# Differential privacy in the 2020 US census, what will it do? Quantifying the accuracy/privacy tradeoff

Samantha Petti and Abraham D. Flaxman, 2019-10-07

# 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 fifteen billion dollars.[1][2] The results will be used for apportioning representation in the US House of Representatives, for dividing federal tax dollars between states, as well as for a multitude of other governmental activities at the national, state, and local level. 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 noise introduced by the process, it also risks reducing the utility of census data—it may produce counts that are substantially noisier than the previous disclosure avoidance system, which was based on a technique called swapping, and relied on the details of the swapping procedure being 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 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 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 , which quantifies how much privacy loss is allowed, meaning how much can one person’s data to affect the analysis.

To be precise, randomized algorithm is -DP if, for each possible output , for any pair of datasets and that are the same everywhere except for on one person’s data,

Differential privacy is a characteristic of an algorithm; it is not a specific algorithm. Algorithms often achieve differential privacy by adding random noise.[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 exactly as enumerated, without any variation caused by adding noise. 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, such as requiring 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 (Noisy Histogram) adds noise from a carefully chosen distribution to the data counts. This produces a set of noisy counts. The noisy counts might have negative counts or violate invariants or other inequalities or be inconsistent with the counts one level up in the geographic hierarchy. The second step (Optimize) adjusts the histogram to be close as possible to the noisy counts, subject to the constraints that all counts be non-negative and 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 (which governs how much noise to add in the Noisy Histogram step), and the entire algorithm achieves -DP for . The 2020 US Census data will have six geographic levels, nested hierarchically: national, state, county, census tracts, block groups, and blocks; but in the 1940 E2E test, only national, state, county, and district levels were included.

### Step One: Noisy Histogram

In the E2E algorithm applied to the 1940s microdata, TopDown added random noise in a flexible way that allowed the user to choose what statistics are the most important to keep accurate. The noise was added to the detailed histogram counts for the level and also to a preselected set of aggregate statistics. Aggregate statistics are sets of histogram count sums specified by some characteristics. For example, the “ethnicity-age” aggregate statistic contains set of four counts: people of 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 E2E test included two such aggregate statistics (internally called “DP queries”): a 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 group-quarters queries and 67.5% spend on the race/ethnicity/age queries.

The epsilon budget of the level governed how much total random noise to add. A further parameterization of the epsilon budget determined how the noise was allocated between the histogram counts and each type of aggregate statistic. We write , where was the budget for the geographic level, was the budget for the detailed queries, and were the budgets for each of the types of aggregate statistics. Then noise was added independently to each count according to the follow distribution:

where denotes the two-tailed geometric distribution,

The noisy counts and noisy aggregate statistics are unbiased estimates with variance , where is the parameter for the geometric noise added. A higher privacy budget means the noise 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 noise added to each histogram count comes from the same distribution, and is independent of all other added noise; the noise 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 noisy data by optimizing a quadratic objective 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 has the property that the value of each variable is as close as possible to the corresponding noisy detailed histogram count or noisy aggregate statistic. This is done in a way that favors closeness for the noisy values constructed by adding noise from a lower variance geometric distribution. The solution to this optimization is not necessarily integral, however, and TopDown uses a second optimization step to round fractional counts to integers.

## Empirical Privacy Loss for quantifying impact of optimize steps

[to come an introduction and justification for the EPL approach—why is EPL expected to be related to epsilon?]

## TopDown options still to be selected

There are 7 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 values, with 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, was split evenly between national, state, county, and enumeration district.
3. What DP Queries to include. In the test, two DP Queries were included: (i) counts stratified by age-group/race/ethnicity (in other words, aggregating over “group quarters” type); and (ii) the group-quarters counts, which tally the number of people free-living as well as in five types of institutional and non-institutional facilities.
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 group quarters; and 67.5% for age-group-/race-/ethnicity-stratified counts.
5. What invariants to include. The test run held the total 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 values, for 4 different seeds to the pseudorandom number generator.

## Our Evaluation Approach

1. We calculated residuals (DP count minus exact 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 to understand relative error as well as the absolute error introduced by TopDown.
2. We calculated a measure of “empirical privacy loss”, inspired by the definition of differential privacy. To measure empirical privacy loss, we approximate the probability distribution of the residuals using Gaussian kernel density estimation with a bandwidth of 0.1, and compare the log-ratio inspired by the definition of -DP algorithms:

* [Some words about why this is different that the worst-case guarantee from epsilon-DP, perhaps based on email exchange with Philip Leclerc.] We hypothesized that the EPL of TopDown will be substantially smaller than the theoretical guarantee of . However, it is possible that it will be *much larger* than , due to the difficult-to-predict impact of including certain invariants.

1. We searched for bias in the residuals from (1), with our hypothesis that the DP counts are positively biased for areas with low diversity. [More detail about the theory behind this hypothesis, and the competing theory that it is about geographic unit size.] For each geographic area, we constructed a “homogeneity index” by counting the cells of the detailed histogram that contained a true count of zero, and we examined the bias (mean residual) of the corresponding counts from TopDown stratified by homogeneity index.

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

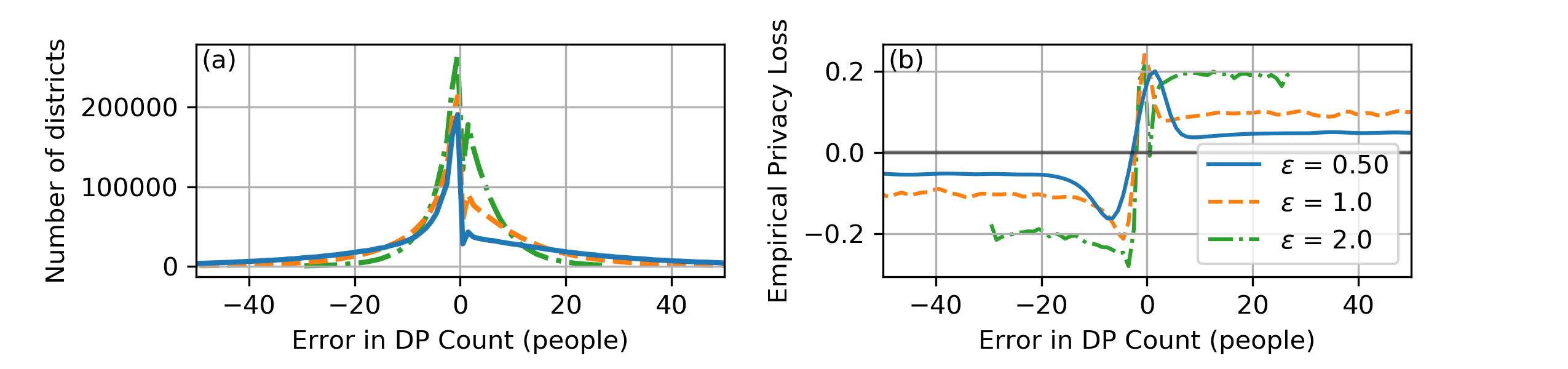
## Error and Privacy of TopDown

We found error in total count (TC) varied as a function of total privacy loss budget. Running TopDown with produced median absolute error in TC of 56 at the enumeration district level and 81 at the county level; produced median absolute error in TC of 15 at the enumeration district level and 24 at the county level; and produced median absolute error in TC of 4 at the enumeration district level and 7 at the county level (Full table in Supplementary Appendix 1). At the state level, there was TC error of , as expected from the state TC invariant. The median TC was 865 for enumeration districts, 18679 for counties, and 1903133 for states.

Error in stratified count (SC) varied similarly; When , the median absolute error in SC at the enumeration district level was 17 people, at the county level was 16 people, and at the state level was 18 people; for , 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 , the median absolute error in SC at the enumeration district level was 2 people, at the county level was 2 people, and at the state level was 2 people. The median SC was 88 for enumeration districts, 47 for counties, and 229 for states. (Figure 1)

We found that the empirical privacy loss was often substantially smaller than the privacy loss budget. For , the empirical privacy loss for TC at the enumeration district level was 0.024 and at the county level was 0.031 (at the state level empirical privacy loss is undefined, since the invariant makes all residuals zero); for , the empirical privacy loss for TC at the enumeration district level was 0.085 and at the county level was 0.081; and for , the empirical privacy loss for TC at the enumeration district level was 0.304 and at the county level was 0.275.

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 . For , the empirical privacy loss for SC at the enumeration district level was 0.288, at the county level was 0.129, and at the state level was 0.067; for , the empirical privacy loss for SC at the enumeration district level was 0.488, at the county level was 0.166, and at the state level was 0.141; and for , the empirical privacy loss for SC at the enumeration district level was 0.450, at the county level was 0.460, and at the state level was 0.531.



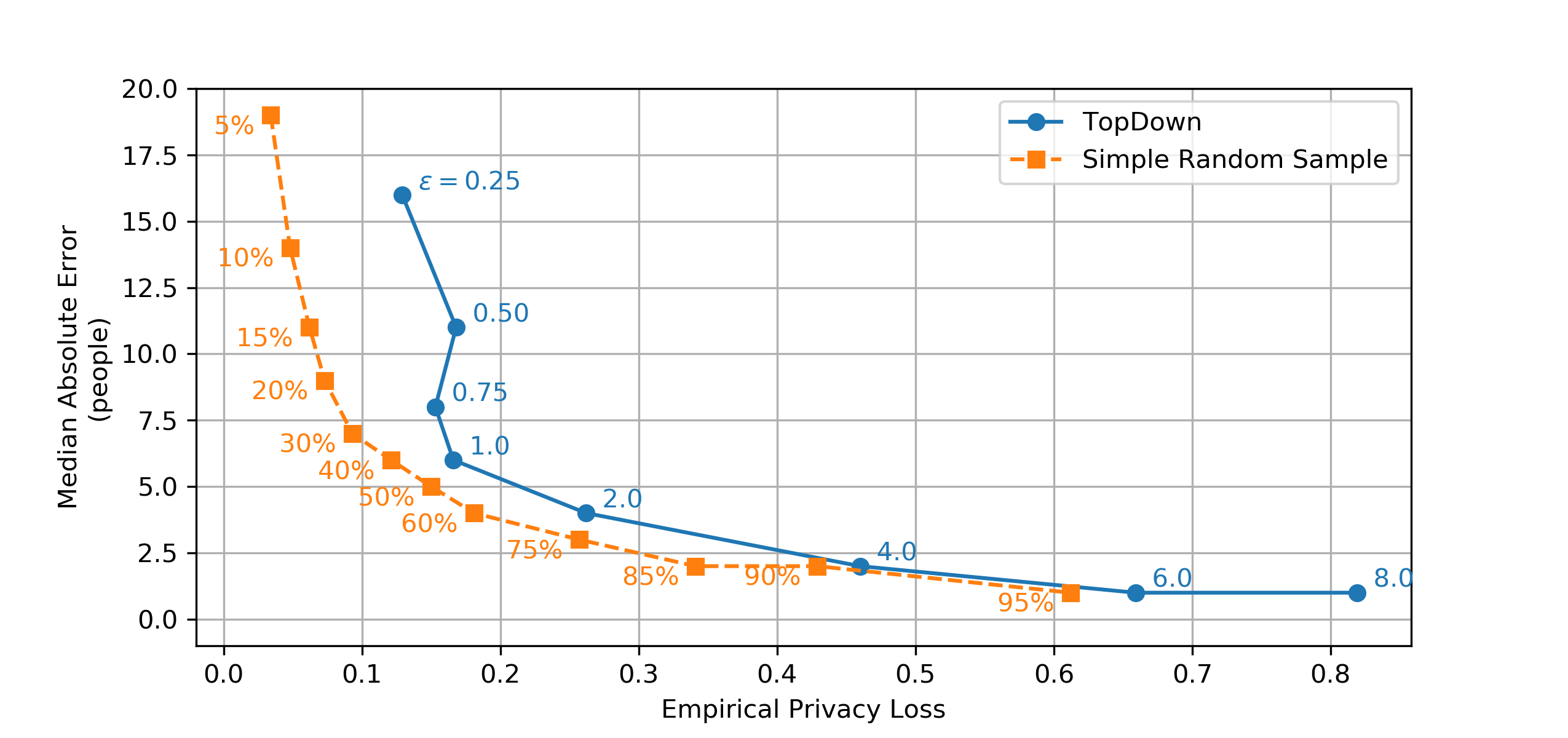
*Figure 1*: Panel (a) shows the distribution of residuals (DP - Exact) for stratified counts at the enumeration district level, stratified by age, race, and ethnicity; and panel (b) shows the empirical privacy loss, where is the probability density corresponding to the histogram in (a), after smoothing with a Gaussian kernel of bandwidth .

## 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 . 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 county 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)

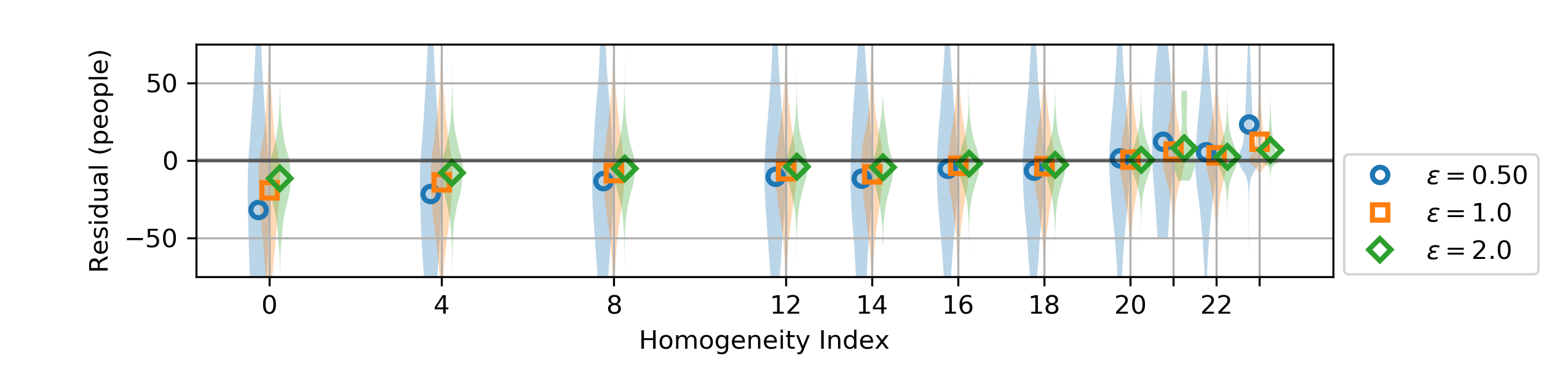


*Figure 2*: The curve with circular markers shows that in TopDown, the choice of controls the tradeoff between MAE and EPL, although for there is not much difference in EPL. The curve with square markers shows the MAE and EPL of Simple Random Sampling for a range of sample sizes, for comparison. For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

## Bias in the noise introduced by TopDown

The bias introduced by TopDown varied with diversity index, as hypothesized. Enumeration districts with homogeneity index 0 (0 empty strata) had TC systematically lower than ground truth, while enumeration districts with 22 empty strata had TC systematically higher. The size of this bias decreased as a function of . Homogeneity index 0 had bias of -52.6 people for , -18.9 people for , and -6.6 people for ; while homogeneity index 22 had bias of 8.7 people for , 3.6 people for , and 1.5 people for .

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 . homogeneity index 0 had bias of -103.7 people for , -33.9 people for , and -10.4 people for ; while homogeneity index 22 had bias of 23.4 people for , 14.5 people for , and 6.0 people for . (Figure 3)



*Figure 3*: The homogeneity index is associated with the residual (difference between the count estimated by TopDown and the true count). This plot shows the association for enumeration districts, and a similar relationship holds at the county level. As increases, the scale of the bias decreases.

# 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 add noise to this data. Such a reader may be open to the possibility that privacy is a valid reason for adding this noise, yet still be concerned about how this noise will affect their next decade of research. Our results visually summarized in Figure 2 can help to understand what this noise will mean: if , for county-level stratified counts, TopDown will be like the noise introduced by working with a 50% sample of the full dataset; if , it will like working with a 75% sample; and if , it will have accuracy matching a 95% sample, which is pretty close to having the full data without noise. 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.

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

For , we found that TopDown introduced near minimal noise and attained empirical privacy loss almost 10 times less than . 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 redistribution of resources from diverse urban communities to segregated rural communities. [More about the hypothesis that bias is due to homogeneity, vs the theory that bias is due to unit size.]

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 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 still 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—attitude surveys have found that calling attention to the privacy and confidentiality of census responses, even if done in a positive manner, reduces the willingness of respondents to answer census questions.[ref]

## 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 exact 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 geometric noise 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 is generalizable to the residuals from the same location when run with different data. 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 -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.

*Acknowledgements*: Thanks to Neil Marquez for suggesting comparing TopDown to simple random sampling.

# References

1. Garfinkel S, Abowd JM, Martindale C. Understanding database reconstruction attacks on public data. Communications of the ACM. ACM; 2019;62:46–53.

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.

3. Ruggles S, Fitch C, Magnuson D, Schroeder J. Differential privacy and Census data: Implications for social and economic research. AEA papers and proceedings. 2019. pp. 403–08.

4. Abowd JM, Garfinkel SL. Disclosure avoidance and the 2018 Census test: Release of the source code. <https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html>; 2019.

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

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.

7. boyd danah. Differential privacy in the 2020 decennial census and the implications for available data products. CoRR [Internet]. 2019;abs/1907.03639. Available from: <http://arxiv.org/abs/1907.03639>

8. Ruggles S, Flood S, Goeken R, Grover J, Meyer E, Pacas J, et al. IPUMS USA: Version 8.0 extract of 1940 census for U.S. Census Bureau disclosure avoidance research [dataset]. <https://doi.org/10.18128/D010.V8.0.EXT1940USCB>; 2018.

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