

THE RISK OF LINKED CENSUS DATA TO TRANSGENDER YOUTH: A SIMULATION STUDY

ABRAHAM D. FLAXMAN AND OS KEYES

Institute for Health Metrics and Evaluation, University of Washington
e-mail address: abie@uw.edu

ABSTRACT. Every ten years, the United States Census Bureau collects data on all people living in the US, including information on age, sex, race, ethnicity, and household relationships. The Census Bureau is required by law to protect these data from disclosure where data provided by any individual can be identified, and, for the first two major releases from the 2020 Decennial Census, it used a novel approach to meet this requirement, the TopDown Algorithm.

We conducted a simulation study to investigate the risk of disclosing a change in how an individual's sex was recorded in successive censuses. In a simulated population based on a reconstruction of the 2010 decennial census of Texas, we compared the number of transgender individuals under 18 identified by linking ~~simulated~~ census data from a simulated 2010 and census and a simulated 2020 census under alternative approaches to disclosure avoidance, including swapping in 2020 (as used in 2010) and TopDown in 2020 (as used for the actual data released from the 2020 enumeration).

Our simulation assumed that in Texas, 0.2% of the 3,009,117 residents who were under the age of 8 in the 2010 census were transgender and would have a different sex reported in the 2020 census, and 23% would reside at the same address, which implied that 1,384 trans youth were at risk of having their transgender identity status disclosed by a reconstruction-abetted linkage attack. We found that without any disclosure avoidance in 2010 or 2020, a reconstruction-abetted linkage attack identified ~~691~~ 307 transgender youth. With 5% swapping in 2010 and 2020, it identified ~~657 individuals, a~~ 5283 individuals, an 8% decrease. With swapping in 2010 and TopDown in 2020 as configured for the actual data release, it identified 194 individuals, a ~~7030%~~ decrease from swapping. By comparing this to a scenario where sex data was replaced with completely random data, we found that the TopDown configuration attains ~~the maximum achievable a~~ baseline maximum level of privacy protection against such an attack, identifying no more individuals than in a scenario where all information on sex was replaced with random noise.

In light of recent legislative and media efforts to prohibit access to trans medicine—particularly for trans youth—our results demonstrate the importance of disclosure avoidance for census data and suggest that the TopDown approach used by the Census Bureau is a substantial improvement compared to the previous approach, achieving ~~the a~~ baseline maximum level of privacy protection ~~possible~~ against such a linkage attack.

Key words and phrases: census, disclosure avoidance, linkage attack.

This research was supported in part by a grant from the Alfred P. Sloan Foundation.

INTRODUCTION

As part of the 2020 decennial census, the US Census Bureau developed a new approach to disclosure avoidance, based on differential privacy, called the TopDown Algorithm (TDA) (Abowd et al. [2022]). The details of this approach have been refined iteratively since it first debuted as part of the 2018 end-to-end test (Garfinkel et al. [2019]). The release of the Demographics and Housing Characteristics (DHC) data in 2023 was the second application of TDA for a data product from the 2020 decennial census, following the release of the Public Law 94-171 redistricting data (released in August 2021). A series of demonstration DHC products from test runs on 2010 decennial data have also been released, in April 2023 (U.S. Census Bureau [2023]), August 2022 (U.S. Census Bureau [2022a]), and March 2022 (U.S. Census Bureau [2022b]), which help us understand plans and trade-offs for some of the TDA options previously enumerated (Petti and Flaxman [2019]).

In support of their work to develop and validate TDA, the Census Bureau has previously released a series of Privacy-Protected Microdata Files (PPMFs) by applying iterations of TDA to the 2010 Census Edited File. The DHC products from March and August of 2022 diverge from this pattern and provide summary tables without releasing a corresponding PPMF. This invites the question of whether the release of a PPMF or reconstruction of microdata from DHC tables might compromise privacy. In this work, we investigated empirically how well TDA protects against disclosure of sensitive information on an individual’s gender identity in DHC data.

Past investigations of demonstration products have focused primarily on the impact of TDA on accuracy of key census-derived statistics, and we agree that there are broad, political implications behind statistical accuracy; the framing of census data informs everything from the shape and number of legislative districts to funding and resourcing for minority groups (see, for example, Thompson [2012]). But this is also true of privacy—accurate representation is not an unalloyed good. For many groups, particularly those who are vulnerable to and have experienced active discrimination by state entities, higher accuracy can also mean higher identifiability and higher *scrutiny*. An example of these concerns surfacing can be seen in the debate surrounding asking questions about citizenship in the decennial census. The 2020 decennial census ended up *not* including such a question, and proposals to include it were ultimately shelved amidst widespread concern from public advocacy groups that the question could be used to identify, surveil, and punish undocumented immigrants, and further lead to reduced engagement with and trust of the census (see Levitt [2019], Baum et al. [2022]). More recent in the public eye (although just as longstanding, as highlighted by Canaday [2009]) are questions of gender (Singer [2015]), on which this investigation is focused.

The last few years have seen heightened scrutiny of transgender (henceforth “trans”) people, with a particular focus on (and moral panic around) trans youth (see Slothouber [2020]). This has included actions by state actors to simultaneously legislate against access to care and equal treatment and use existing mechanisms of government to punish the youth and parents who have become identifiable. Most prominently, the governor of Texas, in Abbott [2022], has directed the state Department of Family and Protective Services to investigate the parents of any trans youth who receives gender-affirming medical care. In order to do so, he advocates drawing on existing systems for child and parent surveillance, including abuse reporting requirements, to identify targets.

As all of this suggests, there are many reasons for us to be cautious around data availability and the pursuit of accuracy as an exclusive goal of work in this area. While accuracy is (obviously) beneficial from a statistical perspective, an absence of privacy simultaneously risks both producing real, material harms for the individuals identified and undermining trust in the census itself and so paradoxically reducing the very accuracy that is aimed for. To demonstrate the importance of factoring identifiability into account—and the necessity of an emphasis on disclosure avoidance in census policy—we used simulation to investigate a risk to privacy, by focusing on the risk of disclosing a child’s transgender status through discordant reporting of binary gender in successive censuses.

METHODS

We used computer simulation to compare the number of trans youth who might be identified in a simulated population under alternative scenarios of disclosure avoidance. Our approach began with a simulated population of size and structure similar to the state of Texas, derived from a reconstruction of the US population on April 1, 2010. To obtain this reconstruction, for each census block, we used integer programming to ~~maximum~~maximize a random objective function for a fully saturated contingency table of stratified person counts (stratified by age, sex, race, and ethnicity) subject to constraints derived from census tables P8, P9, P11, ~~P12~~P12A-N, and P14, which provided person counts for each census block stratified by combinations of age, sex, race, and ethnicity. Table P14 was particularly valuable for this youth-focused reconstruction, because it provided counts stratified by sex and single-year age group for those under age 20.

Since our focus is on linking youth between simulated versions of the 2010 and 2020 decennial censuses, we ~~included simulants from this population constructed our simulated 2020 population by including simulants from our reconstruction~~ who were aged 10 to ~~17 and therefore would be 17~~. By construction, these simulants were under 18 on April 1, 2020, ~~but and were~~ also be born by before April 1, ~~2010–2010~~, and therefore might potentially be linked between the two censuses. We augmented this ~~reconstruction by assigning the simulated population by assigning each~~ simulant’s gender based on tabulated statistics summarizing responses to the Sexual Orientation and Gender Identity (SOGI) module of the Behavioral Risk Factors Surveillance System (BRFSS) collected in 2019 (National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health [2019]). To avoid making the simulation more complicated than necessary, we did not include correlation between gender and any other simulant attribute (such as race, ethnicity, or age).

We initialized each simulant with attributes for age, gender, race, ethnicity, and household, where age was an integer value representing the age in years, gender was a five-valued variable (with values of transgender boy; transgender girl; other transgender; cisgender boy; and cisgender girl), race was a 63-valued variable encoding the possible combinations of the six census racial categories, ethnicity was a two-valued variable for Hispanic/non-Hispanic, and household was an identifier that encoded census geography (state, county, tract, block) as well as housing unit ID.

From this simulated 2020 population, we then simulated time going backwards to 2010 and used these populations to generate simulated census data as follows: we recorded the age in 2020 precisely for each simulant’s reported age in the 2020 census, and then that age minus 10 for each simulant’s reported age in the 2010 census. We used a simple model of the other key demographic factors of births, deaths, in-migration, and out-migration to simulate

how this population might have changed over the decade: since our interest was in linking between censuses, we focused on migration and posited that every household might have moved, making it harder to link. To realize this household mobility, we selected households to stay unmoved from 2010 to 2020 independently, with probability value $p_{\text{stay}} = 23\%$, which we derived from the American Communities Survey (we obtained this value by calculating the sample-weighted proportion of with-children households that had been in residence for at least 10 years in the 2020 five-year Public Use Microdata Sample). We then changed the 2010 address of each non-staying household by selecting a new household for them to move from uniformly at random from all simulated households in Texas that were occupied on Census Day 2010, which offered a simple way to include realistic heterogeneity in population density.

Finally, we simulated the reported value of sex on the 2010 and 2020 decennial census. Our model of reported sex started from the assumption—uncertain though it is—that, in the 2010 census, nearly all of the transgender youth aged zero to 7 had their sex reported based on their sex assigned at birth. We then assumed that some of the simulants who are transgender would report their sex differently in the 2020 census. Based on this premise, we simulated responses on the 2010 and 2020 censuses according to the following cases: for cisgender boy simulants, we recorded their sex as male in 2010 and 2020, and similarly, for cisgender girl simulants we recorded female. For transgender boy simulants, we recorded their sex as female in 2010 and recorded their sex with a value chosen uniformly at random from the set {male, female} in 2020. Similarly for transgender girl simulants, we recorded their sex as male in 2010 and with a value of female in 2020 with probability 50%. For other transgender simulants we recorded their sex as the same value in 2010 and 2020, with the value chosen uniformly at random from the set {male, female}.

We recorded race and ethnicity identically in 2010 and 2020, matching the value of the simulant’s race and ethnicity attributes.

We compared four alternative scenarios of disclosure avoidance: (1) extreme disclosure where names were published, allowing even households that moved to be linked between censuses; (2) tables with no disclosure avoidance, where names were not published, but there was no effort to protect against disclosure by swapping or otherwise perturbing the data in published tables; (3) disclosure avoidance by swapping, where 5% of households were exchanged with another household to protect privacy; and (4) differentially private disclosure avoidance, where the new TDA approach was used to protect against disclosure in published tables. We also included a scenario where we replaced the sex data with completely random data, as a way to understand ~~the maximum achievable~~ a baseline maximum level of privacy protection. We now describe our method of quantifying how many transgender simulants would have their gender identity revealed in each of these scenarios.

Extreme disclosure (Scenario 1): In this scenario (which is purely hypothetical and clearly in violation of law), we assumed that linking on name, age, race, and ethnicity would be able to identify nearly all simulants with discordantly reported values for sex in the 2010 and 2020 censuses. We therefore counted all simulants with differing values reported for sex in 2010 and 2020 to estimate the number of trans youth who would have their gender identity revealed if census microdata including names were released. Prior to beginning analysis, we hypothesized that this would total in the thousands or perhaps even tens of thousands.

No disclosure avoidance (Scenario 2): In this scenario, we assumed that the only simulants at risk of being identified as trans youth by a reconstructed-abetted linkage attack

were those aged 7 and younger in 2010 who had a unique combination of age, race, and ethnicity in their census block (sometimes referred to as “population uniques”). This is a simplifying assumption, however, and in practice there might be risks to individuals who are nearly unique as well. Furthermore, we assumed that individuals who moved residence between the 2010 and 2020 censuses would not have their transgender status revealed, and even individuals who were exposed by a unique combination of attributes in 2010 and did not move by 2020 *might* not have their transgender status revealed, if in-migration to their census block resulted in them no longer having a unique combination of attributes in 2020. The simulants who had a unique combination of age, race, and ethnicity in their (unmoved) census block in 2010 and 2020 could be identified by deterministic record linkage on these attributes. (Or to be more precise, the existence of such a simulant in the census block would be revealed, which would need to be confirmed by further investigation.) We therefore identified all simulants who did not move and had a unique combination of attributes in 2010 and also in 2020, and counted the simulants in this group with differing values reported for sex in 2010 and 2020. ~~This constituted our~~ However, some of these individuals might be protected from disclosure due to tabulations that hide population uniques through aggregation error. To bound the protection afforded by this aggregation error, We created tabulations corresponding to census tables P8, P9, P11, P12A-N, P14 for each census block containing such an individual in 2010 and 2020 and reconstructed microdata using integer programming to see if these individuals were still population uniques after introducing this aggregation error. The number who remained population uniques after introducing aggregation error in this manner constituted our lower bound estimate of the number of trans youth who would have their gender identity revealed by a reconstruction-abetted linkage attack if the tables used for reconstruction were published with no disclosure avoidance measures. We hypothesized that this would total in the hundreds.

Swapping for disclosure avoidance (Scenario 3): We approached this scenario similarly to Scenario 2, but instead of using each simulant’s geography directly in the reconstruction-abetted linkage attack, we first chose a random subset of simulants to have their reported location swapped to somewhere other than their true location. We achieved this with a simple model analogous to the model of migration described above, where we selected some households to report in a location that is not their actual location independently, with probability $p_{\text{swap}} = 5\%$ (we chose this value as a modeling assumption broadly aligned with the publicly available information about the Census Bureau’s approach to disclosure avoidance in the 2010 decennial census). For each of the selected households, we chose a reported location to swap in by selecting a household uniformly at random from all simulated households in Texas on Census Day 2010.

We then identified all simulants who did not move, according to their (possibly swapped) reported location in the 2010 and 2020 censuses, who had a unique combination of age, race, ethnicity, and geography attributes recorded in both censuses, and ~~counted identified~~ the simulants in this group with differing values reported for sex in 2010 and 2020. We then tabulated and reconstructed the individuals in each block with such as simulant, as described in Scenario 2, and counted the simulants who remained population uniques after introducing this aggregation error. This constituted our estimate of the number of trans youth who would have their gender identity revealed by a reconstruction-abetted linkage attack if the tables used for reconstruction were protected by swapping. We hypothesized that this total would be 5% to 10% lower than the total from the no-disclosure-avoidance scenario and therefore would also reveal sensitive information about hundreds of trans youth.

TDA for disclosure avoidance (Scenario 4): For this scenario, we used the Census Bureau’s DHC demonstration product released in April 2023 to generate our estimate of the risk of a reconstruction-abetted linkage attack in this scenario. The final DHC demonstration product included a Privacy-Protected Microdata File (PPMF) for the 2010 decennial census, consisting of a row for each individual and columns for the attributes of age, sex, race, ethnicity, and geography, and we used this for our 2020 simulated data (U.S. Census Bureau [2023]). We used the same reconstructed microdata file described above, which was based on the SF1 tables published as part of the 2010 decennial census (and therefore used swapping to protect against disclosure), and initialized our simulated population in 2020, based on the individuals aged 10 to 17 in this reconstructed microdata. We then simulated the regression of time, going backward from 2020 to the 2010 Census Day, when each simulant would be 10 years younger. We applied our migration model to keep the location in 2010 identical to that in 2020 for only a random fraction of simulants, governed again by the parameter p_{stay} .

As in the other scenarios, we endowed each simulant with a gender attribute, which we calibrated to match to measurements from the 2019 BRFSS SOGI module. This allowed us to use the demonstration DHC as our proxy for the privacy afforded by TDA in 2020 in our assessment of the number of trans youth who would have their gender identity revealed by a reconstruction-abetted linkage attack using data protected by swapping in 2010 and TDA in 2020.

To complete this approach, we identified all simulants who had a unique combination of age, race, and ethnicity in their simulated census block in 2010, and identified which of these simulants matched a unique individual aged 10 years older in the DHC PPMF. For each of these simulants, we then compared the reported sex in the simulated 2010 census with the reported sex in the simulated 2020 census. We counted how many of these links were for simulants who were trans youth. We hypothesized that this number would be at least an order of magnitude smaller than the total from the swapping-for-disclosure-avoidance scenario.

To establish a ~~theoretical minimum~~ baseline on the number of simulants who might be identified by such an attack, we also considered a case where all information on sex was replaced with random noise.

RESULTS

Our simulated population included 25,145,561 individual simulants, matching exactly the 2010 population count for Texas. We focused on the simulants aged 8 to 17, of which we had 3,009,117, and used this to initialize our Census Day population in 2020. Among these simulants, 0.53% were trans, with 0.18% trans boys, 0.23% trans girls, and 0.12% other transgender (closely matching the BRFSS values), which led to 0.20% of the simulants having a different value reported for sex in the 2010 and 2020 censuses. Over the 10 years of simulation, the majority of households moved at least once and only 23% of simulants resided in the same census block in 2010 and 2020 (closely matching the ACS values). Taking these together implied that in our simulation there were $(3,009,117 \text{ youth}) \times (0.2\%) \times (23\%) = 1,384$ trans youth who were at risk of having their gender identity disclosed by this reconstruction-abetted linkage attack.

We found that in our scenario with extreme disclosure, where individual-level data with linkable names were published (Scenario 1), linking between 2010 and 2020 census data to identify individuals with discordantly reported values for sex would identify over 6,000 trans

youth, accounting for 38% of all trans youth in our simulated version of Texas (the remaining 62% were not identified because their reported sex was concordant in both censuses).

In our scenario where tables like those in SF1 or DHC were published precisely as enumerated, without any disclosure avoidance measures applied (Scenario 2), we found that migration and non-uniqueness substantially reduced the number of trans youth whose gender identity was revealed. However, there were still 1,722,991 individuals who were uniquely identified by the age, race, ethnicity, and location in 2010 and 1,413,801 uniquely identified in 2020. In our simulation, we found that if a reconstruction-abetted linkage attack in this scenario was able to attain perfect reconstructed it would identify 691 trans youth. After introducing aggregation error through tabulation and reconstruction, we found that the attack would still identify 307 trans youth.

In our next scenario (Scenario 3), we added swapping-based disclosure avoidance to the tables in Scenario 2 and found that with respect to a reconstructed-abetted linkage attack, swapping acted similarly to a small boost in migration to prevent identifying trans youth. At the 5% swapping level we used in Scenario 3, we found that a reconstruction-abetted linkage attack with perfect reconstruction identified 657 trans youth, a 5. After introducing aggregation error, we found that the attack would still identify 283 trans youth, an 8% reduction from the number identified in Scenario 2.

Our final scenario was designed to represent to the approach taken by the Census Bureau in their 2020 data release. In this scenario, we considered protecting the tables released from the 2010 census with swapping and the tables from the 2020 census with TDA (Scenario 4). We found that this afforded substantially more protection than the other scenarios we considered. We found that TDA was successful in preventing the bulk of the identifications from Scenario 3; in our simulation, a reconstruction-abetted linkage attack identified 194 trans youth when TDA was used for disclosure avoidance on the 2020 tables, a ~~70~~30% reduction in the number identified when swapping was used in Scenario 3. This is actually slightly lower than the number identified purely by chance; when we replaced the reported sex columns in the PPMF by a value chosen uniformly at random, the attack identified 214 trans youth.

1. DISCUSSION

Our simulation results demonstrate the magnitude of the threat that a linkage attack designed to identify trans youth might pose. Were the Census Bureau to publish microdata on the 2010 and 2020 censuses (Scenario 1), it would likely identify the transgender status of over 6,000 trans youth in Texas. In the approach underlying the most recent demonstration data, on the other hand, a reconstruction-abetted linkage attack would likely identify the transgender status of only 194 trans youth in Texas. Since this is less than the number of trans youth identified when we replaced reported sex with a uniformly random value, it shows that TDA as configured for the 2020 DHC attains ~~the maximum achievable our~~ baseline maximum level of privacy protection against this reconstruction-abetted linkage attack, identifying no more individuals than in a scenario where information on sex was replaced entirely by random noise.

This is yet another demonstration of the need for disclosure avoidance in the decennial census. The bulk of previously published investigations into the quality of TDA demonstration products have compared with published results from the 2010 census and often reported differences. But in such comparisons, there is an important limitation because they compare

the (published) results of swapping to the (demonstration) results of TDA applied to the unswapped data. Thus, the conclusion of such a comparison is typically limited to proving that the noise introduced by TDA is different from the noise from swapping. This investigation turns this limitation into a strength, since a reconstruction-abetted linkage attack between 2010 and 2020 decennial censuses *will* be linking data that have been swapped with data that have been protected by TopDown. Modeling in a simulation framework like the approach developed here could potentially also be used in future investigations to more directly compare the noise introduced by swapping to the noise introduced by TDA. This approach could also explore how TDA performs in protecting against identifying other groups, such as children who are adopted or people who are in mixed-race marriages.

Limitations: There are at least three simplifying assumptions in this simulation model that constitute limitations which might be the focus of future work. First, the migration model is quite simplistic, and it is likely that further investigation could more accurately incorporate determinants of migration; the probability that a household has stayed unmoved between decennial censuses is likely to vary by household income, for example, which is an attribute that we did not include in our simulation but could potentially add. Second, our simple model of how sex was reported for trans youth could also be more complex, although it is less clear what sources of data could inform adding this complexity. Third, in this work we assumed that race and ethnicity were unchanged between 2010 and 2020 censuses, but it is likely that evolving conceptions of race and ethnicity have led to some recording of differing values for some individuals, and this would result in some reduction in the number of links in a linkage attack. We conjecture that none of these simplifying assumptions have substantially changed the relative number of trans youth identified in our scenarios, however.

As mentioned in the methods section, our approach using a simulation that moves simulants backwards through time is less familiar and therefore more complicated than we would have liked. However, the approach we used in this work has a strength alluded to above in Scenario 4, because it uses SF1 data that have been perturbed by the swapping approach actually employed by the Census Bureau in the 2010 decennial census, the details of which are not publicly available.

We would also like to emphasize four limitations specific to our model of trans youth. First, our assumption of a uniform probability of markers changing between census years is overly simplistic; we would expect that, in practice, the likelihood of changes is variable depending on both the respondent family’s context and the individual perspective of the child and their parent(s). Second, the limited range of sex options on the census means that many trans youth whose identities fall outside a simplistic binary do not alter their census markers. Third, we did not include any correlation between gender and other simulant attribute, such as race, ethnicity, or age. While we believe that this simplifying assumption is not likely to change our results substantially, a more detailed model could highlight relevant disparities in disclosure risk. Finally, we would expect differences in the amount of geographic mobility and consistency in household structure for trans families writ large. While one response to increasing scrutiny, at least for those with means, is to purposefully move their household, it seems logical that in some cases the risk of increased scrutiny will lead others to purposefully *not* move—to avoid relocating to areas that might be more hostile. The latter of these considerations could suggest this is in fact the *minimal* count of trans people identifiable through the current census approach to data disclosure and, even with the success in simulation of the TDA data disclosure approach, well-intended efforts to increase the ability of Census Bureau instruments to record and represent trans people

(see [White House \[2022\]](#)) could increase the risk of identifiability and harm. Further, as our reference to “means” suggests, there is likely a particular overrepresentation of trans people whose families are financially vulnerable—not only reducing their ability to relocate, but also their ability to resist and survive state pressure.

Although the focus of this piece is on trans *youth*—specifically, those under 18 in both the 2010 and 2020 census, with different sex records in each—it is worth emphasizing that they are not the only people at risk. With the addition of more census tranches (say, 2000, or, going forward, 2030), the range of people at risk of disclosing their transgender status would expand to include trans adults, many of whom, if they have children, are also being targeted for additional scrutiny by state bodies. This is related to an additional risk that *future* data releases by the Census Bureau might be combined with existing data to refine an attack like that investigated in Scenario 4.

Due to data limitations, we had to use computer simulation to conduct this investigation, but it would be possible for the Census Bureau to replicate and expand on analyses such as this one internally, where they can use private data such as the Census Edited File, which is not available to outside researchers. The Census Bureau could reproduce this analysis using its internal data to understand how reality differs from our simulation. We encourage them to share with us how much this risk differs in the true implementation from the risk as modeled in this simulation.

Another direction for future research would be to refine the attacks considered in our scenarios 2-4 to target individuals who are nearly unique to their strata, such as an individual who is one of two people of a given race and age in a census block. This might require innovation in reducing the false discovery rate implicit in the simple attack strategy that we have investigated here. Exploring the risks of methods for reducing the false discovery rate, such as using additional relational information, could be a fruitful direction for future work as well.

We have made a replication archive of this work available online: https://github.com/aflaxman/linked_census_disclosure

ACKNOWLEDGMENT

The authors wish to acknowledge fruitful discussions with danah boyd and Simson Garfinkel.

REFERENCES

- G. Abbott. Letter to Masters, 2022. URL <https://www.documentcloud.org/documents/21272649-abbott-letter-to-masters>.
- J. Abowd, 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. The 2020 Census Disclosure Avoidance System Top-Down Algorithm. *Harvard Data Science Review*, (Special Issue 2), jun 24 2022. <https://hdr.mitpress.mit.edu/pub/7evz361i>.
- M. A. Baum, B. J. Dietrich, R. Goldstein, and M. Sen. Sensitive questions, spillover effects, and asking about citizenship on the US Census. *The Journal of Politics*, 84(3):1869–1873, 2022.
- M. Canaday. The straight state. In *The Straight State*. Princeton University Press, 2009.

- S. Garfinkel et al. 2018 end-to-end test disclosure avoidance system design specification. Technical report, U.S. Census Bureau, 2019.
- J. Levitt. Citizenship and the Census. *Columbia Law Review*, 119(5):1355–1398, 2019.
- National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. 2019 BRFSS survey data and documentation. Technical report, 2019. URL https://www.cdc.gov/brfss/annual_data/annual_2019.html.
- S. Petti and A. Flaxman. Differential privacy in the 2020 US census: What will it do? Quantifying the accuracy/privacy tradeoff. *Gates Open Research*, 3, 2019.
- T. B. Singer. The profusion of things: the “transgender matrix” and demographic imaginaries in US public health. *Transgender Studies Quarterly*, 2(1):58–76, 2015.
- V. Slothouber. (de) trans visibility: moral panic in mainstream media reports on de/retransition. *European Journal of English Studies*, 24(1):89–99, 2020.
- D. Thompson. Making (mixed-) race: census politics and the emergence of multiracial multiculturalism in the United States, Great Britain and Canada. *Ethnic and Racial Studies*, 35(8):1409–1426, 2012.
- U.S. Census Bureau. 2010 demonstration data for the Demographic and Housing Characteristics file (DHC) (v. 2022-08-25). Technical report, 2022a. URL https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-08-25_Summary_File/.
- U.S. Census Bureau. 2010 demonstration data for the Demographic and Housing Characteristics file (DHC) (v. 2022-03-16). Technical report, 2022b. URL https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-03-16_Summary_File/.
- U.S. Census Bureau. 2010 demonstration data for the Demographic and Housing Characteristics file (DHC) (v. 2023-04-03). Technical report, 2023. URL https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/04-Demonstration_Data_Products_Suite/2023-04-03/.
- White House. FACT SHEET: Biden-Harris administration advances equality and visibility for transgender Americans, 2022.