CS 763: Security and Privacy for Data Science Fall 2020

Accuracy Impacts of Differential Privacy in the 2020 US Census

Authors: Keaton Leppanen, Anna Meyer, and Shiyu Yang

The Census Bureau is preparing to implement differential privacy for the first time in the 2020 count. This adaptation poses various technical challenges including implementing differential privacy in a hierarchical setting, finding algorithms that are efficient and provably private, and dealing with invariants. We analyze the Census' algorithm, TopDown, to better understand these technical issues, as well as how the algorithm adds noise to the data and optimizes it to be public-facing. In the second half of the paper, we focus on the empirical results of TopDown. Due to the inherent noisiness of differential privacy, as well as various non-technical constraints, the accuracy of TopDown's outputs vary greatly. We investigate patterns in the results in which certain communities' populations are systematically under-reported. We also attempt to gain a better understanding of the results through experiments. Similarly, we explore the possible underlying causes that contribute to this unequal impact, focusing particularly on the effects of non-technical constraints such as non-negativity and integrality.

1 Introduction

Even though the concept of differential privacy is relatively new, the ubiquity of data collection and the ever-growing threat of data reidentification attacks demonstrate the necessity of enforcing such formal guarantees of privacy when handling personal data. With a recent internal experiment showing that potentially over 45% of US citizens' data is identifiable through a reidentification attack using the 2010 Census data [1], the Census is no exception. In order to combat the potential for such data reidentification, the Census is planning to implement differential privacy for its 2020 count in what they define as the first truly large, centralized implementation of differential privacy. This paper explores the methods which the Census will be employing to achieve this results, the unique challenges the Census faces by implementing differential privacy, as well as the implications of this decision to adopt differential privacy.

1.1 Differential Privacy

Suppose you are deciding whether or not to participate in a survey or other statistical process, such as the US Census. You don't want your privacy to be compromised, though. One way to formalize this wish is to ask that no one can learn anything new about you from the survey results, that they would not have learned if you did not participate in the study. In other words, the response to any database query should be statistically indistinguishable from the query response to a database that differs by exactly one record. This notion of privacy is formalized through differential privacy, which is a property of a randomized data processing algorithm that guarantees privacy with high probability. A query (or algorithm) Q is differentially private if, for every pair of databases D and D' that differ by exactly one record, and for every subset S of the query's domain, equation 1 holds.

$$\mathbb{P}[Q(D) \in S] \leq e^{\epsilon} \times \mathbb{P}[Q(D') \in S] + \delta \tag{1}$$

In this equation, $\epsilon \geq 0$ and $\delta \geq 0$ are tunable parameters, where smaller values of each correspond to stronger privacy guarantees. Differential privacy is typically implemented through adding random noise, such as from the Laplace or Gaussian distributions, to the true output of the algorithm.

1.2 Privacy and the US Census

The US Census Bureau is legally mandated to enumerate the US population every ten years and to keep the results private (both from the general public and from the government, outside of approved uses). In the past, the Census Bureau protected privacy though a variety of methods, including rounding totals, swapping household data between geographic areas, removing outliers, and generating partially synthetic data [8]. The full range of techniques is not public, though, because publishing too many details would pose a privacy risk.

The historical disclosure avoidance techniques deployed by the Census Bureau have worked well so far. At least, as far as we know there have been no real-world privacy breaches from the Census data. However, the ad-hoc techniques used in the past do not provide formal privacy guarantees. Advances in computing power, along with widely-available third-party data, leave Census data vulnerable to reconstruction attacks. In a recent experiment using data from the 2010 Census, the Census Bureau reconstructed individual-level data from the published microdata (anonymized data that is released for a small portion of the population) and statistical tables. They were able to link 45% of the US population to commercially available data and recover personally identifiable information for these people [1]. Of those links, they were able to confirm that 38% were correct (in other words, 17% of the US population was re-identifiable). Furthermore, this re-identification estimate is likely conservative, because the Census Bureau only used legal, commercially available data: private data or illegally accessed data might pose an even larger reconstitution threat [1].

1.3 The Importance of Accuracy

Having a reliable census that accurately reflects population sizes at various levels is important for distributing federal funds, drawing legislative boundaries, informing social science research and more. The differentially private version of the Census must guarantee a certain level of accuracy so that these processes happen fairly, and so that the general public maintains its trust in the Census Bureau.

To expand on the latter point, the census process must be accurate and transparent in order for the public to place trust in the Census Bureau and the data it releases. If the resulting data is inaccurate or hard to use, or if the public does not feel confident in the Census Bureau's data processing abilities, then the reputation and future prestige, funding and response rates might be at risk. At the same time, the Census Bureau needs to ensure privacy. If there were a privacy breach, individuals may be less likely to fill out the census in the future. Additionally, as stated above, maintaining privacy is legally mandated.

More importantly, the data produced by the census must be accurate, especially in terms of population counting, because the reported totals are used to make several important decisions. At its most basic, the data is used for redistricting at the federal, state and local levels. As a result, accurate population counts are critical to making sure that every citizen has equal representation. The Census data also determines how federal funds and government services are allocated to various communities. If counts are inaccurate (in particular, if certain communities are undercounted), this will have a material effect on how schools, social support services, government and infrastructure are funded. Finally, the census data is used by various social and population scientists, both for direct studies and as a tool to help with statistical sampling procedures. Various social and population scientists have expressed concerns with the impact that less accurate, differentially-private census data will have on their research [10][11].

1.4 Structure of This Paper

Next, we will introduce our guiding questions. After that, we describe the Census algorithm, TopDown. Then, we discuss our literature review about empirical results on the impact that TopDown has on the accuracy of the Census. This body of work includes results produced both by Census Bureau researchers and by independent scholars. Finally, we will talk about the results and challenges of running our own experiments.

2 Guiding Questions

Throughout this paper, we focus on the following questions.

1. How does the Census Bureau's differential privacy algorithm work? In what stage of the algorithm is inaccuracy introduced? What about disparate levels of inaccuracy?

2. In what ways does the Census Bureau's differentially private algorithm under-report or over-report the population of certain groups? How large is this effect?

3. Can we tweak the differential privacy algorithm to avoid systematic skews in population size?

3 TopDown Algorithm

3.1 Introduction

When presented with the goal to implement differential privacy for the 2020 Census, the Census Bureau faced several unique challenges that prevented it from simply using a standard method to guarantee privacy. Among these challenges were the scale of the undertaking, legal requirements enforced by policymakers, and the need to maintain consistency with other publicly available datasets and common knowledge. With these challenges in mind, the Census has created The Census Top Down Algorithm (TDA), which is a collection of algorithms that serve as a differentially private mechanism for creating confidentiality preserving, publicly releasable data, and demographic information concerning the US population. In this section we will discuss how the the TDA was designed to deal with these unique specifications and provide a high level overview of how the actual implementation of the algorithm functions.

3.2 Scale

Needing to accommodate tens of billions of differentially private measurements, the Census's 2020 deployment of the TDA will be one of the first truly large-scale deployments of a centralized differential privacy model. As such, the dataset is simply too large for a computer to process in memory, much less in a reasonable time [7]. In order to contend with this issue of scale, the TDA decomposes the task of generating nation-wide data into sub-problems based on geographic levels, and then uses incremental schema extension to combine the results into the final dataset.

Incremental schema extension is the process that allows us to take existing differentially private data and add on additional fields to each record while still preserving the privacy of the data. It is based on the differential privacy property of adaptive composition which allows for the preservation of differential privacy when combining tables by reapplying a differential privacy algorithm with a larger privacy budget [2]. Formally, let S^0 and S be two sets of attributes, where S^0 is a subset of S, and $\tilde{T}^0$ be a differentially private table with attributes S^0. Then by incremental schema extension we can add on additional fields from S to every record in $\tilde{T}^0$, thereby extending $\tilde{T}^0$ and creating a table $\tilde{T}$ that is still differentially private and also contains all the relevant attributes.

In practice, the Census utilizes six geographic levels: National, State, County, Tract (sometimes divided into Tract Group and Tract), Block Group, and Block. The Top Down Algorithm, as its name suggests, begins by generating a high level, differentially private table containing the national level demographic data (Race, Ethnicity, Voting Age, and Housing Type). Then, using incremental schema extension, it generates a new intermediate table representing the state level demographics by adding the additional state attribute to each of the records. This process is repeated for each of the geographic levels, resulting in a final data table where the records have all the demographic and geographic attributes. Not only does this approach address the memory problem, it also provides an opportunity for parallelization to further increases the efficiency of the TDA [7].

3.3 Constraints

The data reported by the Census is used for many important decisions such as apportionment of Representatives in the House of Representatives and funding allocation for numerous entities. As such, the Census has been charged with reporting the accurate values for certain statistics, with no added perturbation. These invariants include total population of each state, number of housing units in each block, and number of occupied group quarters by type in each block, among others. Furthermore, in order to satisfy its goal of being a trustworthy reporter of information, the Census aims to ensure consistency with broader public knowledge in the forms of overlapping datasets and common sense [2]. This further complicates ensuring differential privacy.

In terms of the TDA, these invariants and public knowledge consistency expectations are expressed as a series of constraints that the privatized data must adhere to. One of the most straightforward of these constraints comes from the fact that the total population for each state is required to be held invariant. Therefore, the TDA must ensure that all the counties in a particular state have population totals that, when totalled, equal the total state population. In addition to such invariant based constraints, the Census aims to satisfy a slew of ‘common sense’ constraints such as the number of homeowner spouses not exceeding the number of homeowners, populations not being fractional or negative numbers, and group-quarters that are marked as being gender exclusive (e.g. male and female dormitories) having exclusively residents of that gender [2]. However, with addition of each constraint, there arises implied constraints that must similarly be satisfied. For instance, it is not enough simply to say that the total county populations in a given state sum to the total state population. Given that constraint, we must also now ensure that the cities’ populations sum to a given county’s population, and so on, so forth for each geographic level.

In order to deal with these constraints, the TDA employs Fourier-Motzkin Elimination. When extending a table to include another geography level, TDA looks at the initial set of inequalities the data is required to satisfy in order to be in agreement with public knowledge and invariants. The TDA takes an initial set of inequalities, which the data is required to satisfy to be in agreement with public knowledge and invariants. Then it applies Fourier-Motzkin Elimination to systematically eliminate variables, resulting in a reduced set of desired constraints [2]. For instance, if we were looking at extending the data to include a new location L, we would need to consider what the public already knows about this location. In this example, let’s assume that the number of L’s sub-regions, the total population in each of these regions, the total number of householders in each region, and the total number of voting age persons in each region is public knowledge. From this information we can derive our initial inequalities. Next we would apply Fourier-Motzkin Elimination, which would result in a set of constraints that L must satisfy.

3.4 TopDown Algorithm implementation

At this time many of the final determinations regarding the official 2020 release of the Census' Disclosure Avoidance System, which is built around the Top Down Algorithm, have not been decided upon or are otherwise not publicly known. As such, the information in this section is merely to provide context when considering the TDA's implementation.

3.4.1 Privacy budget allocation

The privacy-loss budget for the 2020 Census has yet to be determined and is ultimately a policy decision. That being said, whatever the final budget turns out to be, it will need to be divided among the products created through the Disclosure Avoidance System and then again shared between the various geographic levels in that product [7].

Recently, the Census provided the public with a demonstration of the TDA in an attempt to replicate the expected products from the 2020 using data from the last Census, the results of which could potentially inform decisions regarding the privacy-loss budget for the 2020 release. In this demonstration, they employed a total privacy budget of $\epsilon = 6.0$. This budget was then subdivided between two products: housing records ($\epsilon = 2$) and person records ($\epsilon = 4$). Each of these then had their budgets shared between their geographic levels by assigning 20% of the budget to both the national and state levels and 12% to the county, group track, track, block group, and block. In a similar demonstrations using a limited dataset gathered in 2018 and the full 1940 Census data, the Census divided the privacy-loss budget across all geographic regions equally [7].

Although an official decision has yet to be made, based on their current demonstrations and publications, the Census is most likely planning to allocate their privacy-loss budget to prioritize person data over housing based data. In addition, the Census seems to want to keep the budget fairly evenly distributed between geographic levels, potentially slightly favoring the national and state levels.

3.4.2 Differential privacy mechanism

Due to the necessity of publishing the invariants, the TDA utilizes bounded differential privacy, a common, albeit weaker, variant on differential privacy involving the release of some unperturbed data points. As such, regardless of the mechanism, TDA will necessarily be weaker than an alternative that adds noise to all released statistics [7]. The exact mechanism by which the Census is planning to add noise to the data is also yet to be announced.

The TDA supports easily interchangable DP mechanisms. In recent demonstrations and all publicly available implementations, the Census has utilized the Geometric Mechanism for adding noise to their data, however, Census researchers have stated that current internal builds of the 2020 implementation of TDA do not rely directly on the Geometric Mechanism, instead incorporating the high-dimensional matrix mechanism. They also strongly suggest that the high-dimensional matrix mechanism will be used in the 2020 release [7].

3.4.3 Algorithm overview

The Top Down Algorithm is organized into 3 main phases: Initialization, Noise Addition, and Post Processing. In the first phase, Initialization, the algorithm calculates the specified invariants for each of the regions. Next, it uses those invariants as well as pre-specified constraints to determine the implied constraints that the algorithm will need to follow. In the middle phase, Noise Addition, the algorithm adds noise to all the levels of the data. Presently this is simply done through the application of the Geometric Mechanism, but this is likely to change, as indicated in Section 3.4.2. Finally, the final phase of the algorithm, Post Processing, takes the private data and recursively applies the previously generated constraints to it by solving a series of least-squares optimization problems. Then, through incremental schema extension, it combines the all the levels of data into a single data product containing the differentially private data [7].

3.4.4 Adapting the official TopDown Algorithm

As part of this project we originally planned to take advantage of the released source code for the prototype of the Disclosure Avoidance System built around the TDA to run several experiments to better understand how the system works. In our experiments, we wanted to measure the effects of varying the total privacy-loss budget as well as its distribution between different geographic levels. However, we experienced a slew of complications: some of the documentation for running the code was incomplete, outdated, or simply inaccurate; the environment to use while running the system was difficult to configure (despite using Amazon Web Services, which is the recommended platform); and major errors due to compatibility issues with external software once we finally got the system to run. As a result, we were unable to complete our experiments.

That being said, the exploration of the actual code used to implement the Disclosure Avoidance System greatly aided in our understanding of how the system functions on a practical level, which complements the theoretical understanding we gained by reading various journal articles describing the algorithm. Working with the code is what also allowed us to discover various specifics about the test runs such as the exact privacy-loss budgets and epsilon distributions used for the official demonstrations.

In addition, when exploring the demonstrations we discovered that for one of the earlier trial runs, the Census Bureau released the results of the algorithm for varying levels of epsilon. This trial run was conducted on the entire 1940 Census data (the latest Census dataset which is available in its entirety to the public) [13]. Analyzing these results isn't exactly analogous to the experiment we wished to execute, since the Census Bureau only varied the total privacy budget while evenly distributing the budget between geography levels. In addition, due to the 1940 Census data being of far smaller scale, vastly different demographics, and using different geographic levels, the specific values of the results are not necessarily reflective of those of the 2020 Census. However, when we examine the results of this experiment we get a fairly good sense of the general effectiveness of the TDA as well as the trade offs between accuracy and privacy. Looking at Figure 1, we can see that the Top Down Algorithm, despite the additional complications due to invariants and constraints, maintains a fairly standard trade-off between accuracy and privacy at all the surveyed geographic levels. We will explore the impact of of the choice of epsilon in greater detail later in the paper.

4 Literature Review

In this section, we discuss various analyses of the TopDown Algorithm's results. These analyses were conducted by a variety of people, including both Census Bureau employees and independent researchers.

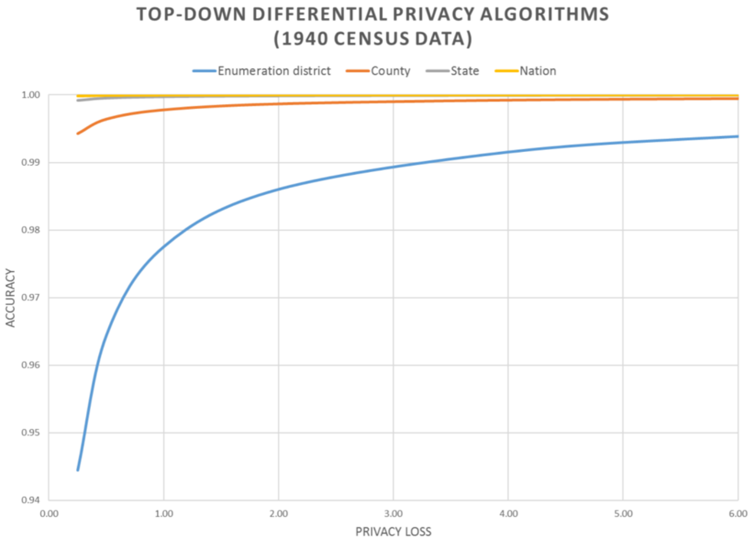

Figure 1: The accuracy of the results obtained by running the TDA on the complete 1940 Census data for epsilons ranging from 0.25 to 6 for each geographic level in the 1940 Census data (these levels differ from the modern day groupings). Figures courtesy of the US Census Bureau [13].

4.1 Empirical results from Census tests on TopDown

In 2018, Census Bureau released a version of the 2010 census data as processed by the proposed 2020 differentially private algorithm. The results spawned backlash from social scientists and others because the amount of noise was more than they had expected [10]. We discuss some of these specific complaints in Section 4.2, but first we analyze the results released by the Census Bureau.

4.1.1 Impact of the choice of epsilon

The choice of epsilon is central to the accuracy/privacy trade-off in any differential privacy use case, and the Census is no exception. At the time of this writing, the overall epsilon for the 2020 Census (and how the privacy budget will be split among different geographic levels) is undecided. The Census Bureau presented results of using various values of epsilon, ranging from 0.01 to 16, to the Data Stewardship Executive Policy committee (the arm of the Census Bureau that will make final decision regarding what value of epsilon to use) and made the results available to the public [12]. The results include density plots of how different counties' populations shifted under the TDA. A few of these plots are shown in Figure 2. We see that for the smallest value of epsilon, 0.01, the noisy reported total is often quite different — on the order of several thousand — from the actual total. A smaller number of counties still have large discrepancies between the actual and reported totals for epsilon values of 0.1 and 0.25. By the time that epsilon is 1 or larger, the difference between most counties' actual and reported total populations is much smaller, on the order of a couple hundred people. When epsilon shrinks further, this pattern continues, as is shown, for example, in the last graph with an epsilon of 8.

Another thing to note is that in all of the graphs in Figure 2, no matter what epsilon is, the mode of the change in population is positive and the left tail of the graph is longer than the right tail. As a result, the average county gains population, but among the counties that lose population, the absolute effect is larger. We will discuss this observation more later on in this section.

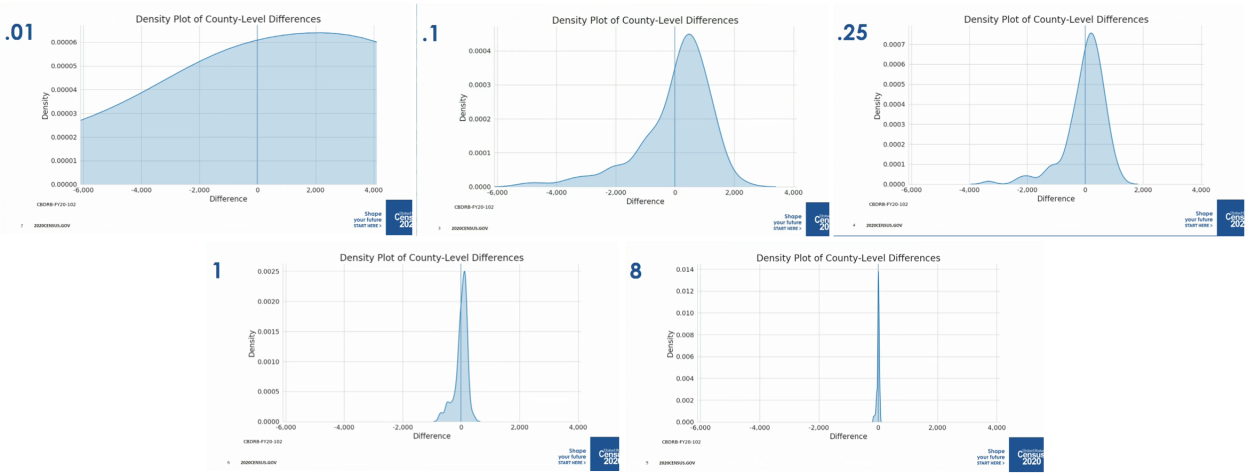

Figure 2: Kernel density plots for how much US counties' total populations differ from the noisy reported totals. The blue number in the upper left corner of each image is the epsilon value. Note that the x-axis in each image is consistent (-6000 to 4000), but the y-axis changes slightly as the density function converges. (Figures courtesy of US Census Bureau [12], session B)

The results become more interesting when we stratify based on county size. The box plots in Figure 3 show that, on average, smaller counties gain population whereas the largest counties lose population. This trend is especially pronounced for small values of epsilon. For example, for $\epsilon = 0.01$, the smallest counties (less than 7,500 people) gain an average of 4,000 residents, or 60% of their total population. These small municipalities may not mind the extra funding that comes along with inflated population counts, but this result is problematic because of accuracy impacts for various social science research, as well as for the corresponding decrease in population and funding for other communities. Looking at the outliers here is worrisome, too: one county increased in size almost six-fold, and two others saw population gains of over 250%. On the other hand, the largest counties lose around 12,000 people, or around 10% of their total population, on average. For the larger values of epsilon (0.5 and 4), the differences are smaller. For $\epsilon = 0.5$, the smallest counties gain about 5% of their total population, but this shrinks to just a bit more than 1% for $\epsilon = 4$. For the largest two-thirds of counties (any county larger than 15,000 people), the absolute change in population is less than 1% for both $\epsilon = 0.5$ and $\epsilon = 4$. We also notice that only the largest population buckets lose population, whereas the majority of counties gain population under TDA. This makes sense in context of our observation about Figure 2, when we saw that most locations gain population, but those that do lose population tend to lose a larger number of people. These box plots show that it is primarily populous counties that are negatively affected by this phenomena.

4.1.2 Impact on demographic information

The choice of epsilon becomes even more difficult (that is, the range in accuracy grows even larger) when considering the smallest subsets of the population, such as breakdowns by age, race, and/or ethnicity. Census Bureau scientists did additional data analysis focused on the state of Virginia, which they selected for its moderate size and well-representative population distribution. As part of their analysis, the Census Bureau presented the results from three counties: Fairfax (the largest county in Virginia, with over 1 million residents), Winchester City (the median-sized county, with 26,000 residents), and Highland (the smallest county, with about 2,300 residents). Figures 4 and 5 show a selection of the results for Fairfax and Highland counties, respectively, with various values of epsilon shown in each image.

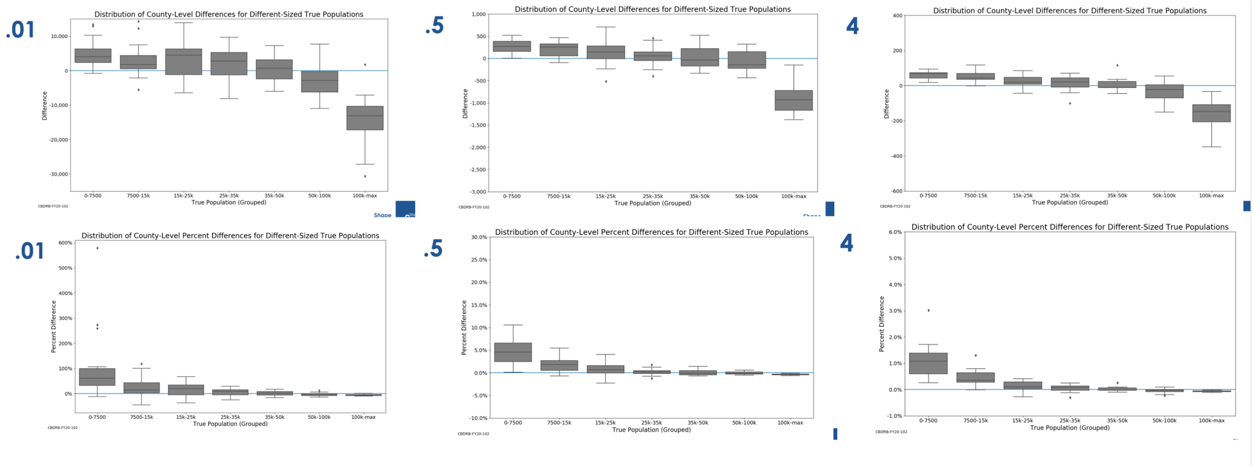

Figure 3: Box plots showing the change in population between actual and noisy counts for various values of epsilon. In each plot, the buckets are 0-7.5k, 7.5-15k, 15-25k, 25-35k, 35-50k, 50-100k, and 100k+. The top row shows the absolute population change. The scales, from left to right, max out at 1000, 1000, and 400. The bottom row shows the percent population change. From left to right, the scales max out at 600%, 30%, and 6%. (Figures courtesy of the US Census Bureau [12] session B)

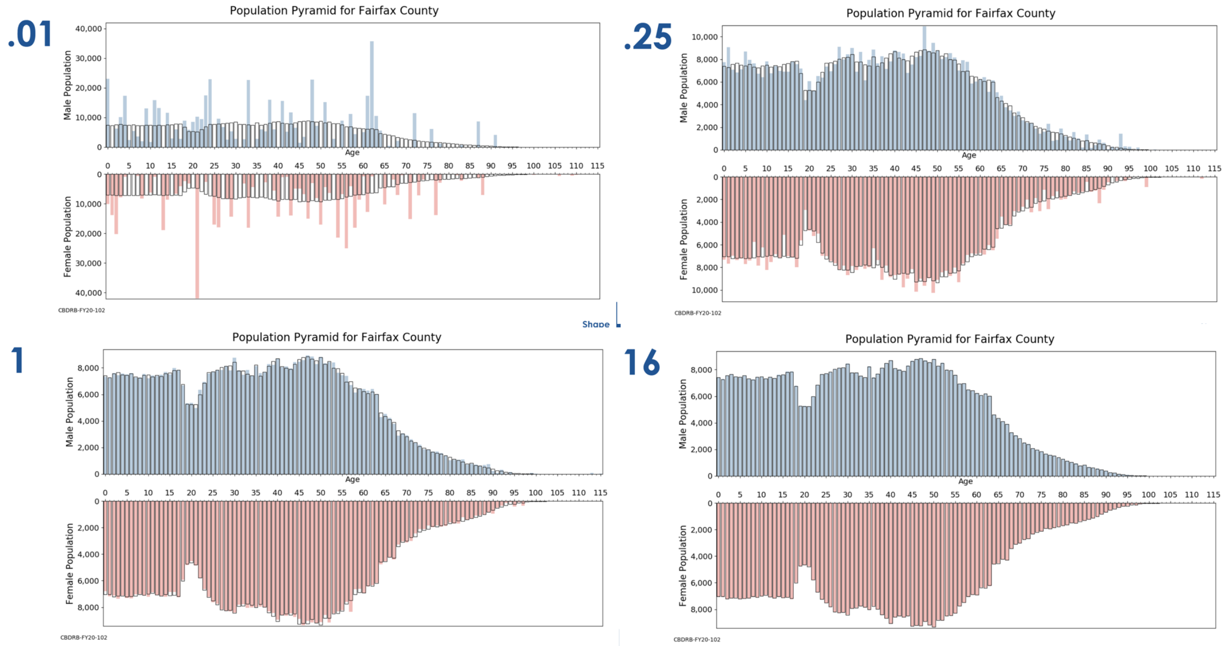

Figure 4: Age and gender population histograms for Fairfax County. The hollow black boxes show the true population, while the blue and red bars show the noisy count of males (top half, in blue) and females (bottom half, in red) at each age, where age is the x-axis and ranges from 0 to 115. Each graph shows the output for a different value of epsilon, which is indicated by the number in the upper left corner. (Figures courtesy of the US Census Bureau [12] session B)

For Fairfax County, we see in Figure 4 that for all but the very smallest epsilon ($\epsilon < 0.25$), the noisy population counts across ages is similar to the true data, as seen by how closely the two sets of histograms overlay each other. In particular, for $\epsilon \geq 1$, the histograms are virtually indistinguishable, at least when looking from a distance. As a result, the data released for Fairfax County could be used for statistical purposes without any trouble, as long as $\epsilon \geq 1$.

On the other hand, the Highland County histograms, displayed in Figure 5, show that the noisy and true counts are quite different from each other. When epsilon is 0.01, all of the county's residents are reported to be only a handful of age and gender combinations (the three most popular being 1-year-old boys, 4-year-old girls, and 51-year-old men). Epsilon values of 0.25 and 1 don't do much better: the reported totals are still very different from the underlying totals. By the time we get to a very large epsilon (16), the two histograms match each other reasonably well. Compared with Fairfax County, though, we can still notice many differences with the original histogram (for example, the noisy counts show no 14-year-old boys or 50-year-old women). The absolute differences in population are not necessarily larger than those in Fairfax County, but due to the smaller overall population, they are more obvious. That being said, the larger percent difference in smaller counties means that any discrepancies will have a larger accuracy impact than an equal-sized absolute population difference in a larger county, and a larger societal cost in the case of undercounting.

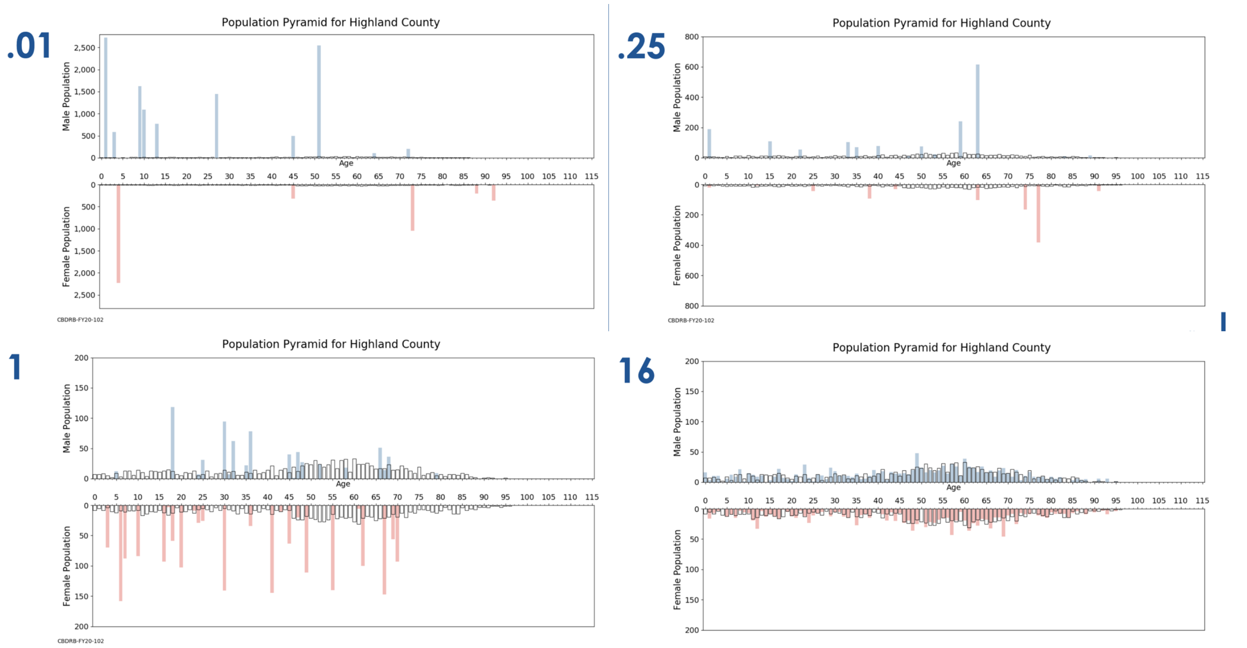

Figure 5: Age and gender population histograms for Highland County. The hollow black boxes show the true population, while the blue and red bars show the noisy count of males and females at each age. Each graph shows the output for a different value of epsilon, which is indicated by the number in the upper left corner. (Figures courtesy of the US Census Bureau [12] session B)

4.2 Empirical Results from Independent Analyses

4.2.1 Homogeneity's impact on reported population counts

Various scholars have conducted their own analyses of the accuracy and privacy of the TopDown algorithm. One such paper attempts to measure the empirical privacy loss of the TopDown algorithm, but in the process also obtains interesting findings about the algorithm's accuracy. Specifically, the authors find that the population of homogeneous areas tends to be over-reported, whereas the population of heterogeneous areas is under-reported on a relative basis. To find this result, they separate out geographic areas based on their homogeneity index, defined as the number of buckets (age/race/ethnicity) that lack a population. Areas with a higher homogeneity index — that is, more empty demographic buckets – have a noisy count that is higher than their true total [9]. On the other hand, areas with a non-zero population in a larger number of demographic buckets will often see their noisy population be slightly less than the true total. The authors hypothesize that this bias shows up due to the non-negativity constraints of the TopDown optimization process [9]. This makes sense when one considers the specific definition of homogeneity used in this paper. Colloquially, we would call a town or census tract homogeneous if a large proportion, say, 90% of the population is the same race, regardless of what the composition of the remaining 10% of the population is. But in the context of this paper, that town would not be considered homogeneous as long as the remaining 10% was comprised of people of various races, ethnicities, and ages. From our knowledge of how people of different ages tend to live near one another, and to a lesser extent, how people of varying races and ethnicities tend to live near each other, we can conclude that most homogeneous areas will have small populations. For example, in a town of 10,000 people, it is very unlikely that there are no white, non-Hispanic men who are 50 years old, whereas this result could easily occur by chance in a town of 100 people. This explanation also makes sense in conjunction with the Census Bureau results we explored in Section 4.1: recall that in Figure 3, the data shows that small towns gain population in the noisy total, whereas larger places' totals get rounded down.

Note that in the TopDown algorithm, a small overall population means that the population in each demographic bucket will also be small or zero. As a result, when we add noise to each entry in the demographic histogram, roughly half of the buckets that were previously zero will now be negative (some of the other entries might be, too). The non-negativity constraints will force all of these to return to zero. By contrast, the starting demographic histogram for a more heterogeneous area will have fewer empty buckets, and therefore fewer negative entries once noise is added. As a result, certain demographic buckets in the homogeneous areas will have their noisy totals rounded up in the optimization stage. This may come at the expense of some demographic buckets in the heterogeneous area that will be rounded down. As a result, the noisy optimized total for the homogeneous area will be larger than its original total, whereas the population for the heterogeneous area will be smaller.

4.2.2 Special populations, population size, urbanization, and reported population counts

In another assessment of the accuracy impacts of the Census's differential privacy algorithm [11], researchers compare 2010 county-level population counts released under the Census's historical disclosure avoidance techniques (i.e., record-swapping, item imputation, whole household imputation, rounding, top- and bottom-coding) and population counts produced with the TopDown algorithm, with a fixed global privacy-loss budget of $\epsilon = 6$, within which personal records use $\epsilon = 4$ and housing records use $\epsilon = 2$. Therefore, any differences between population counts are due to the different disclosure avoidance measures at work.

Overall, the researchers find that the application of differential privacy has especially profound influence on population counts for racial/ethnic minority groups, areas with smaller population, and areas with lower levels of urbanization or adjacency to metropolises. For racial/ethnic minority groups, the resulted change in population counts after applying differential privacy is the smallest for the non-Hispanic white population, compared with every other racial/ethnic group. In contrast, the range of changes in population counts and percentages is much wider for non-Hispanic black and Hispanic groups, sometimes with increases exceeding $1,000\%$ due to the implementation of differential privacy. Furthermore, non-Hispanic blacks experience the steepest reductions in county-level population counts whereas Hispanics experience the largest increases in county-level population counts (Figure 6).

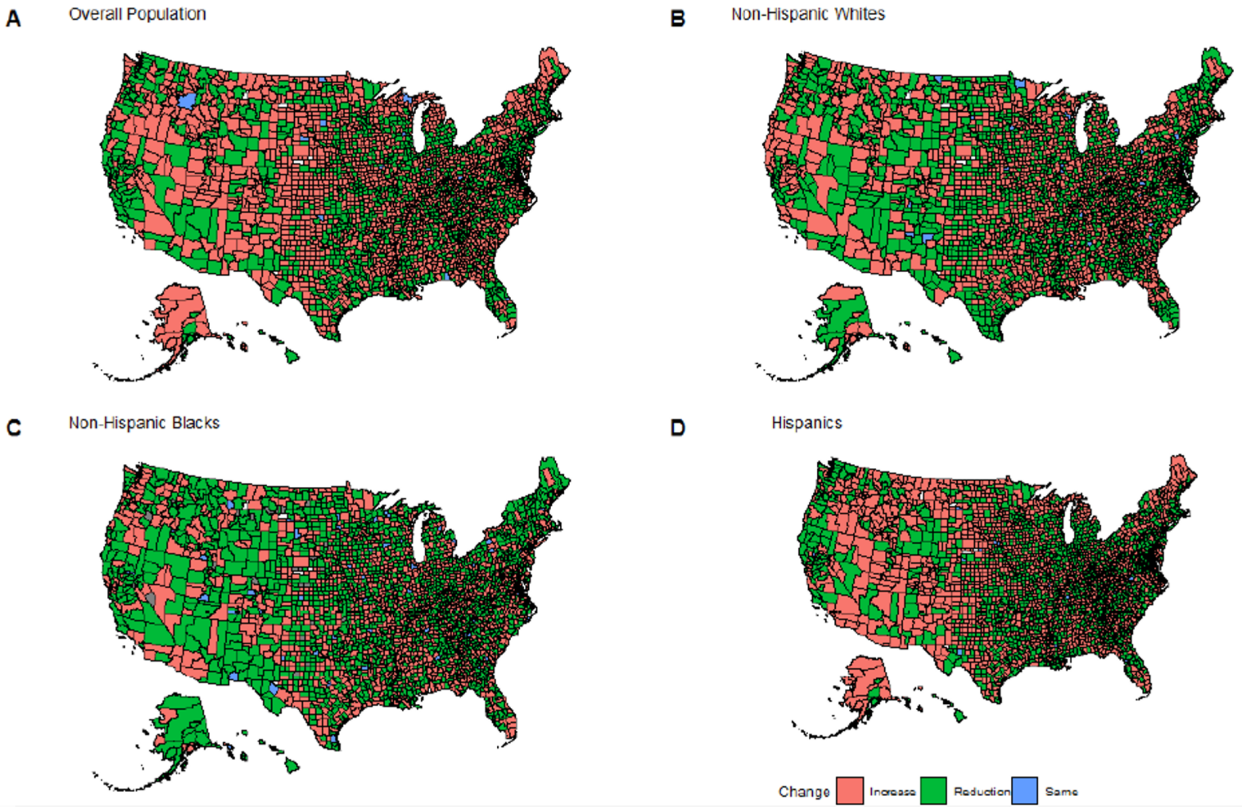

Figure 6: County-level population change for the overall population and three major racial/ethnic group, indicating the population under differential privacy increased (red), remained the same (blue) or decreased (green). Comparisons are between originally published 2010 data and the 2010 data with noise infused by differential privacy (Figures courtesy of Santos-Lozada et al. [11].)

For areas with different population sizes, applying differential privacy to census counts leads to greater population and percentage changes. This effect is especially profound for less populous areas. For the overall population, there is a small increase in population counts in less populous areas. Factoring race/ethnicity into the equation, it is found that implementing differential privacy results in an increase in the Hispanic population in less populous areas. For non-Hispanic whites, there are both increases and decreases in population counts, and the magnitude of change is larger in less populous areas. Similarly, for non-Hispanic blacks, less populous areas show greater levels of variation (Figure 7).

Finally, the researchers evaluate how adopting the differential privacy approach may affect county-level mortality rates. Note that implementing differential privacy can change mortality rates only by changing the denominator, which comes from census population counts, while the numerator is fixed as it comes from vital records. In order to decide whether, and to what extent, the mortality rates calculated using census counts produced with the TopDown algorithm result in artificial increases or deceases compared with rates calculated using counts produced with historical disclosure avoidance techniques, the researchers calculate

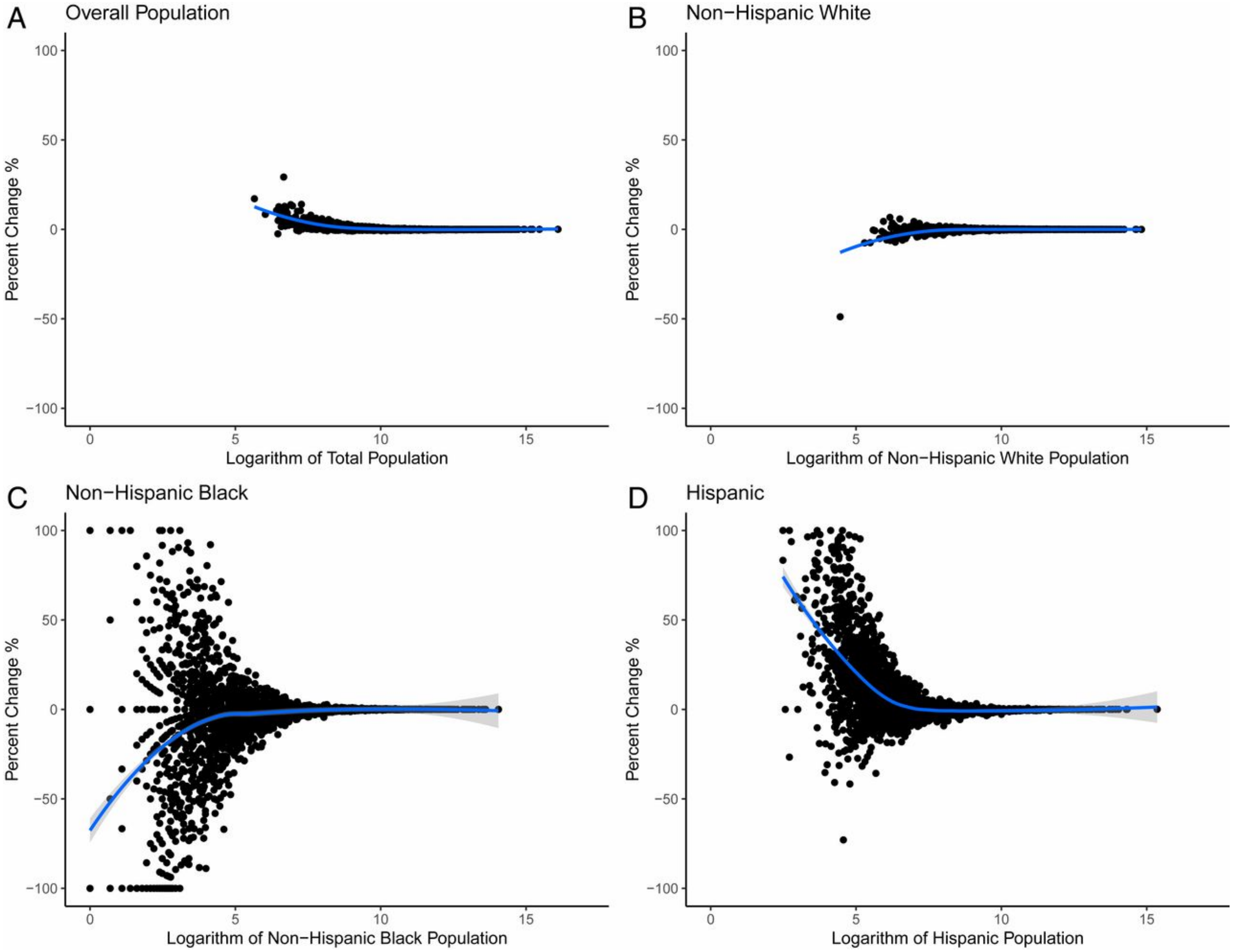

Figure 7: Percent of population change at the county level due to the implementation of differential privacy and 2010 population size (the logarithm scale of 2010 originally published counts): overall population (*A*), non-Hispanic whites (*B*), non-Hispanic blacks (*C*), and Hispanics (*D*). Comparisons are between originally published 2010 data and the 2010 data with noise infused by differential privacy (Figures courtesy of Santos-Lozada et al. [11])

mortality rate ratios (MRR) by dividing 2010 official population counts (M_1) by population counts produced using the TopDown algorithm (M_2), which is then multiplied by 100. Therefore, an MRR greater than 100 means that M_1 is higher than M_2, whereas an MRR less than 100 means that M_1 is lower than M_2. Larger changes in mortality rate estimates are found for areas with lower levels of urbanization or adjacency to metropolitan areas. The magnitude of such changes are in particular larger for non-Hispanic blacks and Hispanics (Figure 8). Additionally, as Figure 8 shows, applying the TopDown procedures to census counts leads to underestimated mortality rates for the Hispanic population in non-metro areas, meaning the size of the Hispanic population is overestimated in these areas.

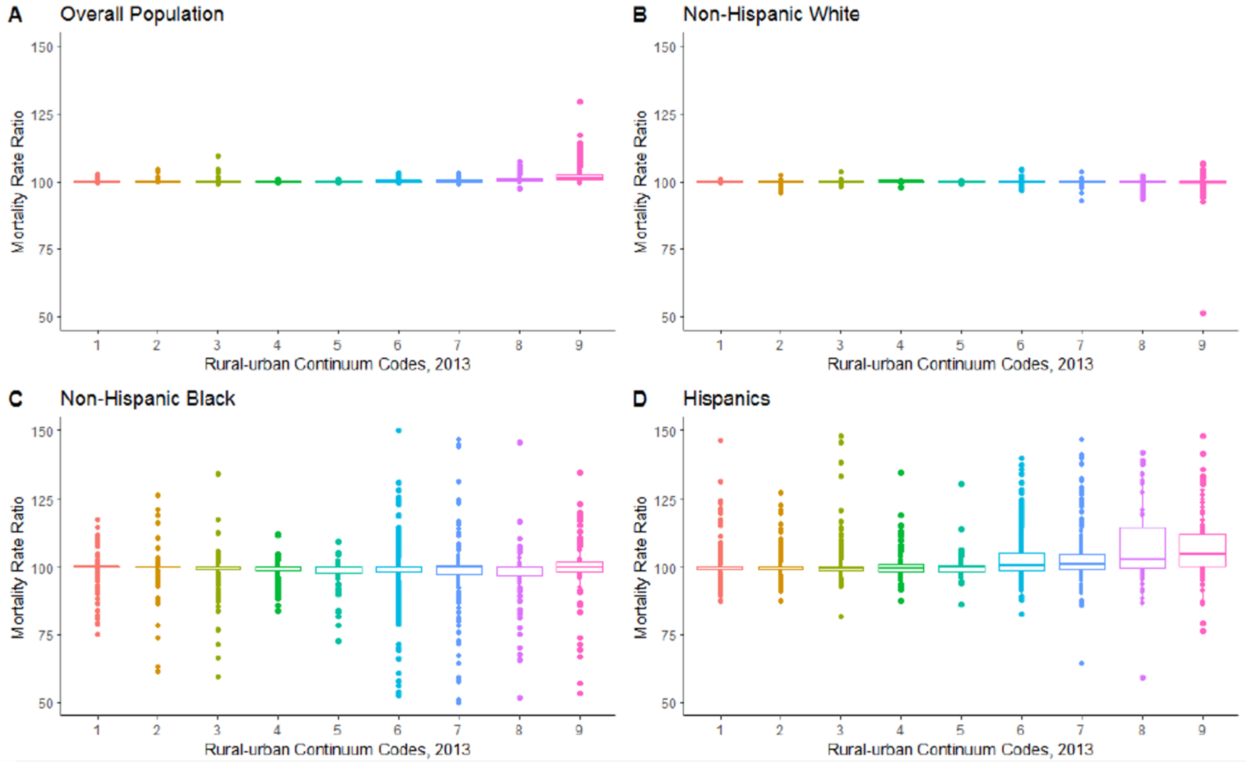

Figure 8: Mortality Rate Ratios for the overall population and three major racial/ethnic groups by 2013 Rural-Urban Continuum Codes (RUCC). MRR comparisons are between originally published 2010 data and the 2010 data with noise infused by differential privacy (Figures courtesy of Santos-Lozada et al. [11]).
**Notes: The USDA Rural-Urban Continuum Codes indicate whether a county is considered: 1. Metro - Counties in metro areas of 1 million population or more, 2. Metro - Counties in metro areas of 250,000 to 1 million population, 3. Metro - Counties in metro areas of fewer than 250,000 population, 4. Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area, 5. Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area, 6. Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area, 7. Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area, 8. Nonmetro - Urban population of 20,000 or more, adjacent to a metro area, 9. Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area.*

Some of these findings have been corroborated by other analyses, as well. For example, in one such independent analysis that looks at the impact of applying differential privacy to counts of the American Indian/Alaska Native tribe populations, DeWeaver finds that the TopDown procedures consistently undercount the American Indian/Alaska Native populations [5]. Furthermore, another analysis finds that the range of change in population counts after applying TopDown is especially wide for less populous American

Indian/Alaska Native reservations or villages [3]. Both studies point out that it is necessary to allocate a nontrivial portion of the privacy loss budget to the American Indian/Alaska Native population, or smaller populations in small areas in general [3][5].

5 Experimental Findings

5.1 A toy experiment

As we saw above, the TopDown algorithm systematically overestimates small populations, and underestimates larger ones. What these data do not tell us, however, is why this disparity occurs: is it an inherent property of differential privacy, a result of the hierarchical structure, or an impact of the optimization process? We initially aimed to explore this question through running the TDA directly, but as described in Section 3.4.4, that did not work out.

Our modified goal with a toy experiment was to try to isolate whether the uneven impact on small versus large communities stems from the "add noise" step or the "optimize" step. To do this, we adapted code written by Abraham Flaxman to measure the empirical privacy loss of the census [6]. The existing code generates a random population sample at various nested population levels. This general setup is a useful model for understanding the basic mechanisms of a hierarchical differentially private dataset, but it assumes the population is approximately evenly distributed across each of the subareas. As a result, running the original code yields generated data where each subarea has a population of about 100 people. Not only does that assumption not carry over to the real world, it also prohibits us from reasoning about the impact that relative population size has on the algorithm's output. We tweaked the synthetic population generation process to cluster the population more tightly in one "large city", several "suburbs" and "small cities", and many small towns and rural areas. The resulting population distribution still is not incredibly realistic, but it effectively captures the contrast between large and small regions.

The provided code has multiple noise generation methods. Of those, we focused in on two: a geometric mechanism that only adds noise to the lowest-level geographic areas, and a hierarchical mechanism that iteratively adds geometric noise at each geographic level, and then smooths the total counts to be consistent with the total count of the parent level. The epsilon that we pass into each function is the same, but we modify the supplied epsilon within the hierarchical mechanism function to ensure that the total privacy budget is equal between the two calculations. We then graphed the true population size against the percent difference with the noisy count. We expected that the version that was calculated hierarchically would show greater percent differences between the exact and noisy counts, but this is not what we found. In Figure 9, we show the percent difference between the actual and noisy counts for each population area. We see that there is no discernible difference in the percent change between the lowest-level geometric mechanism and the hierarchical optimized version.

To dive into these results further, we bucket our data based on exact population size. Figure 10 shows box plots for the first several buckets, with the basic differential privacy result on the left and the hierarchical result on the right. Here, we can visually detect that the middle 50% range is smaller for the hierarchical version. Specifically, the 25th percentile for the regular version is -557% and for the hierarchical version it is -524%. The difference in the 75th percentile is even more extreme: 594% compared to 416%. Some of this disparity may be due to randomness in generating the noisy counts, but repeating the experiment with different random seeds yields similar results.

These results — that is, the lack of difference between the regular and hierarchical differential privacy implementations — suggest that it might be some of the addition constraints, namely non-negativity and integral values, that have a larger impact on accuracy than just the hierarchical consistency constraints.

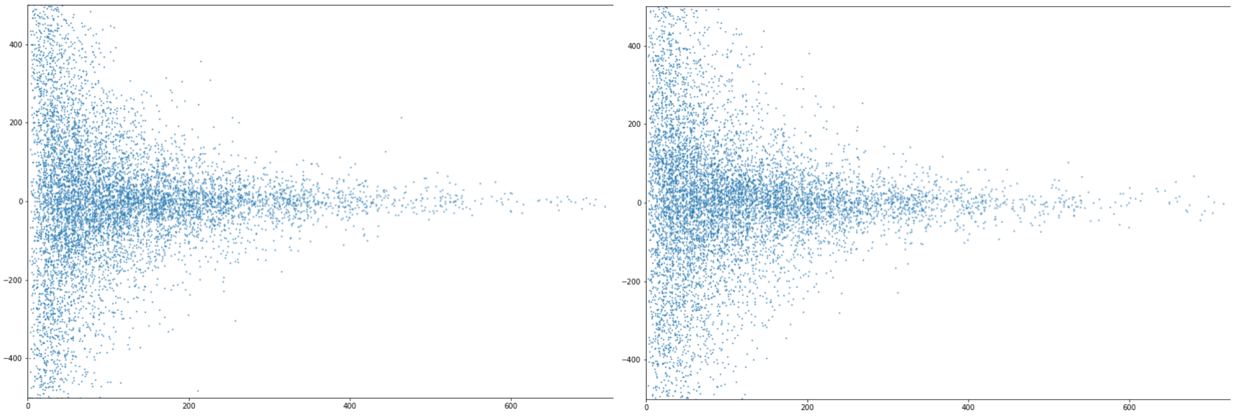

Figure 9: Percent difference in population between the actual and noisy counts. On the left, differential privacy is only applied at the lowest geographic level. On the right, differential privacy is applied at each hierarchical level, and the totals are smoothed for consistency. The x-axis is population (the scale goes from 0 to 700, some larger values are cropped out), and the y-axis is percent difference and goes from -500 to 500.

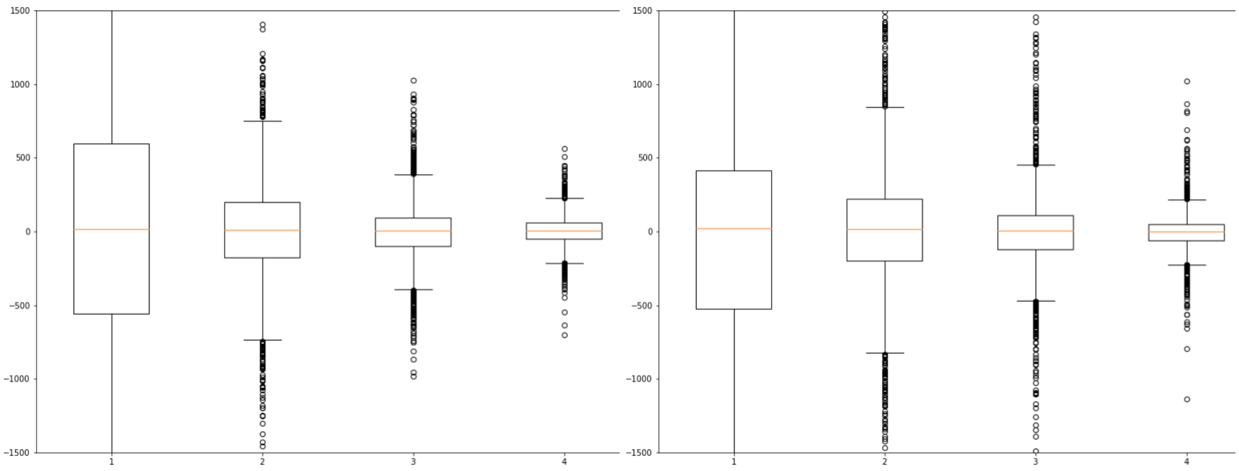

Figure 10: Population buckets and percent difference in population between the noisy and exact counts. On the left, results are from a basic differential privacy mechanism applied only to the lowest geographic area. On the right, the results are from differential privacy applied at each geographic level.

Since this toy model only considers the consistency constraint, it would not show the impact of the other constraints. It has been pointed out by others that the Census could improve accuracy of various statistical applications if they released the non-optimized noisy totals [4]. That data would not make sense to laypeople, but they would not pose any problem for statistical purposes. Additionally, there is no privacy reason why these results should not be public, as the optimization process does not add any additional privacy guarantees.

6 Conclusions and Future Directions

We investigated the design and implementation of the differentially privacy algorithm, TopDown, that will be used to process and release the 2020 US Census data. The algorithm adapts the standard differential privacy approach to work with various constraints, including a hierarchical data structure and invariants. As complicated as it is to get differential privacy working with all the appropriate theoretical bounds in this setting, a larger computational challenge arises when attempting to make the data internally and externally consistent. It is likely through this optimization process that various inaccuracies are exacerbated.

Through our literature review, we found that the TDA provides relatively good accuracy for large populations, or at high geographic levels. However, for small populations (across an entire geographic area, or just stratified by demographics), the accuracy can be quite bad. In particular, small homogeneous populations are generally inflated. As a result, some other populations necessarily must be underreported by the algorithm, and this works out to be more diverse and urban communities. We note that this is problematic: given the high-stakes of many census data applications (e.g., federal funding), having an algorithm that systematically favors certain groups is not acceptable.

Another issue with the Census Bureau's application of differential privacy is that there is a general lack of transparency and public education on both why differential privacy is necessary and what its limitations are. Certain criticisms of the Bureau's use of differential privacy [10] seem to lack basic grounding in privacy and re-identification risks. That is not to say that their concerns about the impacts on their own research (basically, different and less data will be available) are not well-founded or important, but rather it speaks to the general lack of communication between the Census Bureau and stakeholders [4]. Our own experience trying to run the open-source code also speaks to this need for better public communication and documentation. Finally, the continued lack of guidance on what value of epsilon will be used is another shortcoming, because stakeholders have less knowledge about how to prepare for the impacts on their data use cases.

Aside from those policy and communication changes, we also encountered various areas of study that could improve accuracy. Among the most promising was an empirical study [9] that found that the theoretical privacy loss of the TDA is significantly less than the actual privacy loss. If researchers are able to tighten the theoretical bound, this would allow for a larger value of epsilon to be used for the same privacy protection as is currently promised. Similarly, further refinement and adaptation of the techniques used by the Census in the TopDown Algorithm could be useful for other instances where it is desirable to apply differential privacy to large scale data while maintaining consistency with public knowledge.

Overall, the Census has been dealt a challenging hand - preserving the privacy of the gathered information while balancing the necessity for accurate and easy to interpret data. Its decision to use differential privacy is certainly a desirable step in the right direction towards addressing this. In addition, its contributions to the field of differential privacy in the form of novel algorithmic solutions to address constraints, hierarchical data, and the large scale of the data are laudable. However, as explored in this paper, their approach has many unintended and possibly harmful side effects that the Census must be able to, if not mitigate directly, at least clearly communicate their existence and impacts.

References

[1] Abowd, J. M. Tweetorial: Reconstruction-abetted re-identification attacks and other traditional vulnerabilities. http://blogs.cornell.edu/abowd/special-materials/245-2/.

[2] Abowd, J. et al. Census TopDown Algorithm: Differentially private data, incremental schemas, and consistency with public knowledge, 2019.

[3] Akee, R. Population counts on American Indian reservations and Alaska Native villages with and without the application of differential privacy, Dec 2019.

[4] boyd, d. Balancing data utility and confidentiality in the 2020 US Census.

[5] DeWeaver, N. Impact of DP on AI/AN tribes, Dec 2019.

[6] Flaxman, A. D. Github repository (empirical quantification of privacy loss with examples relevant to the 2020 US Census), 2019.

[7] Garfinkel, S. L., and Leclerc, P.

[8] McKenna, L. Disclosure avoidance techniques used for the 1960 through 2010 decennial censuses of population and housing public use microdata samples, 2019.

[9] Petti, S., and Flaxman, A. Differential privacy in the 2020 US Census: what will it do? Quantifying the accuracy/privacy tradeoff.

[10] Ruggles, S., Fitch, C., Magnuson, D., and Schroeder, J. Differential privacy and census data: Implications for social and economic research. In *AEA papers and proceedings* (2019), vol. 109, pp. 403–08.

[11] Santos-Lozada, A. R., Howard, J. T., and Verdery, A. M. How differential privacy will affect our understanding of health disparities in the United States. *Proceedings of the National Academy of Sciences* (2020).

[12] The National Academies of Sciences Engineering & Medicine. Workshop on 2020 Census data products: Data needs and privacy considerations, Dec 2019.

[13] US Census Bureau. Disclosure avoidance and the 2018 census test: Release of the source code, June 2019.