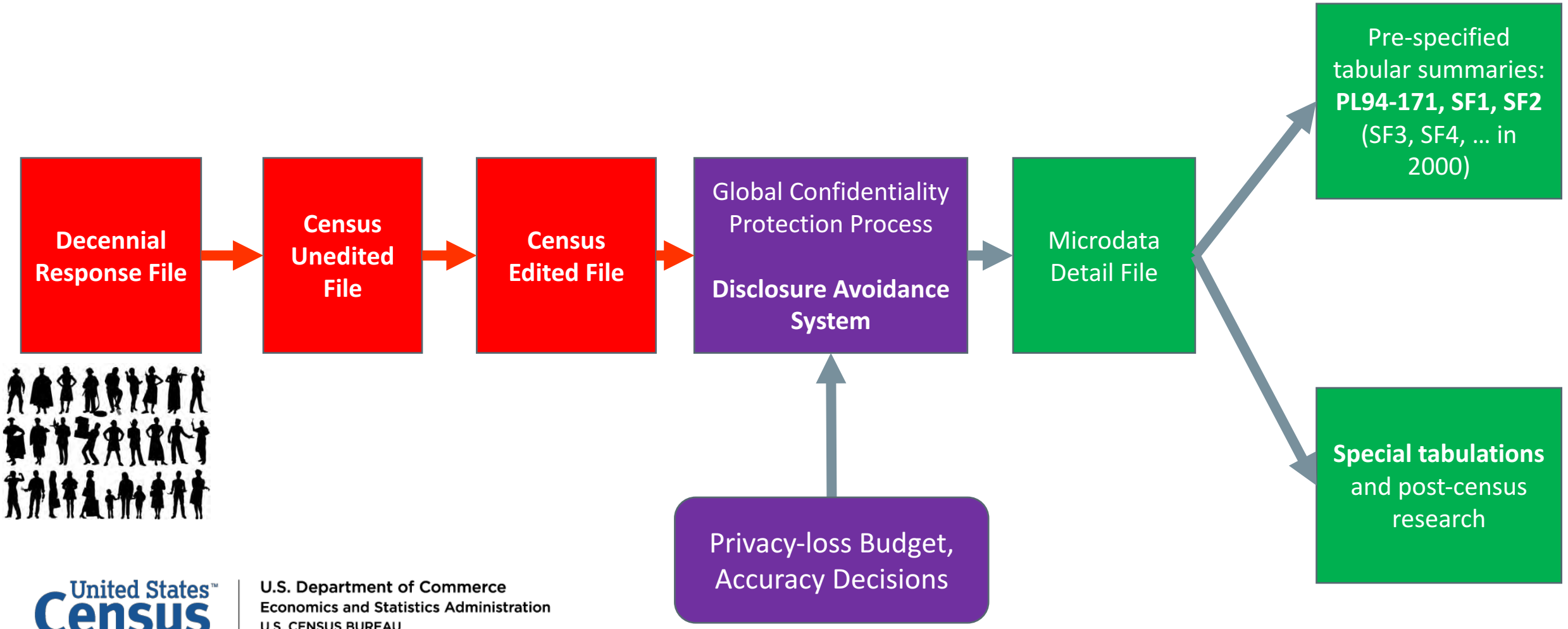
Our plan is to create a “Disclosure Avoidance System” that drops into the Census production system.

Features of the DAS:

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



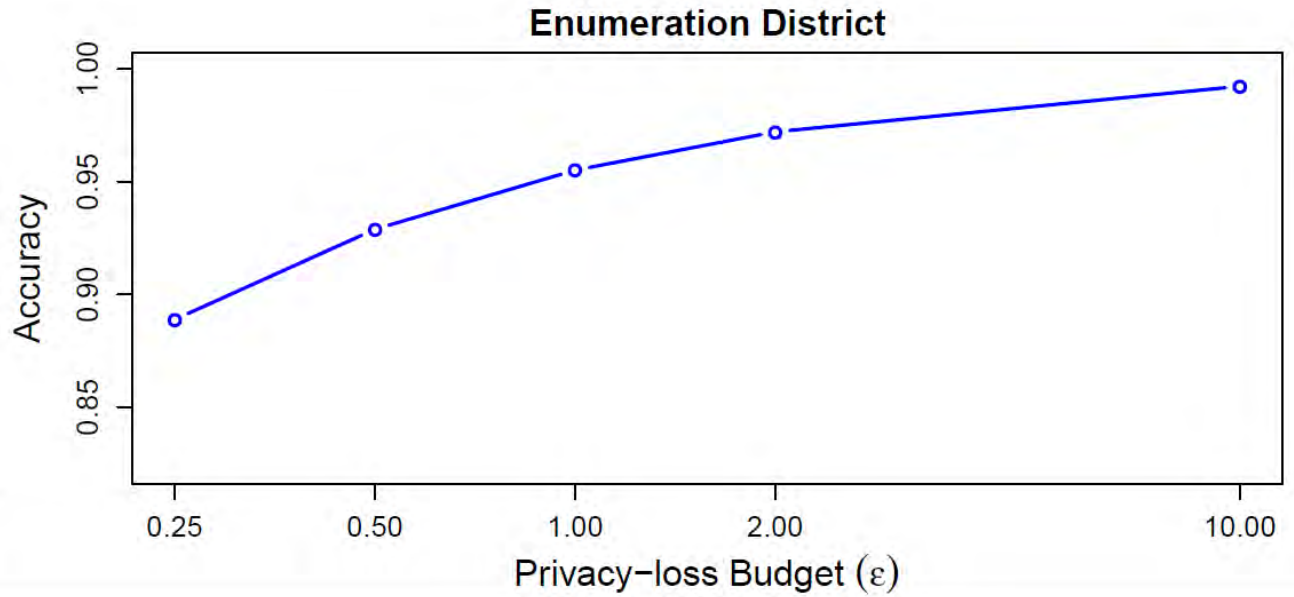
The Disclosure Avoidance System allows the Census Bureau to enforce global confidentiality protections.



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.

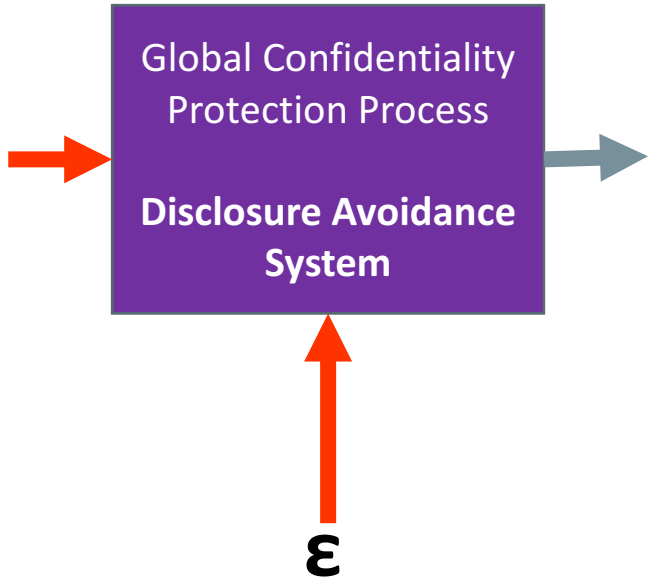


Final privacy-loss budget determined by
Data Stewardship Executive Policy Committee (DSEP)
with recommendation from Disclosure Review Board (DRB)

The Disclosure Avoidance System relies on infusing formally private noise.

Advantages of noise injection 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*

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 are consistent at all levels of geography
- Consistency among query answers

- REINSERT EDITED GRAPHIC FROM EARLIER SLIDE

DON’T TALK ABOUT INVARIANTS, TALK ABOUT TABLE CONSISTENCY.

United States™ Census Bureau

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:

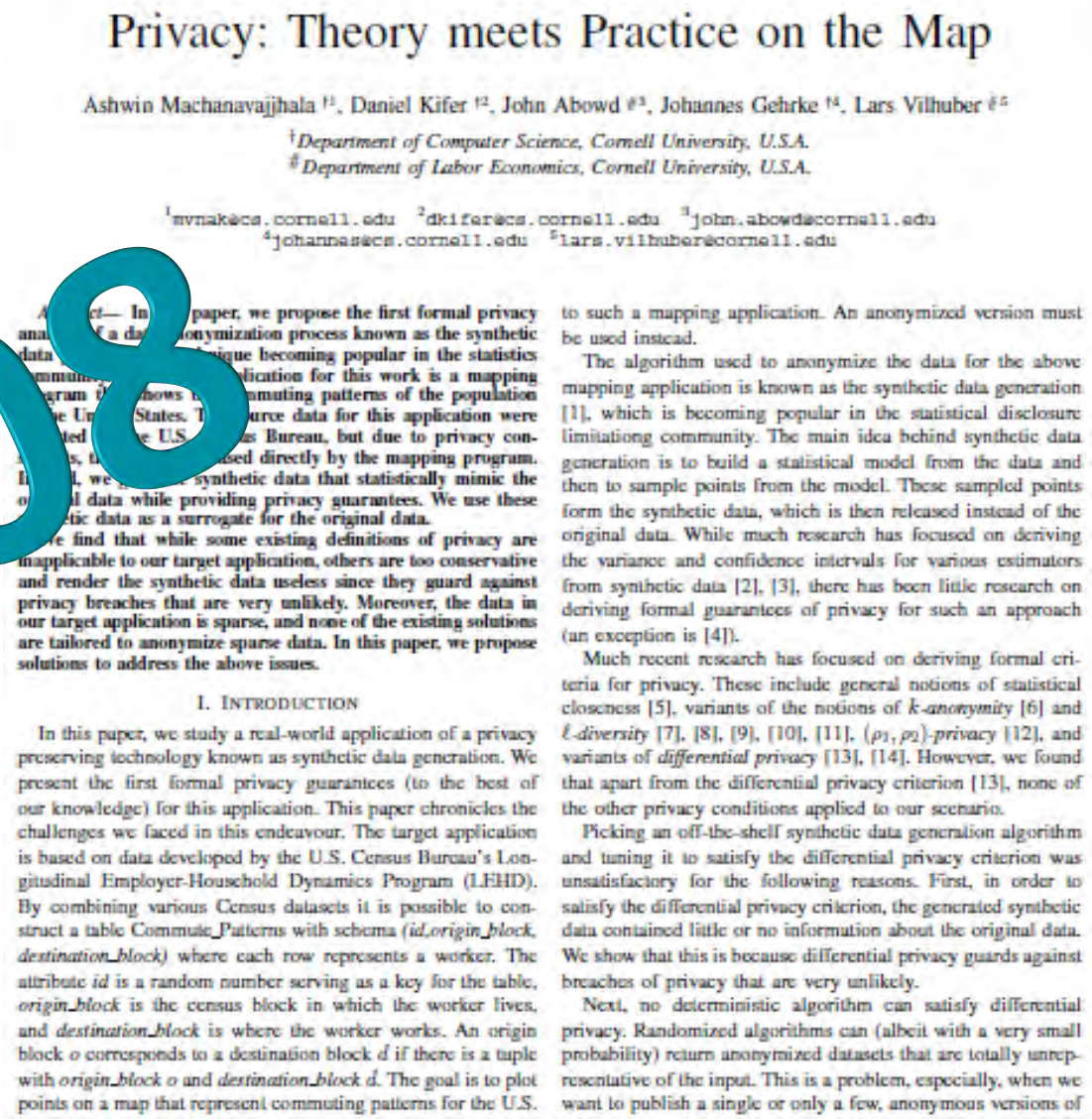
- Differential privacy is not widely known or understood.
- Many data users want highly accurate data reports on small areas.
- Users in 2000 and 2010 didn't know the error introduced by swapping.

Differential Privacy at the US Census Bureau

A Brief History of Differential Privacy at the U.S. Census Bureau

This work was done at Cornell University while Abowd and Vilhuber were on IPA assignments to the Census Bureau. Gehrke is now Technical Fellow at Microsoft. Kifer is now the scientific lead on the 2020 DAS. Machanavajjhala is now a contractual collaborator on the 2020 DAS. Vilhuber is now on IPA assignment to the Census Bureau.

2008



Privacy: Theory meets Practice on the Map

Ashwin Machanavajjhala [1], Daniel Kifer [2], John Abowd [#3], Johannes Gehrke [4], Lars Vilhuber [#5]

[1]Department of Computer Science, Cornell University, U.S.A.
[#]Department of Labor Economics, Cornell University, U.S.A.

[1]mvnak@cs.cornell.edu [2]dkifer@cs.cornell.edu [3]john.abowd@cornell.edu
[4]johannes@cs.cornell.edu [5]lars.vilhuber@cornell.edu

Abstract— In this paper, we propose the first formal privacy analysis of a data anonymization process known as the synthetic data technique becoming popular in the statistics community. The application for this work is a mapping program that shows the commuting patterns of the population of the United States. The source data for this application were collected by the U.S. Census Bureau, but due to privacy constraints, this data cannot be released directly by the mapping program. Instead, we release synthetic data that statistically mimic the original data while providing privacy guarantees. We use these synthetic data as a surrogate for the original data.

We find that while some existing definitions of privacy are inapplicable to our target application, others are too conservative and render the synthetic data useless since they guard against privacy breaches that are very unlikely. Moreover, the data in our target application is sparse, and none of the existing solutions are tailored to anonymize sparse data. In this paper, we propose solutions to address the above issues.

I. Introduction

In this paper, we study a real-world application of a privacy preserving technology known as synthetic data generation. We present the first formal privacy guarantees (to the best of our knowledge) for this application. This paper chronicles the challenges we faced in this endeavour. The target application is based on data developed by the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD). By combining various Census datasets it is possible to construct a table Commute_Patterns with schema *(id,origin_block, destination_block)* where each row represents a worker. The attribute *id* is a random number serving as a key for the table, *origin_block* is the census block in which the worker lives, and *destination_block* is where the worker works. An origin block o corresponds to a destination block d if there is a tuple with *origin_block* o and *destination_block* d. The goal is to plot points on a map that represent commuting patterns for the U.S. to such a mapping application. An anonymized version must be used instead.

The algorithm used to anonymize the data for the above mapping application is known as the synthetic data generation [1], which is becoming popular in the statistical disclosure limitations community. The main idea behind synthetic data generation is to build a statistical model from the data and then to sample points from the model. These sampled points form the synthetic data, which is then released instead of the original data. While much research has focused on deriving the variance and confidence intervals for various estimators from synthetic data [2], [3], there has been little research on deriving formal guarantees of privacy for such an approach (an exception is [4]).

Much recent research has focused on deriving formal criteria for privacy. These include general notions of statistical closeness [5], variants of the notions of *k-anonymity* [6] and *ℓ-diversity* [7], [8], [9], [10], [11], (ρ_1, ρ_2)-*privacy* [12], and variants of *differential privacy* [13], [14]. However, we found that apart from the differential privacy criterion [13], none of the other privacy conditions applied to our scenario.

Picking an off-the-shelf synthetic data generation algorithm and tuning it to satisfy the differential privacy criterion was unsatisfactory for the following reasons. First, in order to satisfy the differential privacy criterion, the generated synthetic data contained little or no information about the original data. We show that this is because differential privacy guards against breaches of privacy that are very unlikely.

Next, no deterministic algorithm can satisfy differential privacy. Randomized algorithms can (albeit with a very small probability) return anonymized datasets that are totally unrepresentative of the input. This is a problem, especially, when we want to publish a single or only a few, anonymous versions of

United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov


https://onthemap.ces.census.gov
Search
OnTheMap
LEHD Home Help and Documentation Reload Text-Only
Start
Base Map
Selection
Save
Load
Feedback
Previous Extent
Hide Tabs
Welcome to OnTheMap!
Start an analysis by using one of the tools below (Search, Import Geography, or Load .OTM file). Hover over the Help icons located throughout the application to see Help tips for using specific functionality. Sections in the control panel can be collapsed or opened by clicking the section title.
2015 Data Now Available (09/25/2017)
Search
Search
Search All Names
Import Geography
Import from KML
Import from SHP
Import from GPS
Load .OTM File
Seattle
San Francisco
San Jose
San Diego
Phoenix
El Paso
Dallas
Austin
Houston
Memphis
Charlotte
Jacksonville
Chicago
Detroit
Boston
New York
Baltimore
1000 km
500 mi
-122.89877, 22.47507
2008: First production implementation of differential privacy worldwide
Privacy Policy | 2010 Census | Data Tools | Information Quality | Product Catalog | Contact Us | Home
Source: U.S.Census Bureau, Center for Economic Studies | e-mail: CES.OnTheMap.Feedback@census.gov

https://onthemap.ces.census.gov

OnTheMap

LEHD Home | Help and Documentation | Reload | Text-Only

Start | Base Map | Selection | Results

Distance/Direction Analysis
Work to Home

Display Settings

Labor Market Segment Filter: All Workers

Year: 2015

Map Controls

Color Key

Thermal Overlay ☑

Point Overlay ☑

Selection Outline ☑

Identify | Zoom to Selection

Clear Overlays | Animate Overlays

Report/Map Outputs

Detailed Report

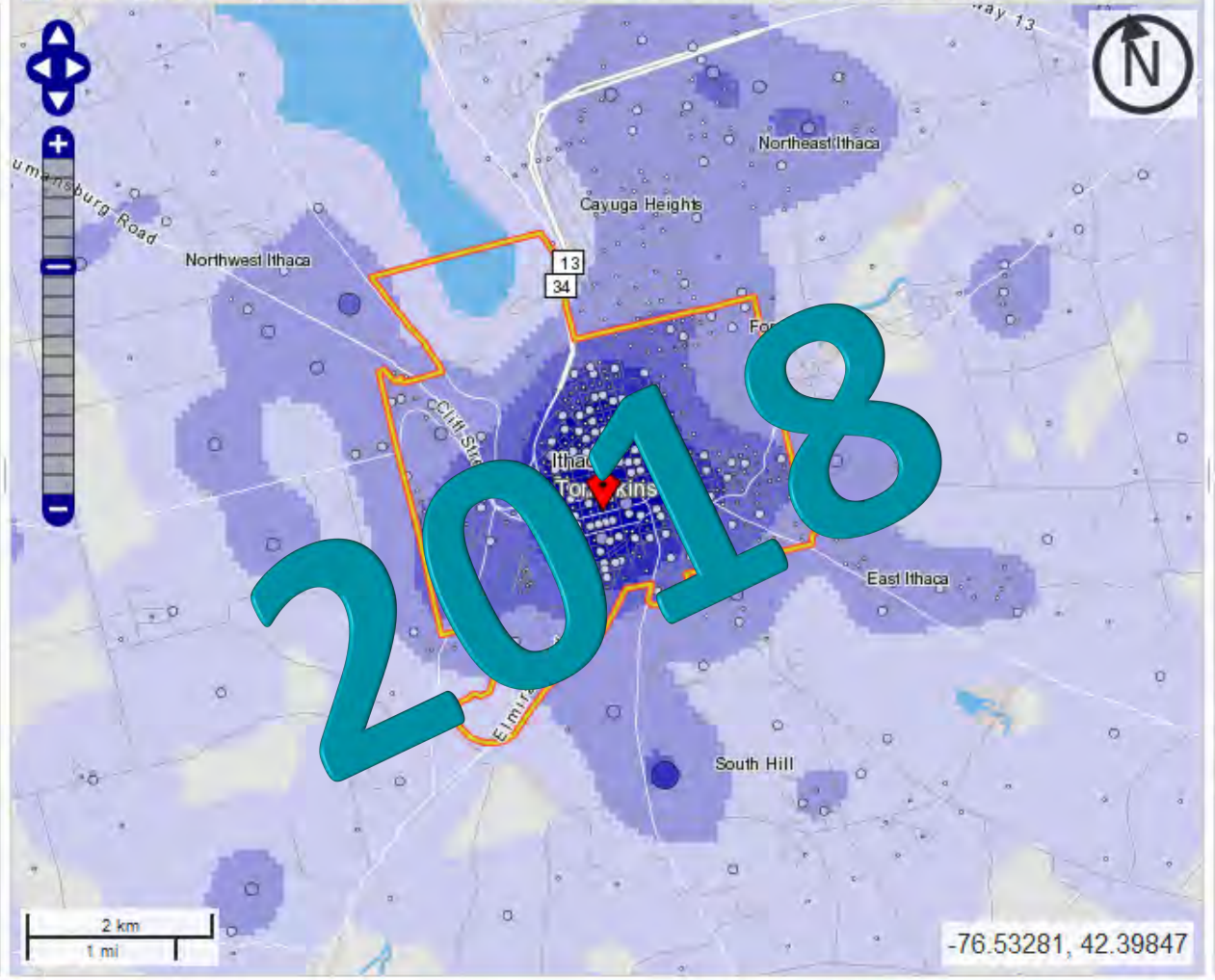
Export Geography

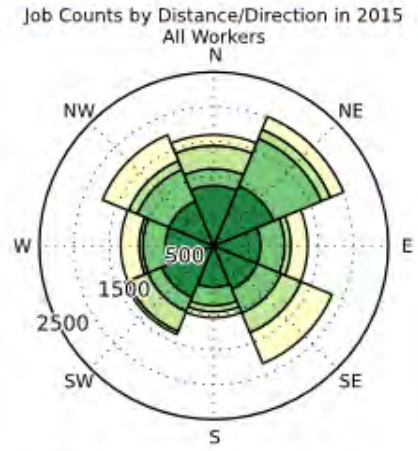
Print Chart/Map

Legends

Change Settings

Save | Load | Feedback | Previous Extent | Hide Tabs | Hide Chart/Report





View as Radar Chart

Jobs by Distance - Work Census Block to Home Census Block

	2015 Count	2015 Share
Total Primary Jobs	12,260	100.0%
Less than 10 miles	5,949	48.5%
10 to 24 miles	2,987	24.4%
25 to 50 miles	1,451	11.8%
Greater than 50 miles	1,873	15.3%

Differential Privacy: A Survey of Deployments

US Census Bureau:
Trusted Curator Model

Longitudinal Employer Household Database (LEHD)

On The Map ε=8.9

"Protecting Graduate Earnings and Employment Outcomes" ε=3

The World:
Local Model

Google Chrome:

- Windows Process Names, Chrome Homepages, etc.
- 135 metrics in total. ε=2 .. 7 ***per metric!***

Apple:

- QuickType suggestions, Emoji suggestions, Lookup Hints, Safari Energy Draining Domains, Safari Autoplay Intent Detection, Safari Crashing Domains, Health Type Usage
- ε=7 (MacOS); ε=14 (iOS) ***per day!***

Microsoft telemetry from Windows 10

Sources: Wired, "How One of Apple's Key Privacy Safeguards Falls Short," Sept. 15, 2017. https://www.wired.com/story/apple-differential-privacy-shortcomings/
AccessNow, "Differential privacy, part 3," Nov. 30, 2017. https://www.accessnow.org/differential-privacy-part-3-extraordinary-claims-require-extraordinary-scrutiny/

In 2017, the Census Bureau announced that it would use differential privacy for the 2020 Census.

There is no off-the-shelf mechanism for applying differential privacy to a national census.

Randomized response (RAPPOR) would introduce far too much noise for any sensible value of ε to be a much statistical value.

- Google and Apple are finding this out.

We cannot simply apply the Laplace Mechanism to tables.

- Our data users expect consistent tables.

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 format privacy mechanisms
- The difficulty of retrofitting legacy statistical products to conform with modern privacy practice

Scientific Issues for the 2020 Census

Hierarchical Mechanisms
We needed a novel mechanism that:

- Assured consistent statistics from US->States->Counties->Tracts
- Provided **lower error** for **larger geographies.**

Invariants
For the 2018 End-to-End test, policy makers wanted exact counts for:

- Number of people on each block
- Number of people on each block of voting age
- Number of residences & group quarters on each block
- These may, however be removed based on what we have learned to-date

Scientific issues for the American Community Survey

The American Community Survey replaced the “Long Form” in 2005.

ACS uses a stratified probability sample of the entire US.

Differential privacy currently does not handle (well):

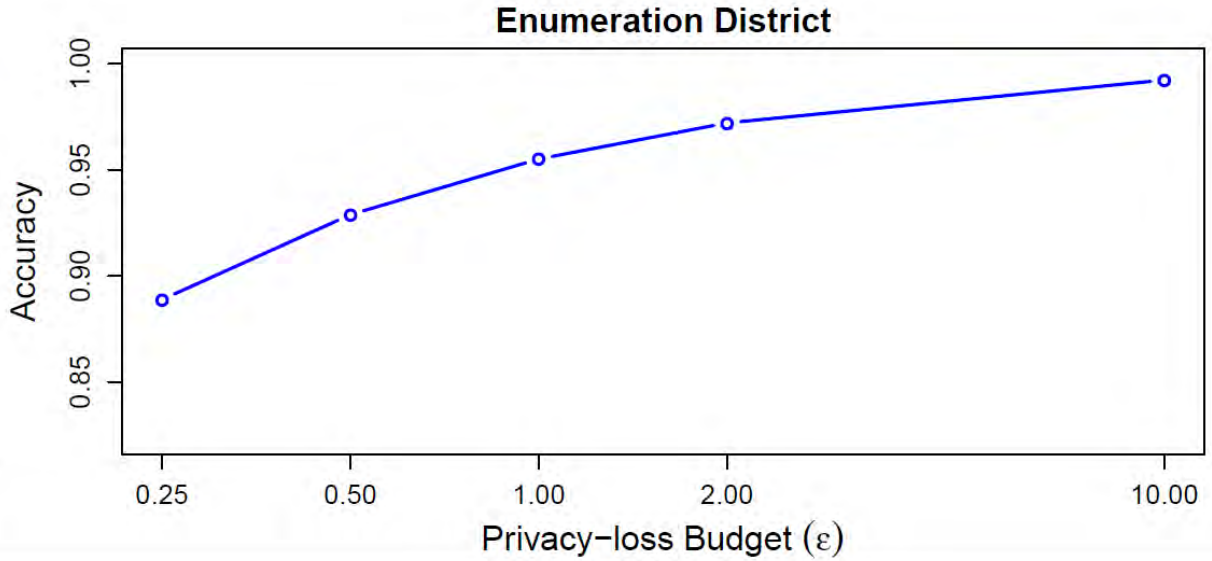
- Weights on survey microdata. (max weight is large/undefined)
- Skip-logic/conditional responses

Scientific Issues: Quality Metrics

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

Options:

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



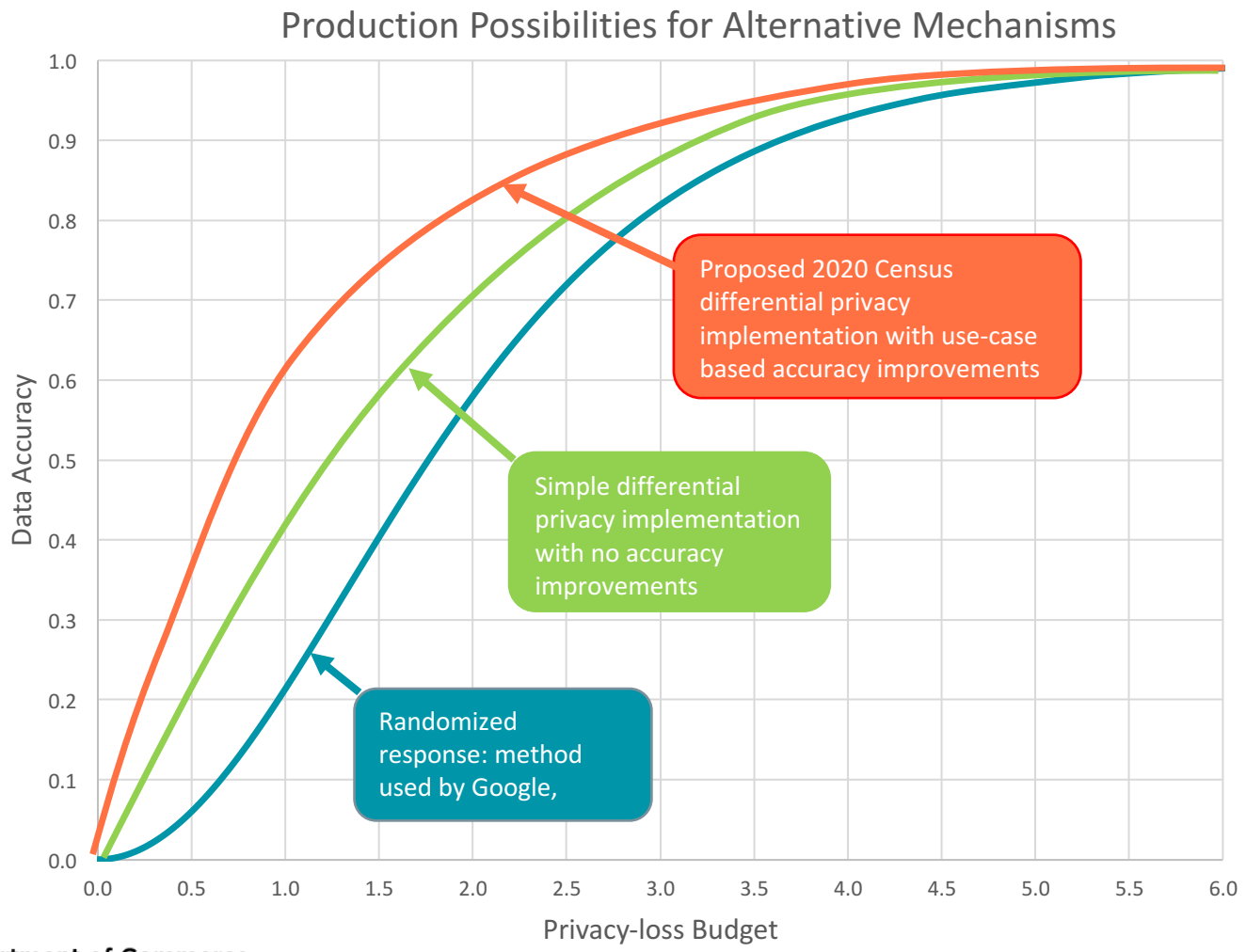
Scientific issues: Equity Issues

Differential Privacy allows us to make some statistics more accurate and others less accurate.

	Males	**Females**
Age < 18	100	150
Age >= 18	150	100

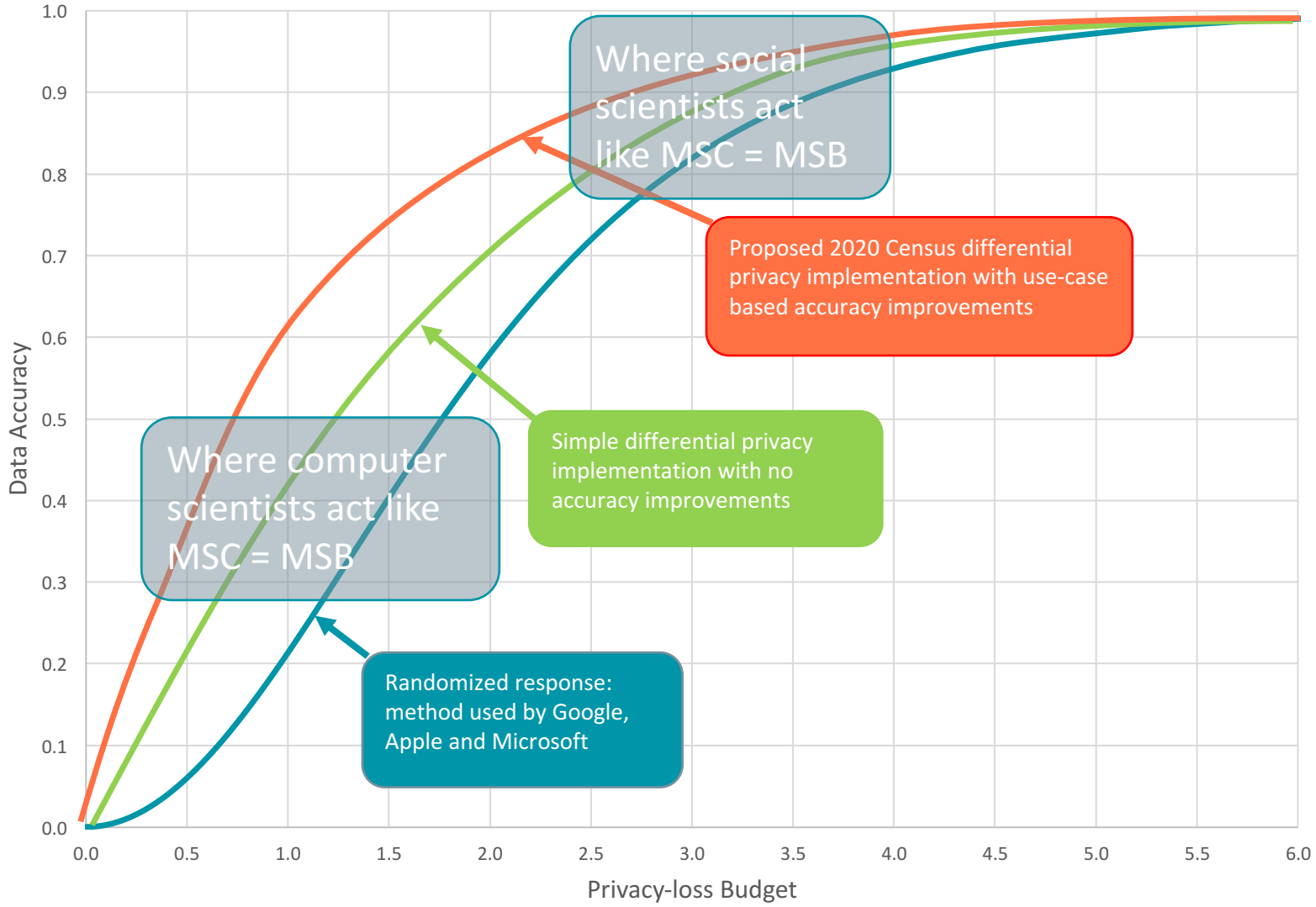
Who decides which is more important? (M/F ? Minor/adult?)

Scientific Issue: Setting Epsilon



We want the most efficient technology, which is the one in red.

Production Possibilities for Alternative Mechanisms



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

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

Recommendations

Repeated Discussions with Decision Makers

Controlled Vocabulary

Integrated Communications Strategy

Questions? Looking for an internship? Feel free to email me at: simson.l.garfinkel@census.gov

References

Dinur, Irit and Kobbi Nissim (2003). “Revealing information while preserving privacy.” in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.

Haney et al. (2017)

Machanavajjhala et al. (2008)