Gates Open Research

RESEARCH ARTICLE

REVISED Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff [version 2; peer review: 2 approved]

Samantha Petti [1], Abraham Flaxman [2]

[1]School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 30332, USA
[2]Institute for Health Metrics and Evaulation, University of Washington, Seattle, Seattle, WA, 98121, USA

v2 **First published:** 04 Dec 2019, **3**:1722
https://doi.org/10.12688/gatesopenres.13089.1

Latest published: 06 Apr 2020, **3**:1722
https://doi.org/10.12688/gatesopenres.13089.2

Abstract

Background: The 2020 US Census will use a novel approach to disclosure avoidance to protect respondents' data, called TopDown. This TopDown algorithm was applied to the 2018 end-to-end (E2E) test of the decennial census. The computer code used for this test as well as accompanying exposition has recently been released publicly by the Census Bureau.
Methods: We used the available code and data to better understand the error introduced by the E2E disclosure avoidance system when Census Bureau applied it to 1940 census data and we developed an empirical measure of privacy loss to compare the error and privacy of the new approach to that of a (non-differentially private) simple-random-sampling approach to protecting privacy.
Results: We found that the empirical privacy loss of TopDown is substantially smaller than the theoretical guarantee for all privacy loss budgets we examined. When run on the 1940 census data, TopDown with a privacy budget of 1.0 was similar in error and privacy loss to that of a simple random sample of 50% of the US population. When run with a privacy budget of 4.0, it was similar in error and privacy loss of a 90% sample.
Conclusions: This work fits into the beginning of a discussion on how to best balance privacy and accuracy in decennial census data collection, and there is a need for continued discussion.

Keywords

Decennial census, differential privacy, TopDown algorithm, empirical privacy loss

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers 1	 2
version 2 (revision) 06 Apr 2020	✓ report	✓ report
version 1 04 Dec 2019	? report	? report

1 **David Van Riper**, IPUMS, University of Minnesota, Minneapolis, USA

2 **Ferdinando Fioretto**, Syracuse University, Syracuse, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Abraham Flaxman (abie@uw.edu)

Author roles: Petti S: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Flaxman A**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: ADF has consulted recently for Kaiser Permanente; Sanofi; Merck for Mothers; Agathos, Ltd; and NORC. SP has no competing interests to disclose.

Grant information: ADF is a recipient of funding from the Bill and Melinda Gates Foundation. The authors received no specific funding to support this work.

How to cite this article: Petti S and Flaxman A. **Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff [version 2; peer review: 2 approved]** Gates Open Research 2020, **3**:1722 https://doi.org/10.12688/gatesopenres.13089.2

First published: 04 Dec 2019, **3**:1722 https://doi.org/10.12688/gatesopenres.13089.1

REVISED Amendments from Version 1

This revised version improves the exposition in many places, thanks to helpful feedback from two peer reviewers, and adds detailed supplementary material on how empirical privacy loss (EPL) functions in the case of three simplified examples, including (1) when EPL is very close to epsilon; (2) when EPL is substantially less than epsilon, due to slack in the inequality in the Sequential Composition Theorem; and (3) when EPL is substantially more than epsilon, due to invariants.

Any further responses from the reviewers can be found at the end of the article

Acronyms

DP - differentially private

E2E - end-to-end

TC - total count

SC - stratified count

MAE - median absolute error

EPL - empirical privacy loss

Introduction

In the United States, the Decennial Census is an important part of democratic governance. Every ten years, the US Census Bureau is constitutionally required to count the "whole number of persons in each State," and in 2020 this effort is likely to cost over 15 billion dollars[1,2]. The results will be used for apportioning representation in the US House of Representatives and dividing federal tax dollars between states, as well as for a multitude of other governmental activities at the national, state, and local levels. Data from the decennial census will also be used extensively by sociologists, economists, demographers, and other researchers, and it will also inform strategic decisions in the private and non-profit sectors, and facilitate the accurate weighting of subsequent population surveys for the next decade[3].

The confidentiality of information in the decennial census is also required by law, and the 2020 US Census will use a novel approach to "disclosure avoidance" to protect respondents' data[4]. This approach builds on Differential Privacy, a mathematical definition of privacy that has been developed over the last decade and a half in the theoretical computer science and cryptography communities[5]. Although the new approach allows a more precise accounting of the variation introduced by the process, it also risks reducing the utility of census data—it may produce counts that are substantially less accurate than the previous disclosure avoidance system, which was based on redacting the values of table cells below a certain size (cell suppression) and a technique called swapping, where pairs of households with similar structures but different locations had their location information exchanged in a way that required that the details of the swapping procedure be kept secret[6].

To date, there is a lack of empirical examination of the new disclosure avoidance system, but the approach was applied to the 2018 end-to-end (E2E) test of the decennial census, and computer code used for this test as well as accompanying exposition has recently been released publicly by the Census Bureau[4,7].

We used the recently released code, preprints, and data files to understand and quantify the error introduced by the E2E disclosure avoidance system when the Census Bureau applied it to 1940 census data (for which the individual-level data has previously been released[8]) for a range of privacy loss budgets. We also developed an empirical measure of privacy loss and used it to compare the error and privacy of the new approach to that of a (non-differentially private) simple-random-sampling approach to protecting privacy.

Methods

Differential privacy definition and history

A randomized algorithm for analyzing a database is differentially private (DP) if withholding or changing one person's data does not substantially change the algorithm's output. If the results of the computation are roughly the same whether or not my data are included in the database, then the computation must be protecting my privacy. DP algorithms come with a parameter ε, which quantifies how much privacy loss is allowed, meaning how much can one person's data to affect the analysis.

To be precise, a randomized algorithm $\mathcal{A}$ is ε-DP if, for each possible output $\mathcal{P}$, for any pair of datasets D and D' that are the same everywhere except for on one person's data,

$$\Pr\left[\mathcal{A}(D)=\mathcal{P}\right]\leq \exp(\varepsilon)\Pr\left[\mathcal{A}(D')=\mathcal{P}\right].$$

Differential privacy is a characteristic of an algorithm; it is not a specific algorithm. Algorithms often achieve differential privacy by adding random variation[5].

The new disclosure avoidance system for the 2020 US Census is designed to be DP and to maintain the accuracy of census counts. To complicate things beyond the typical challenge faced in DP algorithm design, there are certain counts in the census that will be published precisely as enumerated, without any variation added. These invariants have not been selected for the 2020 decennial census yet, but in the 2018 end-to-end (E2E) test, the total count for each state and the number of households in each enumeration district were invariants. There are also inequalities that will be enforced. The E2E test required the total count of people in an enumeration district to be greater or equal to the number of occupied households in that district[9].

TopDown algorithm

At a high level, the census approach to this challenge repeats two steps for multiple levels of a geographic hierarchy (from the top down, hence their name "TopDown"). The first step (Imprecise Histogram) adds variation from a carefully chosen distribution to the stratified counts of individuals. This produces a set of counts with illogical inconsistencies, which we refer to as an "imprecise histogram". For example, counts in the imprecise histogram might be negative, might violate invariants or other inequalities, or might be inconsistent with the counts that are one level up in the geographic hierarchy. The second step (Optimize) finds optimized counts for each most-detailed cell in the histogram, using constrained convex optimization to make them as close as possible to the counts in the imprecise histogram, subject to the constraints that the optimized counts be non-negative, consistent with each other and the higher levels of the hierarchy, and satisfy the invariants and inequalities. These two steps are performed for each geographic level, from the coarsest to the finest. Each level is assigned a privacy budget ϵ_i (which governs how much variation to add in the Imprecise Histogram step), and the entire algorithm achieves ϵ-DP for $\epsilon = \sum_i \epsilon_i$. The 2020 US Census data may have six geographic levels, nested hierarchically: national, state, county, census tracts, block groups, and blocks; but in the 1940 E2E test four levels (national, state, county, and enumeration district) were included.

Step one: Imprecise Histogram. In the E2E algorithm applied to the 1940s microdata, TopDown added random variation in a flexible way that allowed the user to choose what statistics are the most important to keep accurate. The variation was added to the detailed histogram counts for the level and also to a preselected set of aggregate statistics. The detailed histogram counts stratified the population of each geographic by age (two values: under-18-year-olds and 18-plus), race (six values), ethnicity (two values: Hispanic and non-Hispanic), and household/group-quarters type (6 values). The aggregate statistics are sets of histogram count sums specified by some characteristics. For example, the “race/ethnicity/age” aggregate statistic contains 24 counts: people of each of the six racial categories who are also Hispanic ethnicity under age 18, of Hispanic ethnicity age 18 and over, of non-Hispanic ethnicity under age 18, and of non-Hispanic ethnicity age 18 and over.

The aggregate statistics (internally called “DP queries” in the TopDown algorithm) afford a way to choose specific statistics that are more important to keep accurate, and the E2E test included two such aggregates: a household/group-quarters query, which increases the accuracy of the count of each household type at each level of the hierarchy, and a race/ethnicity/age query, which increases the accuracy of the stratified counts of people by race, ethnicity, and voting age across all household/group-quarters types (again for each level of the spatial hierarchy). It also included “detailed queries” corresponding to boxes in the histogram. The detailed queries were afforded 10% of the privacy budget at each level, while the DP queries split the remaining 90% of the privacy budget, with 22.5% spent on the household/group-quarters queries and 67.5% spend on the race/ethnicity/age queries.

The epsilon budget of the level governed how much total random variation to add. A further parameterization of the epsilon budget determined how the variance was allocated between the histogram counts and each type of aggregate statistic. We write $\epsilon_i = h + s_1 + s_2 + ... + s_k$, where ϵ_i was the budget for the geographic level, h was the budget for the detailed queries, and $s_1, ... s_k$ were the budgets for each of the k types of aggregate statistics. Then variance was added independently to each count according to the follow distribution:

$$\text{imprecise detailed histogram count} = \text{precise detailed histogram count} + G(h/2)$$

$$\text{imprecise aggregate stat } j = \text{precise aggregate stat } j + G(s_j/2)$$

where $G(z)$ denotes the two-tailed geometric distribution,

$$\Pr\left[G(z) = k\right] = \frac{(1-\exp(-z))\exp(-z|k|)}{1+\exp(-z)}.$$

The imprecise counts and imprecise aggregate statistics are unbiased estimates with variance $(1 - \exp(-z))^2/(2\exp(-z))$, where z is the parameter for the geometric random variable added. A higher privacy budget means the variance added is more concentrated around zero, and therefore the corresponding statistic is more accurate. Therefore, adjusting the privacy budgets of the various aggregate statistics gives control over which statistics are the most private/least accurate (low fraction of the budget) and the most accurate/least private (high fraction of the budget).

The variation added to each histogram count comes from the same distribution, and is independent of all other added variation; the variance does not scale with the magnitude of count, e.g. adding 23 people to the count of age 18 and older non-Hispanic Whites is just as likely as adding 23 people to the count of age under 18 Hispanic Native Americans, even though the population of the latter is smaller.

Step two: Optimize. In this step, the synthetic data is created from the imprecise detailed histogram counts and aggregate statistics by optimizing a quadratic objective function subject to a system of linear equations and inequalities. The algorithm creates a variable for each detailed histogram count and each aggregate statistic. It adds equations and inequalities to encode the requirements that (i) each count and aggregate statistic is non-negative, (ii) the invariants and inequalities are satisfied, (iii) the aggregate statistics are the sum of the corresponding detailed histogram counts, and (iv) the statistics are consistent with the higher level synthetic data counts (i.e. the total number of people aged 18 and over summed across the counties in a state is equal to the number of people aged 18 and over in that state as reported by synthetic data set constructed in the previous phase). The optimization step finds a solution that satisfies these equations and minimizes the weighted sum of the squared differences between each variable/aggregate of variables and the corresponding imprecise detailed histogram count or imprecise aggregate statistic. This sum is weighted with the weight of each term taken to be proportional to the magnitude of the variation added in step one to create the imprecise count. The solution to this optimization is not necessarily integral, however, and TopDown uses a second optimization step to round fractional counts to integers.

We note that the approach that Census Bureau has taken with the TopDown where imprecise histogram data is optimized based on internal consistency has been developed in a line of research over the last decade to that has focused on obtaining count data that is DP *and* accurate[10–13].

Empirical Privacy Loss for quantifying impact of optimize steps

As described above, the privacy loss of a DP algorithm is quantified by a unitless number, ϵ, that bounds the maximum of the log of the relative change in the probability of an output when one person's data is changed. This bound is typically proven by logical deduction, and for complex DP algorithms, the proof often relies on the Sequential Composition Theorem[5], which states that information derived by combining the output of an ϵ_1-DP algorithm and an ϵ_2-DP algorithm is at most $(\epsilon_1 + \epsilon_2)$-DP. This theorem is an inequality, however, and the inequality might have room for improvement.

It is possible to empirically quantify privacy loss, which has the potential to show that the inequality of the sequential composition theorem is not tight. The brute force approach quantify privacy loss empirically is to search over databases *D* and *D'* that differ on one row to find the event *E* with the largest ratio of probabilities; this is too computationally intensive to be feasible for all but the simplest DP algorithms.

For algorithms that produce DP counts of multiple subpopulations, such as TopDown, it is possible to use the distribution of the residual difference between the precise count and the DP count to derive a proxy of the distribution produced by the brute force approach[14]. The special structure of count queries affords a way to avoid re-running the algorithm repeatedly, which is essential for TopDown, since it takes several hours to complete a single run of the algorithm. Assuming that the residual difference of the DP count minus the precise count is identically distributed for queries across similar areas (such as voting-age population across all enumeration districts), and then instead of focusing on only the histogram counts containing the individual who has changed, we used the residuals for all areal units to estimate the probability of the event we are after:

$$\Pr\left[\text{error}_j = k\right] \approx \left(\sum_{j'=1}^{c} 1\left[\left\{\text{error}_{j'} = k\right\}\right]\right)\Big/ C =: \hat{p}_k,$$

where error_j is the residual difference of DP counts returned by TopDown minus the precise count for that same quantity in the 1940 census, and the $\text{error}_{j'}$ are residuals for *C* other queries assumed to be exchangeable.

To measure the empirical privacy loss (EPL), we approximated the probability distribution of the residuals (DP count minus precise count at a selected level of the geographic hierarchy), which we denote $p^{\text{KDE}}(x)$, using Gaussian kernel density estimation (KDE) with a bandwidth of 0.1, and compare the log-ratio inspired by the definition of ϵ-DP algorithms:

$$\text{EPL}(x) = \log\left(\frac{p^{\text{KDE}}(x)}{p^{\text{KDE}}(x+1)}\right);$$

$$\text{EPL} = \max_{x\in(-\infty,\infty)} \left\{\text{abs}\left(\text{EPL}(x)\right)\right\}$$

See Supplementary Methods Appendix for additional detail on the design and validation of the EPL metric[15].

TopDown options still to be selected

There are seven key choices in implementing TopDown, that balance accuracy and privacy. We list them here, and state how they were set in the 2018 end-to-end test when run on the 1940s Census data:

1. Overall privacy. A range of ϵ values, with {0.25, 0.50, 0.75, 1.0, 2.0, 4.0, 8.0} used in the E2E test run on the 1940 Census Data.

2. How to split this budget between national, state, county, tract, block group, and block. In the test run, ϵ was split evenly between national, state, county, and enumeration district.

3. What aggregate statistics (also known as "DP Queries") to include. In the test, two DP Queries were included: (i) counts stratified by age-group/race/ethnicity (and therefore aggregated over household/group-quarters type); and (ii) the household/group-quarters counts, which tally the total number of people living in each type of housing (in a household, in institutional facilities of certain types, in non-institutional facilities of certain types).

4. At each level, how to split level-budget between detailed queries and DP queries. The test run used 10% for detailed queries, 22.5% for household/group-quarters; and 67.5% for age-group-/race-/ethnicity-stratified counts.

5. What invariants to include. The test run held the total population count at the national and state level invariant.

6. What constraints to include. The test run constrained the total count of people to be greater or equal to total count of occupied households at each geographic level.

7. What to publish. The test run published a synthetic person file and synthetic household file for a range of ϵ values, for four different seeds to the pseudorandom number generator.

Our evaluation approach

1. We calculated residuals (DP count minus precise count) and summarized their distribution by its median absolute error (MAE) for total count (TC) and age/race/ethnicity stratified count (SC) at the state, county, and enumeration-district level. We also summarized the size of these counts from the precise-count versions to understand relative error as well as the absolute error introduced by TopDown.

2. We calculated a measure of empirical privacy loss (EPL), inspired by the definition of differential privacy. To measure EPL, we approximated the probability distribution of the residuals (DP count minus precise count at a selected level of the geographic

hierarchy), which we denote $p^{\mathrm{KDE}}(x)$, using Gaussian kernel density estimation with a bandwidth of 0.1, and compare the log-ratio inspired by the definition of ϵ-DP algorithms:

$$\mathrm{EPL}(x) = \log\left(\frac{p^{\mathrm{KDE}}(x)}{p^{\mathrm{KDE}}(x+1)}\right);$$

$$\mathrm{EPL} = \max_{x \in (-\infty,\infty)} \left\{\mathrm{abs}\left(\mathrm{EPL}(x)\right)\right\}$$

See Supplementary Methods Appendix for additional detail on the design and validation of the EPL metric[15]. We hypothesized that the EPL of TopDown will be substantially smaller than the theoretical guarantee of ϵ, which was proven using the Sequential Composition Theorem, which provides an inequality that is usually not a tight bound[14]. However, it is possible that it will be much larger than ϵ, due to the difficult-to-predict impact of including certain invariants.

3. We searched for bias in the residuals from (1), with our hypothesis that the DP counts are larger than precise counts in spatial areas with high homogeneity and DP counts are smaller than precise counts in areas with low homogeneity. We based this hypothesis on the expected impact of the non-negativity constraints included in the optimization steps of the TopDown algorithm. For each detailed query with a negative value for its noisy count, the optimization step will increase the value to make the results logical, and this reduction in variance must tradeoff some increase in bias. To quantify the scale of the bias introduced by optimization, for each geographic area, we constructed simple homogeneity index by counting the cells of the detailed histogram that contained a precise count of zero, and we examined the bias, defined as the mean of the DP count minus precise count, for these areas when stratified by homogeneity index.

4. We also compared the median absolute error and empirical privacy loss of TopDown to a simpler, but not-differentially-private approach to protecting privacy, Simple Random Sampling (i.e. sampling without replacement) for a range of sized samples. To do this, we generated samples without replacement of the 1940 Census Data for a range of sizes, and applied the same calculations from (1) and (2) to this alternatively perturbed data.

Results

Error and privacy of TopDown

Recall that geographic areas are nested: enumeration districts are contained within counties, which are contained within states. We found error in total count (TC) varied as a function of total privacy loss budget. Running TopDown with ϵ = 0.5 produced median absolute error in TC of 29 at the enumeration district level and 45 at the county level; ϵ = 1.0 produced median absolute error in TC of 15 at the enumeration district level and 24 at the county level; and ϵ = 2.0 produced median absolute error in TC of 8 at the enumeration district level and 13 at the county level (Full table in Extended Data[16]). At the state level, there was TC error of 0.0, as expected from the state TC invariant. The median and 95th percentile of TC from the precise-count data were 865 and 2342 for enumeration districts, 18,679 and 122,710 for counties, and 1,903,133 and 7,419,040 for states.

Error in stratified count (SC) varied similarly; when ϵ = 0.5, the median absolute error in SC at the enumeration district level was 10 people, at the county level was 11 people, and at the state level was 13 people; for ϵ = 1.0, the median absolute error in SC at the enumeration district level was 6 people, at the county level was 6 people, and at the state level was 7 people; and for ϵ = 2.0, the median absolute error in SC at the enumeration district level was 4 people, at the county level was 4 people, and at the state level was 4 people. The median and 95th percentile of SC from the precise-count data were 88 and 967 for enumeration districts, 47 and 17,480 for counties, and 229 and 714,208 for states. (Figure 1)

We found that the empirical privacy loss was often substantially smaller than the privacy loss budget. For ϵ = 0.5, the empirical privacy loss for TC at the enumeration district level was 0.033 and at the county level was 0.035 (at the state level empirical privacy loss is undefined, since the invariant makes all residuals zero); for ϵ = 1.0, the empirical privacy loss for TC at the enumeration district level was 0.064 and at the county level was 0.048; and for ϵ = 2.0, the empirical privacy loss for TC at the enumeration district level was 0.116 and at the county level was 0.094.

This relationship between privacy loss budget and empirical privacy loss was similar for stratified counts (SC) at the enumeration district and county level, but for privacy loss budgets of 1.0 and less, the empirical privacy at the enumeration district level was loss for SC was not as responsive to ϵ. For ϵ = 1.5, the empirical privacy loss for SC at the enumeration district level was 0.200, at the county level was 0.165, and at the state level was 0.104; for ϵ = 1.0, the empirical privacy loss for SC at the enumeration district level was 0.241, at the county level was 0.164, and at the state level was 0.166; and for ϵ = 2.0, the empirical privacy loss for SC at the enumeration district level was 0.280, at the county level was 0.253, and at the state level was 0.300. EPL values for all combinations of ϵ and all geographic levels appear in the Extended Data.

Comparison with error and privacy of simple random sampling

We found that the MAE and EPL of Simple Random Sampling (i.e. sampling uniformly, without replacement) varied with larger sample size in a manner analogous to the total privacy budget in TopDown, for $\epsilon \geq 1$. For a 5% sample of the 1940 Census data, we found median absolute error in TC of 74 at the enumeration district level, 388 at the county level, and 3883 at the state level; a 50% sample produced median absolute error in TC of 17 at the enumeration district level, 90 at the county level, and 932 at the state level; and a 95% sample produced median absolute error in TC of 4 at the enumeration district level, 20 at the county level, and 130 at the state level.

Error in stratified count varied similarly; for a 5% sample, we found median absolute error in SC of 18 at the enumeration district level, 19 at the county level, and 41 at the state level; a 50% sample produced median absolute error in TC of 4 at the enumeration district level, 5 at the county level, and 9 at the state level.

We found empirical privacy loss increased as sample size increased. For a 5% sample, at the enumeration district level, we found EPL of 0.020 for TC and 0.098 for SC, and at the county level, we found 0.035 for TC and 0.034 for SC; a 50% sample produced EPL of 0.079 for TC and 0.318 for SC at the enumeration district level, and 0.082 for TC and 0.150 for SC at the county level; and a 95% sample produced EPL of 0.314 for TC and 1.333 for SC at the enumeration district level, and 0.429 for TC and 0.612 for SC at the county level (Figure 2, Table 1).

Bias in the variation introduced by TopDown

The bias introduced by TopDown varied with homogeneity index, as hypothesized. Enumeration districts with homogeneity index 0 (0 empty cells in the detailed histogram) had TC systematically lower than the precise count, while enumeration districts homogeneity index 22 (the maximum number of empty cells observed in the detailed histogram) had TC systematically higher than the precise count. The size of this bias decreased as a function of ϵ. Homogeneity index 0 had bias of -31.7 people for ϵ = 0.5, -18.9 people for ϵ = 1.0, and -11.6 people for

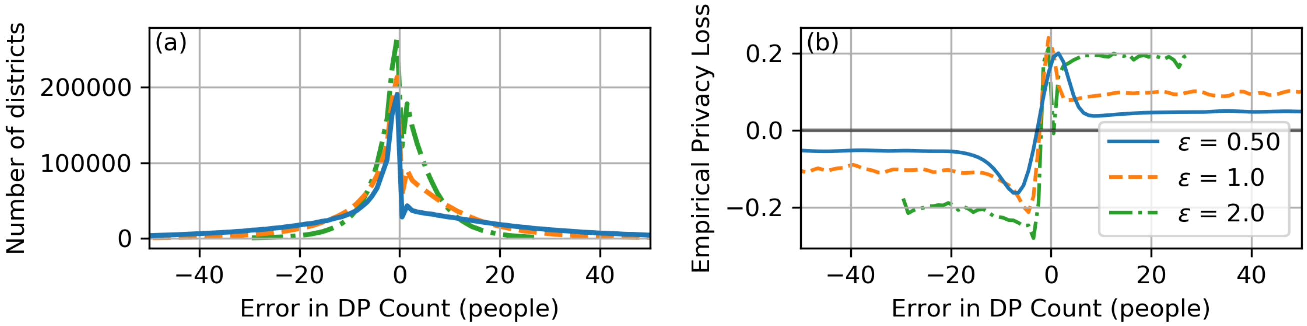

Figure 1. Error distribution and empirical privacy loss for stratified counts at the enumeration district level. Panel (**a**) shows the distribution of residuals (DP - Precise) for stratified counts at the enumeration district level, stratified by age, race, and ethnicity; and panel (**b**) shows the empirical privacy loss function, $\mathrm{EPL}(x) = \log(p^{\mathrm{KDE}}(x)/p^{\mathrm{KDE}}(x+1))$, where $p^{KDE}(x)$ is the probability density corresponding to the histogram in (a), after smoothing with a Gaussian kernel of bandwidth 0.1; the EPL value is the maximum of the absolute value of EPL(x) over all x.

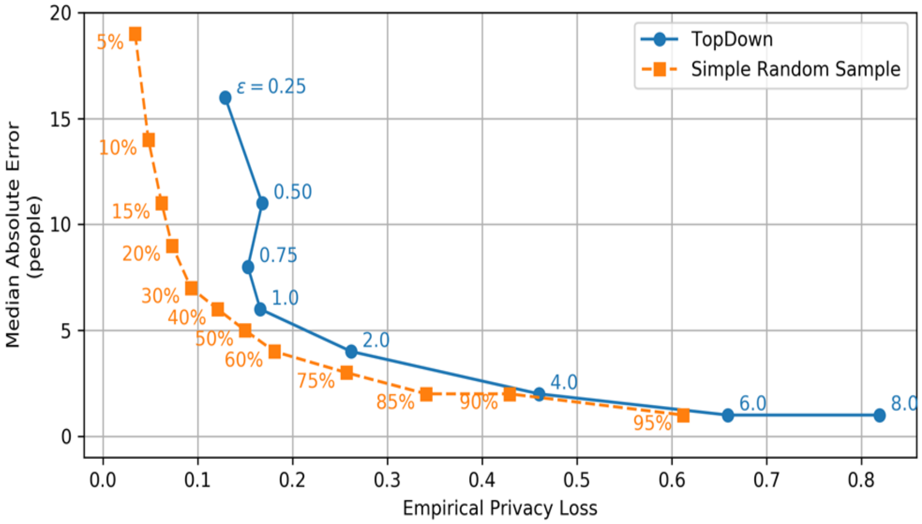

Figure 2. Tradeoff curve of median absolute error and empirical privacy loss of stratified counts at the county level. The curve with circular markers shows that in TopDown, the choice of ϵ controls the tradeoff between MAE and EPL, although for ϵ < 1 there is not much difference in EPL. The curve with square markers shows the MAE and EPL of Simple Random Sampling for a range of sample sizes, for comparison. For example, TopDown with ϵ = 1.0. provides privacy loss and estimation error similar to a sample of 50% of the 1940 census data, while ϵ = 2.0 is comparable to a 75% sample (for counts stratified by age, race, and ethnicity at the county level; different aggregate statistics produce different comparisons).

$\epsilon = 2.0$; while homogeneity index 22 had bias of 5.4 people for $\epsilon = 0.5$, 3.6 people for $\epsilon = 1.0$, and 2.3 people for $\epsilon = 2.0$. (Figure 3)

Counties displayed the same general pattern, but there are fewer counties and they typically have less empty strata, so it was not as pronounced. The size of this bias again decreased as a function of ϵ. Homogeneity index 0 had bias of -59.2 people for $\epsilon = 0.5$, -33.9 people for $\epsilon = 1.0$, and -18.8 people for $\epsilon = 2.0$; while homogeneity index 22 had bias of 21.7 people for $\epsilon = 0.5$, 14.5 people for $\epsilon = 1.0$, and 11.1 people for $\epsilon = 2.0$.

Discussion

We anticipate some readers of this will be social researchers who rely on Census Bureau data for quantitative work, and who have concerns that the Census Bureau is going to reduce the accuracy of this data. Such a reader may be open to the possibility that privacy is a valid reason for reducing accuracy, yet still be concerned about how this will affect their next decade of research. Our results visually summarized in Figure 2 can help to understand the potential change in accuracy: if $\epsilon = 1.0$, for county-level stratified counts, TopDown will be like the uncertainty introduced by working with a 50% sample of the full dataset; if $\epsilon = 2.0$, it will be like working with a 75% sample; and if $\epsilon = 6.0$, it will have accuracy matching a 95% sample, which is pretty close to having the full data without protecting privacy. Such a reader may still want to see an analysis like this run on the 2010 decennial census data, but we hope this will help them rest a little easier about the quality of the data they are relying on for their work.

Table 1. Values of privacy loss, and corresponding proportions of Simple Random Sample (SRS) with most similar median-absolute-error/empirical-privacy-loss profile.

Privacy Budget (ϵ)	Closest SRS sample proportion (%)
1.0	50%
2.0	75%
4.0	90%
6.0	95%

We also expect that some readers will be more drawn to the lower end of the epsilon curve. Just how private is TopDown with $\epsilon = 0.25$, especially when total count at the state-level is invariant? Our results show that all ϵ less than 1.0 have empirical privacy loss around 0.15, independent of ϵ. You can add more and more variation, but, perhaps due to the invariants, that variation does not translate into more and more privacy.

Comparing error in total count or stratified count across levels of the geographic hierarchy reveals a powerful feature of the TopDown algorithm: the error is of similar magnitude even though the counts are substantially different in size. This is because the variation added at each level has been specified to have the same portion of the total privacy budget. It remains to be investigated how alternative allocations of privacy budget across levels will change the error and empirical privacy loss.

For $\epsilon \geq 1.0$, TopDown introduced near minimal variation and attained empirical privacy loss almost 10 times less than ϵ. We also found that this created a quantifiable amount of bias. The bias increased the reported counts in homogeneous districts while decreasing the counts in racially and ethnically mixed districts. The TopDown algorithm may therefore drive some small amount of redistribution of resources from diverse urban communities to segregated rural communities.

Accurate counts in small communities are important for emergency preparedness and other routine planning tasks performed by state and local government demographers, and this work may help to understand how such work will be affected by the shift to a DP disclosure avoidance system.

This work has not investigated more detailed research uses of decennial census data in social research tasks, such as

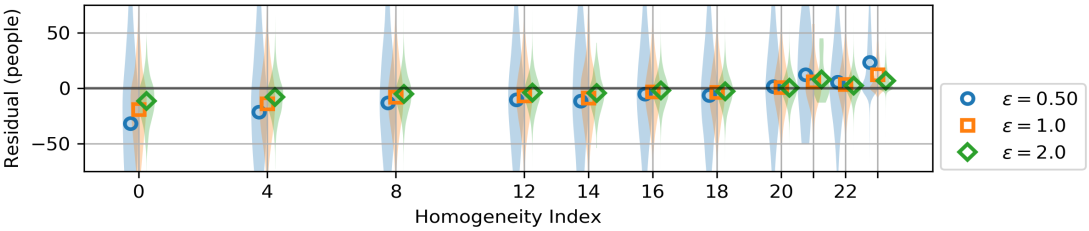

Figure 3. Relationship between homogeneity index and residual for three values of epsilon The homogeneity index, defined as the number of cells with precise count of zero in the detailed histogram, is positively associated with the bias (markers show the mean difference between the DP count estimated by TopDown and the precise count, and shaded area shows the distribution of individual differences). This plot shows the association for enumeration districts, and a similar relationship holds at the county level. As ϵ increases, the scale of the bias decreases. (Enumeration districts attained only a subset of the homogeneity index values between 0 and 23, which is why there are different width gaps between markers. We pooled the residuals for the four runs of TopDown with different random seed.)

segregation research, and how this may be affected by TopDown.

Another important use of decennial census data is in constructing control populations and survey weights for survey sampling of the US population for health, political, and public opinion polling. Our work provides some evidence on how TopDown may affect this application, but further work is warranted.

This work fits into the beginning of a discussion on how to best balance privacy and accuracy in decennial census data collection, and there is a need for continued discussion. This need must be balanced against a risky sort of observer bias—some researchers have hypothesized that calling attention to the privacy and confidentiality of census responses, even if done in a positive manner, could reduce the willingness of respondents to answer census questions, and ongoing investigation with surveys and cognitive testing may provide some evidence on the magnitude of this effect as well as potential countermeasures[17].

Limitations

There are many differences between the 1940 census data and the 2020 data to be collected next year. In addition to the US population being three times larger now, the analysis will have six geographic levels instead of four, ten times more race groups and over 60 times more age groups. We expect that this will yield detailed queries with typical precise count sizes even smaller than the stratified counts for enumeration districts we have examined here. We suspect that impact of this will likely be to slightly decrease accuracy and increase privacy loss, but the accuracy of our hypothesis remains to be seen.

In addition to the changes in the data, additional changes are planned for TopDown, such as a switch from independent geometrically distributed variation to the High Dimensional Matrix Mechanism. We expect this to increase the accuracy a small amount without changing the empirical privacy loss.

In this work, we have focused on the median of the absolute error, but the spread of this distribution is important as well, and in future work, researchers may wish to investigate the tails of this distribution. We have also focused on the empirical privacy loss for specific queries at specific geographic aggregations, and our exploration was not comprehensive. Therefore, it is possible that some other test statistic would demonstrate a larger empirical privacy loss than we have found with our approach. Our approach also assumes that the residuals for different locations in a single run are an acceptable proxy for the residuals from the same location across multiple runs. Although these are certainly different, we suspect that the difference is sufficiently small as to not affect our estimates substantially.

Conclusion

The TopDown algorithm will provide a provably ϵ-DP disclosure avoidance system for the 2020 US Census, and it provides affordances to balances privacy and accuracy. This is an opportunity, but it is not without risks. Taking advantage of the opportunity and mitigating the risks will require that we understand what the approach is doing, and we hope that this analysis of the 2018 E2E test can help build such understanding.

Data availability

Source data

Individual-level data from the 1940 US Census is available from IPUMS https://doi.org/10.18128/D010.V8.0.EXT1940USCB[8].

These data are under Copyright of Minnesota Population Center, University of Minnesota. Access to the documentation is freely available without restriction; however, users must register before extracting data from the website.

The output of the TopDown algorithm when run on the 1940 US Census data is available to download from the US Census Bureau: https://www2.census.gov/census_1940/.

These data are under Copyright of the United States Census Bureau.

Extended data

Zenodo: Extended data for Differential privacy in the 2020 US census, what will it do? Quantifying the accuracy/privacy tradeoff. https://doi.org/10.5281/zenodo.3551215[16].

This project contains a full table of summary counts and errors for a range of levels of geographic hierarchy, stratification, and epsilon.

Zenodo: Supplementary Methods Appendix for Differential privacy in the 2020 US census, what will it do? Quantifying the accuracy/privacy tradeoff: Design and validation of Empirical Privacy Loss (EPL) metric. https://doi.org/10.5281/zenodo.3727242[15].

This project contains additional details on the design and validation of the EPL metric used in this paper.

Extended data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

Scripts to produce all results and figures in this paper are available online: https://github.com/aflaxman/dp_2020_census/.

Archived scripts at time of publication: https://doi.org/10.5281/zenodo.3551217[18].

License: MIT License.

Acknowledgements

Thanks to Neil Marquez (University of Washington) for suggesting comparing TopDown to simple random sampling. Thanks to danah boyd, Cynthia Dwork, Simson Garfinkel, Philip Leclerc, and Kunal Talwar for their helpful comments and discussion of this work.

References

1. Garfinkel S, Abowd JM, Martindale C: **Understanding database reconstruction attacks on public data.** *Communications of the ACM*. 2019; **62**(3): 46–53.
Publisher Full Text

2. United States Government Accountability Office.**Census Bureau improved the quality of its cost estimation but additional steps are needed to ensure reliability.** U.S. G.A.O. 2018.
Reference Source

3. Ruggles S, Fitch C, Magnuson D, *et al.*: **Differential privacy and Census data: Implications for social and economic research.** *AEA papers and proceedings*. 2019; **109**: 403–08.
Publisher Full Text

4. Abowd JM, Garfinkel SL: **Disclosure avoidance and the 2018 Census test: Release of the source code.** 2019.
Reference Source

5. Dwork C, Roth A: **The algorithmic foundations of differential privacy**. *Foundations and Trends in Theoretical Computer Science*. Now Publishers, Inc. 2014; **9**(3–4): 211–407.
Reference Source

6. McKenna L: **Disclosure avoidance techniques used for the 1970 through 2010 Decennial Censuses of Population and Housing.** *Center for Economic Studies, U.S. Census Bureau*. 2018.
Reference Source

7. boyd d: **Differential privacy in the 2020 Decennial Census and the implications for available data products.** CoRR abs/1907.03639. 2019.
Reference Source

8. Ruggles S, Flood S, Goeken R, *et al.*: **IPUMS USA: Version 8.0 extract of 1940 Census for U.S. Census Bureau disclosure avoidance research [dataset].** 2018.
http://www.doi.org/10.18128/D010.V8.0.EXT1940USCB

9. Garfinkel S, others. **2018 end-to-end test disclosure avoidance system design specification.** U.S. Census Bureau. 2019.
Reference Source

10. Hay M, Rastogi V, Miklau G, *et al.*: **Boosting the accuracy of differentially private histograms through consistency.** *Proceedings of the VLDB Endowment. VLDB Endowment*. 2010;1021–32.
Publisher Full Text

11. Li C, Miklau G, Hay M, *et al.*: **The matrix mechanism: Optimizing linear counting queries under differential privacy.** *The VLDB journal*. Springer. 2015; **24**: 757–81.
Publisher Full Text

12. Kuo YH, Chiu CC, Kifer D, *et al.*: **Differentially private hierarchical count-of-counts histograms.** *Proceedings of the VLDB Endowment*. VLDB Endowment. 2018;1509–21.
Publisher Full Text

13. Fioretto F, Van Hentenryck P: **Differential privacy of hierarchical census data: An optimization approach.** *International conference on principles and practice of constraint programming*. Springer. 2019; 639–55.
Publisher Full Text

14. Flaxman AD: **Empirical quantification of privacy loss with examples relevant to the 2020 US Census.** 2019.
Reference Source

15. Flaxman AD, Petti S: **Supplementary Methods Appendix to Differential privacy in the 2020 US census: what will it do?** *Quantifying the accuracy/privacy tradeoff: Design and validation of Empirical Privacy Loss (EPL) metric* (Version v1.0). 2020.
http://www.doi.org/10.5281/zenodo.3727242

16. Petti S, Flaxman AD: **Extended data for Differential privacy in the 2020 US census, what will it do? Quantifying the accuracy/privacy tradeoff.** *Zenodo*. 2019.
http://www.doi.org/10.5281/zenodo.3718648

17. Childs JH, Abowd J: **Update on confidentiality and disclosure avoidance.** U.S. Census Bureau. 2019.
Reference Source

18. Petti S, Flaxman A: **aflaxman/dp_ 2020_census: Replication archive code when paper was resubmitted (Version v1.0.1).** *Zenodo*. 2020.
http://www.doi.org/10.5281/zenodo.3718649

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 11 May 2020

https://doi.org/10.21956/gatesopenres.14301.r28749

✓ **David Van Riper**
IPUMS, University of Minnesota, Minneapolis, MN, USA

I am pleased with the authors' responses to my initial review. Adding extra data in the supplementary materials provides a fuller picture of the analyses they executed, and the clarifications added to the text enhance understanding. I also greatly appreciate the new Supplementary Methods Appendix that provides a more detailed discussion of the EPL measure.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: geography, demography, census data, differential privacy

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 06 May 2020

https://doi.org/10.21956/gatesopenres.14301.r28748

✓ **Ferdinando Fioretto**
Syracuse University, Syracuse, NY, USA

I am happy with the authors' response to my comments and with their new revised paper.

While I would have liked to see a more formal analysis and description of the method evaluated, I also understand the authors' desire to keep the article accessible to a wider audience.

To conclude, I believe that this article could be accepted without further revision.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial Intelligence, Differential Privacy, Optimization

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 03 March 2020

https://doi.org/10.21956/gatesopenres.14238.r28430

?

Ferdinando Fioretto
Syracuse University, Syracuse, NY, USA

Overview

The paper examines the behavior of TopDown, a privacy-preserving algorithm proposed to release differentially private US Census data. The authors examine the privacy, accuracy, and bias trade-off induced by the application of TopDown on the 1940 US Census dataset. The analysis was detailed for various privacy loss levels (i.e., epsilon values) and compared against a simple random sampling approach.

The authors provide a brief overview of Differential Privacy and the TopDown algorithm. Next, they introduce the empirical privacy loss as an empirical quantification of the loss of privacy induced by the application of a differentially private mechanism, and, finally, they provide an extensive evaluation on an application of TopDown on the 1940 US Census data release.

An interesting aspect of this work is the introduction of a novel evaluation metric, called "empirical privacy loss" or EPL. The authors argue that the use of the post-processing strategy adopted by TopDown, that projects the differentially private solution into a feasible space, may reduce the theoretical privacy loss and the experimental evaluation seem to support such claim. In particular, the authors found that the EPL for a given class of counts (total count and stratified count) is smaller than the theoretical privacy loss guaranteed by the algorithm. I have several comments about this metric, reported in the detailed comments section.

I found this work original, in that it provides an extensive evaluation of the privacy, accuracy, and bias trade-off of the Top-Down algorithm. However, I also found the absence of a related work section unusual and would like to point out that there are other works that use optimization techniques to publish accurate count statistics, e.g.:

- Michael Hay, Vibhor Rastogi, Gerome Miklau, Dan Suciu: Boosting the Accuracy of Differentially Private Histograms Through Consistency. PVLDB 3(1): 1021-1032 (2010)[1].
- Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, Vibhor Rastogi: The matrix mechanism: optimizing linear counting queries under differential privacy. VLDB J. 24(6): 757-781 (2015)[2].

and work that pose particular emphasis on Census data:

- Yu-Hsuan Kuo, Cho-Chun Chiu, Daniel Kifer, Michael Hay, Ashwin Machanavajjhala: Differentially Private Hierarchical Count-of-Counts Histograms. PVLDB 11(11): 1509-1521 (2018)[3].
- Ferdinando Fioretto, Pascal Van Hentenryck: Differential Privacy of Hierarchical Census Data: An Optimization Approach. CP 2019: 639-655[4].

It may be useful to discuss some of these proposals.

The paper is well organized and described with a good amount of detail. However, I would have liked to see a more formal description of the TopDown algorithm and of the empirical privacy loss concept. In particular, I believe that describing TopDown using an optimization model would greatly simplify readability and avoid some doubts, such as those I list in my detailed comments. I would also suggest the authors introduce an illustration of the hierarchy utilized by the Census, together with the amount of privacy budget used at each level. This could, for instance, be visualized as a tree, where the root node describes the total counts at the national level, its children describe counts at the state level, and so on. I believe that such an illustration will ease visualizing the process performed by TopDown during Step 2, in order to satisfy the consistency of the problem constraints.

It would also be useful to have a table summarizing the problem constraints. For example, the authors describe equalities constraints, such as those that constrain the aggregate statistics and counts as well as those that force the invariants, and inequality constraints, such as non-negativity and properties over the group sizes.

Detailed Comments:

Section: TopDown algorithm

- The authors provide a helpful overview of the TopDown algorithm, which operates in two steps: Noise addition and Optimization. I believe that the description can be further improved--I found the text to be quite verbose--and would encourage the authors to supply the following information:
    - A table that summarizes the attributes of the histograms to be produced (e.g., counts of each geographic by age, race, ethnicity, household/group quarters) and the aggregate statistics.
    - An illustration highlighting the dependence between counts, and, thus, the constraints arising from these dependencies.

I believe the above can be a helpful aid in the description of the algorithm.

- The authors call "aggregate statistics" as "DP queries". I am not sure why this terminology was selected. At the best of my knowledge, a DP query is simply a function over a dataset that happens to satisfy DP. I would suggest using a different terminology for identifying private aggregates.

- At the end of the third paragraph of **Step One: Imprecise Histogram:** I would have preferred to see a more formal description for the computation of the histogram count and aggregate statistics. For instance, in the current version, it is not clear what is the dimensionality of each query.

- In **Step two: Optimize:** the authors describe how TopDown optimizes the noisy estimates to satisfy the problem constraints. I would strongly suggest using a mathematical model to describe the problem (minimizer and constraints). In the current stage, a reader unfamiliar with the topic may found some sentences confusing. For example, the sentence "finds a solution that […] has the property that the value of each variable is as close as possible to the corresponding imprecise detailed histogram count or imprecise aggregate statistics" may denote that the objective is to minimize some Lp distance between the optimized counts and noisy ones; but for which p? I think that adding a formal model would improve the paper clarity.

Section: Empirical Privacy Loss

- I found the introduction of the empirical privacy loss concept quite interesting. However, I also have a few reservations. First, I think that the formula in this section could be described in more detail. I may have missed something, but I could not find what C correspond to. Also, this formula seems to be hard to compute and I wish the authors have spent a few words on they address such a challenge.

- The notation \hat{p}_k used in the formula \Pr[error …] seems to have the same semantic of notation \hat{p}(x), introduced in point (2) of Section "**Our evaluation approach**". Is this correct, i.e., is it that \hat{p}_k = \hat{p}(k)? If this is the case, then one of the two notations need to be changed for consistency.

- In section **TopDown Options still to be selected:**
  - On point (1): I suggest spacing the epsilon values listed;

  - On point (4): I wonder if the authors have some intuitions on why the test run used more budget for aggregated statistics than for aggregated queries. I believe it would be very insightful to discuss the implications of such budget partitioning.

- In section **Our evaluation approach**:
  - Point (2): I would have liked if the authors could have further elaborated on how the empirical privacy loss is computed. Is it the maximum among all x of ELP(x)?

  - The authors specify that the EPL is computed for the total count and they report a substantially lower loss than the theoretical privacy budget adopted. Since the privacy budget was partitioned among several levels and queries, I wonder if the authors have taken such partitioning into account when computing the final EPL score. I believe this aspect should be discussed in the text.

- Have the authors validated the fidelity of the EPL score on a simple differential privacy application? For instance, I would have liked to see a brief discussion on if this metric is in agreement with the theoretical errors provided by the Laplace mechanism on counting queries (without post-processing).

Results

Error and privacy of TopDown

- The authors explain in detail the results attained in their analysis. I found the reporting of the results at the end of each subsection to be a bit distracting. I suggest the authors introduce one or multiple tables that tabulate the results and only summarize them in the text.

- Additionally, the plots in Figure 1 and the errors describes in the text are for different privacy budget: The figure illustrates the errors for epsilon = 0.5, 1.0, and 2.0, while the text describes the errors for epsilon = 0.25, 1.0, and 4.0. I suggest the authors reporting the results for all the epsilon tested into a table, or to make the description in the text and the figure consistent for the privacy budgets adopted.

- The empirical privacy loss computed was reported for the total count at the enumeration district level and country-level and compared against the privacy budget adopted by the TopDown algorithm. As stated in my comment above, I wonder if this comparison is fair. TopDown seems to partition the privacy budget for different queries, thus leaving the total count queries with substantially less budget than the original total one. I encourage the author to expand on this aspect of the evaluation.

Comparison with error and privacy of simple random sampling

- As for the previous section, I recommend the authors to use a table to tabulate the numerical results described in the last paragraph. In my opinion, it will substantially increase readability.

Bias in the variation introduced by TopDown

- As for the previous section, I suggest the authors tabulate the results of the homogeneity index and bias.

- Are the errors by homogeneity index an average over the sample runs?

References

1. Hay M, Rastogi V, Miklau G, Suciu D: Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*. 2010; **3** (1-2): 1021-1032 Publisher Full Text
2. Li C, Miklau G, Hay M, McGregor A, et al.: The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB Journal*. 2015; **24** (6): 757-781 Publisher Full Text
3. Kuo Y, Chiu C, Kifer D, Hay M, et al.: Differentially private hierarchical count-of-counts histograms. *Proceedings of the VLDB Endowment*. 2018; **11** (11): 1509-1521 Publisher Full Text
4. Fioretto F, Van Hentenryck P: Differential Privacy of Hierarchical Census Data: An Optimization Approach. 2019; **11802**: 639-655 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial Intelligence, Differential Privacy, Optimization

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 20 Mar 2020

Abraham Flaxman, University of Washington, Seattle, Seattle, USA

Overview

The paper examines the behavior of TopDown, a privacy-preserving algorithm proposed to release differentially private US Census data. The authors examine the privacy, accuracy, and bias trade-off induced by the application of TopDown on the 1940 US Census dataset. The analysis was detailed for various privacy loss levels (i.e., epsilon values) and compared against a simple random sampling approach.

The authors provide a brief overview of Differential Privacy and the TopDown algorithm. Next, they introduce the empirical privacy loss as an empirical quantification of the loss of privacy induced by the application of a differentially private mechanism, and, finally, they provide an extensive evaluation on an application of TopDown on the 1940 US Census data release.

An interesting aspect of this work is the introduction of a novel evaluation metric, called "empirical privacy loss" or EPL. The authors argue that the use of the post-processing strategy adopted by TopDown, that projects the differentially private solution into a feasible space, may reduce the theoretical privacy loss and the experimental evaluation seem to support such claim. In particular, the authors found that the EPL for a given class of counts (total count and stratified count) is smaller than the theoretical privacy loss guaranteed by the algorithm. I have several comments about this metric, reported in the detailed comments section.

I found this work original, in that it provides an extensive evaluation of the privacy, accuracy, and bias trade-off of the Top-Down algorithm. However, I also found the absence of a related work section unusual and would like to point out that there are other works that use optimization techniques to publish accurate count statistics, e.g.:

- Michael Hay, Vibhor Rastogi, Gerome Miklau, Dan Suciu: Boosting the Accuracy of Differentially Private Histograms Through Consistency. PVLDB 3(1): 1021-1032 (2010)[1].
- Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, Vibhor Rastogi: The matrix mechanism: optimizing linear counting queries under differential privacy. VLDB J. 24(6): 757-781 (2015)[2].

and work that pose particular emphasis on Census data:

- Yu-Hsuan Kuo, Cho-Chun Chiu, Daniel Kifer, Michael Hay, Ashwin Machanavajjhala: Differentially Private Hierarchical Count-of-Counts Histograms. PVLDB 11(11): 1509-1521 (2018)[3].
- Ferdinando Fioretto, Pascal Van Hentenryck: Differential Privacy of Hierarchical Census Data: An Optimization Approach. CP 2019: 639-655[4].

It may be useful to discuss some of these proposals.

We thank the reviewer for this cogent assessment of our work, as well as for identifying these additional relevant references on the use of optimization techniques in producing differentially *private histograms. We have added references to these in our methods section, following the description of the TopDown algorithm.*

The paper is well organized and described with a good amount of detail. However, I would have liked to see a more formal description of the TopDown algorithm and of the empirical privacy loss concept. In particular, I believe that describing TopDown using an optimization model would greatly simplify readability and avoid some doubts, such as those I list in my detailed comments. I would also suggest the authors introduce an illustration of the hierarchy utilized by the Census, together with the amount of privacy budget used at each level. This could, for instance, be visualized as a tree, where the root node describes the total counts at the national level, its children describe counts at the state level, and so on. I believe that such an illustration will ease visualizing the process performed by TopDown during Step 2, in order to satisfy the consistency of the problem constraints.
It would also be useful to have a table summarizing the problem constraints. For example, the authors describe equalities constraints, such as those that constrain the aggregate statistics and counts as well as those that force the invariants, and inequality constraints, such as non-negativity and properties over the group sizes.

We appreciate the reviewer's suggestion that we include more precise technical details about the optimization model and the geographic hierarchy use by Census, but we would prefer not to over-burden our anticipated readership with this level of detail. Our goal is to give a detailed overview, but only an overview, not a definitive source of the details of TopDown. The Census Bureau has released their own technical documentation of their system (e.g. reference [9] of our text) as well as their source code, and readers who want to get the full details from the definitive source should go to the papers and code from Census for this. Also, the specifics of the geographic hierarchy and the allocation of epsilon between its levels is still work-in-progress at Census Bureau, and while we know precisely what they did for the 2018 end-to-end test, we do expect this to change as they finish the implementation and parameter selection process for the analyzing the 2020 decennial census.

Detailed Comments:

Section: TopDown algorithm

- The authors provide a helpful overview of the TopDown algorithm, which operates in two steps: Noise addition and Optimization. I believe that the description can be further improved--I found the text to be quite verbose--and would encourage the authors to supply the following information:
    - A table that summarizes the attributes of the histograms to be produced (e.g., counts of each geographic by age, race, ethnicity, household/group quarters) and the aggregate statistics.
    - An illustration highlighting the dependence between counts, and, thus, the constraints arising from these dependencies.

I believe the above can be a helpful aid in the description of the algorithm.

As we responded above, we appreciate the reviewer's suggestion to include more precise technical details about the detailed histogram and the aggregate statistics, and the dependencies between the counts, but as we responded above, we believe that these are all still work-in-progress for the 2020 census, and do not want to draw too much attention to the test values that were used in the 2018 end-to-end test.

- The authors call "aggregate statistics" as "DP queries". I am not sure why this terminology was selected. At the best of my knowledge, a DP query is simply a function over a dataset that happens to satisfy DP. I would suggest using a different terminology for identifying private aggregates.

We initially planned to use the term “aggregate statistic” but decided that since the Census Bureau has used the term “DP query” internally, it would be helpful to give our readers this jargon in case they wanted to advocate for allocating privacy budget to specific DP queries that are relevant to their uses of census data. We have edited the text where this jargon is first introduced to better identify it as Census Bureau terminology.

- At the end of the third paragraph of **Step One: Imprecise Histogram:** I would have preferred to see a more formal description for the computation of the histogram count and aggregate statistics. For instance, in the current version, it is not clear what is the dimensionality of each query.

We appreciate the reviewer’s desire for more formal descriptions and more precision in the methods section of this paper, and we regret if our responses are beginning to become repetitive, but we prefer to keep things from becoming overly *formal, as our goal is to reach an audience that has not yet been reached by the more “computer science-y” publications from the Census Bureau that have described their approach. To the reviewer’s example of the dimensionality of each query, there are 2*6*2*6 = 144 counts for each geographic area (the counts are stratified by age (two values: under-18-year-olds and 18-plus), race (six values), ethnicity (two values: Hispanic and non-Hispanic), and household/group-quarters type (6 values)).*

- In **Step two: Optimize:** the authors describe how TopDown optimizes the noisy estimates to satisfy the problem constraints. I would strongly suggest using a mathematical model to describe the problem (minimizer and constraints). In the current stage, a reader unfamiliar with the topic may found some sentences confusing. For example, the sentence “finds a solution that […] has the property that the value of each variable is as close as possible to the corresponding imprecise detailed histogram count or imprecise aggregate statistics” may denote that the objective is to minimize some Lp distance between the optimized counts and noisy ones; but for which p? I think that adding a formal model would improve the paper clarity.

We have adjusted this language to make it clearer that this is a quadratic objective function, i.e. Lp for p = 2. As above, we want to avoid writing out the constrained convex optimization problem in mathematical notation, because we anticipate a substantial portion of our target readership will prefer not *to see it in this form, although we appreciate that there are also those, like the reviewer, who* will *want to see it in this form.*

Section: Empirical Privacy Loss

- I found the introduction of the empirical privacy loss concept quite interesting. However, I also have a few reservations. First, I think that the formula in this section could be described in more detail. I may have missed something, but I could not find what C correspond to. Also, this formula seems to be hard to compute and I wish the authors have spent a few words on they address such a challenge.

We have added supplementary material that provides substantial additional detail on how EPL is defined and how it works.

- The notation \hat{p}_k used in the formula \Pr[error …] seems to have the same semantic of notation \hat{p}(x), introduced in point (2) of Section “**Our evaluation approach**”. Is this correct, i.e., is it that \hat{p}_k = \hat{p}(k)? If this is the case, then one of the two notations need to be changed for consistency.

We have improved this notation to accompany our improved exposition of EPL in the main text and supplementary material.

- In section **TopDown Options still to be selected:**
  - On point (1): I suggest spacing the epsilon values listed;

Good idea, done.

- On point (4): I wonder if the authors have some intuitions on why the test run used more budget for aggregated statistics than for aggregated queries. I believe it would be very insightful to discuss the implications of such budget partitioning.

There is no source of information we know of that justifies the choice Census Bureau made for splitting the privacy budget between the detailed queries and DP queries. We agree that an improved understanding of the tradeoffs between different budgets is an important direction for future research, but we feel that it is beyond the scope of the present paper.

- In section **Our evaluation approach**:
  - Point (2): I would have liked if the authors could have further elaborated on how the empirical privacy loss is computed. Is it the maximum among all x of ELP(x)?

We apologize for the lack of detail in our paper and have added more precision in the main text as well as supplementary material with more detail. The scalar EPL *is the maximum of the absolute value of* EPL(*x*) over all *x.*

- The authors specify that the EPL is computed for the total count and they report a substantially lower loss than the theoretical privacy budget adopted. Since the privacy budget was partitioned among several levels and queries, I wonder if the authors have taken such partitioning into account when computing the final EPL score. I believe this aspect should be discussed in the text.

We intend the EPL to be compared with the overall epsilon, and therefore not to require any consideration of how the privacy budget was partitioned. Indeed, we want to be able to compute EPL for algorithms that are not *provably differentially private, or for algorithms that use many levels or no levels. Perhaps we are misunderstanding the reviewer's question/suggestion.*

- Have the authors validated the fidelity of the EPL score on a simple differential privacy application? For instance, I would have liked to see a brief discussion on if this metric is in agreement with the theoretical errors provided by the Laplace mechanism on counting queries (without post-processing).

We hope that the first example from the new supplementary material will address exactly this request, by calculating the EPL of a Geometric Mechanism with epsilon=0.25. Our Supplementary Appendix Table SA1 also includes a table of EPL values for the Geometric Mechanism a range of epsilon values and random seeds, and provides some evidence that EPL works in this setting:

(a) epsilon value in Geometric DP EPL mean EPL Lower Bound (2.5th Percentile) EPL Upper Bound (97.5th Percentile)

eps	EPL	lower	upper
0.0010	0.0010	0.0008	0.0013
0.0050	0.0048	0.0039	0.0068
0.0100	0.0099	0.0076	0.0130
0.0500	0.0490	0.0390	0.0673
0.1000	0.0980	0.0752	0.1262

0.1500 0.1475 0.1181 0.1941
0.2000 0.1988 0.1521 0.2639

Results

Error and privacy of TopDown

- The authors explain in detail the results attained in their analysis. I found the reporting of the results at the end of each subsection to be a bit distracting. I suggest the authors introduce one or multiple tables that tabulate the results and only summarize them in the text.

We appreciate this suggestion, but we also want to reach audiences that prefer figures and text, as well as those who like tables. We do have a table of all results as supplementary material, and we have added an additional table with results for the SRS comparisons.

- Additionally, the plots in Figure 1 and the errors describes in the text are for different privacy budget: The figure illustrates the errors for epsilon = 0.5, 1.0, and 2.0, while the text describes the errors for epsilon = 0.25, 1.0, and 4.0. I suggest the authors reporting the results for all the epsilon tested into a table, or to make the description in the text and the figure consistent for the privacy budgets adopted.

Thank you for this suggestion. We have updated the text to refer to epsilon values that match the figures. As mentioned above, we also have included values for all epsilons in our supplementary materials.

- The empirical privacy loss computed was reported for the total count at the enumeration district level and country-level and compared against the privacy budget adopted by the TopDown algorithm. As stated in my comment above, I wonder if this comparison is fair. TopDown seems to partition the privacy budget for different queries, thus leaving the total count queries with substantially less budget than the original total one. I encourage the author to expand on this aspect of the evaluation.

We prioritized total count in our figures because of its importance and simplicity, but in our supplementary tables, we also include EPL for counts stratified by age group/race/ethnicity. In the 2018 end-to-end test, there is substantially less stratification than there will be in the 2020 decennial census, and this will be an important area for further research with more recent data.

Comparison with error and privacy of simple random sampling

- As for the previous section, I recommend the authors to use a table to tabulate the numerical results described in the last paragraph. In my opinion, it will substantially increase readability.

We appreciate this suggestion, but we also want to reach audiences that prefer figures and text, as well as those who like tables. We do have a table of all results as supplementary material, and we have added an additional table with results for the SRS comparisons.

Bias in the variation introduced by TopDown

- As for the previous section, I suggest the authors tabulate the results of the homogeneity index and bias.

We don't have a table for this, and we are hesitant to put one in. We think that the figure provides a clearer communication of the pattern (mean residual goes up and to the right as a function of homogeneity index, but the change is attenuated as epsilon gets larger) and the precise numbers are available in the text for any reader who wants this level of precision. Although we greatly appreciate the reviewer's work in commenting on our paper, we have come to see that we just do not value tables of results the way the reviewer does. We hope that this difference in presentation styles does not stand in the way of a producing a strong and scientifically valid paper.

- Are the errors by homogeneity index an average over the sample runs?

The distributions are for errors pooled over the four random seeds (for each epsilon/homogeneity index), and the markers are the average bias across all four runs. We have added to the caption of the figure to make this clear.

- Is the work clearly and accurately presented and does it cite the current literature?

Partly

- Is the study design appropriate and is the work technically sound?

Partly

- Are sufficient details of methods and analysis provided to allow replication by others?

Yes

- If applicable, is the statistical analysis and its interpretation appropriate?

Partly

- Are all the source data underlying the results available to ensure full reproducibility?

Yes

- Are the conclusions drawn adequately supported by the results?

Partly

Thank you for your work reviewing this paper!

Competing Interests: ADF has consulted recently for Kaiser Permanente; Sanofi; Merck for Mothers; Agathos, Ltd; and NORC. SP has no competing interests to disclose.

Reviewer Report 20 December 2019

https://doi.org/10.21956/gatesopenres.14238.r28334

?

David Van Riper
IPUMS, University of Minnesota, Minneapolis, MN, USA

Overview

Using differentially private 1940 census data produced by the US Census Bureau's TopDown algorithm, Petti and Flaxman assess the privacy/accuracy trade-off along multiple dimensions for this algorithm for multiple values of epsilon. The authors analyzed the median absolute error, empirical privacy loss, and bias for the differentially private data. They also compared the median absolute error and empirical privacy loss for differentially private data with data generated through simple random sampling. This is one of the first, if not the first, article assessing the accuracy of decennial census data published through a differentially private algorithm.

Petti and Flaxman provide a good overview of differential privacy and the Census Bureau's TopDown algorithm - a differentially private algorithm for producing decennial census data. They then compare the differentially private 1940 data with the original complete-count 1940 data to assess the accuracy introduce by the TopDown algorithm. They find that error increased as the total privacy loss budget decreased. They also find that empirical privacy loss was smaller than total privacy loss budget. They measure bias introduced by the algorithm and find that bias increases as homogeneity decreases and that bias increases as total privacy loss budget decreases. They conclude that privacy loss does not vary much for epsilon < 1.0, and that the accuracy achieved when using a 50% simple random sample is equivalent to an epsilon of 1.0.

I am intrigued by the empirical privacy loss measure introduced by Petti and Flaxman. Its formula and interpretation mirrors the formula for epsilon-differential privacy. However, I would like to see a more thorough discussion of empirical privacy loss summary statistic reported in the results section of the paper. The authors compare an empirical privacy loss summary statistic with total privacy loss budget on pages 6 and 7 of the paper, but they never explain how the summary statistic was computed. Having that explanation would help me better understand the comparison they make throughout the paper.

The authors compare the empirical privacy loss for a given geographic unit-type of count (total count, stratified count) combination with the overall privacy loss budget. They empirical privacy loss for a given combination is less than the overall privacy loss budget. I wonder if this is the correct comparison to make. The privacy loss budget controls the overall amount of privacy leaked by the publication of all statistics. It is the sum, via sequential composition, of the epsilon fractions assigned to each geographic level-statistic combination. Thus, by definition, the empirical privacy loss associated with a particular geographic level-statistic (e.g., total population count) must be less than the privacy loss budget. I would like to see a fuller discussion of this comparison in the paper. See detailed comment #14 for more details.

I would also additional supplemental datasets (or tables in the paper) with the empirical privacy loss summary statistics for all values of epsilon. The authors report a few values in the text and figures, but having a complete set would allow for a more comprehensive understanding of the relationship between empirical privacy loss and epsilon.

Finally, I strongly recommend that the authors use the same examples in their text as they use in the figures. The text uses epsilons of 0.25, 1.0 and 4.0 and the figures use epsilons of 0.50, 1.0, and 2.0. Making the epsilons consistent between the text and figures will help the reader better understand the analysis.

Detailed comments by section

Methods - TopDown algorithm

- The authors' high level overview (first paragraph in subsection entitled "TopDown algorithm") describe the noise injection (Imprecise Histogram) and optimization steps in the TopDown algorithm. They state that the "second step (Optimize) adjusts the histogram to be close as possible to the imprecise counts". I am uncertain about what histogram the authors refer to in this sentence. Is the histogram based on the original data, or is this the noise-injected detailed histogram? My understanding of the algorithm is that is generates histograms (one for each combination of geographic level and query) from the original data and then injects noise into histograms using the appropriate two-sided geometric distribution. It then passes these noise-injected histograms to the optimization function.

I would like the authors to be more precise in their description of the histogram and the "imprecise counts" in this section.

- The authors state that the 2020 US Census will have six geographic levels nested hierarchically (last sentence of TopDown algorithm paragraph). The Census Bureau allocated privacy loss budget to seven nested geographies (nation, state, county, tract group, census tract, block group, block) for the 2010 demonstration product. The Bureau has not committed to this allocation for

2020 and could still change the allocation strategy. I recommend clarifying that statement to pertain solely to the 2010 demonstration data product.

- In the final clause of the last sentence of the TopDown algorighm paragraph, the authors state that "in the 1940 E2E test, only national, state, county, and district levels were included." I recommend adding the word "enumeration" before district in that clause.

Methods - Step one: Imprecise Histogram

- At the end of first paragraph in this section, the authors describe the "ethnicity-age" aggregate statistic set. The implication of this sentence is that the "ethnicity-age" aggregate statistics set was one pre-selected by Census for noise injection. Census did not choose this aggregate statistic set. The aggregate statistic sets chosen by census were Voting age by Hispanic origin by Race (a 2 x 2 x 6 cell query) and Household/Group quarter (a 6 cell query). I recommend modifying this sentence to describe one of the two pre-selected aggregate statistic sets.

- At the end of the second paragraph, the authors write that "22.5% spent on the group-quarters queries". I recommend changing the fragment to be "22.5% spent on the household/group-quarters queries". The word "household" is important when discussing this DP query. People can either live in household or group quarters, and by definition, households are not group quarters.

Methods - TopDown options still to be selected

- For option 3, I recommend modifying the "(and therefore aggregated over "group quarters types)" to be "(there therefore aggregated over "household/group quarters types)". A household is not a type of group quarter.

- Also in option 3, I recommend modifying the "(ii) the group-quarters counts" to be "(ii) the household/group quarters counts".

- In option 5, add the word "population" between "total" and "count" in the second sentence. Otherwise, readers will not necessarily know which total count to which the authors are referring.

Results - Error and privacy of TopDown

- At the end of first paragraph of this subsection, the authors list the median and 95th percentile of TC for EDs, counties, and states. I think it is important to clarify that these counts are based on the original 1940 census data and not on any of the differentially private 1940 datasets. Since this sentence comes at the end of a paragraph describing median absolute error, readers may assume the medians and 95th percentiles are from a DP dataset. Consider moving that sentence up the start of the paragraph.

- At the end of second paragraph of this subsection, the authors list the median and 95th percentile of SC for EDs, counties, and states. I think it is important to clarify that these counts are based on the original 1940 census data and not on any of the differentially private 1940 datasets. Since this sentence comes at the end of a paragraph describing median absolute error, readers may assume the medians and 95th percentiles are from a DP dataset. Consider moving that sentence up the start of the paragraph.

- The final two paragraphs of this subsection describe the empirical privacy loss for TC and SC for different geographic levels and different epsilons. They describe the EPL for epsilons of 0.25, 1.0, and 4.0 in the text. I would like to have a table, either in the paper or in the extended data product, that lists the EPLs for all values of epsilon and all geographic levels for TC and SC. I wonder how linear the relationship between EPL and epsilon is.

- The authors list a number of EPL values in the final two paragraphs and in the right-hand panel of Figure 1, but I do not know what the EPL value represents. Is it the absolute value of the maximum observed EPL, or is it the range from the maximum to minimum observed EPL value? I would appreciate a more complete discussion of how the authors calculated the value of EPL they plot in Figure 1 and list in the text. The formula on page 5 describes how to compute EPL for a single geographic unit and value of epsilon, but I don't see how that formula extends to the summary statistics reported on page 6.

- Figure 1 plots the error and EPL for epsilon equal to 0.5, 1.0, and 2.0, but the text in the final two paragraphs describes EPL for epsilons of 0.25, 1.0, and 4.0. I strongly recommend making the values in the text and the plot consistent with one another. That consistency will make it easier to interpret the plot in Figure 1.

- The authors compare the empirical privacy loss for a given geographic unit-type of count (total count, stratified count) combination with the overall privacy loss budget. They empirical privacy loss for a given combination is less than the overall privacy loss budget. I wonder if this is the correct comparison to make. The privacy loss budget controls the overall amount of privacy leaked by the publication of all statistics. It is the sum, via sequential composition, of the epsilon fractions assigned to each geographic level-statistic combination. Thus, by definition, the empirical privacy loss associated with a particular geographic level-statistic (e.g., total population count) must be less than the privacy loss budget.

For a given value of epsilon, we can compute the portion of that value that is assigned to each geographic level - query combination. For example, epsilon of 0.25 is divided up as follows:

Geographic levels = 0.25 to each level
Tables = 0.1 (detailed), 0.225 (household-group quarters), 0.675 (voting age - Hispanic - race)

We can multiply the geographic level fraction by the table fractions by epsilon to yield:

Geog level - detailed query = 0.00625 epsilon
Geog level - household group quarters query = 0.0140625 epsilon
Geog level - voting age - Hispanic - race query = 0.0421875 epsilon

These epsilons still do not equate to an epsilon associated with a particular statistic, such as total population count. Given the optimization step and the state-level total population invariant, I'm not sure if we can compute an epsilon value for a particular statistic. But these epsilon values seems like a more appropriate comparison to the empirical privacy loss reported by the authors.

Results - Comparison with error and privacy of simple random sampling

- I would like to have a table of MAE and EPL values for Simple Random Sampling. Consider adding those values to the Extended Data product currently available, or adding another Extended Data product with these values.

- Consider adding a plot of EPL by sample size to supplement or even replace the final paragraph of this subsection. There are a lot of numbers in the final paragraph, and I find it difficult to visualize the relationship between EPL and sampling fraction just by reading the numbers.

- The x-axis for Figure 2 depicts values of Empirical Privacy Loss, but neither the text nor the caption describe how the values were computed. This comment fits with comment 12 - what does the Empirical Privacy Loss summary statistic mean and how was it computed.

Results - Bias in the variation introduced by TopDown

- Figure 3 plots the error and EPL for epsilon equal to 0.5, 1.0, and 2.0, but the text in the first paragraph describes EPL by homogeneity index for epsilons of 0.25, 1.0, and 4.0. I strongly recommend making the values in the text and the plot consistent with one another. That consistency will make it easier to interpret the plot in Figure 3.

- I recommend moving the (Figure 3) parenthetical to the end of the discussion on EPL by homogeneity for enumeration districts. Figure 3 only shows the results for enumeration districts, but the parenthetical comes after the discussion for counties.

- In the paragraph and Figure 3, the authors list a summary statistic for bias by homogeneity index and epsilon. Is the summary statistic the mean or the median?

- Figure 3 displays the violin plot/mean bias for 11 of 23 homogeneity index values. I recommend modifying the figure caption to indicate that the authors are only displaying some of the homogeneity index values on the plot.

- I also recommend modifying the x-axis label to indicate that the homogeneity index values are for enumeration districts. That would help readers immediately understand what geographic units are being plotted.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: geography, demography, census data, differential privacy

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 20 Mar 2020

Abraham Flaxman, University of Washington, Seattle, Seattle, USA

Reviewer 1 Feedback from David Van Riper
Overview
Using differentially private 1940 census data produced by the US Census Bureau's TopDown algorithm, Petti and Flaxman assess the privacy/accuracy trade-off along multiple dimensions for this algorithm for multiple values of epsilon. The authors analyzed the median absolute error, empirical privacy loss, and bias for the differentially private data. They also compared the median absolute error and empirical privacy loss for differentially private data with data generated through simple random sampling. This is one of the first, if not the first, article assessing the accuracy of decennial census data published through a differentially private algorithm.

Petti and Flaxman provide a good overview of differential privacy and the Census Bureau's TopDown algorithm - a differentially private algorithm for producing decennial census data. They then compare the differentially private 1940 data with the original complete-count 1940 data to assess the accuracy introduce by the TopDown algorithm. They find that error increased as the total privacy loss budget decreased. They also find that empirical privacy loss was smaller than total privacy loss budget. They measure bias introduced by the algorithm and find that bias increases as homogeneity decreases and that bias increases as total privacy loss budget decreases. They conclude that privacy loss does not vary much for epsilon < 1.0, and that the accuracy achieved when using a 50% simple random sample is equivalent to an epsilon of 1.0.

I am intrigued by the empirical privacy loss measure introduced by Petti and Flaxman. Its formula and interpretation mirrors the formula for epsilon-differential privacy. However, I would like to see a more thorough discussion of empirical privacy loss summary statistic reported in the results section of the paper. The authors compare an empirical privacy loss summary statistic with total privacy loss budget on pages 6 and 7 of the paper, but they never explain how the summary statistic was computed. Having that explanation would help me better understand the comparison they make throughout the paper.

Thank you for your interest in this construct. We have added a substantially more detailed exposition of EPL in a Supplementary Methods Appendix.

The authors compare the empirical privacy loss for a given geographic unit-type of count (total count, stratified count) combination with the overall privacy loss budget. The empirical privacy loss for a given combination is less than the overall privacy loss budget. I wonder if this is the correct comparison to make. The privacy loss budget controls the overall amount of privacy leaked by the publication of all statistics. It is the sum, via sequential composition, of the epsilon fractions assigned to each geographic level-statistic combination. Thus, by definition, the empirical privacy loss associated with a particular geographic level-statistic (e.g., total population count) must be less than the privacy loss budget. I would like to see a fuller discussion of this comparison in the paper. See detailed comment #14 for more details.

In any epsilon-DP algorithm, we expect that the EPL that we have defined will be less than the privacy loss budget overall. However, there is not a formal guarantee that the privacy loss at a geographic level of TopDown will be less than the overall epsilon (due to invariants), and we believe that it is useful to see that EPL suggests that in practice the privacy loss at non-invariant *levels is lower than the overall epsilon. Even for a DP version, such as TopDown with no invariants, although it is logical to suspect that the inequality from sequential composition has slack, this is also not guaranteed. We have tried to compare some of these claims in simplified settings in the supplementary material we have added about EPL.*

I would also additional supplemental datasets (or tables in the paper) with the empirical privacy loss summary statistics for all values of epsilon. The authors report a few values in the text and figures, but having a complete set would allow for a more comprehensive understanding of the relationship between empirical privacy loss and epsilon.

We have added an EPL column to the table in the supplementary appendix.

Finally, I strongly recommend that the authors use the same examples in their text as they use in the figures. The text uses epsilons of 0.25, 1.0 and 4.0 and the figures use epsilons of 0.50, 1.0, and 2.0. Making the epsilons consistent between the text and figures will help the reader better understand the analysis.

Thank you for this feedback. We have harmonized the text and figures to all use the epsilon values of 0.5, 1.0, and 2.0, because these provide a clearer visual comparison and epsilon 2.0 produces a noise level in finest-grained detailed queries most similar to that used in the census tract detailed queries of the 2010 demonstration product.

Detailed comments by section

Methods - TopDown algorithm

- The authors' high level overview (first paragraph in subsection entitled "TopDown algorithm") describe the noise injection (Imprecise Histogram) and optimization steps in the TopDown algorithm. They state that the "second step (Optimize) adjusts the histogram to be close as possible to the imprecise counts". I am uncertain about what histogram the authors refer to in this sentence. Is the histogram based on the original data, or is this the noise-injected detailed histogram? My understanding of the algorithm is that is generates histograms (one for each combination of geographic level and query) from the original data and then injects noise into histograms using the appropriate two-sided geometric distribution. It then passes these noise-injected histograms to the optimization function.

  I would like the authors to be more precise in their description of the histogram and the "imprecise counts" in this section.

The reviewer’s understanding matches with ours, and we have edited the phrasing in this paragraph to try to make this clearer.

- The authors state that the 2020 US Census will have six geographic levels nested hierarchically (last sentence of TopDown algorithm paragraph). The Census Bureau allocated privacy loss budget to seven nested geographies (nation, state, county, tract group, census tract, block group, block) for the 2010 demonstration product. The Bureau has not committed to this allocation for 2020 and could still change the allocation strategy. I recommend clarifying that statement to pertain solely to the 2010 demonstration data product.

We have backed off this claim, by changing “will” to “may”. At this point, it seems possible that the hierarchy will even be substantially different than what was used in the public demonstrations so far.

- In the final clause of the last sentence of the TopDown algorithm paragraph, the authors state that "in the 1940 E2E test, only national, state, county, and district levels were included." I recommend adding the word "enumeration" before district in that clause.

Added.

Methods - Step one: Imprecise Histogram

- At the end of first paragraph in this section, the authors describe the "ethnicity-age" aggregate statistic set. The implication of this sentence is that the "ethnicity-age" aggregate statistics set was one pre-selected by Census for noise injection. Census did not choose this aggregate statistic set. The aggregate statistic sets chosen by census were Voting age by Hispanic origin by Race (a 2 x 2 x 6 cell query) and Household/Group quarter (a 6 cell query). I recommend modifying this sentence to describe one of the two pre-selected aggregate statistic sets.

Good idea.

- At the end of the second paragraph, the authors write that "22.5% spent on the group-quarters queries". I recommend changing the fragment to be "22.5% spent on the household/group-quarters queries". The word "household" is important when discussing this DP query. People can either live in household or group quarters, and by definition, households are not group quarters.

Good point. We have also propagated this change through the rest of the document.

Methods - TopDown options still to be selected

- For option 3, I recommend modifying the "(and therefore aggregated over "group quarters types)" to be "(there therefore aggregated over "household/group quarters types)". A household is not a type of group quarter.

Agree.

- Also in option 3, I recommend modifying the "(ii) the group-quarters counts" to be "(ii) the household/group quarters counts".

Agree.

- In option 5, add the word "population" between "total" and "count" in the second sentence. Otherwise, readers will not necessarily know which total count to which the authors are referring.

Agree, good point.

Results - Error and privacy of TopDown

- At the end of first paragraph of this subsection, the authors list the median and 95th percentile of TC for EDs, counties, and states. I think it is important to clarify that these counts are based on the original 1940 census data and not on any of the differentially private 1940 datasets. Since this sentence comes at the end of a paragraph describing median absolute error, readers may assume the medians and 95th percentiles are from a DP dataset. Consider moving that sentence up the start of the paragraph.

We have attempted to clarify this without restructuring the paragraph, and also added a similar clarification to the methods section.

- At the end of second paragraph of this subsection, the authors list the median and 95th percentile of SC for EDs, counties, and states. I think it is important to clarify that these counts are based on the original 1940 census data and not on any of the differentially private 1940 datasets. Since this sentence comes at the end of a paragraph describing median absolute error, readers may assume the medians and 95th percentiles are from a DP dataset. Consider moving that sentence up the start of the paragraph.

We have attempted to clarify this without restructuring the paragraph, and also added a similar clarification to the methods section.

- The final two paragraphs of this subsection describe the empirical privacy loss for TC and SC for different geographic levels and different epsilons. They describe the EPL for epsilons of 0.25, 1.0, and 4.0 in the text. I would like to have a table, either in the paper or in the extended data product, that lists the EPLs for all values of epsilon and all geographic levels for TC and SC. I wonder how linear the relationship between EPL and epsilon is.

Good idea. There is a table of extended data already, but we failed to include a column of EPL data in it! We have added this column, and since your wondering got us wondering, we examined this relationship and found that it goes relatively linearly up-and-to-the-right. Unfortunately, we cannot figure out how to include this figure in our response.

- The authors list a number of EPL values in the final two paragraphs and in the right-hand panel of Figure 1, but I do not know what the EPL value represents. Is it the absolute value of the maximum observed EPL, or is it the range from the maximum to minimum observed EPL value? I would appreciate a more complete discussion of how the authors calculated the value of EPL they plot in Figure 1 and list in the text. The formula on page 5 describes how to compute EPL for a single geographic unit and value of epsilon, but I don't see how that formula extends to the summary statistics reported on page 6.

We have added supplementary material with a more detailed exposition of the EPL value, as well as including a short, but more precise definition in the methods section. However, some busy readers are sure to look first at the figures, so we have also attempted to clarify here that the EPL value is the maximum of the absolute value of the log of p^KDE(x) / p^KDE(x+1) over all x.

- Figure 1 plots the error and EPL for epsilon equal to 0.5, 1.0, and 2.0, but the text in the final two paragraphs describes EPL for epsilons of 0.25, 1.0, and 4.0. I strongly recommend making the values in the text and the plot consistent with one another. That consistency will make it easier to interpret the plot in Figure 1.

Thank you for this feedback. We have harmonized the text and figures to all use the epsilon values of 0.5, 1.0, and 2.0, because these provide a clearer visual comparison and epsilon 2.0 produces a noise level in finest-grained detailed queries most similar to that used in the census tract detailed queries of the 2010 demonstration product.

- The authors compare the empirical privacy loss for a given geographic unit-type of count (total count, stratified count) combination with the overall privacy loss budget. They empirical privacy loss for a given combination is less than the overall privacy loss budget. I wonder if this is the correct comparison to make. The privacy loss budget controls the overall amount of privacy leaked by the publication of all statistics. It is the sum, via sequential composition, of the epsilon fractions assigned to each geographic level-statistic combination. Thus, by definition, the empirical privacy loss associated with a particular geographic level-statistic (e.g., total population count) must be less than the privacy loss budget.

Although this assertion is true in traditional epsilon-DP, as discussed above, TopDown complicates it by introducing invariants (so, for example, the EPL of total population count at the state level is infinity). At county or enum_dist level, is EPL less than epsilon because the levels actually combine to compromise privacy? Or is it less because the deductive proof that the algorithm is epsilon-DP relies on application of the "sequential composition theorem" which is an inequality that might include substantial slack? We have not modified the main text to attempt to elaborate on this, but we have added supplementary material with simplified examples of when EPL is very close to epsilon; when it is less, due to slack; and when it is more, due to invariants.

- For a given value of epsilon, we can compute the portion of that value that is assigned to each geographic level - query combination. For example, epsilon of 0.25 is divided up as follows:

  Geographic levels = 0.25 to each level
  Tables = 0.1 (detailed), 0.225 (household-group quarters), 0.675 (voting age - Hispanic - race)

  We can multiply the geographic level fraction by the table fractions by epsilon to yield:

  Geog level - detailed query = 0.00625 epsilon
  Geog level - household group quarters query = 0.0140625 epsilon
  Geog level - voting age - Hispanic - race query = 0.0421875 epsilon

  These epsilons still do not equate to an epsilon associated with a particular statistic, such as total population count. Given the optimization step and the state-level total population invariant, I'm not sure if we can compute an epsilon value for a particular statistic. But these epsilon values seems like a more appropriate comparison to the empirical privacy loss reported by the authors.

Thank you for this useful exposition, and we considered pivoting to these sized values in the supplementary material that explores how our EPL is constructed. Were we to focus only on the Imprecise Histogram portion of TopDown and only consider the detailed queries at a single geographic level, we would find an EPL that precisely matches the 0.00625 epsilon you have calculated. We believe that the value of the EPL construct is to provide an alternative to the epsilon value deduced with formal logic in cases like TopDown, where the sequential composition and invariants together might yield an empirical privacy loss that is smaller or *larger than the proven epsilon. We eventually decided to focus on a somewhat larger epsilon in the supplementary material because we thought that this made the exposition and visuals was clearer.*

Results - Comparison with error and privacy of simple random sampling

- I would like to have a table of MAE and EPL values for Simple Random Sampling. Consider adding those values to the Extended Data product currently available, or adding another Extended Data product with these values.

Good idea. We have added another supplementary table with these values.

- Consider adding a plot of EPL by sample size to supplement or even replace the final paragraph of this subsection. There are a lot of numbers in the final paragraph, and I find it difficult to visualize the relationship between EPL and sampling fraction just by reading the numbers.

We are hesitant to add this plot, because we want to focus the reader's attention on the comparison of SRS EPL and TopDown EPL. Is including the supplementary table from the point above sufficient to address this request? Then the interested reader will have all the numbers necessary to make this plot for themselves. It looks like this:

- The x-axis for Figure 2 depicts values of Empirical Privacy Loss, but neither the text nor the caption describe how the values were computed. This comment fits with comment 12 - what does the Empirical Privacy Loss summary statistic mean and how was it computed.

We have added supplementary material with a more detailed exposition of the EPL value, as well as including a short, but more precise definition in the methods section, and adding text to the caption of Figure 1, so we hope that now it is clearer what we have computed here.

Results - Bias in the variation introduced by TopDown

- Figure 3 plots the error and EPL for epsilon equal to 0.5, 1.0, and 2.0, but the text in the first paragraph describes EPL by homogeneity index for epsilons of 0.25, 1.0, and 4.0. I strongly recommend making the values in the text and the plot consistent with one another. That consistency will make it easier to interpret the plot in Figure 3.

We have changed the epsilon values described in the text to match the figure, as above.

- I recommend moving the (Figure 3) parenthetical to the end of the discussion on EPL by homogeneity for enumeration districts. Figure 3 only shows the results for enumeration districts, but the parenthetical comes after the discussion for counties.

Good idea, we have done this.

- In the paragraph and Figure 3, the authors list a summary statistic for bias by homogeneity index and epsilon. Is the summary statistic the mean or the median?

It is the arithmetic mean. We have added to the text in the methods section and figure caption to make this clearer.

- Figure 3 displays the violin plot/mean bias for 11 of 23 homogeneity index values. I recommend modifying the figure caption to indicate that the authors are only displaying some of the homogeneity index values on the plot.

We have added to the caption to better explain this potentially confusing point---we have not suppressed any homogeneity index values, rather there are no enumeration districts with value zero in precisely 19 cells of the detailed histogram.

- I also recommend modifying the x-axis label to indicate that the homogeneity index values are for enumeration districts. That would help readers immediately understand what geographic units are being plotted.

Good idea, we have done this.

- Is the work clearly and accurately presented and does it cite the current literature?

Yes

- Is the study design appropriate and is the work technically sound?

Yes

- Are sufficient details of methods and analysis provided to allow replication by others?

Yes

- If applicable, is the statistical analysis and its interpretation appropriate?

Partly

- Are all the source data underlying the results available to ensure full reproducibility?

Yes

- Are the conclusions drawn adequately supported by the results?

Partly

Thank you for your work reviewing this paper!

Competing Interests: ADF has consulted recently for Kaiser Permanente; Sanofi; Merck for Mothers; Agathos, Ltd; and NORC. SP has no competing interests to disclose.