The 2020 Decennial Census TopDown Disclosure Limitation Algorithm

A Report on the Current State of the Privacy Loss-Accuracy Trade-off

Philip Leclerc

On behalf of and with the support of the 2020 DAS development team

U.S. Census Bureau

CNSTAT

December 11, 2019

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

Shape your future START HERE >

United States Census 2020

The Census Bureau Re-Identification Experiments Using the 2010 Census

What we did

Database reconstruction for all 308,745,538 people in 2010 Census

Link reconstructed records to commercial databases: acquire PII

Successful linkage to commercial data: putative re-identification

Compare putative re-identifications to confidential data

Successful linkage to confidential data: confirmed re-identification

Harm: attacker can learn self-response race and ethnicity

What we found

For all 308,745,538 reconstructed records, census block and voting age (18+) were correctly reconstructed in all 6,207,027 inhabited blocks

Block, sex, age (in years), race (OMB 63 categories), ethnicity reconstructed:

- Exactly: 46% of population (142 million of 308,745,538)
- Allowing age +/- one year: 71% of population (219 million of 308,745,538)

Block, sex, age linked to commercial data to acquire PII

- Putative re-identifications: 45% of population (138 million of 308,745,538)

Name, block, sex, age, race, ethnicity compared to confidential data

- Confirmed re-identifications: 38% of putative (52 million; 17% of population)

For the confirmed re-identifications, race and ethnicity are learned correctly, although the attacker may still have uncertainty

CBDRB-FY20-100

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)

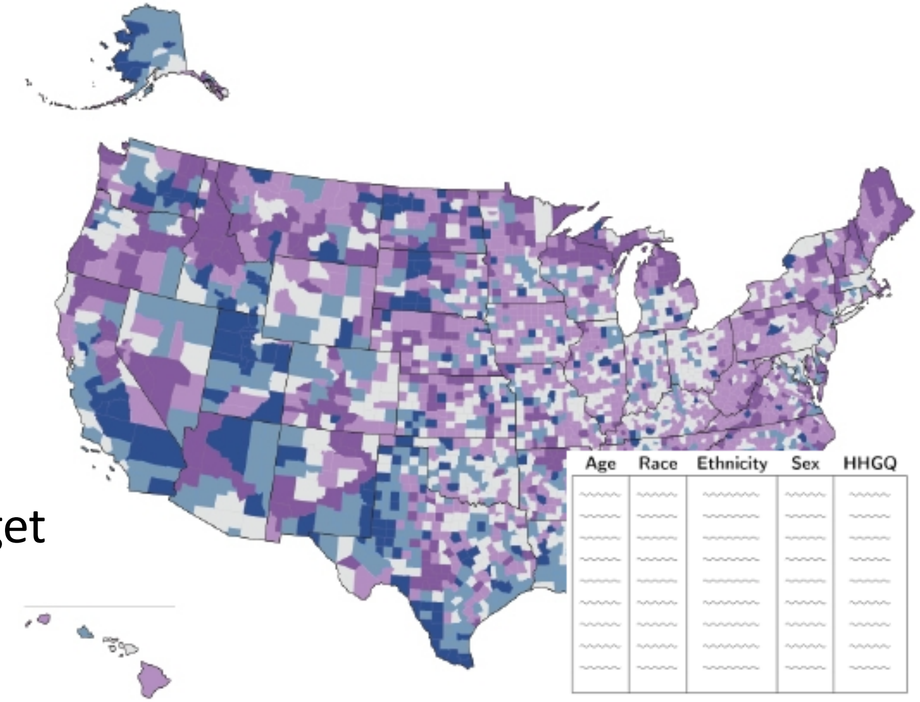
Utility/Accuracy for pre-specified tabulations

- 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 $\to 0$ as privacy loss $\epsilon \to \infty$

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)

Census TDA:
Choosing a Privacy-Loss Budget

Picking Ɛ Requires Understanding Both Privacy & Accuracy

- Given an implementation of TDA, how can we help policy-makers choose an ε (and related parameters)?
- We have employed 2 approaches to help explain the privacy implications of ε:
  - Mathematical guarantees: what is the worst that could happen?
  - Optimistic empirical analyses: how does a specific reconstruction-abetted re-identification attack behave at each Ɛ?
- Mathematical guarantees hold for all possible attackers, compute, data, algorithms
- Empirical analyses are optimistic: things could be worse with more data, attackers, compute! But they provide a direct comparison to the internal attack that motivated the Census Bureau to use formal privacy

Worst-case Guarantees Control Risk Relative to a Private Baseline

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).

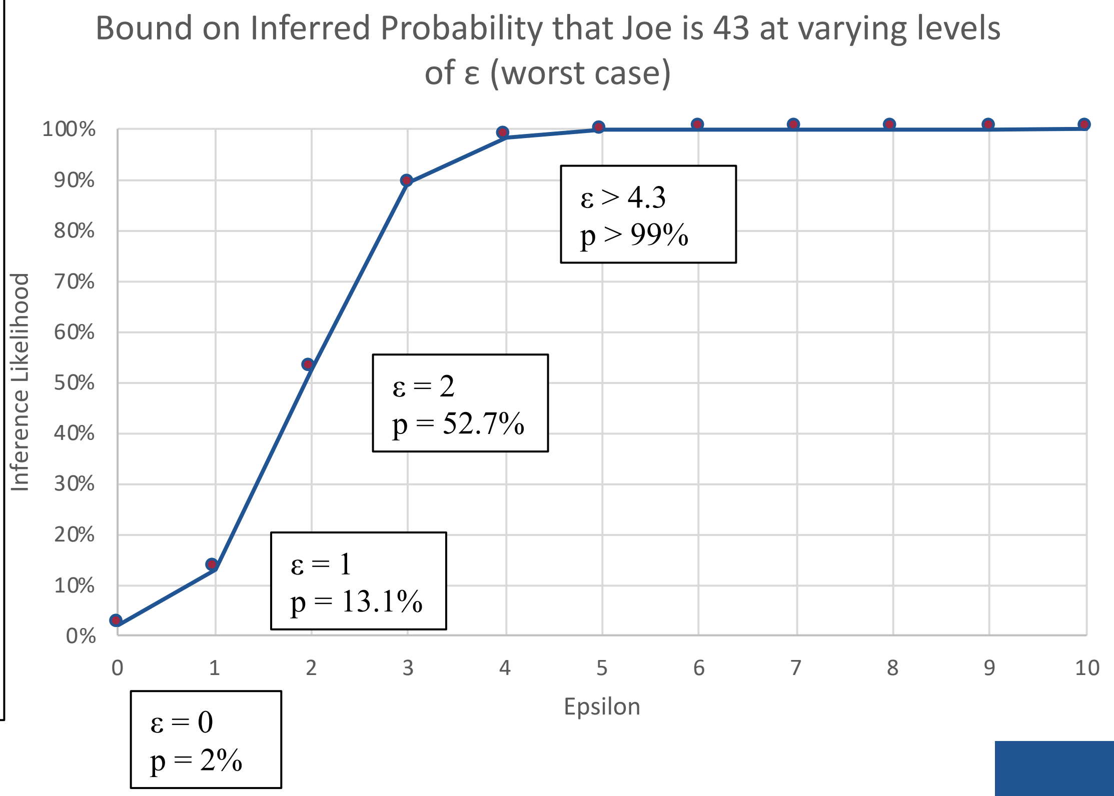
The Worst Case: A Concrete Example

Can Sara determine (some) Joe's exact age?

The Private Baseline: Suppose Joe submits erroneous information for the Census, so that Census publications cannot possibly reflect Joe's data – we take this as our private baseline scenario. In this scenario, Sara will still be able to predict with some probability that Joe is 43 years old; for the sake of illustration, suppose Sara's probability that Joe is 43 in this scenario is 2%. *Importantly, Sara can arrive at this inference even though Joe's data wasn't used at all!*

In the real world, where Joe (hopefully!) does provide accurate information, then some information about him will "leak" through the publication of data products. This new information can improve Sara's estimate; this improvement we interpret as privacy-eroding, since it can only occur because Joe provided his actual data.

ε controls the maximum possible improvement in Sara's inference when Joe submits real versus fake data. In this way, ε quantifies privacy loss.



NOTE: this theoretical guarantee holds even if Sara has infinite computing resources, infinitely powerful algorithms, and has arbitrary prior information that she can combine with the published Census tabulations.

Policy-makers Set the Privacy Loss Budget

- For Census's recently released 2010 Demonstration Data Products[1], Census's Data Stewardship Executive Policy Committee reviewed empirical accuracy metrics, interpretations of the privacy guarantee, & chose $\epsilon_{Persons}$ and ϵ_{HHs} to balance these competing concerns
- For this iteration of this process, accuracy data were produced with runs carried out on Virginia (a compromise between run-time & complexity/scale)
- In the next few slides we'll share the same accuracy metrics the DAS TDA development team provided to support DSEP's decision-making (additional metrics were also provided by Census Population & Demographics experts)

1: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html CBDRB-FY20-101

Accuracy Metrics: A Key Bit of Notation

- To define our error metrics, we'll use notation like $H_{MDF}(j, g)$, read as: the count of persons in a histogram *H* in the *MDF* of type *j* for geographic unit *g*
- The histogram object is flexible: it could be the cross-product of all of our variables (500K-1.23M cells), but it could also be a smaller "sub-"histogram. For example, we will use the Sex-by-Age histogram, which has shape $2 \cdot 116$ (one count for each combination of Sex and the 116 possible levels of Age)
- We typically take sums or average over all geounits in a specified geolevel (e.g. all tracts) or over all record-types *j* in the given histogram, with exceptions where indicated

For The 2010 Demonstration Data Products, We Used 2 Primary Metrics [1]

- The first metric was 1-TVD ("one minus average Total Variation Distance")
- We computed this as:
  - Given data as a multi-dimensional histogram (containing counts of records of distinct types, indexed consistently) in the CEF, H_{CEF}, & in the MDF, H_{MDF}, with $|H_{CEF}| = N$ the true national population, do
  - $1 - TVD(H_{CEF}, H_{MDF}) = 1 - \frac{\sum_g \sum_j |H_{MDF}(j,g) - H_{CEF}(j,g)|}{2\,N}$
- 1-TVD has some notable properties:
  - Is bounded within [0,1]
  - Can be very heuristically understood as "the proportion of table entries that were exactly as enumerated"
  - As defined here, tends to emphasize more populous geounits

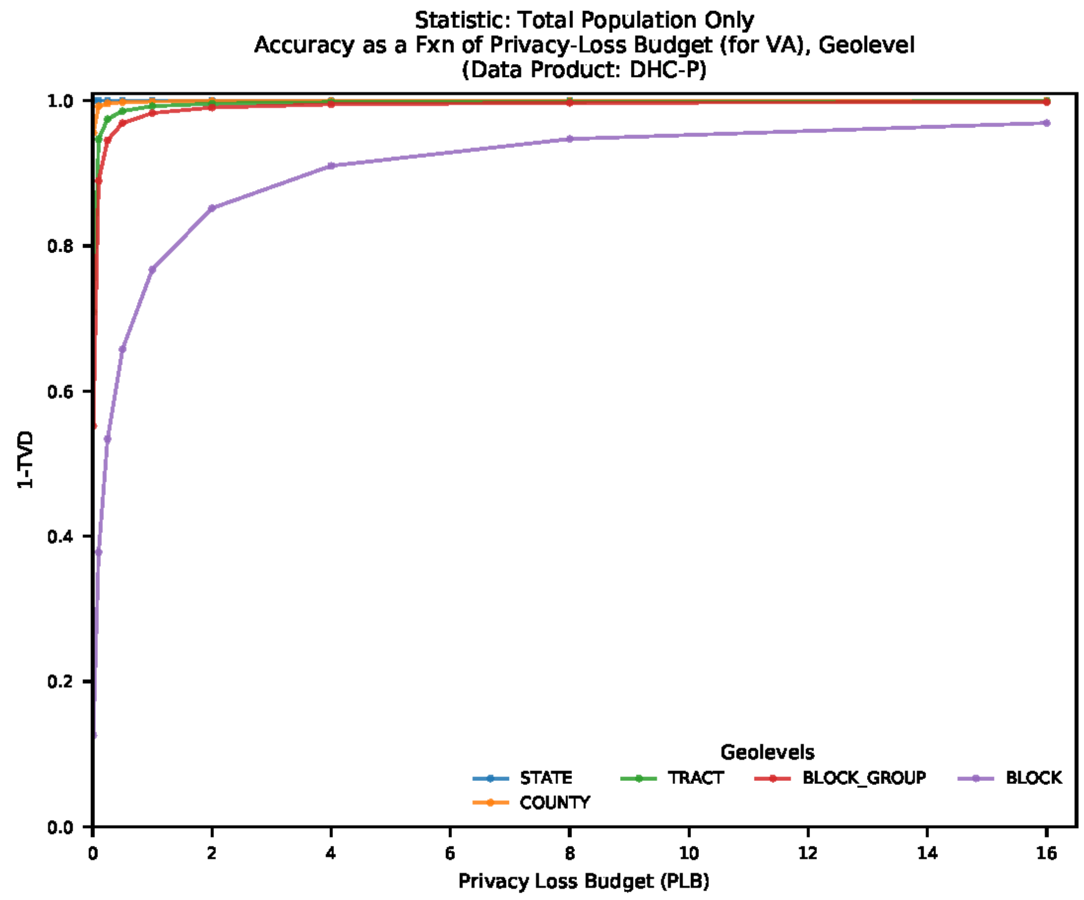
Shape
your future
START HERE >

United States®
Census
2020

For The 2010 Demonstration Data Products, We Used 2 Primary Metrics [2]

- The second metric was an L1 error over quantiles, a measure of difference in the shape of two distributions. We computed this as:
  - Given a target set of attribute-levels T (e.g., T=Male) to be crossed with Age, **drop** any geographic unit *g* that had either $H_{CEF}(T,g) = 0$ or $H_{MDF}(T,g) = 0$
  - For the remaining geounits $g \in G' \subset G$, set $q_{P,g}(T,q)$ to be the *q*th percentile of the distribution of ages for persons in *g* in product P with properties matching *T* (e.g., median age of men in the CEF for geounit *g*). Then do:
  - $L1\left(q_g(T,p)\right) = AVG_{g\in G'}\left(|q_{CEF,g}(T,p) - q_{MDF,g}(T,p)|\right)$
- This metric was exclusively used for the Sex-by-Age sub-histogram. It allows for statements like, *"On average, the median Age in a Tract for Males (Females) was off by XXX years"*

Persons: Total Population 1-TVD [1 of 5]



Generally, 1-TVD performance is better for tabulations with fewer counts per geographic unit. Total Population, for example, contributes just a single count per geounit. (CBDRB-FY20-103)

New Experiments: How does our re-identification attack fare on MDFs produced by TDA?

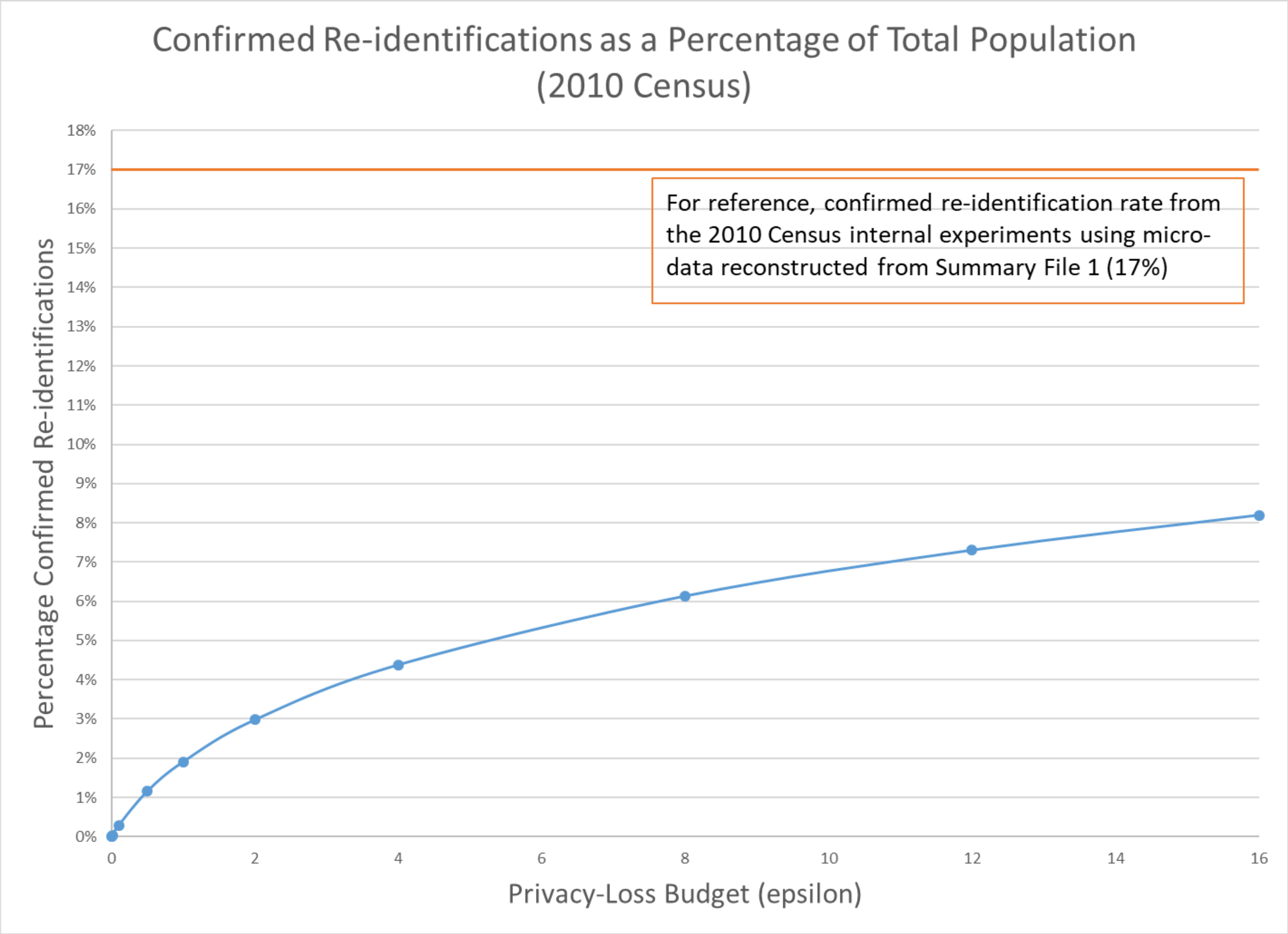
New Experiments

Using exactly the same re-identification strategy, analyze the national differentially private microdata for persons at different privacy-loss budgets from 0 to 16

We used PLB of 4 for the differentially private person-level microdata compute the 2010 Demonstration Data Products from DHC-P..

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

CBDRB-FY20-103



CBDRB-FY20-103

Shape your future START HERE >

United States Census 2020

In case you have follow-up questions/comments…

Philip Leclerc

Mathematical Statistician

Center for Enterprise Dissemination-Disclosure Avoidance

Philip.Leclerc@census.gov