# Introduction

In the United States, the Decennial Census is an important part of democratic governance. Every ten years, the US Census Bureau is consititutionally 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 decenial census is also constitutionally mandated, and the 2020 US Census will use a novel approach to “disclosure avoidance” to protect respondents’ data.[4] This approach builds on Differential Privacy (DP), a mathematical definition of privacy and privacy loss 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 discloure avoidance system, which was based on a technique called swapping, and relied on the detailed of the swapping procedure being secret.[6]

To date, there is a lack of empirical examination of DP in census DAS, but the approach was applied to the 2018 end-to-end test of the decennial census, and computer code used for this test as well as accompaning exposition has recently been released publicly by the Census Bureau.[4][7]

We used the recently released code, preprints, and data files to quantify the error introduced by the E2E disclosure avoidance system when Census Bureau used it to guarantee differential privacy for the 1940 US Census (for which the full data has previously been released) at 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 DP 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 -differentially private 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. For census counting tasks such a producing histograms of the total count of people in each state, or counts of people stratified by census block, age, sex, race, and ethnicity, differential privacy is often implemented by adding noise to the counts.

The new disclosure avoidance system for the 2020 US Census is designed to be DP and to mantain 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 where 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. We will refer to the invariants and other inequalities collectively as the “public properties” of the database.

*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. The second step (Optimize) finds the tuned-up histogram that minimizes a quadratic objective function, subject to constraints from a system of linear equations and inequalities that represent the invariants, inequalities, consistency with the DP counts from one level higher, and non-negativity. The solution to this constrained optimization is as close to the noisy counts as possible while also satisfying internal consistency. The final output of the TopDown algorithm is a synthetic data set that has data counts matching the values that minimize the constrained optimization. This satisfies -DP and also the invariants and inequalities using an approach that affords detailed control of how the privacy budget is distributed between and within levels of the hierarchy.

The census data has six geographic levels, nested hierarchically: national, state, county, census tracts, block groups, and blocks. The census’s DP algorithm uses a top-down approach to create the synthetic data; steps one and two are performed six times, from the coarsest to the finest level. Each level is assigned a privacy budget and the entire algorithm is provably -DP for

At a specific geographic level (say census tracts) the true data has the form of a histogram; a set of boxes each labeled with a geographic unit (e.g. census tract one), a race combination (e.g. Black), one sex (e.g. Female), and one age (e.g. 46). Although the 2020 census will include more variables, the 1940 data run with E2E test code included the following: race (6 mutually exclusive categories), ethnicity (non-Hispanic and Hispanic), age group (under-18 and 18-and-over), and group quarters (2 categories). The number in the box, which we call a histogram count, is the number of people in the geographic unit with the features of the label (e.g. the number of 18-and-over non-Hispanic White women in enumeration district 107). Step one adds geometrically distributed random noise to numbers in each box according to the privacy budget at the level . This noisy data is unsatisfactory because the noisy counts (i) are sometimes negative, (ii) do not satisfy the public properties, and (iii) are inconsistent with the synthetic data produced at the coarser level (e.g. the sum of the noisy counts in all the boxes corresponding to a census tracts within Cook county may not equal the number of people in Cook County reported in the synthetic data produced in the previous level.) Step two solves an optimization problem which adjusts the counts in boxes so that they are non-negative integers, satisfy the invariants and inequalities, are consistent with the synthetic data produced at the coarser level, and are as close as possible to the noisy counts.

### 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 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. Census Bureau researchers have discussed plans to include each value that will appear in a tabular summary in the set of aggregate statistics. 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 histogram counts, and were the budgets for each of the types of aggregate statistics. Then noise was added independently to each histogram count and aggregate statistic according to the follow distribution:

where denotes the geometric distribution,

Note 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).

Note that the noise added to each histogram count comes from the same distribution; the noise does not scale with the magnitude of count, e.g. adding one hundred people to the count of age 18 and older non-Hispanic Whites is just as likely as adding one hundred people to the count of age under 18 Hispanic Native Americans, even though the population of the latter is smaller.

Although the E2E test used independent geometric noise for each detailed query and DP query at each level, the version of TopDown for the 2020 Census DAS will likely use the High Dimensional Matrix Mechanism [8], which may reduce the variance of the noise.

### 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 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 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 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 the TopDown algorithm uses a second optimization step to round fractional counts to integers. In this optimization, the linear equations and inequalities are the same as from the previous optimization, and the objective function is changed to minimize , where each corresponds to a (potentially non-integer) detailed query count given in the synthetic data and required to take an integer value of 0 or 1, where implies should be rounded down and implies that should be rounded up.

## TopDown options still to be selected

The 7 key policy choices, and 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: age/race/ethnicity (i.e. aggregating over group quarters) and gq (i.e. number free-living and number not)
4. At each level, how to split level-budget detailed DP. The test run used 10% for detailed queries, 22.5% for group quarters; and 67.5% for age/race/ethnicity.
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 and summarized their distribution by its median absolute error (MAE) for total count and age/race/ethnicity stratified count at the state, county, and enum\_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 calculate 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 kernel density estimation, and compare the log-ratio inspired by the definition of -DP:

* 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 search for bias in the residuals from (1), with our hypothesis that the DP counts are positively biased for areas with low diversity. For each geographic area, we constructed a “homogeneity index” by counting the cells of the detailed query 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 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

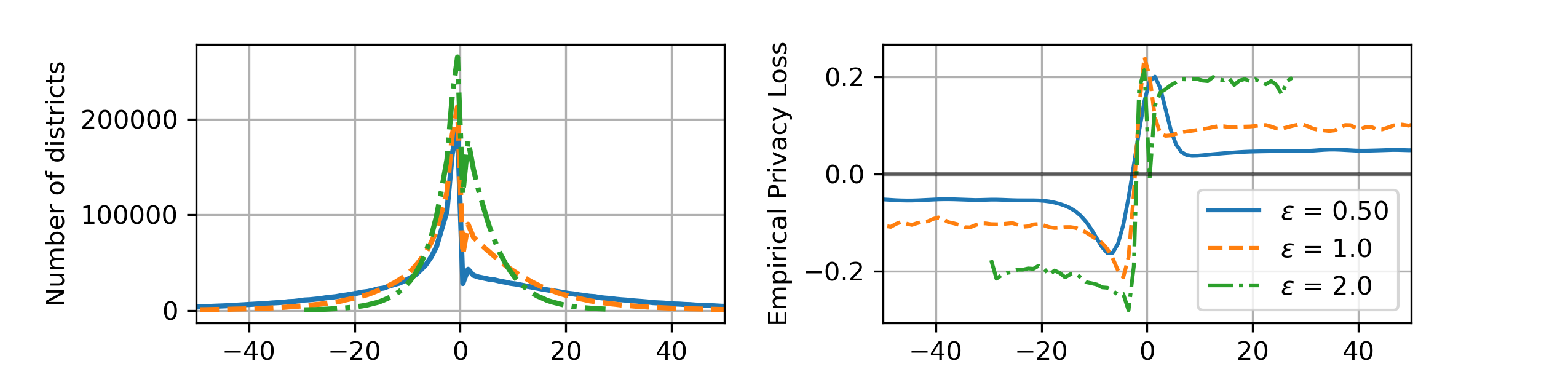
## Error and Privacy of TopDown

We found error in total count varied as a function of total privacy loss budget. For 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. (Figure 1)

Error in stratified count 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.

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 error 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.



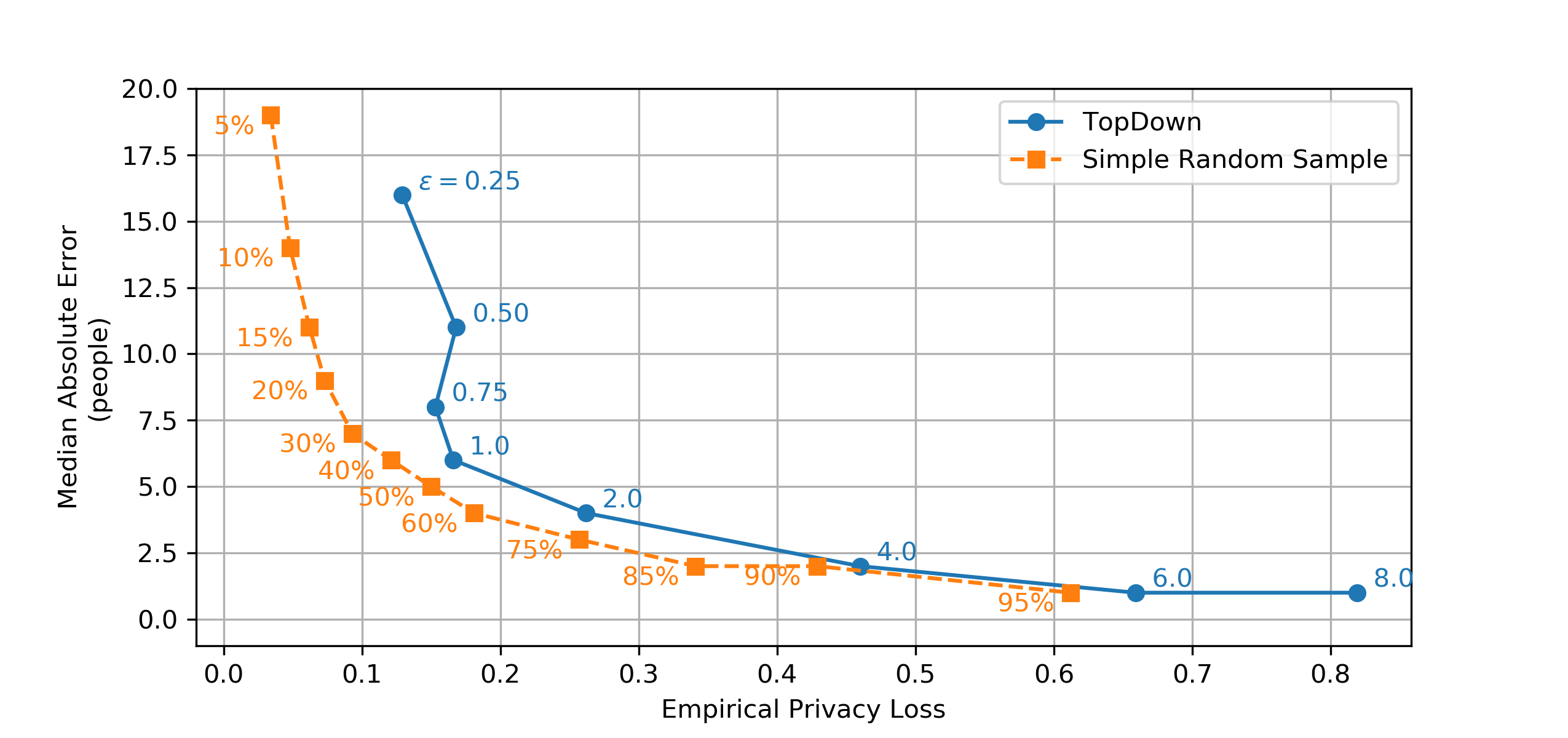
*Figure 1*: Panel (a) shows the distribution of error (DP - True) for stratified counts at the enumeration district level, stratified by age, race, and ethnicity; and panel (b) shows the empirical privacy loss, where is the probability density corresponding to the histogram in (a), after smoothing with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

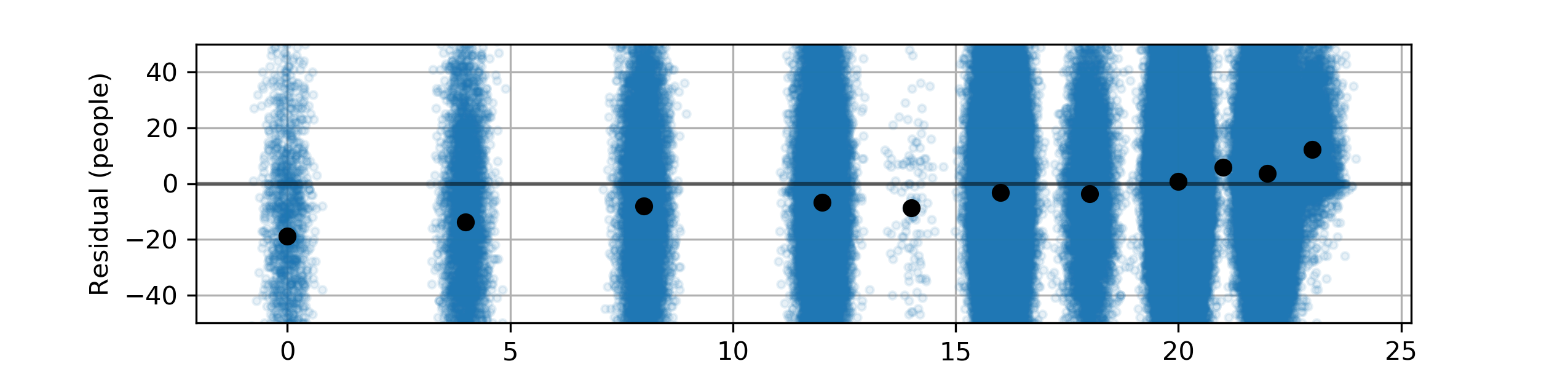


*Figure 2*: The curve with circular markers shows that in TopDown, the choice of controls the tradeoff between MAE and EPL, although for there is not much difference in EPL. The curve with square markers shows the MAE and EPL of Simple Random Sampling for a range of sample sizes, for comparison. For 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 curves).

## 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; this plot shows .

# Discussion

For , TopDown introduced near minimal noise and attained empirical privacy loss almost 10 times less than , but 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; since the errors are similar in *absolute* scale, cities of sufficient size will likely not notice who they have lost, but rural districts likely *will* notice or at least benefit from the population count (and appropriations) that they have gained. The TopDown algorithm will likely drive some redistribution of resources from diverse urban communities to segregated rural communities.

The small communities that are likely to have upward bias in their TopDown counts will be the ones small enough to benefit from the quality Assurance processes that have been implemented in past censuses, such as the Count Question Resolution program, and the results in this paper can help anticipate and plan for this process.

Accurate counts in small communities are important for emergency preparedness and other routine planning tasks performed by state and local goverment 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. On the other hand, human subject research requires informed consent (Belmont Principles); de-identified data is not HSR, but if it is re-identifiable, it should not be considered de-identified, should it?

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. This work provides some evidence on how TopDown may affect this work, 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, reduce willingness to answer census questions.[ref]

## Limitations

There are many differences between the 1940 census data and the 2020 data to be collected next year. Number of geographic levels, number of strata to be included in detailed queries.

Additional changes are being planned, HDMM instead of geometric mechanism, for example.

Our approach to quantifying error focused on the median absolute error, and there are important tails to this distribution as well.

Our empirical privacy loss is not comprehensive, and there is the possibility that some other perturbation or some other test statistic would reveal a larger 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, it is likely that the difference is sufficiently small as to not affect our estimates substantially.

*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. Office USGA. 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, others. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science. Now Publishers, Inc. 2014;9:211–407.

6. McKenna L, others. Disclosure avoidance techniques used for the 1970 through 2010 decennial censuses of population and housing. 2018.

7. boyd. 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. Chen R, Xiao Q, Zhang Y, Xu J. Differentially private high-dimensional data publication via sampling-based inference. Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining. ACM; 2015. pp. 129–38.