2020 Decennial Census: Formal Privacy Implementation Update

Philip Leclerc, Stephen Clark, and William Sexton

Center for Disclosure Avoidance Research

U.S. Census Bureau

Presented at the DIMACS/Northeast Big Data Hub Workshop on Overcoming Barriers to Data Sharing including Privacy and Fairness, Rutgers University, October 24, 2017

Roadmap

- Decennial & Algorithms Overview (P. Leclerc)
- Structural Zeros (W. Sexton)
- Integrating Geography: Top-Down vs Bottom-up (S. Clark)
- Questions/Comments

We are part of a team developing formally private mechanisms to protect privacy in the 2020 Decennial Census.

- Output will be protected query responses converted to microdata
- Microdata privacy guarantee is differential privacy conditioned on certain invariants (with interpretation derivable from Pufferfish)
- For example, total population, number of householders, number of voting age persons are invariant

The Decennial Census has many properties not typically addressed in the DP literature.

- Large scale with a complex workload
  - Fewer variables but larger sample than most Census products
  - Still high-dimensional relative to DP literature
  - Low and high sensitivity queries, multiple unit types
- Microdata that have legal integer response values is required by the tabulation system
- Evolving/distributed evaluation criteria (on-going discussion with domain-area experts)
  - Which subsets of the workload are most important?
  - How should subject-matter expert input be used to help leadership determine the weights of each subset of the workload?
  - How should the algorithms team allow for interpretable weighting of workload subsets?

The Decennial Census has many properties not typically addressed in the DP literature.

- Geographic hierarchy (approximately 8 million blocks)
- Modestly to extremely sparse histograms
  - Histograms are flat arrays with one-for-one map to all possible record types
  - Generated as Cartesian product of each variable's levels; impossible record types then removed
- Some quantities/properties must remain invariant
- Households/persons DP microdata must be privately joined: the data are *relational*, not just a single table

We intend to produce DP microdata, not just DP query answers.

- Microdata is the format expected by upstream processes
- Microdata are familiar to internal domain experts and external stakeholders
- Compact representation of query answers, convenient for data analysis
- Consistency between query answers by construction

Census leadership will determine the privacy budget; we will try to make tradeoffs as palatable as possible.

- The final privacy budget will be decided by Census leadership
- Our aim is to improve the accuracy-privacy trade-off curve
- We must provide interpretable “levers/gears” for leadership’s use in budget allocation

We tried a number of cutting-edge DP algorithms & identified best performers.

- Basic building blocks
  - Laplace Mechanism
  - Geometric Mechanism
  - Exponential Mechanism
- Considered, tested, under consideration
  - A-HPartitions
  - PrivTree
  - Multiplicative Weights Exponential Mechanism (/DualQuery)
  - iReduct/NoiseDown
  - Data-Aware Workload-Aware mechanism
  - PriView
  - Matrix Mechanism (/ GlobalOpt)
  - HB Tree

We tried a number of cutting-edge DP algorithms & identified best performers.

- Currently competitive for low-sensitivity, modest-dimensional tables
  - Hierarchical Branching “forest”
  - Matrix Mechanism (/ GlobalOpt)
- *None of these methods gracefully handle DP joins*

To enforce exact constraints, we explored a variety of post-processing algorithms.

- Weighted averaging + mean consistency / ordinary least squares
  - Closed form for per-query a priori accuracy
  - Does not give integer counts
  - Does not ensure nonnegativity
  - Does not incorporate invariants
  - Fast with small memory footprint

To enforce exact constraints, we explored a variety of post-processing algorithms.

- Nonnegative least squares
  - No nice closed form for per-query a priori accuracy
  - Does not give integer counts
  - Scaling issues (scipy/ecos/cvxopt/cplex/gurobi/…other options?)
  - Small consistent biases in individual cells become large biases for aggregates
  - Only incorporates some invariants
  - Fast with small memory footprint

To enforce exact constraints, we explored a variety of post-processing algorithms.

- Mixed-integer linear programming
  - No closed form for per-query a priori accuracy
  - Gives integer counts
  - Ensures nonnegativity
  - Incorporates invariants
  - Slow with large memory footprint

To enforce exact constraints, we explored a variety of post-processing algorithms.

- General linear + quadratic programming (LP + QP), iterative-proportional fitting
  - No closed form for per-query a priori accuracy
  - Gives integer counts (assuming total unimodularity)
  - Ensures nonnegativity
  - Incorporates (most) invariants
  - Fast with small memory footprint (but still bottlenecked by large histograms)
- *None of these methods gracefully handle post-processing joins*

We still don’t know the dimensionality for the 2020 census, but we have a pretty good idea.

- The demographic person record variables are age, sex, race/Hispanic, relationship to householder
- Age ranges from 0 to 115 inclusively
- Sex is male or female
- Race will likely include Hispanic in 2020
- Major Race Categories: WHT, BLK, ASIAN, AIAN, NHPI, SOR plus also likely HISP, MENA
- We also consider combinations of races
  - WHT and BLK and NHPI
- Relationship: 19 plus maybe foster child

Obviously adding categories increases dimensionality. We believe our computation limits are reached at dim = 3 million.

- 17 x 2 x 2 x 116 x 63 = 496,944 (2010)
- The following are plausible requirements for 2020:
  - 19 x 2 x 116 x 127 = 559,816 (added relationships, combined HISP)
  - 19 x 2 x 116 x 255 = 1,124,040 (added MENA)
  - 20 x 2 x 116 x 255 = 1,183,200 (added foster child)

The dimensionality of low-sensitivity household tables presents a computational conundrum.

- 14 key variables in 2010:
  - Age of Own Children / of Related Children (4 / 4 levels)
  - Number of People under 18 Years excluding Householder, Spouse, Partner (5 levels)
  - Presence of People in Age Range (including/excluding) Householder, Spouse, Partner (32 / 4 levels)
  - Presence of Non-Relatives / Multi-Generational Households (2/ 2 levels)

The dimensionality of household tables presents a computational conundrum.

- 14 key variables in 2010 (cont):
  - Household type / size (12 / 7 levels)
  - Age / sex / race of householder (9 / 2 / 7 levels)
  - Hispanic or Latino householder (2 levels)
  - Tenure (2 levels)

Generation of a histogram yields a maximum dimensionality of 1,734,082,560.

- This is roughly 3,500 times larger than the demographics dimensionality from 2010
- Likely intractable to generate DP microdata and handle post-processing
- Structural zeros provide some alleviation

A structural zero is something we are “certain” cannot happen even before the data is collected.

- Data are cleaned (edit and imputation) before DP is applied
  - If edit and imputation team makes something impossible, we can’t reintroduce it
- Demographic structural zeros:
  - Householder and spouse/partner must be at least 15 yrs old
  - Child/stepchild/sibling must be under 90 yrs old
  - Parent/parent-in-law must be at least 30 yrs old
  - At least one of the binary race flags must be 1
- Household structural zeros:
  - Every household must have exactly one householder
  - Child cannot be older than householder
  - Difference in age between spouse and householder

For demographic tables, structural zeroes aren't necessary to make the problem tractable but we still like them.

- Reducing dimensionality simplifies solution space for optimization.
- Assuming 20 x 2 x 116 x 255 histogram, how much does it help?
  - 5 x 2 x 15 x 255 = 38,250 (householders, spouses, partners under 15)
  - 2 x 2 x 30 x 255 = 30,600 (parent/parent-in-law under 30)
  - 1 x 2 x 95 x 255 = 48,450 (foster children over 20)
  - Total number of structural zeros = 212,160
  - About an 18% reduction

The reduction in dimensionality for household tables is substantial but will it be enough?

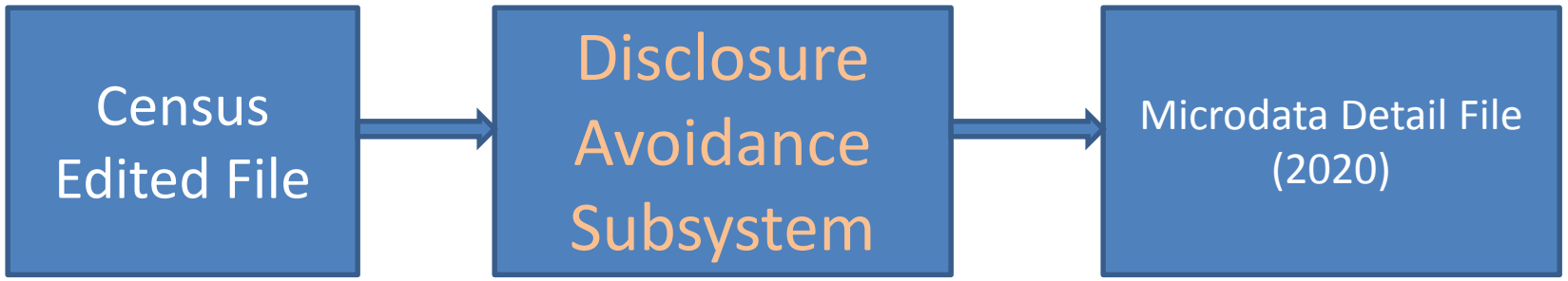
- By conditioning on household size alone, we reduce the dimensionality to 586,741,680. This is approximately a 3-fold reduction
- The interactions between age of own children and age of related child give further improvements which yield an upper bound of 297,722,880
- Additional reductions from structural zeros yield an approximation of about 60 million

There are several acronyms we want to introduce.

- CUF = “Census Unedited File” = respondent data
- CEF = “Census Edited File” = data file after editing
- MDF = “Microdata Detail File” = data file after disclosure controls are applied
- DAS = “Disclosure Avoidance Subsystem” = subsystem used to preserve privacy of data while maintaining usability of data
- 18E2ECT = “2018 End-to-End Census Test” = a test used to prepare Decennial systems for the actual 2020 Decennial Census

The Disclosure Avoidance Subsystem implements the privacy protections for the Decennial Census.

- Operates on the edited Census records
- Designed to make Census records safe to tabulate



We preserve privacy of data with better techniques, while maintaining data usability for stakeholders.

- Legacy techniques do not quantify the privacy loss inherent in publication releases
- New DP techniques allow measurable control over privacy loss incurred in generating the MDF from the CEF
- Our general approach for each geographic unit:
  - Generate DP queries using the CEF
  - Generate microdata that conforms closely to DP query answers

We have a complex geographic hierarchy.

- 8 million blocks
- Nation, state, county, tract, block group, block
- Sequential composition between levels
- Parallel composition within each level
- Two natural ways to traverse this hierarchy
  - Top-down (nation down to block)
  - Bottom-up (block up to nation)

The top-down approach starts with a national population and imputes geography.

- First we generate national DP microdata
- We then take DP queries over the CEF at the state level
- We assign national DP people to states based on state DP queries
- Repeat this process down to the block level
- Assignment to lower levels of geography must respect exactly known counts
- We plan to use this approach for the final 2020 production run

The bottom-up approach starts with a block population and aggregates up to the national level.

- First we generate block DP microdata using the geometric mechanism
- Post process block DP microdata to respect invariants
- We then aggregate to get block group DP microdata
- Repeat this process up to the national level
- We plan to use this approach for the 18E2ECT, which will generate response data only in the Providence, RI test area

Post-processing is a notable bottleneck.

- Post-processing is necessary to enforce invariants
- QP/LP optimizers encounter numerical and runtime issues at histogram sizes around 3 million
- We hope to leverage Spark to improve post-processing scalability, but the QP/LP algorithms are naturally sequential
- Some ideas for improving scalability:
  - Imitate regression trees to decompose optimization problems
  - Imitate branch-and-bound on TUM LPs with massive parallelism
  - Traditional decomposition techniques (e.g. Bender's)

We still have much to accomplish.

- Perform a bottom-up run for the E2E
- Perform a full top-down run on the Public Law 94-171, redistricting data, workload using real data (and simulated/external data)
- Generate tools for stakeholders to get a better feel for how DP affects accuracy
- Use Spark to scale DP and post-processing methods to larger tables

Questions/Comments
(Thanks for Attending!)

Philip Leclerc, Stephen Clark, and William Sexton
E-Mail: Philip.Leclerc@census.gov, Stephen.Clark@census.gov,
and William.N.Sexton@census.gov