# 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.[ref GAO-18-635 Census Bureau Improved the Quality of Its Cost Estimation but Additional Steps Are Needed to Ensure Reliability] 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.[1]

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.[ref TopDown draft? something better?] 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.[ref Dwork and Roth, Privacy Book?] 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.[2]

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.[refs to census pubs, danah boyd’s whitepaper]

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* 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. It follows logically from the definition that achieving -differential privacy in this manner requires adding noise with variance of at least that of a symmetric geometric distribution with parameter . [Additional details here? in Appendix?]

The new disclosure avoidance system for the 2020 US Census is designed to be differentially private and to mantain the accuracy of census counts. To complicate things beyond the typical challenge faced in differentially private algorithm design, there are certain counts in the census that will be published exactly as enumerated, without any variance from adding noise. These *invariants* have not been selected for the 2020 decennial census yet, but in the 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.

*TopDown algorithm goal and high-level description.* 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 the name TopDown). The first step adds randomness to the data counts in a way that satisfies -differential privacy. This produces a set of counts, . The noisy counts in might have negative counts and violate the invariants and inequalities. The second step finds the argument 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 values that optimize this constrained optimization are as close to the noisy data as possible while also satisfying the public properties and requiring that all counts are positive integers and other internal consistency constraints. 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 -differential privacy and the invariants and inequalities and affords detailed control of how the privacy budget is spent between and within levels of the hierarchy.

The census data has six different geographic levels: national, state, county, census tracts, block groups, and blocks. The census’s differentially private algorithm uses a top-down approach to create the synthetic data; steps one and two are performed six times in order from the coarsest to the finest level. Each level is assigned a privacy budget and the entire algorithm will satisfy -differential privacy where

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 only included race (6 mutually exclusive categories), ethnicity (non-Hispanic and Hispanic), age group (under-18 and 18-and-over), and group quarters (6 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 step.) Step two solves an optimization problem which adjusts the counts in boxes so that they are positive, 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 (and are integers).

### Step One: Adding random noise.

The E2E DAS 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-sex" aggregate statistic contains set of four counts: Hispanic men, Hispanic women, non-Hispanic men, and non-Hispanic women. 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 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 age across all household/group quarters types (again for each level of the spatial hierarchy). 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 governs how much total random noise to add. A further parameterization of the epsilon budget dictates how the noise will be allocated between the histogram counts and each type of aggregate statistic. We write , where is the budget for the geographic level, is the budget for the histogram counts, and are the budges for each of the types of aggregate statistics. Then noise is added independently to each histogram count and aggregate statistic as follows:

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 18 year-old non-Hispanic White Men is just as likely as adding one hundred people to the count of 78 year-old Hispanic Native Men, even though the population of the latter is smaller. [This example doesn’t exist in the 1940s because we have fewer variables. I am not sure if I should restrict to their variables for the examples. I think it makes sense for the examples to be more like the 2020 census.]

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 [ref], which may reduce the variance of the noise.

### Step Two: Optimization.

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 men summed across the counties in a state is equal to the number of men in the 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 itegral, 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:

1. Overall privacy
2. How to split this budget between national, state, county, tract, block group, and block
3. At each level, how to split level-budget detailed DP
4. What DP Queries to include
5. What invariants to include
6. What constraints to include
7. What to publish

## Our Evaluation Approach

1. Calculate errors and their variance (for total count and age/race/ethnicity stratified count for state, county, and enum\_district). We also summarized the size of these counts to understand relative error as well as the absolute error introduced by TopDown.
2. Calculate “empirical privacy loss” (which we have just invented for the purposes of this paper; need to describe it here) (for same groupings)

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. [ref invariants paper from Census Bureau]

1. Search for bias, with our hypothesis that it appears differentially with respect to diversity of spatial units.)

We compare the bias, variance, and privacy to a simpler, but not-differentially-private approach to protecting privacy: simple random sample of X% total for a range of X.

# Results

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. The empirical privacy loss for SC was roughly twice the empirical privacy loss for TC and the county and enumeration district level. 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 AROUND HERE

Compared to 50% sample, … (Figure 2)

FIGURE 2 AROUND HERE

The bias introduced by TopDown varied with diversity index, as hypothesized. Enumeration districts with only X empty strata had TC and SC systematically lower than ground truth, while enumeration districts with X empty strata had TC and SC systematically higher. The size of this bias decreased as a function of , from tc\_bias\_enum\_dist\_X\_0\_25 for to tc\_bias\_enum\_dist\_X\_4\_00 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 , from tc\_bias\_county\_X\_0\_25 for to tc\_bias\_county\_X\_4\_00 for . (Figure 3)

FIGURE 3 AROUND HERE

# Discussion

TopDown introduced near minimal noise, but created a quantifiable amount of bias. Bias disproportionately affected small, homogeneous districts; cities of sufficient size will not notice who they have lost, but rural districts likely *will* notice the population count (and appropriations) that they have gained. The new DAS affords redistribution of resources from diverse urban communities to segregated rural communities.

Quality Assurance and the correct count process.

Emergency preparedness and other routine tasks.

Research tasks, e.g. segregation research and how it may be hampered. 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?

Survey weights.

Need for continued discussion.

## Limitations

To Come

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

## References

1. 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.

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