[Proposal] Quantifying the Risk of Linked Census Data to Residents of Subsidized Housing

Diana Qing[1], Ryan Steed[2], and Zhiwei Steven Wu[2]

[1]University of California, Berkeley
[2]Carnegie Mellon University

January 25, 2023

Abstract

As the U.S. Census Bureau implements its controversial new disclosure avoidance system, researchers and policymakers debate the necessity of modernized privacy protections for public data. Protecting the confidentiality of respondents is crucial for conducting an accurate U.S. census. In particular, respondents in subsidized housing may deliberately not mention children or other household members for fear of being found in violation of their lease or local housing code. Our preliminary results show that under traditional disclosure avoidance methods, this concern is founded. By linking 2010 Decennial Census data with external survey data, we demonstrate a simple, cheap integer programming attack that identifies at least one household living in violation of occupancy guidelines in as many as 1 in 11 blocks—at minimum 1 in 32 blocks—with housing subsidized under Section 8 in 2010. But, this same attack fails to consistently recover those violations when applied to noisy statistics retroactively protected by the 2020 disclosure avoidance system. We propose several additional experiments to substantiate these initial results and evaluate the effectiveness and disparate impact of the Census Bureau's new disclosure avoidance plans in preventing this kind of attack. The results of these experiments will provide a concrete empirical demonstration of a privacy risk that is already top-of-mind for census respondents—a valuable example for policymakers seeking to design a trustworthy, accurate census.

1 Introduction

In response to concerns about reconstruction-abetted attacks on census participants' confidential responses (Abowd, 2021; Garfinkel et al., 2018; Hawes, 2020), the United States Census Bureau updated its disclosure avoidance system for the 2020 Decennial Census to adopt differential privacy (Abowd et al., 2022), a modern formal privacy standard (Dinur and Nissim, 2003; Dwork et al., 2006). But because the new disclosure avoidance system requires injecting noise into raw statistics and post-processing these noisy measurements, many data users fear the change will interfere with research (Ruggles et al., 2019), redistricting (noa, 2021), voting rights protection (Ochoa and Minnis, 2021), and other key applications of census data. Others question whether the changes are even necessary or effective (Ruggles et al., 2019; Kenny et al., 2021). Title 13 prohibits the Census Bureau from disclosing personally identifiable information in its data products. Detractors suggest that differential privacy "goes above and beyond" this legal requirement (Mervis, 2020).

But the reasons to keep census data confidential *do* reach above and beyond legality. Respondents are less likely to respond truthfully (or respond at all) if they are worried about how their responses might be used against them. We investigate one specific concern: that block-level summary statistics about household composition could be used to identify households living in violation of their leases. In transcript from a 2020 workshop on privacy and census participation, researcher danah boyd recounts conversations where citizens in New York City and Detroit raised specific concerns about responding to the census (Cork et al., 2020, p. 130):

> In New York City, [danah boyd] said that there are many people living in Section 8 public housing who are terrified of giving full information to the city (or the census) about who is living in their homes, for fear of being kicked out. These are people who are concerned about being subject to constant surveillance: video surveillance in the buildings, concerns about surveillance of electric and water

> utility usage. They fear that the information that they provide might be directly used against them, but the prospect of being reidentified from public statistical information is terrifying as well.

Concerns about privacy have direct impacts on data quality, especially for hard-to-count groups. O'Hare (2015) cites multiple reports of respondents deliberately not mentioning their children on census forms for fear of reprisal from landlords, immigration agencies, social services agencies, or custodial parents.

Our work. We show that for families receiving housing assistance under Section 8, these concerns are founded. Using tables from the 2010 Census Summary File 1—which will be migrated in 2020 to the Demographic & Housing Characteristics (DHC) file—we develop two approaches to rigorously determine whether it is possible to identify low-income households residing in Section 8 housing who are living in violation of housing occupancy standards. Households living in violation of occupancy standards face the risk of eviction, and even if a landlord chooses not to evict, knowing that their responses in the Census can be linked back to their property may deter these households from participating wholly in the census. Our preliminary results (§ 4) suggest that an attacker could combine household sizes from the 2010 Census with bedroom counts and statistics on population demographics from HUD survey data and solve a simple integer program to detect if a given census block must contain a household in violation of the standard HUD occupancy limits, all within 30 seconds on a standard laptop.

In 2010, this attack could have identified 647 blocks (more than 1 in 11 blocks with Section 8 properties in 2010) with at least one household in violation of Section 8 occupancy standards and as many as 3,448 households (about 1 in 130 households in Section 8 properties) living in violation of HUD guidelines. Even without any demographic information, an attacker attempting to fit just the smallest household sizes in the block into Section 8 bedrooms can still identify 205 blocks with at least one violation (more than 1 in 32 blocks with Section 8 properties in 2010) for a total of as many as 1,051 households in violation.

However, when we apply this same heuristic to 2010 statistics protected by 2020 differentially private disclosure avoidance measures—the 2010 Demographic and Housing Characteristics (DHC) Demonstration Data released in August 2022 (US Census Bureau, 2022)—our results suggest an attacker could discover less than 61% of the originally discovered blocks in violation. Furthermore, on these demonstration data, an attacker using the integer programming approach would be unable to create valid reconstructions for more than 9 in 10 blocks, identifying only around 3% of the blocks originally in violation. The number of blocks with households subsidized under Section 8 grew by more than 6% from 2010 to 2020—should the 2020 Decennial Census prove vulnerable a similar attack, even more households are at risk.

Our proposal. We propose several additional experiments to substantiate these initial results and evaluate the impact of disclosure avoidance plans for the 2020 Decennial Census Demographic & Housing Characteristics file (§ 5). First, we plan to analyze whether protections provided by the 2020 disclosure avoidance system against this kind of attack are equally afforded to the most marginalized groups (e.g., people with disabilities). Second, we plan to improve the validity of our results by improving the quality of our reconstructions with public use microdata samples, varying the amount of outside knowledge available to attackers, and validating our results more thoroughly on a sets of synthetic microdata. Time allowing, we also hope to compare the effectiveness of our attack against swapping, the primary method for disclosure avoidance, by applying both random swapping and the Census Bureau's differentially private disclosure avoidance algorithm, TopDown (Abowd et al., 2022), to the same set of ground-truth synthetic microdata.

These experiments will provide further empirical evidence for the risks of linked census data, absent updated disclosure avoidance measures. Most notably, the Census Bureau reports that when linked with commercial data, just a portion of the 2010 Census data was sufficient to reconstruct and re-identify the personal characteristics of millions of individuals (Abowd, 2021; Hawes, 2020). Other research shows that vulnerable groups are particularly at

risk—for example, in a public comment on differentially private demonstration data for the DHC file, Flaxman and Keyes (2022) suggest that a reconstruction-abetted linkage attack could be used to identify transgender children in Texas. The results of this proposal will provide a clear example of a concrete privacy risk that is already top-of-mind for census respondents, and assess the efficacy of planned disclosure avoidance techniques for protecting against modern linkage attacks.

2 Data

We examine 6,543 census blocks with Section 8 properties in the 50 states, the District of Columbia, and Puerto Rico. We link two publicly available datasets: (1) Block-level summary statistics from 2010 Summary File 1, which consists of data collected during the 2010 Decennial Census (US Census Bureau, 2011); (2) Block-level property survey data (US Department of Housing and Urban Development, 2010) from the U.S. Department of Housing & Urban Development (HUD) on subsidized housing units leased to low-income households. For our preliminary results, we focus on just the 2010 Section 8 New Construction and Substantial Rehabilitation Program (Section 8 NC/SR), which is now known as Project-Based Section 8—about a quarter of the households included in the dataset. Later, we will also include public housing, Section 236, and other multi-family privately owned housing. This survey (hereafter "HUD survey") is typically administered by the building owner or housing agency representative (US Department of Housing and Urban Development, 2023). Because the survey is still running, our attacks could be applied to the 2020 Decennial Census when block-level statistics from the tables we use become available.

To test the efficacy of the 2020 disclosure avoidance system, we also apply our methods to the DHC 2010 Demonstration Data released in August 2022 (US Census Bureau, 2022). These demonstration files result from the latest version of the Census Bureau's new, differentially private TopDown algorithm (Abowd et al., 2022) applied to the 2010 Decennial Census

records. This dataset can be viewed as the counterfactual set of statistics that would have resulted had the Census Bureau applied the 2020 disclosure avoidance system to produce the 2010 version of Summary File 1.

We focused on a small set of tables from 2010 Summary File 1 (and likewise the 2010 Demonstration Data), including information on the number of households, the number of housing units, the number of households that identified as each race and ethnic group, the number of children (defined as those younger than 18) and number of adults (defined as those 18 or older), and the number of households of sizes 1 through 7+, all at the block level.[1] From HUD, we selected statistics on the number of housing units, the number of households, the percentage of housing units with 0 or 1 bedroom, 2 bedrooms, or 3+ bedrooms, and demographic information on what percentage of households identify as a certain race, ethnicity, age group, disability status, income level, etc. Only some households in each Section 8 property responded to the HUD survey—we conservatively assume that these are the only Section 8 households in the block, and that they live in the largest units reported in the HUD survey.[2]

Using Census Geocoder (US Census Bureau, 2023), we grouped each property in the HUD survey dataset (hereafter "Section 8 properties") by census block, ignoring Section 8 properties for which no block matching the listed property postal address could be found (see Appendix A). Then, we aggregated the Section 8 property statistics at the block level and linked them block-by-block with the tables from Summary File 1.

[1]Specifically, we used data from these 6 tables from Summary File 1: (1) P1 Total Population; (2) P5 Hispanic or Latino Origin by Race; (3) P15 Hispanic or Latino Origin of Householder by Race of Householder; (4) P16 Population in Households by Age; (5) H3 Occupancy Status; (6) H13 Household Size.

[2]There are also 2,837 blocks where the percentage of 0 or 1 bedroom, 2 bedroom, and 3 bedroom units could not be precisely rounded to a set of integers summing to the number of reporting Section 8 households. To be as conservative as possible, for blocks where the number of households is greater than the number of units, we assume the $N_{\text{unassigned}}$ largest Section 8 households are not in violation, where $N_{\text{unassigned}}$ denotes the difference between the number of Section 8 households and the number of Section 8 units. Typically, $N_{\text{unassigned}} \leq 1$.

3 Methods

Following privacy concerns expressed by tenants (O'Hare, 2015; Cork et al., 2020), we imagine an attacker interested in learning which households subisidized under Section 8 are living in violation of occupancy standards. HUD tenant selection and leasing guidelines *require* landlords to match applicants to units based on household size (US Department of Housing and Urban Development, 2013, Chs. 3 & 5). Tenants are assigned to units on the assumption that their households meet occupancy standards, which may even be included explicitly in their lease (US Department of Housing and Urban Development, 2017). Respondents may rightly fear that if a landlord or property manager discovers that their household exceeds these occupancy standards, they will be evicted from their homes. To test whether this is possible within the 2010 Census protections, we test two simple approaches an attacker might use to identify blocks with households in violation: a "greedy" heuristic baseline and a more rigorous integer programming approach.

Though exceptions are possible (e.g., if bedrooms in the unit are unusually small) HUD suggests that at a minimum property managers adopt a "two heartbeats per room" policy—no more than two persons should be required to occupy a bedroom (US Department of Housing and Urban Development, 1987, 2013). Thus, b bedroom units have an occupancy limit of $2b$ people. The only exception to this rule are studios with 0 bedrooms, which have an occupancy limit of 1 person. Since our HUD survey data includes options for only 0 or 1 bedroom, 2 bedrooms, and 3+ bedrooms, we make a conservative assumption that any household assigned to live in a 0 or 1 bedroom unit can have a maximum of 2 people. We also conservatively assume that any-sized household can fit into a unit with 3+ bedrooms, since the bedroom count could be arbitrarily large.

3.1 "Greedy" Baseline

While we have information on the number of Section 8 and non-Section 8 households in each block, we do not have information on the *sizes* of the Section 8 and non-Section 8 households—these are the values an attacker aims to reconstruct. As a simple baseline, we imagine that the Section 8 properties in a block are filled with all the smallest households (those least likely to violate occupancy limits), while the largest households in the block live outside the Section 8 properties. Using this conservative heuristic, an attacker need only know the household sizes from Summary File 1 and the number of bedrooms in the housing units in the Section 8 properties. To test this heuristic, we implement a simple algorithm (detailed in full in Appendix B). Starting with the largest-sized Section 8 households, the attacker fits as many large households into units with 3+ bedrooms as they can, then moves onto 2 bedroom units once there are no more 3+ bedroom units, then moves onto 0 or 1 bedroom units once there are no more 2 or 3+ bedroom units. If it is the case that a household is unable to fit into a unit without violating the occupancy limits and there are no larger units available for this household to live in, then the attacker concludes that a household of this size must be in violation. This approach allows the attacker to identify, at a conservative minimum, any blocks with households in violation of occupancy standards.

3.2 Integer Programming

Our greedy heuristic allows us to identify blocks that we were certain contained violations, but it is unlikely that all the smallest-sized households in each block are Section 8. Thus, in each block, to ensure the households that are assigned to be Section 8 more accurately reflect the demographic characteristics of the Section 8 property that is provided in the HUD survey data, we develop a second approach where we add constraints on the size, race, ethnicity, and age composition of the Section 8 households using summary statistics from Summary File 1 and the HUD survey data (for a full description of the variables and summary statistics used, see Appendix C.) We define an integer program constrained by these statistics, where

the variables to be reconstructed are the demographic characteristics, household size, and bedroom count of each household in the block.

For our initial results, we additionally constrain the size of each Section 8 household by the number of bedrooms in the unit according to HUD's "two heartbeats per room" occupancy guideline. Then, we solve the integer program with Gurobi (Gurobi Optimization, LLC, 2023) and identify blocks for which no possible reconstruction of the household-level microdata could satisfy these occupancy limits. We specified and solved this integer program twice—once with and once without occupancy limits—for all blocks with Section 8 properties in United States in less than 24 hours on a server with 32 8-core, 2-thread, 2.9GHz CPUs and 256GB RAM.

4 Initial Results

4.1 "Greedy" Baseline

After running our approach on all 52 US states and territories, we found that in the best case scenario (using our heuristic approach, which includes only the household sizes from Summary File 1 and bedroom counts from HUD), an attacker could determine that 205 blocks (3.13% of blocks with Section 8 properties) contain at least one household in violation of occupancy limits. Across the US, there are 1,051 households (0.23% of 459,160 Section 8 households total) that an attacker could not possibly fit into the Section 8 properties without violating the occupancy limits (using our heuristic algorithm). The majority (1,047) are living in studios or one bedroom units.[3]

When we run the same approach on the 2010 Demonstration Data—which includes retroactively injected noise from the Census Bureau's new differentially private 2020 dis-

[3]637 households have a household size of 3, 194 households have a household size of 4, 162 households have a household size of 5, 35 households have a household size of 6, and 23 households have a household size of 7+. The skew towards more households in violation being of size 3 is a result of our assignment algorithm assigning smaller-sized households to housing units last.

closure avoidance system—the number of blocks an attacker could identify as in violation using this heuristic increases to 761. Of the 205 blocks originally found to be in violation, only 124 (16.3% precision, 60.5% recall) remain in violation when the attacker uses the noisy demonstration data. Without additional information, an attacker attempting to leverage the noisy statistics has no way of knowing which 124 blocks of the 761 blocks suspected to be in violation are the same as the ones identified using the original statistics.

4.2 Integer Programming

If an attacker uses a more rigorous integer program, which includes additional statistics on total number of people, race, ethnicity, and age, the number of blocks for which an attacker would not be able to find a reconstruction compliant with occupancy limits—indicating at least one household in violation—jumps to 647 (9.89% of blocks with Section 8 properties). It may be quite easy to tell which households are in violation with a simple visit to the property—39 of these blocks have less than 15 households in a Section 8 property; 156 have less than 30.

For 214 of those blocks (3.27% of all blocks with Section 8 properties), our HUD survey data report more people living in Section 8 properties than could possibly fit in the largest household sizes reported in the 2010 Decennial Census. On its face, this result seems to clearly indicate an occupancy limit violation in one or more households in the block, but it could be the case that these properties grew or shrunk between the census response date and the HUD survey response date, or (more unlikely) that people were more likely to report additional occupants to their building owner than a census worker. Errors in the reported addresses in the HUD survey data might also account for this discrepancy (see Appendix A).

Still, an attacker would not be able to find a compliant reconstruction for the 433 other blocks (6.62% of blocks with Section 8 properties) where the reported number of Section 8 occupants *does* fit in the census household size distribution. In just these blocks, our results suggest an attacker could reconstruct as many as 3,448 households living in violation

of occupancy limits (0.75% of all reporting Section 8 households).[4] As in our heuristic baseline, the majority (3,069) are households living in studios or one bedroom units.

When we construct the same integer programs using the noisy 2010 demonstration data instead of the original 2010 census statistics, we find that an attacker would be unable to construct valid microdata for nearly 93% of all blocks, even if the attacker uses only the statistics from Summary File 1. For those blocks, the attacker cannot apply the same integer programming attack. In the remaining blocks—the blocks for which we can construct valid microdata using the noisy statistics—there are only 41 blocks in violation of occupancy limits (51% precision), and only 21 are also among the original 647 violations (3.2% recall).

5 Proposed Research

Our initial results suggest that residents' fears that their census responses could reveal a violation of their lease or housing code may be founded, but that the new disclosure avoidance system provides some empirical assurance. We propose several additional experiments to substantiate these results and relate this privacy risk to current disclosure avoidance plans for 2020 Decennial Census data.

Comparing modern and traditional disclosure avoidance measures. If time allows, we will also test our methods by fully replicating the TopDown algorithm that will be applied to the 2020 DHC file (Abowd et al., 2022) and comparing it to swapping, the primary disclosure avoidance method applied to the 2010 Decennial Census (McKenna, 2018). While the exact swapping implementation used by the Census Bureau has not been released, we will follow Christ et al. (2022) in implementing a general approach close to the Census Bureau's general description (McKenna, 2018). Since the 2010 Summary File 1 already includes swapping in the race and ethnicity tables that could occlude our initial integer program results, we will implement both methods on a set of synthetic microdata (described below)

[4]This is the total number of households in violation in the first reconstruction found; the true number may be higher or lower. See § 5 for a proposed experiment to improve this estimate.

that can serve as a baseline for validation. There are two benefits to using simulations on synthetic data. First, the violations we find in the 2010 Summary File 1 may be affected by swapping, if we rely on swapped tables—with raw synthetic microdata, we can more closely validate this approach. Second, the DP demo product only represents one run of the differentially private randomized mechanism—while we can generalize over all the blocks in the nation, we would also like to run multiple trials of the mechanism to see how results vary for the same block.

Improving our reconstructions. In our initial results, we aim only to estimate the number of blocks with at least one household in violation of occupancy limits. However, we would also like to analyze the reconstructed households themselves. While our current method stops at the first compliant reconstruction, we propose to search for the *most likely* reconstruction by pairing a likelihood function with the same integer program. We will construct this likelihood function using Public Use Microdata Samples (PUMS) (US Census Bureau, 2014), also from the 2010 Decennial Census. By comparing possible reconstructions with a 10% sample of microdata from each area, we can analyze just the most likely reconstructions. With these reconstructions, we can answer more detailed questions about what kinds of households are affected by this privacy risk.

Validating with simulation. While we can claim with confidence that—conditional on the accuracy of our data—an attacker can positively identify blocks or reconstructed households in violation of HUD occupancy guidelines, there may be additional households in violation that this method does not detect (false negatives). To validate our results, and provide more detailed information that would otherwise be potentially risky for the real households in our data, we propose to run our method on a set of simulated microdata. Varying the rate and composition of households in violation, we will measure the performance of our heuristic and integer program attacks and provide an indication of how our reconstructions might be used to re-identify real households. For example, while we currently report the *number* of reconstructed households in violation, a simulated attack would allow us to

also report the *percentage* of violating households detected by each method—in other words, the likelihood that any violating respondent would be revealed if they accurately report all the members of their household.

Introducing variations in outside knowledge. Our current methods presume that attackers have access to population demographics in Section 8 housing from HUD survey data or another source (e.g., internal administrative data). However, it may be the case that an attacker (for instance, a landlord) does not know with certainty tenants' demographics, and we would not expect tenants to accurately report that information to the building owner. We plan to solve our integer program again without these variables as a third treatment to see if results change compared to our baseline and current approach. Another benefit to this approach is that it removes the effects of swapping in the 2010 Decennial Census, which only affects the race and ethnicity tables used in our preliminary results (McKenna, 2018).

Measuring disparate impact Not only are we interested in quantifying how many blocks and households are at risk of re-identification across the U.S., but we also seek to understand whether certain demographic populations might be more heavily impacted by inadequate privacy protection in Census data products. For example, Figure 1 shows the distribution of various demographics in the violating and non-violating blocks. With better quality reconstructions and more detailed validation, we can determine whether plans for disclosure avoidance protect the groups most vulnerable to privacy harms, repeating the analysis in Figure 1 for our results on the demonstration and synthetic data. For example, when we compare blocks with violations with blocks without violations, we notice that blocks with households living in violation of occupancy limits tend to have a slightly higher percentage of Section 8 residents with a disability, according to the HUD survey data.

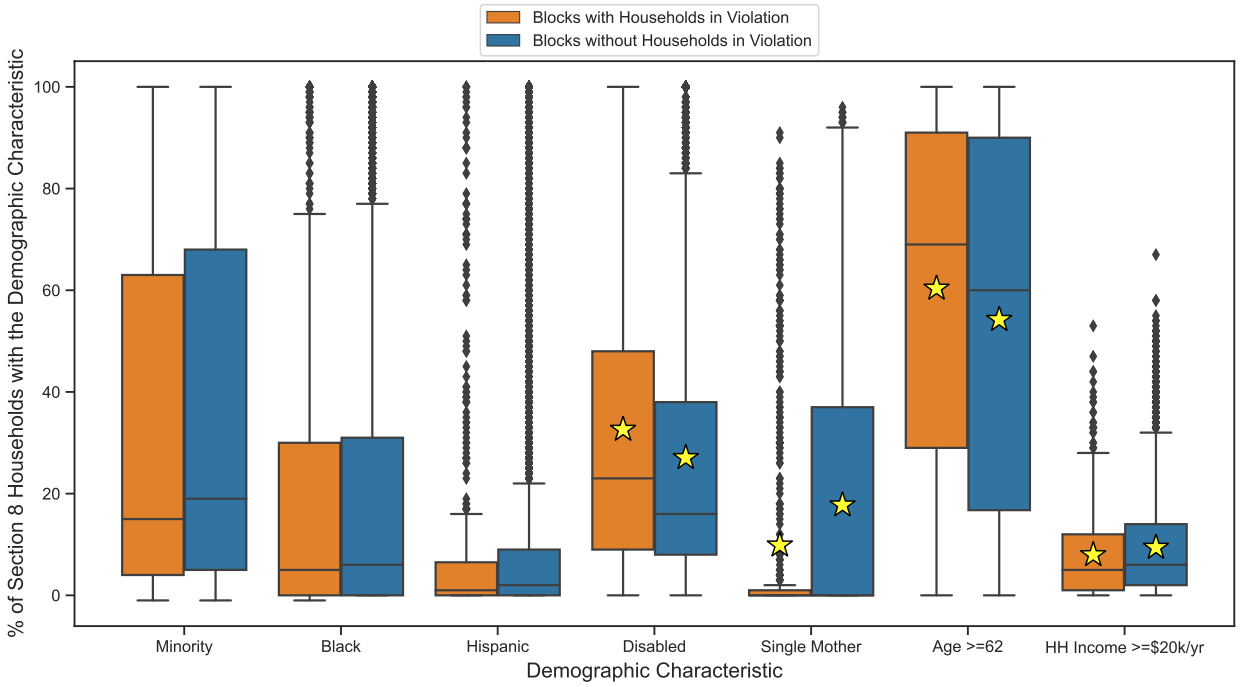


Figure 1: Distribution of demographic characteristics—as a percentage of individuals living in Section 8 properties in each block—in blocks with detected violations and in blocks without, using the originally published 2010 Decennial Census data. Stars represent the mean, and only appear for demographics where there is a statistically significant ($p < 0.05$) difference between the mean for blocks with violations and blocks without according to a two-sample t-test.

6 Discussion

By linking 2010 Decennial Census data with HUD survey data, we demonstrate a possible attack that could reveal households living in violation of HUD occupancy limits in as many as 1 in 11 census blocks with Section 8 properties. In less than 30 seconds on a standard laptop, we can link 2010 Decennial Census and HUD survey data to detect if a given census block must contain a household in violation of the standard HUD occupancy limits. Our preliminary results suggest that in 2010, this attack could have identified 647 blocks (more than 1 in 11 blocks with Section 8 properties) and as many as 3,448 households (about 1 in a 130 households in Section 8 properties) living in violation of HUD guidelines.When we apply the same attacks to 2010 statistics protected by 2020 differentially private disclosure avoidance measures, an attacker could recover less than 61% of the blocks in violation origi-

nally discovered with our heuristic method, and only 3% of those discovered with our integer programming approach. Our preliminary results suggest that this attack fails to consistently recover those violations when applied to noisy statistics retroactively protected by the 2020 disclosure avoidance system.

Building on our initial results, our planned experiments will test a) how disclosure avoidance measures planned for the DHC could affect these risks, compared to traditional measures, and b) which demographic groups are most affected. Quantifying the empirical risk to marginalized groups is crucial for assessing the efficacy of current and future disclosure avoidance systems. The results of this study will constitute a motivating example for policymakers aiming to design a more trustworthy and accurate census.

Acknowledgments

Thanks to danah boyd, Abraham Flaxman, Terrance Liu, Sarah Radway, and Manish Raghavan for their guidance and feedback on this research so far. Thanks especially to Manish Raghavan, whose code was invaluable for our exploratory experiments and a model for our codebase. Thanks also to Sarah Young for helping us locate the survey data used in this study.

References

(2021, September). Alabama v. U.S. Dep't of Commerce. 546 F. Supp. 3d 1057 (M.D. Ala. 2021).

Abowd, J. (2021, April). Declaration in State of Alabama et al. v. Department of Commerce et al. Case No. 3:21- CV-211-RAH-ECM-KCN, Middle District of Alabama, Eastern Division, United States District Court.

Abowd, J., R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala, B. Moran, W. Sexton, M. Spence, and P. Zhuravlev (2022, June). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review* (Special Issue 2).

Christ, M., S. Radway, and S. M. Bellovin (2022, April). Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the U.S. Census. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1564–1564. IEEE Computer Society. ISSN: 2375-1207.

Cork, D. L., C. F. Citro, and N. J. Kirkendall (2020). Panel Discussion on Key Privacy Issues: Privacy and Census Participation. In *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*, Washington, DC, pp. 129–133. The National Academies Press.

Dinur, I. and K. Nissim (2003). Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '03*, San Diego, California, pp. 202–210. ACM Press.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi and T. Rabin (Eds.), *Theory of Cryptography*, Lecture Notes in Computer Science, Berlin, Heidelberg, pp. 265–284. Springer.

Flaxman, A. D. and O. Keyes (2022, June). The Risk of Linked Census Data to Transgender Children: A Simulation Study. In *Summary of Public Feedback Received on the 2010 Demonstration Data Product - Demographic and Housing Characteristics File*. Section: Government.

Garfinkel, S. L., J. M. Abowd, and C. Martindale (2018, January). Understanding Database Reconstruction Attacks on Public Data. *ACMQueue 16*(5), 28–53. Accepted: 2020-12-06T22:17:34Z.

Gurobi Optimization, LLC (2023). Gurobi Optimizer Reference Manual.

Hawes, M. B. (2020, April). Implementing Differential Privacy: Seven Lessons From the 2020 United States Census. *Harvard Data Science Review 2*(2).

Kenny, C. T., S. Kuriwaki, C. McCartan, E. T. R. Rosenman, T. Simko, and K. Imai (2021). The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census. *Science Advances 7*(41), eabk3283. _eprint: https://www.science.org/doi/pdf/10.1126/sciadv.abk3283.

McKenna, L. (2018, November). Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing. The Research and Methodology Directorate, U.S. Census Bureau.

Mervis, J. (2020, February). Researchers finally get access to data on Facebook's role in political discourse. *Science*.

Ochoa, S. A. and T. A. Minnis (2021, April). Preliminary Report: Impact of Differential Privacy & the 2020 Census on Latinos, Asian Americans and Redistricting. Report, MALDEF & Asian Americans Advancing Justice.

O'Hare, W. P. (2015). *The Undercount of Young Children in the U.S. Decennial Census*. SpringerBriefs in Population Studies. Cham: Springer International Publishing.

Ruggles, S., C. Fitch, D. Magnuson, and J. Schroeder (2019). Differential Privacy and Census Data: Implications for Social and Economic Research. *403 AEA Papers and Proceedings 109*, 403–408.

US Census Bureau (2011). 2010 Census Summary File 1.

US Census Bureau (2014). 2010 Census Public Use Microdata Sample File (PUMS).

US Census Bureau (2022, August). DAS Development Update 2022-08-25. Section: Government.

US Census Bureau (2023). Census Geocoder.

US Department of Housing and Urban Development (1987). Tenant Selection. In *Public Housing Occupancy Handbook*, Volume REV-2 of *Handbook*, pp. 5–1–5–22. U.S. Department of Housing and Urban Development Office of Public and Indian Housing.

US Department of Housing and Urban Development (2010). Picture of Subsidized Households.

US Department of Housing and Urban Development (2013, November). *HUD Handbook 4350.3: Occupancy Requirements of Subsidized Multifamily Housing Programs*. Number 4350.3 in Handbook.

US Department of Housing and Urban Development (2017, June). Model Lease for Subsidized Programs (90105a).

US Department of Housing and Urban Development (2023). About IMS/PIC - Public and Indian Housing.

A Linking the 2010 Census to HUD Survey Data

Since our HUD survey data did not contain information about which census block each property belongs to, we identified the block, tract, and county of each property using the Census Geocoder (US Census Bureau, 2023). We filtered out 1,528 properties where the Census Geocoder was unable to find a county, tract code, and block code. There were also 337 blocks that had more than one Section 8 property—for these blocks, we created a single Section 8 partition by aggregating the summary statistics across properties within a block. Since we identify violations by comparing household size to the number of bedrooms in the unit the household resides in, creating superproperties ensures that each Section 8 unit can only have one Section 8 household living in it. We also ignored 2,125 blocks where the number of households in the block provided by Summary File 1 was less than the number of occupied Section 8 units reported by HUD, as well as 220 blocks where the number of people reported living in Section 8 properties was greater than the number of people reported living in the block in Summary File 1. For these properties, the address of the property in our HUD dataset is in a different block than the households themselves—for example, when the leasing company's building is located in a different block from the housing property. Since we do not have access to shape files or unit-level coordinates to more accurately determine which block each housing property belonged to, we leave out these blocks and assume the rest are properly addressed.

B "Greedy" Baseline

For a given block, suppose there are N households sorted in ascending order by household size. Households $1, 2, \ldots, N_s$ are assigned to be Section 8 while households $N - N_s, N - N_s + 1, \ldots, N$ are assigned to be non-Section 8. Since households are ordered by increasing household size, the N_s smallest households in the block are Section 8 while the remaining $N - N_s$ largest households are non-Section 8.

Suppose there are U Section 8 housing units in the block sorted in descending order by number of bedrooms in the unit b_i, and each of these units i does not yet have a household assigned to live in it. Going through the Section 8 households in descending order by household size p_i ($N_s, N_s - 1, \ldots, 2, 1$) then for each household i, check if p_i can "fit" into b_0, the number of bedrooms in unit U_0, the unit with the most bedrooms among all remaining unassigned Section 8 units. p_i fits into b_j if the pair adhere to the "two heartbeats per room" HUD occupancy standard, $p_i \leq 2b_i$.

- If $p_i \leq 2b_0$, such that household i is not in violation, then we assign household i to live in U_0, removing unit U_0 from the list of unassigned units that households can be assigned to live in. Then U_0 is reassigned to be the next largest unassigned unit.

- If $p_i > 2b_0$, household i is considered to be in violation since there are no larger Section 8 units for i to fit into. This is true because we always try to fit households into the unassigned Section 8 unit with the most number of bedrooms, U_0, so if a household is unable to fit into even the largest unit, the household must be in violation. We then add household i to a running list of households in violation, which allows us to keep

track of the total number of households deemed in violation in the block. Unit U_0 remains an unassigned unit available to another Section 8 household.

For any block where $N > U$ (typically due to rounding error when converting the percentage of each bedroom size to a whole integer), instead of starting with household N_s, we start with the household at index $N_s - U$ in $N_s, N_s - 1, \ldots, 2, 1$.

C Integer Programming

For a given block, suppose there are N households, and that households $1, 2, \ldots, N_s$ are located in a Section 8 property. For each household i, we aim to reconstruct the following variables:

- $p_i \in \{0, 1, \ldots,$SF1[P]$\}$: the number of people in household i, which we conservatively assume can be no more than the total block population SF1$[P]$ recorded in Summary File 1.
- $r_i^j \in \{0, 1\}$: a binary indicator of whether household i contains a person of race/ethnicity $j \in J$ (e.g. White and non-Hispanic, White and Hispanic, etc.)
- $c_i \in \{0, 1, \ldots, p_i\}$: the number of people under 18 in household i.
- $b_i^k \in \{0, 1\}$: a binary indicator of whether the household lives in a unit with $k \in \{\leq 1, = 2, \geq 3\}$ bedrooms.

We select the following summary statistics from Summary File 1 from the 2010 Decennial Census (SF1) and the 2010 Section 8 HUD survey (HUD):

- SF1$[N_{p=x}] \; \forall \; x \in \{1, 2, \ldots, 6\}$: the number of households of size x reported in Summary File 1.
- SF1$[P]$: the block population reported in table P1 in Summary File 1.
- HUD$[P]$: the total number of residents reported in the HUD survey.
- SF1$[N_j] \; \forall \; j \in J$: the number of households with a person of race/ethnicity j reported in Summary File 1.
- HUD[Householder of race j] $\forall \; j \in J_{-h}$: the number of households with a householder of non-Hispanic race j reported in the HUD survey.
- HUD[Hispanic householder]: the number of households with a Hispanic householder reported in the HUD survey.
- SF1$[N_c]$: the number of people under 18 SF1$[N_c]$ reported in Summary File 1.
- HUD[1 adult with children], HUD[2 adults with children]: the number of households with 1 adult with children or 2 adults with children, respectively, reported in the HUD survey.

- HUD[Units with k bedrooms] $\forall\, k \in \{\leq 1, = 2, \geq 3\}$: the number of households with x bedrooms reported in the HUD survey.

We constrain the entire block with Summary File 1 statistics:

- $\left(\sum_{i=1}^{N} \mathbb{I}(p = x) = \text{SF1}[N_{p_i=x}]\right)\ \forall\, x \in \{1, 2, \ldots, 6\}$: the total number of households of size x matches Summary File 1.
- $\sum_{i=1}^{N} p_i \geq \text{SF1}[P]$: the total number of people in the block is at least the block population reported in Summary File 1. (We lower bound this and similar statistic in case of nonresponse or other inconsistency between Summary File 1 and the HUD survey.)
- $\left(\sum_{i=1}^{N} r_i^j \geq \text{SF1}[N_j]\right)\ \forall\ j \in J$: the total number of households with a person of race/ethnicity j is at least the number of households with a person of race/ethnicity j reported in Summary File 1.
- $\sum_{i=1}^{N} c_i = \text{SF1}[N_c]$: the total number people under 18 matches Summary File 1.

We also constrain the first N_s households by the statistics reported in the HUD Section 8 survey:

- $\sum_{i=1}^{N_s} p_i \geq \text{HUD}[P]$: the total number of people in Section 8 properties in the block is at least the total number of residents reported in the HUD survey.
- $\left(\sum_{i=1}^{N_s} r_i^j \geq \text{HUD}[\text{Householder of race } j]\right)\ \forall\ j \in J_{-h}$: the total number of Section 8 households with a non-Hispanic person of race $j \in J_{-h}$ is at least the number of households with a householder of race j reported in the HUD survey.
- $\sum_{i=1}^{N_s} \prod_{j=1}^{J_h} r_i^j \geq \text{HUD}[\text{Hispanic householder}]$: the total number of Section 8 households with a person of Hispanic origin is at least the number of households with a Hispanic householder reported in the HUD survey.
- $\sum_{i=1}^{N_s} \mathbb{I}(c_i > 0) = \text{HUD}[\text{1 adult with children}] + \text{HUD}[\text{2 adults with children}]$: the total number of Section 8 households with children under 18 matches the total number of households with children reported in the HUD survey.
- $\left(\sum_{i=1}^{N_s} b_i^k \geq \text{HUD}[\text{Units with } k \text{ bedrooms}]\right)\ \forall\ k \in \{\leq 1, = 2, \geq 3\}$: the number of Section 8 households with x bedrooms is at least the number of households with x bedrooms reported in the HUD survey.

Finally, we constrain the bedroom size in Section 8 units according to the "two heartbeats per room" rule: $\sum_{i=1}^{N_s} b_i^{\leq 1} \mathbb{I}(p_i > 2) = 0$ and $\sum_{i=1}^{N_s} b_i^{=2} \mathbb{I}(p_i > 4) = 0$. If a reconstruction satisfying all these constraints does not exist, the block must contain a household in violation of this occupancy limit.