Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the U.S. Census

Miranda Christ*
Department of Computer Science
Columbia University
https://www.cs.columbia.edu/~mchrist

Sarah Radway*
Department of Computer Science,
Fletcher School of Law & Diplomacy,
Tufts University†
https://sites.tufts.edu/sradway/

Steven M. Bellovin
Department of Computer Science
Columbia University
https://www.cs.columbia.edu/~smb

2022 IEEE Symposium on Security and Privacy (SP) | 978-1-6654-1316-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/SP46214.2022.9833668

***Abstract*—There has been considerable controversy regarding the accuracy and privacy of de-identification mechanisms used in the U.S. Decennial Census. We theoretically and experimentally analyze two such classes of mechanisms, swapping and differential privacy, especially examining their effects on ethnoracial minority groups.**

We first prove that the expected error of queries made on swapped demographic datasets is greater in sub-populations whose racial distributions differ more from the racial distribution of the global population. We also prove that the probability that m unique entries exist in a sub-population shrinks exponentially as the sub-population size grows. These properties suggest that swapping, which prioritizes unique entries, will produce poor accuracy for minority groups.

We then empirically analyze the impact of swapping and differential privacy on the accuracy and privacy of a demographic dataset. We evaluate accuracy in several ways, including methods that stress the effect on minority groups. We evaluate privacy by counting the number of re-identified entries in a simulated linkage attack. Finally, we explore the disproportionate presence of minority groups in identified entries.

Our empirical findings corroborate our theoretical results: for minority representation, the utility of differential privacy is comparable to the utility of swapping, while providing a stronger privacy guarantee. Swapping places a disproportionate privacy burden on minority groups, whereas an ϵ-differentially private mechanism is ϵ-differentially private for all subgroups.

1. Introduction

An urgent need for comparative analysis of the data de-identification methods of swapping and differential privacy has emerged with the U.S. Census Bureau's controversial decision to transition from swapping to differential privacy in the 2020 Census. After swapping as used in the 2010 U.S. Census was shown to be vulnerable to reconstruction attacks [1], the U.S. Census Bureau adopted differential privacy, which affords stronger privacy guarantees. This decision sparked an ongoing debate about the utility of differentially private census data.

Policy decisions are rooted in demographic data. From the allocation of economic assistance funding to infrastructure and voting implementation, the data collected, analyzed, and released by groups such as the U.S. Census Bureau affects the lives of all residents. The use of demographic data allows for highly targeted and effective decision-making by governments and by industry worldwide. However, accurate large-scale datasets of personal information come at a price: the price of privacy.

The privacy-utility trade-off is the relationship between data accuracy and privacy. It represents the notion that with increased privacy, decreased utility necessarily follows, and vice versa. This phenomenon is important to consider in the context of demographic data. Specifically, privacy is integral to maintaining a safe society for minority groups. Consider the role of census data in the establishment of Japanese internment camps: U.S. Census Bureau block data was used to target Japanese-Americans for imprisonment during World War II [2]. Census data, with sensitive fields intact, provides easy access to identifying information about specific individuals, and has the power to inflict significant harm on all represented communities.

Furthermore, privacy is necessary for achieving a high response rate and obtaining accurate responses. As noted in Baldrige v. Shapiro, "an accurate census depends in large part on public cooperation[,]" relying on "assurances that information furnished to the Secretary [of Commerce] by individuals is to be treated as confidential" [3].

Groups such as the U.S. Census Bureau lessen this identification risk by de-identifying datasets. The most recent method of de-identification chosen for use in the U.S. census is differential privacy. Differential privacy (DP) is a relatively new formal privacy definition, spanning a class

*These authors contributed equally to this work.
†Work done while at Columbia University and Tufts University.

DOI 10.1109/SP46214.2022.00135

of de-identification algorithms that serve as a means to satisfy the evolving need for anonymization and accuracy. Throughout this paper, we examine the legitimacy of DP's application to demographic data, specifically regarding minority populations within a dataset.

There has been significant opposition regarding the adoption of DP in the U.S. census. Critics claim that DP produces less accurate population counts, and diminishes minority representation. The accuracy of demographic data is especially important, as its use cases are sensitive: U.S. census data is used for funding integral assistance programs including Medicaid, Head Start, SNAP, and block grant programs for community mental health services [4]; under-representation can lead to under-funding of national and block-based assistance programs. Groups such as the National Congress of American Indians (NCAI) have cited concern over DP's impact on small populations, and have pushed back on its use [5]. The criticisms of groups such as NCAI are valid, in that by nature, DP does introduce inaccuracy into a dataset; DP generally offers less utility for smaller population subsets than for larger population subsets. However, these criticisms do not consider whether this phenomenon occurs in the de-identification methods that preceded DP.

We compare DP and its predecessor: the data perturbation method used in the 2010 U.S. Census, known as swapping. First, we prove that the expected error in count queries performed on a dataset de-identified by random swapping is higher for subpopulations with distributions further from the global population distribution, e.g., diverse block groups within more homogenous states. We then prove that the probability that a fixed number of unique entries exist in a subpopulation decreases exponentially as the subpopulation size increases. Thus, we expect the accuracy of swapping to degrade as subpopulation size shrinks, and we expect swapping methods that prioritize unique entries to be swapped to exhibit more extreme accuracy degradation in small subpopulations.

We then empirically evaluate the accuracy and privacy of datasets produced by both swapping-based and DP mechanisms. Our implementations are largely inspired by the U.S. Census Bureau, due to their importance as a user of DP for demographic data. We use a geometric mechanism in the central differential privacy model, similar to a simplified version of the U.S. Census Bureau's TopDown Algorithm without post-processing. We implement two standard approaches to swapping, including both a randomized and similarity-threshold based metric for swap selection; we run these swapping mechanisms for various similarity thresholds and with various approaches to randomness. We measure the accuracy of these mechanisms by computing histograms by race over each dataset type, and comparing these histograms to the true race histograms. We compare histograms using three metrics: a variant of KL divergence, mean squared error, and the number of minority race-ethnicity groups whose populations significantly decreased. We analyze the privacy of the swapped data by simulating a linkage attack and counting the number of identified entries.

We show both theoretically and empirically that when DP and swapping are implemented at an acceptable privacy level, their utility with regard to minority representation is comparable for demographic datasets similar to and similar in dimensionality to the U.S. census redistricting data. We present the accuracy and privacy levels found across a wide range of ϵ values and swap rates, for block groups in counties of various demographic makeups.

2. Background and Related Work

Data swapping was introduced by Dalenius and Reiss in [6]. It involves selecting pairs of rows in a database and, for each pair, exchanging the values in a subset of their columns. The number of entries that are swapped is parameterized by the *swap rate*: a higher swap rate means more exchanges and more privacy.

Differential privacy [7] is a rigorous privacy definition. Broadly speaking, a de-identification mechanism is differentially private if for any two datasets differing at only one row, the distribution of the mechanism's output is roughly the same. The amount that the distributions may differ is parameterized by the privacy parameter, ϵ. DP is achieved using randomized mechanisms, which typically introduce structured noise (e.g., additive noise from a Laplacian distribution) to de-identify data.

In 2020, the U.S. Census Bureau switched from using traditional disclosure avoidance methods (primarily swapping) to using the differentially private TopDown Algorithm (TDA) [8]. This decision has been criticized by data users concerned about a decrease in data utility due to inaccuracy introduced by DP [9], along with minority groups concerned about population undercounts [5], [10]. A lawsuit was filed by the State of Alabama against the U.S. Department of Commerce and the U.S. Census Bureau; in part of this lawsuit, the State of Alabama argued that the inaccuracy introduced by the use of DP in the 2020 U.S. Census will make the data unusable for redistricting. Alabama lost on purely legal grounds; the data quality issue was not addressed.[1]

It is important to note that while the choice of de-identification method for the 2020 U.S. Census is up for debate, the U.S. Census Bureau's move away from swapping as implemented in the 2010 U.S. Census was imperative. Since the law mandates that census data not be individually identifying [11], the 2010 implementation of swapping, which was shown to be vulnerable to reconstruction attacks [1], is no longer viable. When many queries (i.e., aggregate statistics) over confidential data are released, such as in the U.S. census, a threat to privacy is database reconstruction, where an attacker uses the answers to the queries to rebuild the confidential database. Dinur and Nissim [12] showed that for a dataset represented using n bits, noise of magnitude $\Omega(\sqrt{n})$ must be added to prevent reconstruction attacks. In

1. https://www.brennancenter.org/our-work/court-cases/alabama-v-us-dept-commerce lists all major filings, rulings, etc. As of September 9, 2021, Alabama had withdrawn its suit and not refiled.

practice, de-identification methods, likely including those of the U.S. Census Bureau before 2020, often do not use this level of noise, suggesting that many publicly available de-identified datasets are vulnerable to such attacks. This vulnerability was confirmed for 2010 U.S. Census data by the U.S. Census Bureau in 2018, when their team of researchers launched a reconstruction attack on the publicly available census data and found that 46% of their reconstructed entries exactly matched the corresponding original entries [13, page 32].

We are not the first to examine the impact of DP mechanisms on minority subpopulations in the context of the U.S. census. In particular, several studies use the U.S. Census Bureau's 2010 demonstration data, which includes the 2010 U.S. Census data de-identified using a version of their TopDown Algorithm, the mechanism designed by the U.S. Census Bureau to produce differentially private tabular summary statistics. These studies compare the 2010 TopDown-Algorithm-produced data to the official 2010 U.S. Census data, which was de-identified using traditional disclosure avoidance methods (i.e., their swapping algorithm). In [14], Santos-Lozada et al. compare mortality rates computed using the 2010 TDA-produced data and the official 2010 U.S. Census data. They find higher levels of discrepancy in mortality rates for minority categories (e.g., Non-Hispanic Black) as compared to more populous categories (e.g., Non-Hispanic White). In [15], Hauer and Santos-Lozada find substantial distortion in COVID-19 mortality rates computed using differentially private data for small population groupings. Again, they use the 2010 U.S. Census data, which was de-identified using swapping, as a baseline.

Their results show that swapping and the differentially private TopDown Algorithm produce *different* data, but they do not show which method yields more accurate data. A direct comparison of the impact of swapping and DP mechanisms on U.S. census data is challenging since the unmodified data cannot be released due to statutory requirements [11]. We bypass this issue by generating synthetic data as our ground truth data and de-identifying this data using both swapping and DP.

Other works examine the TopDown Algorithm's general fitness for use, not necessarily focusing on minority groups. In [16], Wright and Irimata simulate the TopDown Algorithm with $\epsilon = 4$ on Rhode Island data and find increased variability in smaller subpopulations. In [17], Cohen et al. develop an experimental framework for analyzing TopDown Algorithm's impact on redistricting data, propose improvements for the Census Bureau and data users, and find that some concerns about usability can be overcome. Garfinkel et al. [18] attribute some of the challenges surrounding the adoption of DP to data users' unfamiliarity with DP, and they highlight a need for improved communication between data users and the Census Bureau. In [19], Kenny et al. argue that the TopDown Algorithm is both too inaccurate to be used for redistricting and not sufficiently private. However, this paper has been contested by DP experts [20].

The effect of DP on minority groups has been examined in other contexts as well. Bagdasaryan, Poursaeed, and Shmatikov [21] show that for neural networks trained using DP-SGD, the accuracy cost is greater for minority groups, and any accuracy disparity in non-DP training is exacerbated when DP is applied. Xu, Du, and Wu [22] give a modified differentially private stochastic gradient descent algorithm, dubbed DPSGD-F, that mitigates this cost disparity. This line of research is interesting and relevant, but not easily applicable to our setting: DP applied during the training of machine learning models will face different challenges than DP applied to histograms over demographic data.

Separately, there has been work examining the impact of data swapping. Hawes and Rodríguez [23], using 1980 U.S. Census data, show that even when 50% of households are swapped, 12.96% of the population can be re-identified. Kim [24] shows that swapping methods that prioritize at-risk individuals degrade the manifestation of jointly distributed variables in the de-identified data. Ramchandran et al. [25] simulate a linkage attack to correlate American Community Survey and Public Use Microdata Sample data, which were de-identified using swapping and released by the U.S. Census Bureau.

We are the first that we know of to directly compare the effects of swapping to the effects of DP on minority underrepresentation across a wide range of ϵ values and swap rates. We do so by generating synthetic microdata using 2010 U.S. Census data, de-identifying our synthetic data using swapping and DP mechanisms, and comparing the accuracy and privacy levels afforded by these methods at varying swap rates and ϵ values. We use a DP mechanism similar to the Census Bureau's Top-Down Algorithm, but without U.S.-census-specific post-processing. Thus, our results apply more generally to demographic data of a similar structure.

De-identifying demographic data in a way that preserves minority representation is relevant to contexts beyond the U.S. census. For example, Ito et al. [26] compare several DP mechanisms for computing official population statistics in Japan. In Israel, survey data released to the public by Israel's Central Bureau of Statistics was re-identified by a group of students at Tel Aviv University [27]. While we use the U.S. census as a case study, our results help inform the global issue of preserving minority representation while ensuring sufficient privacy levels in publicly available demographic datasets resembling the U.S. Decennial Census. We perform additional analyses to show how our methods do and don't generalize to higher-dimensional data.

2.1. Definitions

We include the definition of ϵ-differential privacy from [7]:

Definition 1 (ϵ-differential privacy [7])**.** A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is ϵ-differentially private if for all $S \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $||x - y||_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(y) \in S]$$

One accuracy measure used in this paper is the μ-smoothed Kullback-Leibler divergence (D^{μ}_{KL}) from Tantipongpipat et al. [28]:

Definition 2 (D^{μ}_{KL} [28])**.** For small $\mu > 0$, the μ-smoothed KL divergence between an original distribution P and an altered distribution Q is

$$D^{\mu}_{KL}(P||Q) := \sum_{x \in supp(P)} (P(x)+\mu) \log\left(\frac{P(x)+\mu}{Q(x)+\mu}\right)$$

3. Theoretical Analysis

We first prove that if entries are chosen uniformly at random to be swapped, the expected error of a counting query performed over a subpopulation is higher for subpopulations whose distributions differ more from the global population distribution.

In this section, by random swapping with swap rate κ, we mean the de-identification method in which each row is chosen to be swapped with probability κ. If a row r is designated to be swapped, r is replaced with the values of another row r' chosen uniformly at random from the entire dataset. While some swapping methods replace only a subset of the values rather than the entire row, these results apply to such methods, since we can restrict the database to include only the attributes that are swapped. In practice, most implementations *swap* r and r'. Since we examine this method's effect on only a subset of the dataset, if r is in this subset and r' is not, it is not relevant that r' is replaced with r unless r' was also in this subset. If our subset is small, this happens rarely. Thus we can treat r as being replaced for the sake of simplifying the analysis while still maintaining a realistic model.

We examine the effect of random swapping on the accuracy of counting queries (e.g., the size of the Asian population) on a small subset of the dataset.

Let $\mathcal{D}$ be a dataset. Let $S \subseteq \mathcal{D}$ be a subset of the rows in $\mathcal{D}$. In U.S. census data, S may represent a block, for example. Let S' denote the dataset obtained by taking the output of the swapping mechanism on input $\mathcal{D}$, and restricting it to the rows in S. If S represents a block in a dataset, S' is that block's data after swapping is carried out. Suppose we are interested in the number of rows that fall into some category; e.g., we wish to know how many residents in our block are Asian. Let $c(\cdot)$ denote the number of rows in database $\cdot$ that are in category c.

Theorem 1 (Random Swapping and Counts)**.** Let $n = |S|$ be the number of rows in S, $\alpha = \frac{c(S)}{|S|}$ be the fraction of rows in S in category c, and $\beta = \frac{c(\mathcal{D})}{|\mathcal{D}|}$ be the fraction of rows in the entire dataset $\mathcal{D}$ in category c. Then:

$$\mathbb{E}\left[\left|\frac{c(S')}{|S'|} - \frac{c(S)}{|S|}\right|\right] = \frac{|n(\alpha+\kappa(\beta-\alpha)) - n\alpha|}{n} = |\kappa(\beta-\alpha)|$$

The proof is by linearity of expectation and algebraic manipulation. For interested readers, we have included it and the proof of Theorem 2 in Appendix A.

This difference increases with $|\beta-\alpha|$. Thus, subpopulations with different distributions than the overall dataset will have a higher accuracy loss of count statistics compared to subpopulations distributed similarly to the overall dataset. For example, random swapping will significantly diminish the size of the Asian population in a predominately Asian block if the proportion of Asian people is much greater within the block than in the whole database. It also increases with the swap rate κ, meaning as more entries are swapped, this difference becomes more pronounced.

Next, we analyze the probability that a subpopulation S of size n has at least m unique rows. A row is unique if there exists no other row in the dataset with that row's attribute combination.

We model the data as having each row drawn i.i.d. from some underlying joint distribution $A_1, \ldots, A_k$ over the attributes with support $\mathcal{A}$. Since the rows are drawn independently, we can let R be a random variable representing an arbitrary row and let $P_i := \Pr_{A_1,\ldots,A_k}[R = r_i]$ denote the probability that R equals a given row r_i.

Theorem 2 (Unique Rows and Population Size)**.** Let $\mathcal{E}_1$ be the event that any m of the n rows in the dataset are unique. Let $\mathcal{E}_2$ be the event that the first m rows in the dataset are unique. Then the following hold:

$$\text{(1) } \Pr[\mathcal{E}_2] = \sum_{\{r_j\}_{j\in[m]} \subseteq \mathcal{A}} m! \left(\prod_{i=1}^{m} P_i\right)\left(1 - \sum_{i=1}^{m} P_i\right)^{n-m}$$

$$\text{(2) } \Pr[\mathcal{E}_2] \leq \Pr[\mathcal{E}_1]$$

$$\text{(3) } \Pr[\mathcal{E}_1] \leq \sum_{\{r_j\}_{j\in[m]} \subseteq \mathcal{A}} n^m \left(\prod_{i=1}^{m} P_i\right)\left(1 - \sum_{i=1}^{m} P_i\right)^{n-m}$$

The proof is straightforward, by computing the relevant probabilities and applying a union bound for (3). See Section A.2.

As n grows, the $\left(1 - \sum_{i=1}^{m} P_i\right)^{n-m}$ term will decrease faster than the n^m term increases (recalling that m here is a constant), and the $\left(\prod_{i=1}^{m} P_i\right)$ term will stay fixed. $\Pr[\mathcal{E}_2]$ also decreases exponentially in n. Thus, we have bounded the probability that any m rows are unique between two functions that decrease exponentially in n.

Asymptotically, large groups have an exponentially lower probability of having m unique rows than small groups for any fixed m. Thus any de-identification method whose inaccuracy level increases with the number of unique entries will swap more entries and consequently introduce more inaccuracy in small populations. This applies to swapping as implemented in the U.S. census: in the 2000 U.S. Census, "the probability of swapping was increased to those cases where disclosure risk was thought to be higher such as cross-tabulations of key variables, smaller blocks, and also households that contained unique races in that census block" [13, Page 28]. The implementation in the 2010 U.S. Census was "largely similar."

While useful for analyzing the general behavior of this phenomenon, asymptotics are not as informative for specific values of n. For this reason, we also evaluate these methods empirically.

4. Empirical Evaluation Method

We empirically evaluate the accuracy and privacy impact of swapping and differentially private mechanisms. Our analysis focuses on minority representation in the context of the privacy-utility trade off; we evaluate the under-representation of minority groups following the use of mechanisms resembling those of the U.S. census, run on synthetic demographic data.

4.1. Creating Data

Exact U.S. census data records can only be publicly released after 72 years [29], thus we created synthetic data that preserve key characteristics of modern communities.

The smallest unit of organization of U.S. census data is a *block* [30]; the next-smallest unit is a called a *block group*. Block groups are clusters of blocks, with populations totaling between 600-3,000 [31]. The U.S. Census Bureau recommends that data users aggregate blocks together for improved accuracy [32]; thus we chose to create our data in block group format.

We created synthetic data with populations of 600-1,500 per block group, based upon published block and county data from the 2010 Census. We standardized our block group sizes to lessen the impact of block group size on results, as we primarily examine the impact of minority group status on accuracy and privacy.

We selected nine counties from across the U.S., of varying degrees of diversity. We retrieved published data regarding the age and sex of county residents from the 2010 Census, as well as household size and tenure data of residents from the 2019 American Community Survey. For our race data, we used published 2010 Census block data: we randomly selected blocks within each county, combining data from multiple blocks when necessary to create block groups of our desired size. Through doing so, we aimed to preserve the idiosyncratic nature of race data on a low level, data that we worried would be lost at county scale.

While our data does not exactly equal any block group's true values, it closely resembles U.S. Decennial Census data, and we treat it as the ground truth data for the remainder of our work. This allows us to evaluate the effectiveness of the de-identification mechanisms using the same ground truth data. Each individual is represented by exactly one row in our dataset. Each attribute (age, sex, Hispanic Y/N, race, household size, household tenure) is represented by a column (table in Appendix B for convenience). Thus our ground truth dataset for a given block group has n rows and 6 columns, where n is the number of people in that block group. Our data closely mirrors the fields collected in the U.S. Decennial Census; see Appendix B and [33].

4.2. Swapping Implementations

The exact swapping implementation used by the Census Bureau has not and will not be released, as the Census Bureau has shown that this will allow for potential re-identification [34]. In order to account for this uncertainty, we examine two general approaches to swapping: one that swaps data with a *random* record, and one that swaps data with a *similar* record, defined by a similarity threshold. We believe our work to be fair generalizations of the methods likely implemented by the Census Bureau.

Based upon the work of Kim [24], we created an adapted version of swapping for our synthetic data, requiring the selection of a swap rate (the percentage of the data to be swapped), and a variable to swap on. The data about an individual intended to be swapped can come from inside of the dataset or from a separate dataset. In the context of the U.S. census, the data should come from a different census block in the same county or state. We chose to swap from a one-million member synthetic data sample of the appropriate state, including race, age, etc. We created these synthetic state datasets by sampling from the true distributions of the state for each attribute query.

In choosing entries to be swapped, we prioritize unique dataset entries, as these require the greatest intervention to preserve privacy, and such prioritization has been suggested in documents published by the Census Bureau [13].

All of our swapping approaches entail the exchange of an entry marked for swapping with another entry within the same state. For comprehensiveness, we use several implementations varying on the choice of entry to swap with. We ran each of these implementations at varying swap rates, ranging from .01 (swapping 1% of the data) to 1.0 (swapping 100% of the data) for each of the blocks. Note that although swapping affects two entries, we always swap an entry within our block group with an entry outside the block group. Thus, the swap rate is the percentage of entries in the block group that have been swapped.

Random Swapping. Our *true random* approach involves exchanging the data marked for swapping with a uniformly random row from the true state distribution.

Our *pseudo-random* approach involves drawing a uniformly random entry from the subset of the state dataset with matching (or nearly matching) values for the attributes of age and sex. We refer to this as *swapping on* the age and sex attributes, similar to Algorithm 2's implementation of swapping on the number of individuals in the household in the work of Kim [24]. This means that when swapping an entry, we preserve the true values of its age and sex characteristics, and pull an individual from our state distribution that possesses the same values for these characteristics, with a race randomly drawn from the state's distribution.

Similar Swapping. *Similar swapping* involves swapping an entry with another entry within some similarity threshold. If there are multiple entries within this threshold, we choose one uniformly at random; if none fit, we expand the threshold until matches are available.

We define similarity notions for race, age, and sex. For race (with Hispanic origin encoded), we define similarity based on the frequencies of the races in the United States

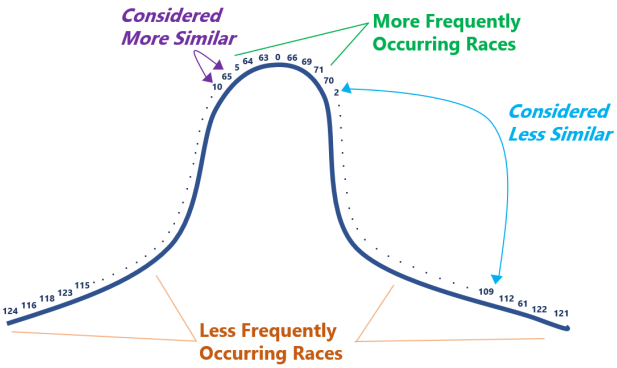

Figure 1: Method of frequency-based race ordering.

as a whole. A populous race is similar to another populous race but not to a lower population race. This distribution is outlined in Figure 1.

For age, the threshold represents the number of years by which an entry's age can differ from the age of the entry with which it is swapped. For sex, a similarity threshold of 1 or greater simply means that sex is not a match constraint. For example, we could set a similarity threshold of 1; this would mean that a row with age 21, sex of female, and race encoding 64 could be swapped with an individual of age 21 ± 1 (20, 21, 22), sex M or F, and a race of 64, 5, or 63 (see Figure 1). In our runs, we swapped on age, sex, and race; we ran mechanisms at similarity thresholds of 0, 1, and 3. In the remainder of the work, we refer to these mechanisms as 0-Similar, 1-Similar, and 3-Similar. We chose these thresholds to capture a range of possible implementations, as will be examined in Section 4.4 and Section 4.5.

4.3. Differential Privacy Implementation

We implemented a simpler version of the TopDown Algorithm, which omits post-processing and uses fewer queries. We designed our version to be easier to analyze and more easily generalizable to applications beyond the U.S. census.

In our mechanism, we first compute counting queries over subsets of the population. We construct our queries to simulate similar queries used in the 2010 Demonstration Data Product [35, Table 2.1]. In our queries, we use age buckets of varying sizes. For example, age buckets of size 15 means individuals are grouped into age ranges 0-14, 15-29, etc. Our mechanism is parameterized by the number of age buckets; for the given grouping into buckets, it outputs the following noisy queries. For each combination of (age bucket) x (Hispanic Y/N) x (Census race 0-62) x (household size 1-4), we compute the size of the population with these attribute values. We then add noise drawn from a Geometric distribution to each of these counts, using the Geometric class from IBM's Differential Privacy Library [36], [37]. Our mechanism sometimes yields negative values, which we map back to 0 when measuring accuracy. This post-processing does not affect the privacy of our mechanism.

We publish these noisy counts for each query of this form; in this model, noise is added to the output of each query on the dataset, and the output of the mechanism is a series of queries and answers. The set of queries to be made must be known during the data de-identification process, and noise is added only when queries are made. For example, if the mean of a set of values is queried, a mechanism in the centralized model may compute the true mean and add noise to the output.

Our mechanism adds noise to each of these queries. Our additive noise value k is drawn with the following probability, where $\alpha = e^{\epsilon}$: $\mathsf{Geo}(\alpha) = \left(\frac{\alpha-1}{1+\alpha}\right) \alpha^{-|k|}$.

Since the subsets over which our counting queries are performed are disjoint, and the geometric mechanism operating on each query is ϵ-differentially private [38], [39], our mechanism is ϵ-differentially private.

This implementation reflects a baseline for TDA capability. We are limited by data and processing power, and are thus incapable of exactly replicating TDA. The Census Bureau ran TDA on 21 AWS r5.24xlarge instances, at $6.048 per hour [40], [41]. To calculate one run for one epsilon value, all instances needed to be run for almost 25 hours, costing $3,175. Our work would require 5,000 of these runs (as we ran 200 epsilon values 25 times), which would increase our costs to millions of dollars.

TDA computes many overlapping counting queries at various levels of granularity (e.g., block level, county level, state level, country level), leveraging this hierarchy of queries to improve accuracy [42]. We would thus expect similar or improved accuracy from TDA compared to our algorithm, given TDA's ability to draw from higher level query data and adjust. In addition to using a wider array of queries, TDA involves extensive post-processing to comply with U.S. census-specific requirements. For example, the total state counts must be exact. We perform only minimal post-processing: when computing the D_{KL}^{μ}, we round negative population counts to 0. Though we do not reproduce TDA exactly, our implementation captures the general behavior of TDA, and generalizes to demographic data beyond the U.S. census.

4.4. Accuracy

In order to evaluate the accuracy of data produced by various DP and swapping mechanisms, we compute the accuracy of a histogram of the de-identified race data. This histogram is formatted 2 (Hispanic) x 63 (census race). For the original and swapped data, this histogram is computed by counting the number of individuals in each of these categories. For the differentially private data, this histogram is computed by aggregating the counts of the individuals in each of these categories, summing across all age buckets and household sizes. We treat the histogram as a vector of length 126 for accuracy evaluation.

Metrics. For each de-identified dataset (i.e. at each swap rate and each epsilon value and age bucket count), we use

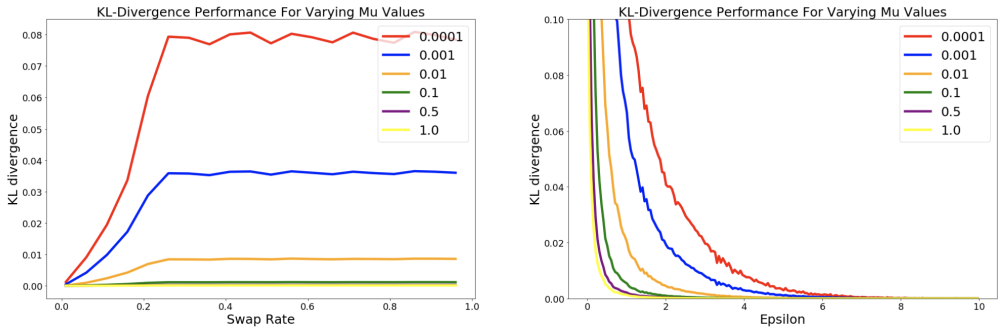

Figure 2: D_{KL}^{μ} results for various mu values; 0-threshold similar swapping (left) and binary DP mechanism (right), Alameda.

three accuracy metrics to specifically convey representation of minority populations. The first is the mean squared error (MSE) of the histogram of the counting queries over the de-identified dataset, as compared to the equivalent histogram over the original dataset. The second is the μ-smoothed Kullback-Leibler divergence (D_{KL}^{μ}) from Tantipongpipat et al. [28], which is a tunable distance measure between two probability distributions, where the parameter μ tunes how heavily low-probability events are weighted. As when computing the MSE, we first compute the empirical distributions of races in the de-identified data and the original data. We then compute their D_{KL}^{μ} distance. Whereas MSE treats the accuracy of all groups equally (regardless of size), D_{KL}^{μ} allows for us to understand mechanism impact on minority populations: in our application, low-probability events correspond to minority groups. Therefore, lower values of μ mean smaller groups have a larger impact on the D_{KL}^{μ}, resulting in higher D_{KL}^{μ} values as seen in Figure 2. Thus, we chose .0001 for our value of μ, to emphasize minority populations. A smaller D_{KL}^{μ} value represents better accuracy; a small MSE value also represents better accuracy.

Finally, we compute the number of minority race-ethnicity groups whose population significantly decreased under the de-identification mechanism. A population qualifies as significantly decreased if its de-identified size is $\leq 75\%$ of its true size. This metric is meant to reflect the number of groups to which funds are severely underallocated because of the de-identification process.

In the graphs that follow, we compute our accuracy metrics for several implementations of our DP and swapping mechanisms. Our similar swapping mechanism is parameterized by the similarity threshold; that is, how close we require two swapped entries to be when swapped. Across the board, we observe that a smaller threshold results in higher accuracy. We note that across our accuracy measurements, the Threshold-0 swapping variant maintains good accuracy. This implementation is meant to show a baseline for how well we can expect swapping to do. Since each row is swapped with a very similar (identical, if possible) row, accuracy is largely preserved. However, we show later that the privacy is poor, since the data changes minimally.

Our DP mechanism is parameterized by the number of age buckets. Recall that our DP de-identified data consists of the population size of each combination of (age bucket) x (Hispanic Y/N) x (census race 0-62) x (household size 1-4). When there are 45 age buckets, each including 2 ages, this data is much more granular. Since our mechanism adds noise to the population count of each attribute combination, more noise is added when there are more age buckets. The accuracy of our DP mechanisms is worse for smaller age group sizes, since having more age buckets means more queries to which noise is added. We note that our queries with two age buckets are most similar to the Census Bureau's redistricting data, which uses a binary voting age variable.

D_{KL}^{μ} and Diversity. We define minority ethnoracial groups as those making up less than 10% of their block group. For all county block groups and mechanisms, D_{KL}^{μ} produced similar levels of accuracy for counties of similar diversity. This is consistent with the use of D_{KL}^{μ}. As can be seen in Figure 2, greater minority emphasis (a lower μ value) results in a higher D_{KL}^{μ} value. This reflects worse accuracy of minority group representation under de-identification mechanisms, suggesting that block groups with larger minority populations will experience worse accuracy. However, we will examine the accuracy of both swapping and DP; noting that while both mechanism types produce similar minority-emphasized accuracy values, DP is far more consistent across varying diversity levels in practice.

Figures 3, 4, and 5 display the D_{KL}^{μ} accuracy trends for every mechanism, for block groups of high diversity (Alameda), medium diversity (Fayette), and low diversity (Jefferson). From Theorem 1, we expect the error in random swapping to be greater in groups whose racial distributions differ more from the state racial distribution. The trend present in Figure 3 confirms this: the highest diversity block groups, whose distributions differ significantly from their overall states, show the highest error from swapping. This trend is additionally present for similar swapping, as shown in Figure 4. Our swapping mechanisms thus produce more accurate datasets for groups where there were fewer individuals of minority races than for groups with more individuals belonging to minority races.

In DP data, however, noise is added independently of the block group's racial distribution; therefore, we expected

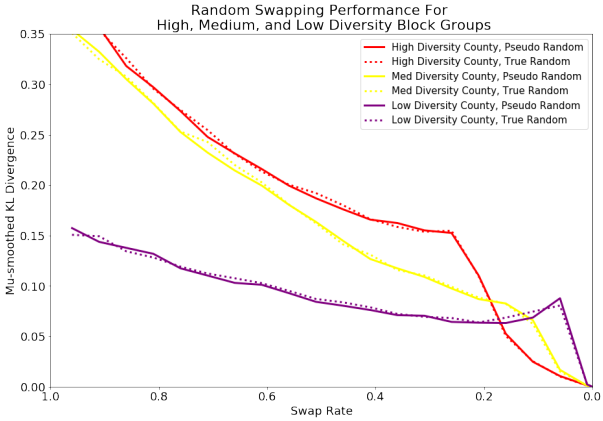

Figure 3: D_{KL}^{μ} results for varying diversity: random swapping, all swap rates.

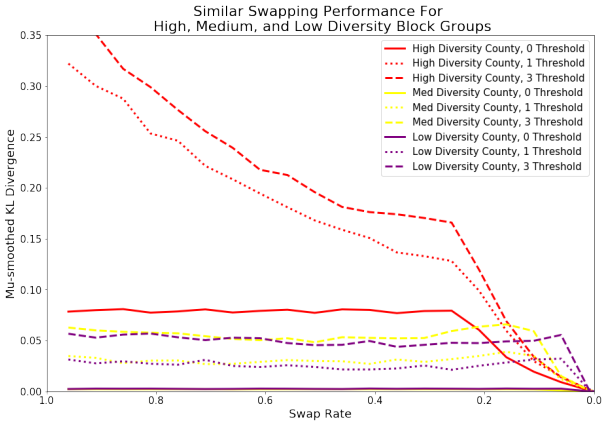

Figure 4: D_{KL}^{μ} results for varying diversity: similar swapping, all swap rates.

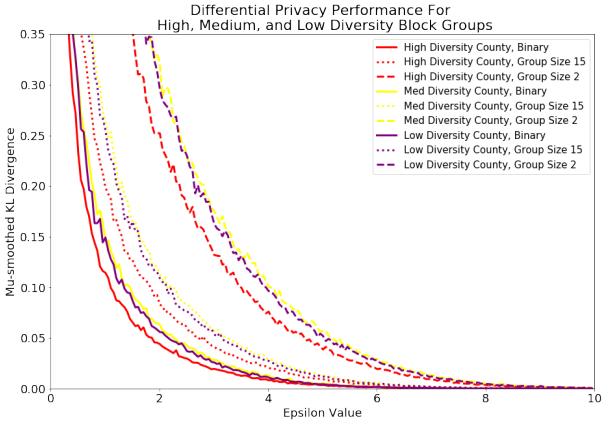

Figure 5: D_{KL}^{μ} results for varying diversity: DP, all ϵ values.

its accuracy to be more consistent for all groups. This is supported in our findings, as can be seen in Figure 5, where accuracy does not vary based upon block group diversity.

Selecting Parameter Ranges For Comparison. We use the U.S. Census Bureau's choice of privacy parameters to inform what we consider reasonable epsilon values and swap rates. The Census Bureau has stated that for the 2020 redistricting data, they will use epsilon values of $\epsilon = 2.47$ for housing data and $\epsilon = 17.14$ for the persons file, giving a total global privacy-loss budget of $\epsilon = 19.61$ [43], and swap rates implemented ranging from 5% to 50% [44]. Our queries contain both attributes present in the housing file (e.g., household size) and attributes present in the persons file (e.g., age). They are inspired by the second-to-last query in [35, Table 2.1(b)], with the addition of household size and the removal of citizenship, which was removed from the 2020 Census. We additionally implement more granular age queries, using up to 45 age buckets, rather than grouping individuals only by voting age. We let our graphs vary from $\epsilon = 2$ to $\epsilon = 10$, to capture a range of interesting and reasonable values that may be assigned in practice to our queries. Since we do not use TDA exactly, and our queries differ slightly from those of the U.S. Census Bureau, our ϵ value does not correspond exactly to theirs.

Take note that for Figures 6, 7, 8, 9, and 10, the swap rates and epsilon values shown on the top and bottom x-axes are not directly correlated (i.e., an epsilon of 2 is not equal to a swap rate of .5). We do not intend to draw a direct comparison; rather, we contextualize these parameters using our empirical accuracy and privacy results.

D_{KL}^{μ} Results. Figures 6 and 7 display D_{KL}^{μ} results for all swapping and DP mechanisms under these parameter bounds, for two counties of varying diversity: Alameda (high diversity) and Washington (low diversity). Block groups of medium diversity (such as Fayette) show results between the extremes of Alameda and Washington, as we would expect. Similar swapping mechanisms produce accuracy levels similar to our DP mechanism with 2, 6, and 45 age buckets.

Observe that the swapping curves in Figures 6 and 7 change drastically in slope at swap rates of roughly 0.25 and 0.05 respectively. We call this the 'uniqueness threshold' and further discuss its significance later, in the context of privacy.

These results stand regardless of the attributes in the counting query. While our graphs display queries over Hispanic x race groups, queries over (age bucket) x Hispanic x race x (household size) show a similar trend: see Figure 8. Although the accuracy of queries over all four attributes is worse than for queries over just Hispanic x race for both mechanisms, the relative accuracies of swapping and DP remain consistent.

MSE Results. In order to validate our findings regarding mechanism impact on minority dataset members, we evaluated the MSE specifically for the minority subset of our

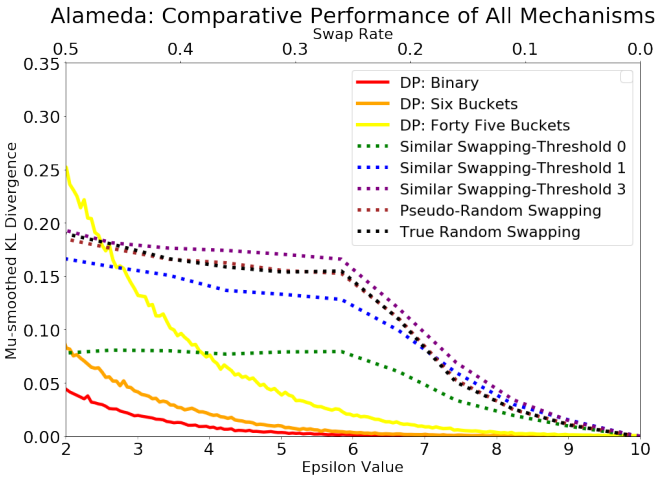

Figure 6: D_{KL}^{μ}: Alameda (high diversity), all mechanisms.

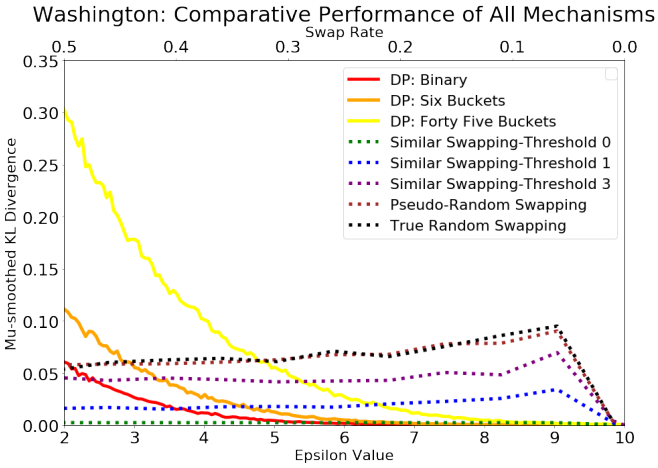

Figure 7: D_{KL}^{μ}: Washington (low diversity), all mechanisms.

datasets (those belonging to races with frequencies below 10%). As can be seen in Cibola County in Figure 9, these results support our findings using D_{KL}^{μ}: for all reasonable parameters, swapping and DP produce comparable accuracy. As in D_{KL}^{μ} as well, DP asymptotically approaches zero error, while swapping's trend is less predictable, and varied greatly across counties and implementations. This similarity between results from D_{KL}^{μ} and a second measure, MSE, provides greater confidence in our findings, and suggests the strength of DP in achieving acceptable accuracy for minority groups.

Additionally, to confirm that our trends for minority population segments were not misrepresentative of overall mechanism behavior, we evaluated the MSE over the global population of each block group.

For all block groups, the MSE for the DP mechanisms was similar. For all group sizes, DP approaches zero MSE for the given range of epsilon values. The MSE was additionally promising for similar swapping mechanisms, but less consistent. The MSE for random swapping varied

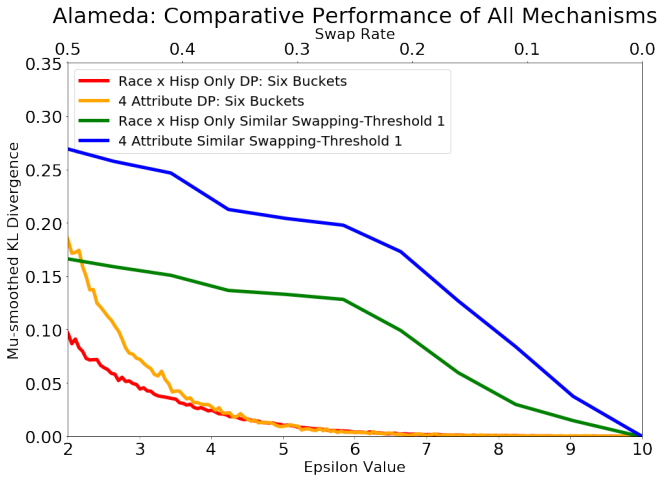

Figure 8: Accuracy of race x Hisp. vs. 4-attribute queries.

greatly across counties, was unpredictable for various swap rates within each county, and was far from zero.

Generally, the MSE for swapping was lower for block groups of lower diversity and higher for block groups of higher diversity, as shown in Figure 9 for Jefferson and Alameda Counties respectively. This trend was not present for DP, where the MSE was relatively consistent across block groups, regardless of diversity. This is the same phenomenon suggested by Theorem 1 and present in the D_{KL}^{μ} figures.

The MSE results for all block groups confirm our confidence in the representativeness of our mechanisms: the parameter ranges where we see reasonable performance resemble those chosen by the Census Bureau. As we see in Jefferson and Alameda block groups in Figure 9, around an epsilon value of 2, the value provided as a benchmark from the Census Bureau, we begin to see good accuracy levels.

Underallocation. We measured the number of (Hispanic) x (census race) groups whose populations significantly decreased due to de-identification. The trend is similar to the one observed in D_{KL}^{μ} and MSE, and it is shown for Cibola County in Figure 10.

Takeaways. Our findings suggest that for the use case of the U.S. Census Bureau, accuracy is comparable for both swapping and DP. In our analysis using MSE, D_{KL}^{μ}, and underallocation, for both the general population and of minority populations, DP's accuracy was comparable, if not better, for reasonable epsilon values and relevant age group sizes. Additionally, DP's accuracy trend was consistent across diversity levels, and its misrepresentation of diverse groups was less severe than that of swapping.

4.5. Privacy

Different levels of perturbation correlate with different levels of privacy. In our DP mechanism, the privacy param-

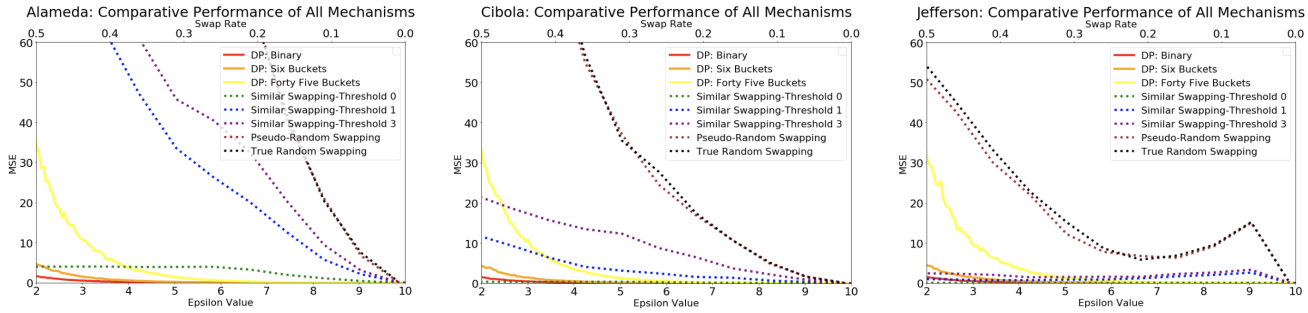

Figure 9: MSE results for Alameda, Cibola, and Jefferson Counties.

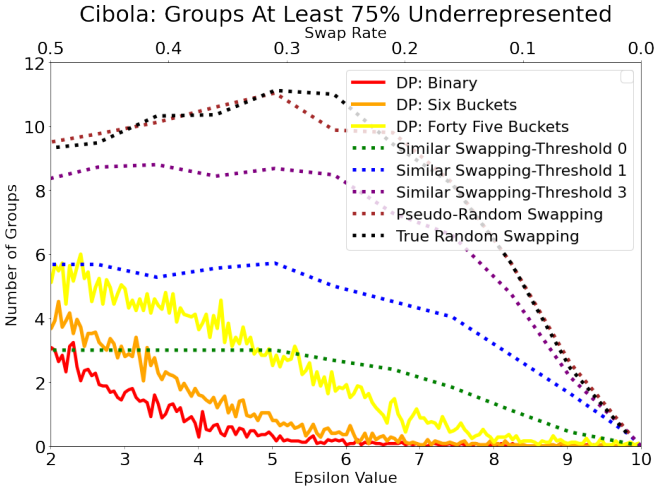

Figure 10: Underallocation: number of minority groups with significant population loss. Cibola (high diversity).

eter ϵ also parameterizes the Geometric distribution from which the noise is drawn. A lower value of ϵ means more noise and a stronger privacy guarantee. Thus for our DP mechanisms, this relationship between noise level, given by the geometric distribution, and privacy level, given by the ϵ-DP guarantee, is nicely defined. This relationship is less predictable for swap rate and privacy; therefore, we empirically evaluate the privacy of our swapping mechanisms.

Measuring Privacy Through Database Linkage. To measure privacy for the swapped datasets, we use the success rate of a simulated database linkage attack, in which an attacker matches entries from our de-identified dataset to a publicly available dataset. This is a powerful and relevant privacy attack. Narayanan and Shmatikov [45] famously correlated a de-identified Netflix database with publicly available IMDb profiles by matching users' viewing histories. The public IMDb profiles often included names, which Narayanan and Shmatikov linked to supposedly private Netflix watch histories. Though this may seem innocuous, your viewing history may reveal your sexual orientation or political leanings. Thus, we used this attack to correlate our de-identified data to a public dataset in a similar manner.

Our attack models an attacker trying to learn specifically whether an individual is Hispanic. We use Algorithm 1 from [25], which Ramachandran et al. used to correlate American Community Survey and Public Use Microdata Sample data. For each block group, we start with our ground truth data for that block group. We create a synthetic dataset P by restricting the ground truth data to include exactly the age, race, and household size of every individual in that county. These fields are those relevant for executing our attack, but you can think of P as also containing a more sensitive attribute, such as income. For each dataset D de-identified via swapping, we run Algorithm 1 from [25] on inputs P and D to count the number of confirmed matches. A match is a row in the public dataset that matches exactly one row in the de-identified dataset with respect to age bucket, census race, and household size. A confirmed match is a match where the row in the public dataset represents the same individual in the de-identified dataset, and whose Hispanic value is the same in the de-identified dataset and the ground truth dataset. We call these confirmed matches identifiable or identified.

Privacy of DP Data. This re-identification algorithm fails for datasets de-identified using our DP method. Recall that the de-identified data produced by our DP mechanism is a histogram of counting queries describing the population size of each combination of (age bucket) x (Hispanic) x (census race) x (household size). There is no notion of rows here, so we cannot run the attack as described.

One might instead measure the success of an attack that attempts to learn whether an individual is Hispanic by comparing the size of the Hispanic and non-Hispanic groups matching their other attributes. However, this attack is at its core statistical inference, and it produces good results even for individuals whose data is not present in any of the datasets. This is unfair—guessing whether someone is Hispanic based on whether their neighbors are Hispanic should not constitute a privacy violation. DP experts in [20] discuss further why such attacks are not individually identifying. We also considered creating rows according to the empirical distribution over the attributes given by the DP histogram, and running the same attack we used on the

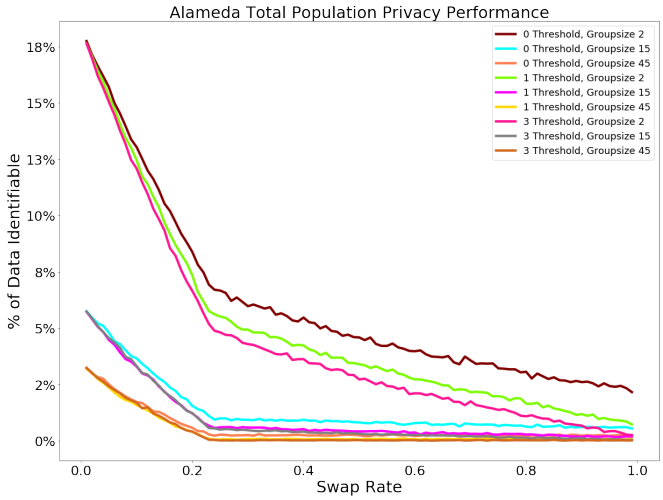

Figure 11: Confirmed match percentages for total population, Alameda.

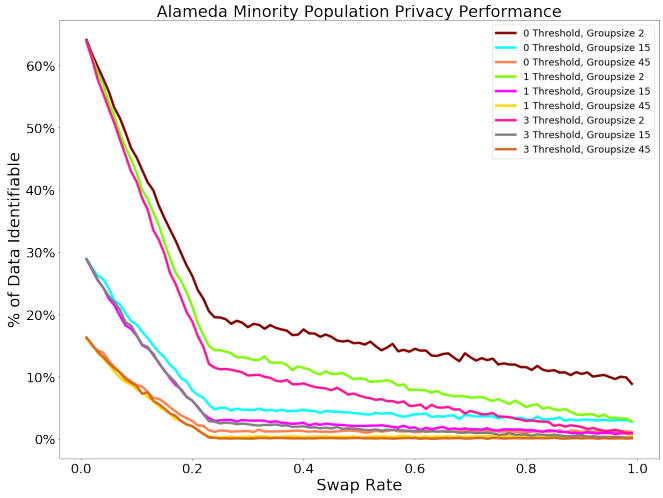

Figure 12: Confirmed match percentages for ethnoracial minority segment of population, Alameda.

swapped data for this new dataset. However, this method is also subject to the statistical inference issue.

Overall Results. Figure 11 displays the total number of identifiable entries in the Alameda block group for varying swap rates. Results for the random swapping implementation were very similar. These identified entries represent individuals matched uniquely from the public dataset to the de-identified dataset.

The number of identified entries exhibits weak privacy for smaller swap rates, and significantly stronger privacy as the swap rate increases past a given value; however, the relationship between this swap rate and these privacy values is unpredictable. We additionally observe this 'uniqueness threshold' in privacy protection is approximately at the swap rate where the block group runs out of unique entries to swap. Since the number of unique entries ranges from 10 to 461 across our block groups, the swap rate at which all unique entries have been swapped out varies greatly from block group to block group.

As we interpret these results, it is helpful to establish a rate of permissibility for dataset re-identification. While there is no federal mandate or recommendation regarding a privacy threshold for demographic data specifically, we may look to the regulation of medical data via HIPAA. Works surrounding HIPAA's Safe Harbor Act, such as [46] [47], reference a "nationally accepted standard of re-identification risk" of .04%. Medical data is, by nature, different from census data, but it is important to remember that census demographic data may be used in database matching with secondary, even more personal data, in the same way that the demographic data contained in these HIPAA datasets reveals personal medical data. We may keep this value of .04% in mind in our own demographic dataset privacy evaluation.

After the aforementioned 'uniqueness threshold', many block groups have a significantly lower rate of identification, often below 2%. However, this is not the case for diverse counties such as Alameda and Hawaii: at a swap rate of 0.5, these county block groups face similar swapping identification rates as high as 12%. Even at a swap rate of 1.0 (where all data has been swapped), some similar swapping mechanisms were still producing identification rates more than 100 times the precedented rate of .04 %. This nonzero identification rate occurs even when all data has been swapped, because entries are swapped with similar entries, i.e., (in a simplified case) an entry representing the only individual with age 20-30 will likely remain the only entry in this age range after similarity swapping has been carried out. For this same reason, group size has significant impact on privacy: it is far more likely that a 45-year-old will be the only entry in the age group 44-45, as opposed to the age group 30-45. This is why in Figure 11, the identification rate is significantly higher for mechanisms evaluated at smaller group sizes. This suggests that there is a significantly higher threat to privacy for groups of high diversity, as they will contain more unique entries.

Minority Impact. We further evaluated the privacy impact of swapping mechanisms specifically for minority groups. In order to do so, we set a threshold of less than 10% of the population to define a minority ethnoracial group, and determined the number of identifiable entries from minority groups.

Figure 12 displays the findings for the Alameda County block group, the same block group shown in Figure 11. There is a significantly higher threat to privacy for the minority groups in the county—the risk is nearly five times higher that a minority will be identified than for the Alameda block group as a whole.

This is still the case for largely homogeneous groups—take, for example, Washington County. Overall identification rates were low, maxing out at about 4% for 0 swap rate. This is because the block group had a population fitting our definition of minority of 38, compared to its total population of 1,090. However, the identification rate for this

minority population subset is significantly higher, at roughly 30%, 55%, and 80% for age group sizes of 2, 15, and 45 respectively for 0 swap rate. For the age groups of 15 and 45, roughly 10% of this minority subset remains identifiable even at a swap rate of 0.8. It is clear that database matching poses a significant risk to privacy of swapped datasets, but further, that individuals from minority groups in a dataset face the brunt of this burden.

5. Overall Findings

Our findings show that there is a significant risk of re-identification posed by swapping mechanisms. The Census Bureau must limit its range of swap rates to only those with permissible privacy: for all of our swapping mechanisms, a wide range of low swap rates yield poor privacy.

Thus, these lower swap rates are not usable. At swap rates with acceptable privacy levels, the accuracy is poor. **When swapping is implemented at a swap rate that yields acceptable privacy, DP has comparable or improved accuracy for all reasonable values of epsilon.** Thus, although we do not determine which swap rates and epsilon values give equal privacy, our findings indeed support DP.

To demonstrate, we look back to our swapping data in Figure 6 for Alameda: in Figure 11, we observe unacceptable privacy ($> 0.5\%$ identifiable) for swap rates lower than 0.2. Thus, we are limited to using swap rates greater than 0.2, for which the D_{KL}^{μ} of swapping is above .07. This D_{KL}^{μ} value is significantly impaired in comparison with DP, which generally remains under a value of .07 within its entire range of reasonable ϵ values (with the exception of 45 buckets, which is unnecessarily granular compared to the census queries). The same pattern is true for our underrepresentation metric in Figure 10; thus, more minority groups will suffer significant resource underallocation under swapping than under DP.

In fact, we see that under swapping, minority groups suffer worse accuracy and privacy than their majority counterparts. For example, for the Cibola block group, we see in Figure 12 that a significant number of race-ethnicity groups remains at least 75% underrepresented at any swap rate higher than 0.2. This is not the case for DP, for which the number of severely underrepresented groups approaches zero by an epsilon value of 5.5 for the two mechanisms of realistic bucket sizes (15 and 45). Since concerns surrounding DP largely surrounded the impact on minority groups, it is notable that DP seems to produce better results.

Under swapping mechanisms, diverse communities (groups with more unique entries) suffer worse accuracy and privacy. Furthermore, for swapping, accuracy comes at the expense of privacy, but not in a predictable manner. It is challenging to choose a swap rate that provides adequate accuracy and privacy, especially for counties with large segments of the population being from minority groups. DP is better in this sense, providing a mechanism that can reliably achieve reasonable privacy and accuracy simultaneously, producing consistent results across groups of varying diversity.

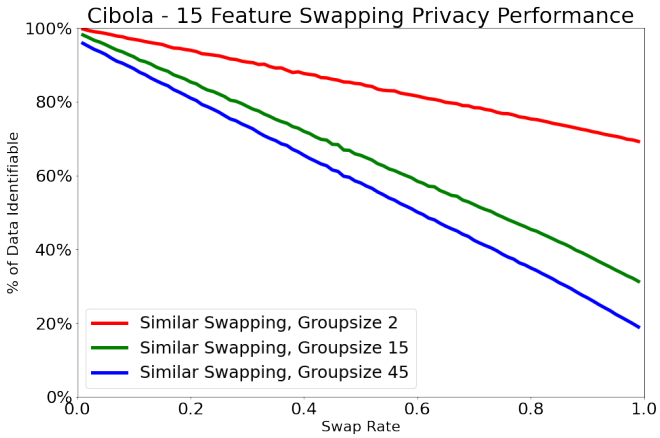

Figure 13: Privacy of similar swapping for a Cibola block group dataset of 15 attributes.

6. Limitations and Further Work

Our work is only intended to draw conclusions about datasets similar in size and structure to the U.S. Decennial Census data. For questionnaires such as the American Community Survey (ACS), with significantly more attributes, a more fine-grained mechanism will be necessary to represent DP's capability.

We have seen that swapped data, especially data pertaining to minority groups, is susceptible to linkage attacks, even for Decennial Census data with relatively few attributes. This risk is even more pronounced for more detailed data. The number of unique individuals increases with the number of attributes, limiting us to a much higher swap rate. Furthermore, when individuals are unique in the entire dataset, not just their geographic subregion, swapping provides no protection against identification. We ran our swapping suite on a 15-attribute dataset to confirm this phenomenon. For Cibola, a high diversity county, the accuracy is similar to the accuracy of swapping for our smaller dataset, since we measure the same queries. The inflection point where accuracy plateaus is at a much higher swap rate of 0.8, since many more rows are unique in this dataset. While the accuracy is good, we see in Figure 13 that privacy is abysmal: even at a swap rate of 100%, no swapping variant reaches below 15%. On such detailed data, naive swapping cannot achieve any satisfactory privacy guarantee.

Because of computation costs, we analyze the variance of our DP mechanism rather than simulating it. The accuracy of our naive mechanism degrades significantly for large datasets, since the amount of noise increases with the number of attribute combinations: for a dataset with c attribute subcombinations per (race) x (Hisp), the variance given our mechanism with parameter ϵ of each race x Hispanic population count is $\frac{2ce^{\epsilon}}{(1-e^{\epsilon})^2}$. Thus the variance of each (race) x (Hisp) population count in our 15-attribute dataset is $4.94 \times 10^9 \frac{e^{\epsilon}}{(1-e^{\epsilon})^2}$ for our mechanism with parameter ϵ. The derivations of these expressions are given in Appendix A.3 for interested readers. In practice, the data steward would

make fewer, possibly overlapping queries instead of one query for each attribute combination, especially as the number of attributes increase. The necessary amount of noise depends on the chosen queries but should be vastly smaller than the amount of noise introduced when computing the count of every attribute combination. For example, the race histogram of a dataset with any number of attributes could be computed with DP simply by adding geometric noise to each race's population count, yielding a variance of only $\frac{2e^{\epsilon}}{(1-e^{\epsilon})^2}$. Our naive method makes every possible query to apply to all use cases and thus introduces the maximum amount of noise; specific use cases will likely require fewer queries and vastly less noise.

Further work could compare DP and swapping on more detailed datasets, with a more sophisticated DP mechanism. Implementing DP for more detailed data is notoriously challenging, and the U.S. Census Bureau has has announced that they will not implement DP for the ACS until at least 2025. Upon release of this mechanism, we may examine efficacy of their DP implementation for the ACS' more complex data. Additionally, we used a single attack for all swapping implementations. Intuitively, the swapping implementation with a similarity threshold of 0 should afford less privacy than that with a similarity threshold of 3, since swaps in the first version introduce less variation. A more tailored attack may capture this relationship between threshold and privacy, and we suspect that the low-threshold similarity swaps result in even more privacy loss than our attack shows.

7. Conclusion

It is necessary to evaluate the impact of DP before widespread implementation—especially when its use case is as significant as the U.S. census. Ensuring that this analysis is effective and representative has been challenging, in part due to limited data access and processing constraints. We show that when swapping is implemented at a swap rate necessary for acceptable privacy, DP has comparable or improved accuracy *for all reasonable values of epsilon*.

We directly compare DP and swapping, using the same ground truth dataset for all of our implementations. We are the first to perform this direct comparison for a wide range of swap rates and ϵ values. Previous publicly available work has either treated 2010 Census data as the ground truth [14], [15], when in reality it was modified via swapping; or it has considered only few ϵ values and swap rates [16]. While it is possible that more comprehensive analysis was carried out by the U.S. Census Bureau, this analysis is likely confidential due to privacy concerns. We believe this paper puts into question current works critiquing DP's use in cases like the U.S. census. Our findings show that when swapping is implemented at a sufficient swap rate to remove identifiable entries, DP shows comparable, often improved, accuracy for all reasonable values of epsilon.

The privacy guarantees afforded by DP may promote greater census participation, in turn yielding higher accuracy. As we examined in Section 4.5, swapping poses a significant threat of large-scale identification by database matching. This threat particularly impacts minority groups: even at a swap rate of 1.0, similar swapping mechanisms still produce minority identification rates as high as 12% (Figure 12). Swapping places a disproportionate privacy burden on minority groups, whereas an ϵ-differentially private mechanism is ϵ-differentially private *for all subgroups*.

As we have seen with census data's involvement in identifying Japanese individuals for internment [2], protecting the privacy these minority groups is especially important. Because, unlike swapping, DP guarantees little and quantifiable change in overall data by the participation of one, its use better protects these individuals' privacy and may convince individuals previously hesitant to participate.

In addition to privacy, we similarly examined how these mechanisms impacted minority group accuracy. As suggested by Theorems 1 and 2 and verified empirically, minority groups are more likely unique and prioritized for swapping. This results in minority entries being swapped away from our block groups, yielding worse accuracy for more diverse counties as expected from Theorem 1. We find that accuracy rapidly decreases until a swap rate is reached where all unique entries have been swapped, at which point the accuracy decreases more slowly. This dramatic cliff for minority groups in swapping is not present with epsilon values for DP, since the noise added is data-independent.

Given DP's similar accuracy and greater privacy relative to swapping, our work supports DP for use cases such as the U.S. Decennial Census. There are further concerns to examine; as mentioned, we ran a simplified version of TDA due to processing constraints, and we used synthetic data due to privacy constraints. Furthermore, our work does not consider substantial variation of county size—our block groups consist of a limited variation in size, which captures typical block groups but does not encompass outliers present in the true census. We intend for this work to serve as an analytical evaluation of these modern de-identification tools, and we hope that it will inspire further analytical work examining the merits of DP in demographic data de-identification.

Acknowledgments

We would like to thank Simson Garfinkel, Rachel Cummings, danah boyd, Cynthia Dwork, and the referees for their assistance and many helpful comments. We would also like to thank Susan Landau, Tal Malkin, and Mihalis Yannakakis for their encouragement and support.

This research was supported in part by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research under award number DE-SC-0001234, a Google Faculty Grant, NSF grants CNS-1923528 and CCF-2107187, a grant from the Columbia-IBM center for Blockchain and Data Transparency, a grant from JPMorgan Chase & Co., and a grant from LexisNexis Risk Solutions. Any views or opinions expressed herein are solely those of the authors listed.

References

[1] J. Abowd, "Staring-down the database reconstruction theorem," https://www.census.gov/content/dam/Census/newsroom/press-kits/2018/jsm/jsm-presentation-database-reconstruction.pdf, 2018.

[2] J. Minkel, "Confirmed: The U.S. Census Bureau gave up names of Japanese-Americans in WW II," *Scientific American*, March 2007. [Online]. Available: https://www.scientificamerican.com/article/confirmed-the-us-census-b/

[3] Baldrige v. Shapiro, 455 U.S. 345 (1982).

[4] United States Census Bureau, "Importance of the data." [Online]. Available: https://2020census.gov/en/census-data.html

[5] National Congress of American Indians, "Differential privacy and the 2020 U.S. Decennial Census: Impact on American Indian and Alaska Native data." [Online]. Available: https://www.ncai.org/prc/2020_Census_and_AIAN_data_FINAL_9_11_2019.pdf

[6] T. Dalenius and S. P. Reiss, "Data-swapping: A technique for disclosure control," *Journal of Statistical Planning and Inference*, vol. 6, no. 1, pp. 73–85, 1982. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0378375882900581

[7] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[8] J. Abowd, R. Ashmead, G. Simson, D. Kifer, P. Leclerc, A. Machanavajjhala, and W. Sexton, "Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge," https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf.

[9] S. Ruggles, C. Fitch, D. Magnuson, and J. Schroeder, "Differential privacy and census data: Implications for social and economic research," in *AEA Papers and Proceedings*, vol. 109, 2019, pp. 403–08.

[10] G. Wezerek and D. Van Riper, "Changes to the census could make small towns disappear," *The New York Times*, Feb 2020.

[11] "13 U.S.C. § 9," https://www.law.cornell.edu/uscode/text/13/9.

[12] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 202–210. [Online]. Available: https://doi.org/10.1145/773153.773173

[13] G. Long, "Formal privacy methods for the 2020 census," https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/privacy-methods-2020-census.pdf.

[14] A. R. Santos-Lozada, J. T. Howard, and A. M. Verdery, "How differential privacy will affect our understanding of health disparities in the United States," *Proceedings of the National Academy of Sciences*, vol. 117, no. 24, pp. 13 405–13 412, 2020. [Online]. Available: https://www.pnas.org/content/117/24/13405

[15] M. E. Hauer and A. R. Santos-Lozada, "Differential privacy in the 2020 census will distort COVID-19 rates," *Socius*, vol. 7, p. 2378023121994014, 2021.

[16] T. Wright and K. Irimata, "Variability assessment of data treated by the topdown algorithm for redistricting," *Statistics*, p. 02, 2020.

[17] A. Cohen, M. Duchin, J. Matthews, and B. Suwal, "Census topdown: The impacts of differential privacy on redistricting," in *2nd Symposium on Foundations of Responsible Computing*, 2021.

[18] S. L. Garfinkel, J. M. Abowd, and S. Powazek, "Issues encountered deploying differential privacy," in *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 2018, pp. 133–137.

[19] C. T. Kenny, S. Kuriwaki, C. McCartan, E. T. Rosenman, T. Simko, and K. Imai, "The use of differential privacy for census data and its impact on redistricting: The case of the 2020 us census," *Science advances*, vol. 7, no. 41, p. eabk3283, 2021.

[20] M. Bun, D. Desfontaines, C. Dwork, M. Naor, K. Nissim, A. Roth, A. Smith, T. Steinke, J. Ullman, and S. Vadhan, "Statistical inference is not a privacy violation," https://differentialprivacy.org/inference-is-not-a-privacy-violation/, June 2021.

[21] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," *Advances in Neural Information Processing Systems*, vol. 32, pp. 15 479–15 488, 2019.

[22] D. Xu, W. Du, and X. Wu, "Removing disparate impact on model accuracy in differentially private stochastic gradient descent," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1924–1932.

[23] M. Hawes and R. A. Rodríguez, "Determining the privacy-loss budget," https://www2.census.gov/about/partners/cac/sac/meetings/2021-05/presentation-research-on-alternatives-to-differential-privacy.pdf.

[24] N. Kim, "The effect of data swapping on analyses of American Community Survey data," *Journal of Privacy and Confidentiality*, vol. 7, no. 1, 2015.

[25] A. Ramachandran, L. Singh, E. Porter, and F. Nagle, "Exploring re-identification risks in public domains," in *2012 Tenth Annual International Conference on Privacy, Security and Trust*. IEEE, 2012, pp. 35–42.

[26] S. Ito, T. Miura, H. Akatsuka, and M. Terada, "Differential privacy and its applicability for official statistics in Japan – a comparative study using small area data from the Japanese population census," in *Privacy in Statistical Databases*, J. Domingo-Ferrer and K. Muralidhar, Eds. Cham: Springer International Publishing, 2020, pp. 337–352.

[27] A. Ziv, "Israel's 'anonymous' statistics surveys aren't so anonymous," *Haaretz*, January 7 2013. [Online]. Available: https://www.haaretz.com/surveys-not-as-anonymous-as-respondents-think-1.5288950

[28] U. Tantipongpipat, C. Waites, D. Boob, A. A. Siva, and R. Cummings, "Differentially private mixed-type data generation for unsupervised learning," *arXiv preprint arXiv:1912.03250*, 2019.

[29] "44 U.S.C. § 2201," https://www.law.cornell.edu/uscode/text/44/2201.

[30] K. Rossiter, "What are census blocks?" July 2011, https://www.census.gov/newsroom/blogs/random-samplings/2011/07/what-are-census-blocks.html.

[31] U. S. C. Bureau, "Census.gov glossary," https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_4, 2019.

[32] R. Jarmin, "Redistricting data: What to expect and when," https://www.census.gov/newsroom/blogs/director/2021/07/redistricting-data.html, 07 2021.

[33] "U.S. Census FAQ," Population Reference Bureau, May 2019. [Online]. Available: https://www.prb.org/resources/u-s-2020-census-faq/

[34] J. Mervis, "Can a set of equations keep US census data private," *Science*, vol. 10, 2019.

[35] National Academies of Sciences, Engineering, and Medicine, *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*, D. L. Cork, C. F. Citro, and N. J. Kirkendall, Eds. Washington, DC: The National Academies Press, 2020. [Online]. Available: https://www.nap.edu/catalog/25978/2020-census-data-products-data-needs-and-privacy-considerations-proceedings

[36] N. Holohan, S. Braghin, P. Mac Aonghusa, and K. Levacher, "Diffprivlib: the IBM differential privacy library," *arXiv preprint arXiv:1907.02444*, 2019.

[37] IBM, "Ibm/differential-privacy-library," https://github.com/IBM/differential-privacy-library.

[38] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," *SIAM Journal on Computing*, vol. 41, no. 6, pp. 1673–1693, 2012.

[39] E. Shi, T. H. Chan, E. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data," in *Proc. NDSS*, vol. 2. Citeseer, 2011, pp. 1–17.

[40] "Amazon EC2 R5 instances," https://aws.amazon.com/ec2/instance-types/r5/, 2021.

[41] "Amazon EMR pricing," https://aws.amazon.com/emr/pricing/, 2021.

[42] M. Hawes and M. Ratcliffe, "Understanding the 2020 census disclosure avoidance system." [Online]. Available: https://www2.census.gov/about/training-workshops/2021/2021-0-13-das-presentation.pdf

[43] "Census Bureau sets key parameters to protect privacy in 2020 census results," https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html, 2021.

[44] "Determining the privacy-loss budget: Research into alternatives to differential privacy," https://www2.census.gov/about/partners/cac/sac/meetings/2021-05/presentation-research-on-alternatives-to-differential-privacy.pdf, 2021.

[45] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008, pp. 111–125.

[46] N. C. on Vital and H. Statistics, "National committee on vital and health statistics ad hoc work group for secondary uses of health data," 2007 August. [Online]. Available: https://ncvhs.hhs.gov/transcripts-minutes/transcript-of-the-august-23-2007-ncvhs-ad-hoc-workgroup-for-secondary-uses-of-health-data-hearing/

[47] V. Janmey and P. L. Elkin, "Re-identification risk in HIPAA de-identified datasets: The MVA attack," *AMIA Annual Symposium Proceedings*, vol. 2018, pp. 1329–1337, 2018.

Appendix A. Proofs of Theorems

A.1. Proof of Theorem 1

Proof.

$$\begin{aligned}\mathbb{E}[c(S')] &= \sum_{\substack{r\in S\\ r \text{ in cat.}}} \left((1-\kappa)+\kappa\cdot\frac{c(\mathcal{D})}{|\mathcal{D}|}\right) + \sum_{\substack{r\in S\\ r \text{ not in cat.}}} \kappa\cdot\frac{c(\mathcal{D})}{\mathcal{D}} \\ &= c(S)\cdot\left(1-\kappa+\kappa\frac{c(\mathcal{D})}{|\mathcal{D}|}\right) + (|S|-c(S))\kappa\frac{c(\mathcal{D})}{|\mathcal{D}|} \\ &= n(\alpha+\kappa(\beta-\alpha))\end{aligned}$$

where $n = |S|$ is the number of rows in S, $\alpha = \frac{c(S)}{|S|}$ is the fraction of rows in S in category c, and $\beta = \frac{c(\mathcal{D})}{|\mathcal{D}|}$ is the fraction of rows in the entire dataset $\mathcal{D}$ in category c. Thus, the expected difference between $\frac{c(S')}{|S'|}$ and $\frac{c(S)}{|S|}$ is

$$\begin{aligned}\mathbb{E}\left[\left|\frac{c(S')}{|S'|}-\frac{c(S)}{|S|}\right|\right] &= \frac{1}{n}\left|n(\alpha+\kappa(\beta-\alpha))-n\alpha\right| \\ &= |\kappa(\beta-\alpha)|\end{aligned}$$

□

A.2. Proof of Theorem 2

Proof. Let $A_1, ..., A_k$ be random variables modeling each of k attributes. They may be jointly distributed. Let $\mathcal{A}$ denote the support of $(A_1, ..., A_k)$. Each element in $\mathcal{A}$ is a combination of attribute values. Let the i^{th} row in S be represented by a random variable R_i. The probability that the first m rows in the dataset are unique is:

$$\sum_{\substack{\{r_1,...,r_m\}\\ \subseteq\mathcal{A}}} \left(m!\prod_{i=1}^{m}\Pr[R_i=r_i]\prod_{i=m+1}^{n}\Pr\left[R_i\notin\{r_j\}_{j\in[m]}\right]\right)$$

The sum is over all sets of m unique attribute combinations in $\mathcal{A}$; that is, sets $\{r_1, ..., r_m\} \subseteq \mathcal{A}$ where $r_i \neq r_j$ for $i \neq j$. For each such set, there are $m!$ ways for the first m rows to be assigned to these combinations such that no two rows have the same combination. The first m rows with values $\{r_1, ..., r_m\}$ are unique if none of the remaining $n-m$ rows take on any of these values.

Let $\mathcal{E}$ denote the event that the first m rows are unique. Since the rows are drawn independently, we can let R be a random variable representing an arbitrary row and let $P_i := \Pr_{A_1,...,A_k}[R = r_i]$ denote the probability that R equals a given row r_i. We can simplify the above expression: $\Pr[\mathcal{E}] = \sum_{\{r_1,...,r_m\}\subseteq\mathcal{A}} \left(m!\prod_{i=1}^{m}P_i\right)\left(1-\sum_{i=1}^{m}P_i\right)^{n-m}$.

We use this to upper and lower bound the probability that at least m rows are unique. Since the rows are chosen i.i.d, for any set of m rows, the probability that those rows are unique is $\Pr[\mathcal{E}]$. Thus by a union bound, the probability that *any* m rows are unique is at most $\Pr[\text{ any } m \text{ rows are unique }] \leq \binom{n}{m}\Pr[\mathcal{E}]$.

The probability that any m rows are unique is at least the probability that the first m rows are unique. Putting this together, we have $\Pr[\mathcal{E}] \leq \Pr[\text{ any } m \text{ rows are unique }] \leq \binom{n}{m}\Pr[\mathcal{E}]$.

For a fixed m, both $\Pr[\mathcal{E}]$ and $\binom{n}{m}\Pr[\mathcal{E}]$ decrease exponentially in n, since $\binom{n}{m} \leq \frac{n^m}{m!}$:

$$\binom{n}{m}Pr[\mathcal{E}] \leq \sum_{\{r_1,...,r_m\}\subseteq\mathcal{A}} n^m\left(\prod_{i=1}^{m}P_i\right)\left(1-\sum_{i=1}^{m}P_i\right)^{n-m}$$

□

A.3. DP Variance

We simulate adding geometric noise to each attribute combination. For each combination of race and ethnicity, we sum the noisy values of the subcategories. The number of values for each attribute are as follows. Age: # buckets; house size: 4; household tenure: 2; heat: 9; housing type: 9; military: 2; move in date: 6; property value: 7; rent: 8; number of rooms: 9; year built: 15; selected monthly owner costs: 7. Thus for a given combination of race and ethnicity, there are $(\#\text{ buckets}) \times 4 \times 2 \times 9 \times 9 \times 2 \times 6 \times 7 \times 8 \times 9 \times 15 \times 7 = 411,505,920 \times (\#\text{ buckets})$ categories. For $(\#\text{ buckets}) = 6$, this is $c = 2469035520 = 1.47 \times 10^9$

categories. The variance of a random variable with the geometric distribution with parameter α is The variance is

$$\sum_{k=-\infty}^{\infty} k^2 \frac{\alpha-1}{1+\alpha}\alpha^{-|k|} = \frac{2(\alpha-1)}{1+\alpha}\sum_{k=0}^{\infty} k^2\alpha^{-k}$$
$$= \frac{2(\alpha-1)}{1+\alpha}\frac{\alpha(\alpha+1)}{(\alpha-1)^3}$$
$$= \frac{2\alpha}{(\alpha-1)^2}$$

which is $\frac{2e^\epsilon}{(1-e^\epsilon)^2}$ for $\alpha = e^\epsilon$. Thus the variance of our mechanism is

$$\frac{2ce^\epsilon}{(1-e^\epsilon)^2} \approx 4.94 \times 10^9 \frac{e^\epsilon}{(1-e^\epsilon)^2}$$

We include a table of the variance of the (race) x (Hisp) population counts for datasets with varying numbers of c attribute combinations per (race) x (Hisp) bucket. The given values of variance are for $\epsilon = 4$.

Our mechanism for our smaller Decennial Census dataset has $c = 8$ for 2 age buckets, $c = 24$ for 6 age buckets, and $c = 180$ for 45 age buckets.

# attribute combinations	Variance
1	0.038
8 (2 age buckets)	0.304
10	0.380
24 (6 age buckets)	0.912
50	1.901
100	3.801
180 (45 age buckets)	6.841
100	38.011
general x	$0.038 \times x$

Appendix B. Data Format

Age	Sex	Race	Hispanic	Household Size	Household Tenure
(0-90)	M, F	(0-63)	Y/N	(1, 2, 3, 4+)	Rent, Own

As shown in the table above, our dataset include all of the features collected about individuals and households in the U.S. Decennial Census ([33], see "What information does the census collect? What questions does the census ask?").

Our query, [Age x Sex x Race x Hispanic x HouseholdSize], was chosen based upon true Census queries. The National Academies' Census Data Products outlines the queries from the 2010 Demonstration Data Products[35]. Our selected query is more comprehensive than all those listed, combining the queries listed in 2.1b and 2.1c. Thus, this query can be used to make any of the other provided queries, and provides a complete de-identification mechanism. We suggest that using a less comprehensive query would further improve DP's accuracy. While we only look at accuracy from the perspective of race, we swap/add noise to all features except for household tenure.

The counties that we used for our block groups include:

- **High diversity:** Alameda County, CA; Hawaii County, HI; Cibola County, NM
- **Medium diversity:** Grand Forks County, ND; Fayette County, GA; Nantucket County, MA
- **Low diversity:** Jefferson County, MO; Armstrong County, PA; Washington County, VT