An Examination of the Alleged Privacy Threats of Confidence-Ranked Reconstruction of Census Microdata

David Sánchez[1][0000−0001−7275−7887], Najeeb Jebreel[1][0000−0002−4911−3802], Krishnamurty Muralidhar[2][0000−0002−7239−7356], Josep Domingo-Ferrer[1][0000−0001−7213−4962], and Alberto Blanco-Justicia[1][0000−0002−1108−8082]

[1] Department of Computer Science and Mathematics, CYBERCAT–Center for Cybersecurity Research of Catalonia, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Catalonia
[2] Department of Marketing and Supply Chain Management, Price College of Business, University of Oklahoma, 307 West Brooks, Adams Hall Room 10, Norman, OK 73019, U.S.A.
{david.sanchez,najeeb.jebreel,josep.domingo,alberto.blanco}@urv.cat, krishm@ou.edu

Abstract. The threat of reconstruction attacks has led the U.S. Census Bureau (USCB) to replace in the Decennial Census 2020 the traditional statistical disclosure limitation based on rank swapping with one based on differential privacy (DP), leading to substantial accuracy loss of released statistics. Yet, it has been argued that, if many different reconstructions are compatible with the released statistics, most of them do not correspond to actual original data, which protects against respondent reidentification. Recently, a new attack has been proposed, which incorporates the confidence that a reconstructed record was in the original data. The alleged risk of disclosure entailed by such confidence-ranked reconstruction has renewed the interest of the USCB to use DP-based solutions. To forestall a potential accuracy loss in future releases, we show that the proposed reconstruction is neither effective as a reconstruction method nor conducive to disclosure as claimed by its authors. Specifically, we report empirical results showing the proposed ranking cannot guide reidentification or attribute disclosure attacks, and hence fails to warrant the utility sacrifice entailed by the use of DP to release census statistical data.

Keywords: U.S. Decennial Census 2020, Statistical Disclosure Limitation, Reconstruction, Differential Privacy, Reidentification

1 Introduction

The U.S. Decennial Census is the world's most prominent census data release, accounting for more than 330 million people. The policies implemented by the

U.S. Census Bureau (USCB) —the statistical agency in charge of compiling, managing, and releasing the U.S. Census data— carry therefore a great influence on the decisions made by many other statistical agencies.

For the Decennial Census 2020 release, the USCB decided to replace the statistical disclosure limitation (SDL) methodology they had been using in previous editions (based on data swapping) with a method based on differential privacy (DP). The reason for such a change was the alleged disclosure risks resulting from potential record reconstruction attacks on the prior 2010 Census release [1,10]. These works claim it is possible to accurately reconstruct the individual records from the released census statistics and that this reconstruction entails serious privacy threats to citizens. To prevent accurate reconstruction, a DP-based method consisting in adding random noise to the released statistics was implemented in the 2020 release [2].

DP is a robust privacy-enhancing method when properly implemented [7], but it also introduces significant logical inconsistencies and inaccuracies in the protected data due to its random and perturbative nature [21,8]. In fact, in many cases the 2020 Census DP-protected data have been manipulated to the point where relationships between variables are not possible on the ground: there are blocks with more households than household population, blocks with no population in households yet having occupied house units, or blocks with non-zero population but with no adults [16]. In this respect, the quality of block-level data in the 2020 release was so poor that the USCB acknowledged that block-level data should not be used for any meaningful analysis [24]. Given that census data are crucial for research and social decision-making, one can understand the consternation of potential users [13,19,11,22].

Several authors have demonstrated that the claims of disclosure risk ensuing from the reconstruction attacks in [1] and [10] were vastly overstated [19,17]. A major issue is that there are usually a lot of different reconstructions compatible with the released output, which makes it impossible for the attacker to know which is the good reconstruction (the one corresponding or closest to the original data) [17,18].

After assessing the situation, the USCB decided to abstain from using DP to protect the American Community Survey for the foreseeable future since "it's also not clear that differential privacy would ultimately be the best option" [4].

However, a recent work by Dick *et al.* claims it is possible to rank records reconstructed by an attacker to reflect how likely they are to belong to the original data [5]. The authors propose a *confidence-ranked reconstruction attack* (**CRR**) that reconstructs records from the published census statistics and ranks them by how frequently they appear in multiple reconstructions. The authors assert that their ranking can be used by an adversary to conduct a variety of targeted attacks on individuals because the highest-ranked reconstructed records have a high chance of appearing in the original data. As a conclusion, they raise "sober warnings on the privacy risks of releasing precise aggregate statistics of a dataset" and note that "the only defenses against [reconstruction attacks] are to introduce imprecision in the underlying statistics themselves, as techniques like

differential privacy do". This is precisely what the USCB did in the 2020 Census release [12].

Despite the issues of the 2020 Census protection, CRR renewed the interest in using DP in census releases. Specifically, in a subsequent article entitled "Database Reconstruction Does Compromise Confidentiality", [12] —by the current Chief Scientist at the USCB and her immediate predecessor— fully endorsed Dick *et al.*'s conclusions, both regarding the privacy threats of CRR, and the advice to use DP to protect the released statistics.

Due to the (observed) adverse consequences of using DP in Census releases (or in any data release, since that was not the scenario DP was designed for [7,3]), the flaws of previous reconstruction attacks, and the influence that the USCB's decisions may have on other statistical agencies or social science in general, the claims by Dick *et al.* deserve detailed scrutiny.

Contributions and plan

In this paper, we empirically demonstrate the inability of the CRR attack to threaten privacy in any meaningful manner. In particular, we show that: (i) the highest ranked records according to Dick *et al.*'s confidence ranking are also the most common and, hence, the most inherently protected against reidentification; (ii) rare or unique records, whose reconstruction might put the corresponding individuals at risk of disclosure, go unnoticed by the proposed confidence ranking; and (iii) the inaccuracy of the reconstruction and the large diversity of the non-existent records it generates (records that do not exist in the original data) render attribute disclosure attacks ineffective.

The rest of the paper is organized as follows. Section 2 reviews the CRR attack. In Section 3 we detail how we replicated Dick *et al.*'s experiments. Sections 4 and 5 assess the actual reidentification and attribute disclosure risks implied by the CRR attack. Finally, we present some conclusions.

2 Reviewing the confidence-ranked reconstruction attack

In and of itself, reconstruction poses no privacy risk unless *it is accurate*. In the following, we demonstrate that the confidence-ranked reconstruction that Dick *et al.* propose is ineffective both at accurately reconstructing Census records and at detecting, even approximately, records at risk.

The primary purpose of the CRR is to attach a confidence level to each reconstructed record that measures how likely it is for that record to appear also in the data they use as ground truth. This ground truth consists of synthetic microdata published by the USCB that closely resemble the real 2010 Census microdata. By "casting the reconstruction problem as an instance of large-scale, non-convex optimization, along with a subsequent step to convert non-continuous (*e.g.*, categorical) features back to their original schema" [5], they rank reconstructed record prototypes by how frequently they appear in multiple reconstructions. A *record prototype* is a record type with some *multiplicity* (that is, the number

of repetitions) in the microdata. We talk about prototypes rather than actual records because CRR is incapable of ascertaining the multiplicity of those prototypes and, therefore, *it cannot produce an accurate reconstruction of the synthetic microdata.*

Nevertheless, the authors interpret the rank of the reconstructed records as a measure of risk, because the empirical results they report show that the top k ranked (*i.e.*, most frequent) reconstructed prototypes were present in the synthetic data set (D) in a large proportion. More specifically, they argue that the most confident/highest ranked records they reconstruct are at risk because they are those that are most likely to appear in the original data and, therefore, can be the target of a variety of privacy attacks, including "identity theft".

Dick *et al.* measure the effectiveness of their CRR attack as the proportion of the top k ranked prototypes (*i.e.*, the most frequently reconstructed ones) that were present in D. According to their results, this proportion was near 1 for small values of k. However, since the measured proportion considers the number of record prototypes in D (without counting their repetitions), rather than the actual number of records (counting repetitions), the authors are neglecting the multiplicity of each prototype in D, which is key to privacy: a prototype appearing, say, 10 times in D means that 10 individuals share the same record values for the considered attributes and, therefore, those individuals are intrinsically protected against reidentification (they are 10-anonymous, in terms of the k-anonymity privacy model [20], and the probability of successfully reidentifying one of them is at most 1/10).

To better understand the importance of this aspect and to illustrate the practical ineffectiveness of the CCR attack, in the following, we exactly replicate the experiments done by Dick *et al.* on the same data, and report the number of repetitions in D of their reconstructed prototype (which the authors did not do).

3 Replicating Dick et al.'s experiments

Our experiments have been done using the code and settings provided by Dick *et al.*[3]. The additional code we have written for our experiments is also available for reproducibility [4]. Specific details follow.

Data Set: We used the same subsets of synthetic U.S. Census microdata as ground truth. This dataset was released by the USCB and closely resembles the real 2010 Census microdata in terms of statistical characteristics. Specifically, we used the 2020-05-27 vintage Privacy-Protected Microdata File (PPMF) [23], which consists of 312,471,327 rows representing synthetic responses for individuals in the 2010 Decennial Census. The columns in the PPMF include attributes such as the respondent's home location (state, county, census tract, and census block), housing type, sex, age, race, and Hispanic or Latino origin.

[3] `https://github.com/terranceliu/rap-rank-reconstruction`
[4] `https://github.com/NajeebJebreel/CRR-analysis`

Statistical Queries: Dick *et al.*'s reconstruction was executed on the tables employed by the USCB for their internal reconstruction attack on the 2010 Census data. Each table defines a set of statistical queries that are performed on the (protected) Census microdata. These queries involve specifying column names, subsets of column domains, and census block or tract identifiers. The objective is to count the number of microdata rows that satisfy the specified criteria. The tables include P1 (total population), P6 (race), P7 (Hispanic or Latino origin by race), P9 (Hispanic or Latino and Not Hispanic or Latino by race), P11 (Hispanic or Latino and Not Hispanic or Latino by race for the population 18 years and over), P12 (sex by age for selected age categories), P12 A-I (sex by age for selected age categories iterated by race), PCT12 (sex by single year age), and PCT12 A-N (sex by single year age iterated by race). The P tables are released at the block level, while the PCT tables are released only at the census tract level.

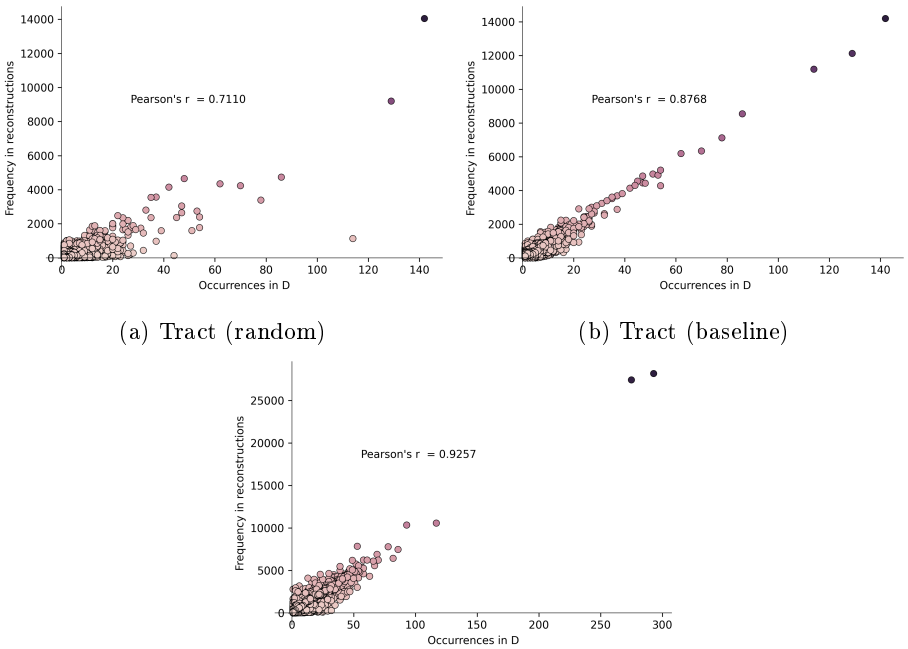
Experiments: We used subsets of the PPMF, which comprised all the rows belonging to specific census tracts or blocks. For tracts, we used the same random sample employed by Dick *et al.* On the other hand, blocks were selected according to the following (non-random) criteria described by the authors [5]: for each state, choose the block closest in size to the mean block size as well as the largest block; in addition, choose blocks closest in size to M/C, where M is the maximum block size in the state and $C \in \{2, 4, 8, 16\}$. This resulted in a total of 50 tracts and 300 blocks. Then, we reconstructed 100 data sets for each original data set, by employing both *baseline* and *random* initialization methods, while adhering to the same configuration and parameters. Baseline initialization follows the distribution of D —assumed to be publicly known— to set up the algorithm's parameters, whereas random initialization leverages no prior information.

4 Results and discussion

Having replicated Dick *et al.*'s experiments, we investigated the correlation between the multiplicities of records in D and their corresponding rank in the reconstructed data. Our aim was to determine whether highly ranked records threaten the privacy of individuals as Dick *et al.* claim.

To do so, we counted the number of occurrences of each record prototype in D and we compared it with its frequency in the reconstructions (which is the foundation of the confidence ranking). Figure 1 reports this comparison both at the tract and block levels and both with random and baseline initializations. We also report results for tracts without considering the block attribute (random initialization), which Dick *et al.* removed to ease tract reconstructions.

Our results clearly show that records with larger multiplicity in D have a higher frequency (rank) in the CRR. This dependency is very strong in all cases (Pearson's correlation r between 0.71 and 0.99), but especially for block random or baseline initialization (where $r > 0.96$). As a matter of fact, in all cases, the record with the largest multiplicity in D was also the most frequent/top-ranked prototype in the CRR. Since the confidence ranking is based on the frequency



(a) Tract (random)

(b) Tract (baseline)

(c) Tract without block (random)

(d) Block (random)

(e) Block (baseline)

Fig. 1: Comparison between the multiplicities of record prototypes in D and their frequency in CRR.

of record prototypes in multiple reconstructions, the top-ranked prototypes are the k most common prototypes in D; that is, those with the greatest number of repetitions and, thus, those intrinsically most private. Therefore, all records top-ranked by CRR (*i.e.*, those claimed to be at risk) are intrinsically protected against reidentification because "our privacy is protected to the extent that we blend in with the crowd" [9].

Conversely, rare or unique records in D, which are actually vulnerable to reidentification because of their rarity, do not appear among the top-ranked k. Hence, records really at risk stay paradoxically undetected by CRR.

By examining the distribution of D, we see that the percentage of unique records is 10.17% at the block level and 1.85% at the tract level. Hence, 89.83% are non-unique at the block level and 98.15% are non-unique at the tract level. Also, the multiplicity of the most common records is in the order of dozens at the tract level and in the order of hundreds at the block level. In terms of the well-known k-anonymity model [20], this means that unequivocal reidentification of any of those 89.83% and 98.15% records is not possible and that the probability of correct reidentification is as low as $1/140$ at tract level ($1/300$ when dropping the block attribute), and $1/800$ at the block level. From these results, it can be seen that CRR would still (wrongly) claim high rates if D was already protected via k-anonymity, *i.e.*, all prototypes in it had multiplicity $k > 1$. Therefore, it is clear that CRR does not capture the reidentification risk in D.

Notice that, since tracts are subdivided into blocks, one would expect the multiplicity in tracts to be larger than in blocks. However, the figures above show the opposite. These counterintuitive results are because we exactly replicated Dick *et al.*'s data sampling, where the sampling of blocks is independent from the sampling of tracts (both are done at the state level). More importantly, whereas tracts were randomly sampled, the largest and mean blocks per state were deterministically chosen. This yields blocks that tend to be larger than tracts. No reason is given in [5] for the different (non-random) criteria they employed to choose blocks. One motivation for choosing large blocks is that block-level data become more homogeneous because large blocks correspond to dense populations that tend to have inhabitants with similar profiles. This results in a greater amount of common records, which are those the confidence ranking scores higher.

Given the strong positive correlation between the multiplicity of prototypes and their confidence ranking, one might argue that the lowest-ranked record prototypes may be used to detect unique records in D and, therefore, indicate a privacy threat. In the following, we argue that this is not the case because of two reasons.

First, we must consider how Census data were protected in the 2010 release, on which all reconstruction attacks in the literature are based. Protection in that release was performed via *data swapping*. According to the USCB [25], the selection of the records to be swapped was highly targeted to the records with the most disclosure risk, that is, those that were unique in their block based on a set of key demographic variables. Also, the probability of being swapped had

an inverse relationship with block size. Data from these households at risk were swapped with data from other households that had similar characteristics on a certain set of variables but were from different geographic locations.

According to the description above, the swapping methodology was targeted toward certain records that were defined as at risk of disclosure (*i.e.*, vulnerable records). Even though the exact details on how vulnerable records were determined are not public, the selection criterion was based on their uniqueness in their block [26]. Given that the statistics employed for CRR come from the swapped/protected data (and not the *original* data), we can deduce that reconstructed unique records are more likely to be protected (that is, swapped) records than original records.

Second, in order to be able to reconstruct (even protected) unique records in D, this should be done exhaustively and unequivocally. For this to be possible with CRR, records that are rare in the reconstructed data should also appear as rare in D with high certainty. To see how likely this is to happen, we calculated the following proportions:

- number of prototypes that occur once in D and occur once in the reconstructions over the total number of prototypes that occur once in the reconstructions;
- number of prototypes that occur twice in D and occur once or twice in the reconstructions over the total number of prototypes that occur once or twice in the reconstructions;
- number of prototypes that occur three times in D and occur once, twice, or three times in the reconstructions over the total number of prototypes that occur once, twice, or three times in the reconstructions.

Table 1 reports the above proportions as percentages, and they are extremely small. In other words, the certainty that a rare reconstructed prototype occurs in D is almost zero.

Table 1: Percentage of rare record prototypes in reconstructions that occur once, twice, or three times in D. Results are given for tracts, for tracts without block, and for blocks.

Level	Tract			Tract (w/o block)			Block		
Occur. D/ freq. CRR	1/1	2/≤2	3/≤3	1/1	2/≤2	3/≤3	1/1	2/≤2	3/≤3
Random	0.09%	0.08%	0.05%	0.15%	0.16%	0.15%	0.20%	0.15%	0.12%
Baseline	0.09%	0.04%	0.02%	N/A	N/A	N/A	0.15%	0.06%	0.03%

To further illustrate the ineffectiveness of CRR as a reconstruction method, in Table 2 we report the percentage of reconstructed prototypes that did not occur in D (w.r.t. the total number of record prototypes in D).

Table 2: Percentage of reconstructed record prototypes that did not occur in D

Level	Tract	Tract (w/o block)	Block
Random	4805.3%	703.2%	820.1%
Baseline	2223.6%	N/A	449.9%

We can see that Dick *et al.*'s "optimization-based" method generates a very large number of non-existent record prototypes, which is 4 to 48 times larger than the number of prototypes in D. As discussed above, this results in very large uncertainty when reconstructing the rarest records in D, because there is a very high probability that the reconstructed prototypes do not exist in the original data.

On the other hand, by counting the percentage of records in D that did not appear in any of the reconstructions, we obtain the non-negligible figures reported in Table 3.

Table 3: Percentage of record prototypes in D that did not appear in the reconstructions

Level	Tract	Tract (w/o block)	Block
Random	8.16%	1.97%	3.31%
Baseline	3.07%	N/A	1.01%

5 On attribute disclosure risk

In a yet more recent article [6], the authors argued that the main risk posed by CRR is actually *attribute disclosure*, that is, unequivocal inference by the adversary of confidential attribute values of known individuals, even without being able to reidentify them in the data set. This threat was not mentioned in their initial paper describing CRR [5], which only discussed privacy attacks related to identity disclosure. In that new paper, the authors fell short of supporting this new claim with any theoretical or empirical evidence. They just conjectured attribute disclosure through the following single fabricated example referred to a different dataset (the American Community Survey (ACS)):

> Suppose that it is known that there are two 46-year-old married men with a particular racial designation and level of educational attainment within the [ACS] dataset. These features might be publicly known since they are not features the individuals intend to hide. Even if these two individuals match on all other features as well —that is, they share the same record prototype— if we are able to learn it, then we have learned

> facts about both of them (*e.g.*, their citizenship status, their income, etc.) that they may not have wanted to share. It does not matter that we cannot determine which record corresponds to which individual (the question does not even make sense, as the records are identical) since we have learned private information about both.

For the described privacy threat to lead to unequivocal inference of confidential attribute values on the known individuals, it is not enough to accurately generate the corresponding record prototype. The following must also hold:

1. The original data set must not contain any other records sharing the same values for the public attributes with the known individuals (in the above quotation, age, marital status, race, and education) but having different values for the confidential attributes (in the above quotation, citizenship status and income).
2. The original data set should be an exhaustive sample of the population to which the known individuals belong. Otherwise, we cannot be sure whether the known individuals are or are not present in the data set and, therefore, whether the potentially reconstructed records matching their public attributes correspond to them. Even though this holds for the Decennial Census (as long as the known individuals live in the U.S.), it is not the case for other non-exhaustive datasets, such as the American Community Survey (ACS) that Dick *et al.* used to illustrate the attribute disclosure threat.
3. CRR must not generate record prototypes sharing the same values for the public attributes but having diverse values for the confidential attributes.

Even though 1) and 2) may hold in practice, we have shown in Section 4 above that CRR was not only inaccurate but that it generated an enormous diversity of non-existent records (see Table 2). Therefore, 3) will not hold, and the diversity of the (erroneous) confidential values in the reconstructed records will prevent unequivocal attribute disclosure. In fact, it is well documented that diversity of confidential attributes effectively prevents attribute disclosure in k-anonymous-like data releases [15,14].

6 Conclusions

We have shown that, despite the claims of [5] and [6], the proposed CRR attack is ineffective to i) detect the most privacy-sensitive records, ii) guide targeted reidentification attacks, or iii) guide attribute disclosure attacks. Moreover, we have shown that the "optimization-based" method by Dick *et al.* fails to reconstruct a significant proportion of original records while generating an enormous amount of records that do not exist in the original data set. The latter adds a very large uncertainty to disclosure inferences, whether aimed at reidentification or attribute disclosure. Having demonstrated that the CRR attack implies no privacy risks, we can conclude that it cannot be used to justify the use of (utility-damaging) DP methods —contrary to what [5] and [2] claim.

Acknowledgments

This research was funded by the the European Commission (project H2020-871042 "SoBigData++"), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez), MCIN/AEI/ 10.13039/501100011033 and "ERDF A way of making Europe" under grants PID2021-123637NB-I00 "CURLING" and PRE2019-089210, and INCIBE and European Union NextGenerationEU/PRTR (project "HERMES" and INCIBE-URV Cybersecurity Chair).

References

1. Abowd, J.: Declaration of John M. Abowd. Case no. 3:21-CV-211-RAH-ECM-KCN (2021)
2. Abowd, J., Hawes, M.: Confidentiality protection in the 2020 U.S. Census of Population and Housing. Annual Review of Statistics and Its Application **10**, 119–144 (2023)
3. Blanco-Justicia, A., Sánchez, D., Domingo-Ferrer, J., Muralidhar, K.: A critical review on the use (and misuse) of differential privacy in machine learning. ACM Computing Surveys **55**(8), 1–16 (2023)
4. Daily, D.: Disclosure avoidance protection for the American Community Survey (2022), `https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acs.html`, accessed 3 May 2023
5. Dick, T., Dwork, C., Kearns, M., Liu, T., Roth, A., Vietri, G., Wu, Z.S.: Confidence-ranked reconstruction of census microdata from published statistics. Proceedings of the National Academy of Sciences U.S.A. **120**(8), e2218605120 (2023)
6. Dick, T., Dwork, C., Kearns, M., Liu, T., Roth, A., Vietri, G., Wu, Z.S.: Reply to Sánchez et al.: Multiplicity does not protect privacy. Proceedings of the National Academy of Sciences U.S.A. **120**(8), e2304263120 (2023)
7. Domingo-Ferrer, J., Sánchez, D., Blanco-Justicia, A.: The limits of differential privacy (and its misuse in data release and machine learning). Communications of the ACM **64**(7), 33–35 (2021)
8. Dove, I.: Applying differential privacy protection to ONS mortality data. Pilot study (2021), `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/applyingdifferentialprivacyprotectiontoonsmortalitydatapilotstudy`, accessed 23 May 2023
9. Gehrke, J., Hay, M., Lui, E., Pass, R.: Crowd-blending privacy. In: Advances in Cryptology – CRYPTO 2012. p. 479–496 (2012)
10. Hawes, M.: Reconstruction and reidentification of the Demographic and Housing Characteristics file (DHC) (2022), `https://www2.census.gov/about/partners/cac/sac/meetings/2022-09/presentation-reconstruction-and-re-dentification-of-dhc-file.pdf`, accessed 14 March 2023
11. Hotz, V., Bollinger, C., Komarova, T., Manski, C., Moffitt, R., Nekipelov, D., Sojourner, A., Spencer, B.: Balancing data privacy and usability in the federal statistical system. Proceedings of the National Academy of Sciences U.S.A. **119**(31), e2104906119 (2022)

12. Keller, S., Abowd, J.: Database reconstruction does compromise confidentiality. Proceedings of the National Academy of Sciences U.S.A. **120**(12), e2300976120 (2023)
13. Kenny, C., Kuriwaki, S., McCartan, C., Rosenman, E., Simko, T., Imai, K.: 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) (2021)
14. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: 23rd IEEE International Conference on Data Engineering (ICDE'07). pp. 106–115 (2007)
15. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: L-diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data **1**(1), 3–es (2007)
16. Menger, G.: Using 2020 Census data (2021), `https://appliedgeographic.com/2021/09/using-2020-census-data/`, accessed 9 May 2023
17. Muralidhar, K.: A re-examination of the Census Bureau reconstruction and reidentification attack. In: Privacy in Statistical Databases - PSD 2022. pp. 312–323. Springer (2022)
18. Muralidhar, K., Domingo-Ferrer, J.: Database reconstruction is not so easy and is different from reidentification. Journal of Official Statistics **39**(3), 381–398 (2023)
19. Ruggles, S., Riper, D.V.: The role of chance in the Census Bureau database reconstruction experiment. Population Research and Policy Review **41**, 781–788 (2022)
20. Samarati, P.: Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering **13**(6), 1010–1027 (2001)
21. Santos-Lozada, A., Howard, J., A.M.Verdery: How differential privacy will affect our understanding of health disparities in the united states. Proceedings of the National Academy of Sciences U.S.A. **117**(24), 13405–13412 (2020)
22. Schneider, M.: Researchers ask Census to stop controversial privacy method (2022), `https://www.usnews.com/news/business/articles/2022-08-08/researchers-ask-census-to-stop-controversial-privacy-method`, accessed 15 May 2023
23. U.S. Census Bureau: Developing the DAS: Demonstration data and progress metrics (2020), `https://www.census.gov/programs-surveys/decennial-census/decade/2020/planningmanagement/process/disclosure-avoidance/2020-das-development.html`
24. U.S. Census Bureau: Disclosure avoidance for the 2020 Census: An introduction (2021), `https://www2.census.gov/library/publications/decennial/2020/2020-census-disclosure-avoidance-handbook.pdf`, accessed 15 May 2023
25. Zayatz, L., Lucero, J., Massell, P., Ramanayake, A.: Disclosure avoidance for Census 2010 and American Community Survey five-year tabular data products. Tech. Rep. RRS2009-10, Census Bureau (2009), `https://www.census.gov/content/dam/Census/library/working-papers/2009/adrm/rrs2009-10.pdf`, accessed 9 March 2023
26. Zayatz, L., Lucero, J., Massell, P., Ramanayake, A.: Disclosure avoidance for Census 2010 and American Community Survey five-year tabular data products (2009), `https://www.census.gov/content/dam/Census/library/working-papers/2009/adrm/rrs2009-10.pdf`, accessed 9 May 2023