Census TopDown Algorithm (TDA):
A Primer on Its Structure, Properties, and Parameters

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

NAC/CSAC Working Groups on Differential Privacy
June 10, 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.

Census TDA: Requirements and Properties I

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 x relgq x sex x age x Hispanic x cenrace
- This product yields a “histogram” (fully saturated contingency table)
- With shape: ≈ 8M x 42 x 2 x 116 x 2 x 63 = ≈ 8M × 1,227,744

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)

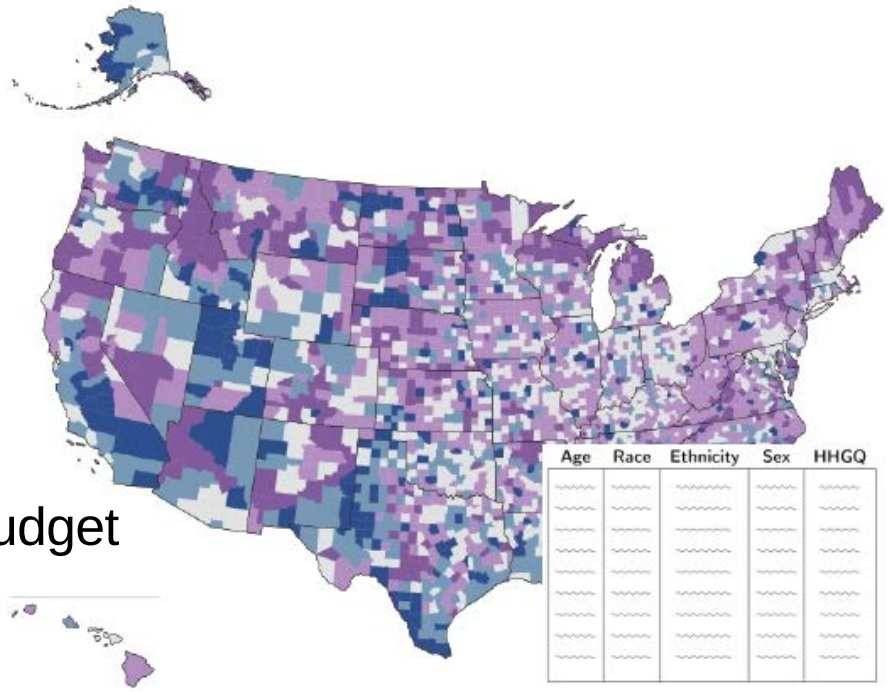
Utility/Accuracy for pre-specified tabulations

- 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 7 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!)

Note: With the adoption of multi-pass post-processing, TDA produces the histograms for each geographic level in a series of passes. (see slide 7)

Benefits of TDA

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

Multi-pass Post-processing

To address the sparsity issue, which introduced substantial distortions in the 2010 Demonstration Data Products, TDA processing is now performed in a series of passes.

At each geographic level, the algorithm constructs histograms for a subset of queries in a series of passes for that level, constraining the histogram for each pass to be consistent with the histogram produced in the prior pass.

Pass 1: Total Population
Pass 2: Tabulations supporting PL94-171 Redistricting Data
Pass 3: Tabulations supporting Population Estimates and most demographic use cases
Pass 4: The remainder of the DHC tabulations

Privacy-loss Budget Allocation
(as implemented in 5/27/2020 Detailed Summary Metrics)

By Geographic Level:

Nation	20%
State	20%
County	12%
Tract Group	12%
Tract	12%
Block Group	12%
Block	12%

By Query Set:

DHC-Person Allocations	
total	30%
hhgq	15%
votingage * hispanic * cenrace	29%
age * sex * hispanic * cenrace	25%
detailed	1%

Query Size
(for full Group I product implementation)

DHC-Person Queries (per geographic level)	
total	1
relgq	42
votingage * hispanic * cenrace	252
age * sex * hispanic * cenrace	29,232
detailed	1,227,744

DHC-Household Queries (per geographic level)	
total	1
hisp * race	14
tenvacgq	34
sex * hisp * race * hhtype_dhch	14,616
elderly * sex * hhtype_dhch	4,176
hhage * hhtype_dhch * sex	9,396
detailed	1,052,352

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