UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

Title

What is Reconstruction and Reidentification? Illustrations from the 2010 US Census Tabular Data Release

Krish Muralidhar, University of Oklahoma, Norman, OK 73072, USA (krishm@ou.edu)
Josep Domingo-Ferrer, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia (josep.domingo@urv.cat)

Abstract

In a recent court filing, the then associate director for research and methodology and chief scientist of the U.S. Census Bureau (USCB) claimed that: "Somewhere between 52 and 179 million person who responded to the 2010 [U.S] Census can be correctly re-identified …" If true, this claim speaks poorly of the USCB's efforts to protect respondent confidentiality in the 2010 Census. Fortunately, these claims turn out to be exaggerated. We present a careful re-examination of the data to show that the claims of disclosure of identity of the respondents are overstated. By linking the USCB's definitions of reconstruction and reidentification and the traditional statistical disclosure definitions of value and identity disclosure, we show that these claims are based on a misunderstanding of these definitions.

1 Introduction

In 2018, the US Census Bureau (USCB) announced that differential privacy has been adopted as the methodology for protecting data release from the 2020 Decennial Census (Abowd 2018b). This reflected a change in the methodology from the one used in 2010 (data swapping) to a new, more secure method (differential privacy). This announcement was followed by multiple releases of demonstration products illustrating the new disclosure avoidance system (DAS). Analysis of the DAS led to considerable consternation among users of Census data and two lawsuits due to the poor quality of the protected data. The *only* justification for this change in policy was the claim that the tabular data released from the 2010 decennial census was subject to a high level of disclosure of the identity of individuals who responded to the census. The specific claim, based on the analysis performed by a group of researchers within the USCB who we refer to as the Census Bureau Reidentification Team (CBRT), is provided in Abowd (2021, Appendix B):

> If the external data on name, address, sex, and age are comparable to the 2010 Census, then the attacker will putatively re-identify 238 million persons (77% of the 2010 Census resident U.S. population). Confirmed re-identifications will be 179 million (58% of the same population). This means that with the best quality external data, relative to the 2010 Census, as many as 179 million persons could be correctly re-identified using the attack strategy outlined here.

This is an astounding statement which essentially proclaimed to the world that the USCB failed to protect the 2010 Decennial Census data that was collected under explicit guarantee of anonymity. Between 2018 and 2021, this claim (58% confirmed reidentification) was repeated in multiple presentations made by the CBRT staff (e.g., Abowd 2018c; Hawes 2022). Yet, no details of the reidentification procedure were released during this period. In 2021, in response to a suit from the State of Alabama, Dr. Abowd, the Chief Scientist at the USCB provided details of the reidentification experiment (Abowd 2021) as a part of his deposition.

Since then, two studies (Ruggles and Van Riper 2021, Muralidhar 2022) have raised serious questions about the validity of the reidentification claims made by CBRT. Despite repeated requests of additional information regarding the reidentification results, no additional information has been forthcoming. Instead, CBRT have now modified their reidentification procedure (Hawes 2022), *now claiming that the reidentification of 75.5% of the respondents can be confirmed*. Albeit unwittingly, the description of the new procedure provides further insight into the whole reidentification process. In this study, using the CBRT description of the new procedure, we show that the claims of reidentification (both old and new) are completely unjustified and must be corrected in the interest of scientific accuracy.

2 Reconstruction, Reidentification, and Confirmation

CBRT claims of confirmed reidentification are based on an attack strategy consisting of three steps:

(1) Reconstruction: Using the 2010 tabular data release[1], reconstruct the microdata (individual level records) containing the variables (Block, Sex, Age, Race, Ethnicity) for the entire population.
(2) (Putative) Reidentification: Link these records to an external source file containing the variables (Name, Address, Age, Sex). When a link is found, append the Protected Identification Key (PIK) created using (Name and Address) to the reconstructed records to result in the reconstructed (PIK, Block, Sex, Age, Race, Ethnicity).

[1] See Abowd 2021 (Appendix B, page 1) for the list of the tables used in the reconstruction. Also note that the reconstruction focused exclusively on recreating individual level data and no effort was made to reconstruct household related information (Abowd 2021a, Appendix B, para 12, page 5).

(3) Confirmation (of Reidentification): Compare the putatively reidentified reconstructed data to the original data in the Census Edited File (CEF) to confirm reidentification.

2.1 Reconstruction

The original reconstruction attack "was carried out by constructing a system of equations consistent with the published tables listed above that, once solved, could then be converted into microdata." (Abowd 2021, Appendix B, para 4, page 2) According to CBRT, this reconstruction attack resulted in agreement with the original data in CEF of: (a) 46% for an exact match on Age and (b) 71% for a match on (Age $\pm$ 1) year. These results were criticized by Muralidhar (2022) who showed that even for a single Age Group, for a single sex, in a single tract, there is a practically infinite number of possible reconstructions, and claims of reconstruction accuracy based on a single reconstruction are of little value.

Recently, CBRT have presented a modified version of their reconstruction attack (Hawes 2022). In their modified version, they have resorted to reconstructing the microdata as (Block, Sex, Age Group, Race, Ethnicity) for all records over age 21. In other words, CBRT have completely given up trying to reconstruct individual year of age for those over the age of 21. This is exactly the simple procedure suggested by Ruggles et al (2018) well before CBRT adopted it. Also note that the original criticism of alternate reconstructions in assigning individual year of age to specific records (Muralidhar 2022) still holds for reconstructions relative to those records age 21 and under.

With the new procedure, CBRT reports a 91.8% agreement rate between the reconstructed microdata and CEF, much higher than the 71% reported originally. Hawes (2022) also reports that the agreement rate for records with age 21 and under is only 46.5%. Given the overall agreement rate is 91.8%, this implies that the agreement rate for those over age 21 should have been practically 100%.[2] There is nothing surprising about this. The *only uncertainty* in record level reconstruction is the Age variable, since it is provided only as Age Group in the block level tabular data which must then be converted to individual year of age. The earlier reconstruction procedure attempted to do this, and the new reconstruction procedure does not. All CBRT has done with the new procedure is to *relax the requirements for the agreement rate and claim a higher agreement rate*.

CBRT also claims that 65% of the blocks have "zero variability" (Hawes 2022), implying that the problem of alternate reconstructions has been eliminated. But this is extremely misleading! Variability in reconstruction comes primarily from reconstructing the unknown Age variable, while all the other variables are available directly from the table. For all records above 21 years of age (approximately 70% of the entire population), CBRT has foregone reconstructing individual year of age in the new reconstruction procedure. Hence, "65% of the blocks have zero variability" does not mean that there is no variability; it simply means that CBRT has stopped measuring variability.

2.2 Reidentification

The reidentification procedure (reidentification attack) is described in Abowd (2021, Appendix B, para 18, page 7). In this procedure, every record in the reconstructed data is compared to the external source data on (Block, Sex, Age). In the first pass, an exact match on Age is used, followed by a second pass looking for match on (Age $\pm$ 1 year). The records for which a match is found are labeled "putative reidentifications".

[2] We are unable to reconcile this 91.8% overall (exact and binned) agreement rate and the 46.5% agreement rate for ages 21 and under, provided in Hawes (2022). Unfortunately, no details were given.

CBRT used two different external source files. The first one comprised of data gathered from commercial databases originally used to verify the accuracy of the 2010 census data. The second used the CEF itself (but without the race and ethnicity variables) as a surrogate external source file. Abowd (2021) contends that using the abbreviated CEF file as the external file represents the worst-case risk scenario where the adversary has data of the highest possible accuracy (the true original data). In this study, the abbreviated CEF will be used as the external source file. This an extremely conservative approach favoring CBRT claims.

To illustrate putative reidentification, consider Block 1014, Tract 201, Autauga County, Alabama.[3] There are five White and one Black-African American non-Hispanic males in the (35 – 39) age group in this block. Table 1 presents the reconstructed data and the data from the external source file. The reconstructed records are labeled using the numbers 1 to 6 and the respondents from the external source using the letters A to F. Since we are assuming that CEF will serve as a surrogate for the external source, there must be six males with individual year of age between 35 and 39, unless one or more records have been swapped. For simplicity, assume that the data has not been swapped.

Table 1. Reconstructed data and data from the external source file.

PIK	Sex	Age		Rec. #	Sex	Age Group	Race	Ethnicity
A	Male	37		1	Male	(35 - 39)	White	Not Hispanic
B	Male	38		2	Male	(35 - 39)	White	Not Hispanic
C	Male	36		3	Male	(35 - 39)	White	Not Hispanic
D	Male	38		4	Male	(35 - 39)	White	Not Hispanic
E	Male	38		5	Male	(35 - 39)	White	Not Hispanic
F	Male	36		6	Male	(35 - 39)	BAfA	Not Hispanic

Comparing the data from the external source and the reconstructed data, it is easy to see that there is a match in the external source for every record in the reconstructed data, resulting in 100% putative reidentification. But that is to be expected. Since the external source is CEF (sans race and ethnicity), *there will always be a match on (Age and Sex) between the reconstructed records and CEF in the same block*. This is true not only for this block, but for all Age Groups (above 21 years) and for both sexes.

The exception to the previous statement are those (Block, Sex, Age) combinations where records have been swapped prior to the tabular data being released. The Census Bureau has never released the exact percentage of records that were swapped but has indicated that it is a small number (perhaps less than 5%). Hence, it is likely that for most blocks, putative reidentification for all records above age 21 is close to 100%.

Using the original reconstruction procedure and CEF as the external source file, CBRT report that 238,175,305 records (77% of the US population) were putatively identified. The putative identifications using the new reconstruction procedure are even higher (Hawes 2022). In reporting the recent results, CBRT define the population as only those with PIK, which reduces the total population to 279,179,329 (Abowd 2021a, Appendix B, Table 2). The putative reidentification for this population is now 97% (approximately 271 million). This translates to approximately 88% of the total population (with and without PIK), an increase of 10% (approximately 32 million) over the original claim.

[3] The choice of this county is only a matter of convenience (it happens to be the first tract, in the first county in the first state). We can find similar examples across the board.

But is this truly reidentification? Reidentification "occurs when a one-to-one relationship can be established between a record in the released data and a specific entity" (Chen and Keller-McNulty 1998). Hence, to claim reidentification, it is necessary to establish a unique, one-to-one correspondence between a particular reconstructed record and a particular respondent from the external source of identified data.

The data in Table 1 indicates that, for the variables (Sex and Age) used to match the reconstructed records and the external source data, the records in the reconstructed data are *indistinguishable*. Gehrke et al (2012) say the following about indistinguishability: "if an individual *t* blends in a crowd of *k* people in the database, then the mechanism essentially does not release any information about individual *t* beyond the general characteristics of the crowd of *k* people; in particular, the mechanism does not release any personal information that is specific to individual *t* and no one else." The data in Block 1014 also satisfies *k*-anonymity (with $k = 6$) for quasi-identifiers (Sex, Age) since "every combination of values of the quasi-identifiers can be indistinctly matched to at least *k* individuals." (Samarati and Sweeney 1998). Under these conditions, it is impossible to assign a specific identity to a specific record. *Any record can be assigned the identity of any respondent*. An intelligent adversary would realize that claims of reidentification are easily refuted in this scenario.

Yet, even though the six records in Table 1 are indistinguishable, the confirmation procedure used by the CBRT will result in confirming the identity of a minimum of four (and a maximum of six) of these records. We now show the reason for this counter-intuitive result.

2.3 Confirmation

The confirmation procedure is described in Abowd (2021a, para 21, page 8). In this procedure, the variables from the putative identification (PIK, Block, Sex, Age, Race, Ethnicity) are compared to the same variables in CEF, first matching on exact year of Age followed by (Age $\pm$ 1). Table 2 presents the hypothetical CEF for males in the (35 – 39) age group in Block 1014, Tract 201 in Autauga county, along with the PIK-appended reconstructed records.[4]

Table 2. Hypothetical CEF for males in the (35-39) age group in Block 1014, Tract 201 in Autauga county, along with the PI-appended reconstructed records.

PIK	Sex	Age	Race	Ethnicity		PIK	Rec. #	Sex	Age Group	Race	Ethnicity
A	Male	37	BAfA	Not Hispanic		A	1	Male	(35 - 39)	White	Not Hispanic
B	Male	35	White	Not Hispanic		B	2	Male	(35 - 39)	White	Not Hispanic
C	Male	36	White	Not Hispanic		C	3	Male	(35 - 39)	White	Not Hispanic
D	Male	39	White	Not Hispanic		D	4	Male	(35 - 39)	White	Not Hispanic
E	Male	38	White	Not Hispanic		E	5	Male	(35 - 39)	White	Not Hispanic
F	Male	36	White	Not Hispanic		F	6	Male	(35 - 39)	BAfA	Not Hispanic

[4] Note that assigning individual year of age does not affect the results since any individual year of age could be assigned to any record.

It is easy to verify that for Table 2, four of the identities are confirmed (records 2, 3, 4, 5) while the identities of the remaining two records are not confirmed (records 1 and 6 which have incorrect Race). At first glance, this seems like a perfectly reasonable approach for confirming reidentification.

Since any reconstructed record could be assigned any identity, we now present an alternate reconstruction for the same data. In Table 3 the PIK assigned to each record in the reconstructed data is different from the one in Table 2. Yet we still have the same result of four identities being confirmed (1, 3, 4, 5) and two identities not being confirmed (2, 6). Using every possible assignment of identity during the reidentification state, we can show that the confirmation procedure will confirm, either four of six records (when the identity of the Black respondent is incorrectly assigned) with probability 5/6 or six of six records (when the identity of the Black respondent is correctly assigned) with probability 1/6.

Table 3. Alternate reconstruction

PIK	Sex	Age	Race	Ethnicity		PIK	Rec. #	Sex	Age Group	Race	Ethnicity
A	Male	37	BAfA	Not Hispanic		B	1	Male	(35 - 39)	White	Not Hispanic
B	Male	35	White	Not Hispanic		A	2	Male	(35 - 39)	White	Not Hispanic
C	Male	36	White	Not Hispanic		D	3	Male	(35 - 39)	White	Not Hispanic
D	Male	39	White	Not Hispanic		F	4	Male	(35 - 39)	White	Not Hispanic
E	Male	38	White	Not Hispanic		C	5	Male	(35 - 39)	White	Not Hispanic
F	Male	36	White	Not Hispanic		E	6	Male	(35 - 39)	BAfA	Not Hispanic

Ironically, the record whose identity is least likely to be confirmed is that of the Black respondent since *this is the only record that is distinguishable from all the other records in this block*. The probability that the identity of this record will be confirmed is 1/6, which represents the true probability of correctly identifying any record based on comparing (Sex, Age).

The situation is much worse when we consider larger blocks. In general, let n_1 and n_2 represent the number of records belonging to the majority (Race, Ethnicity) combination and the number of records of all others in any given (Block, Sex, Age Group) (age > 21). In the worst case, assume all n_2 minority individuals are labeled as majority individuals, which also means that n_2 among the majority individuals are labeled as a minority individuals. Thus, $2n_2$ reidentifications among the $n_1 + n_2$ will be incorrect. In this case, the proportion of confirmed reidentifications using the CBRT procedure is $Maximum\{0, (n_1 - n_2)/(n_1 + n_2)\}$.

For homogenous blocks, n_1 is likely to be much higher than n_2, and the CBRT procedure is likely to result in confirming the reidentification of a large proportion of records. The factor $n_1 - n_2$ is called the inflation factor of confirmed reidentification. If a given (Block, Sex, Age Group) consisted of 99 White non-Hispanic individuals and one individual of some other (Race, Ethnicity), the confirmed reidentification rate is inflated by a factor of 98. This is precisely what Ruggles and Van Riper (2021) predicted, for exactly this reason. This also explains the high confirmed reidentification rates observed for large blocks.

Ironically, the USCB's own Research and Methodology Directorate has suggested a procedure to overcome this problem. This procedure is described in McKenna (2019) as follows (emphasis added):

> For microdata, such reidentification studies are performed by looking for unique combinations of variables in the microdata that are thought to be identifying, looking for externally available data sets that contain the same variables, and then linking data records in the two data sets using the linkage variables. Finally, it is necessary to verify the proposed matches by comparing the suppressed identities in the microdata with the identities in the external data set to see if the matches are true matches or false matches. This last comparison step is vital, because often survey records are unique within the sample but not in the population (Ramachandran, 2012).

The advantage of this procedure is that it eliminates random matches from being classified as true matches. Let us consider the data in Table 2 where CEF on the left represents the true data and the reconstructed data (with identity appended) represents the external data. Consider the first record from the reconstructed data. Using the two-step procedure suggested by the Research and Methodology Directorate, the probability that this record has the same PIK in CEF as the randomly assigned PIK in the augmented reconstructed record is 1/6. Hence, the probability that a correct match is found is 1/6. This is true for all records in this block. The probability of correct confirmation is 1/6, the same probability that the correct PIK is assigned to the correct individual.

3 Discussion and Conclusion

The fallacy in the reidentification and confirmation procedure used by CBRT lies in the simple fact that it is not confirming identity; it is only confirming whether (Race, Ethnicity) have been correctly assigned. Hence, *even if the same reconstructed record is assigned different identities, the confirmation procedure will confirm all these different identities as the correct identity if the (Race, Ethnicity) are assigned correctly*. Thus, it is not evaluating *identity disclosure, it is evaluating value disclosure*. This is confirmed by the illustrative example from Table 1.

Lambert (1989) provides clear definition of both identity and value disclosure (bold emphasis added):

> In an *identity disclosure*, or *identification*, **a respondent is linked to a particular record in a released file**. Identification, **sometimes called re-identification**, is equivalent to inadvertent release of an identifiable record.
>
> An *attribute disclosure* occurs when the intruder believes something new has been learned about the respondent. An attribute disclosure may occur with or without an identification.

The discussion in the previous section makes it very clear that the CBRT procedure falls squarely under the description of value disclosure – the intruder believes that something new (the value of Race and Ethnicity variables) has been learned about the respondent without necessarily identifying them. The acceptance of the above definitions of identity and value disclosure is almost universal in the disclosure risk literature.

This does not mean that CBRT is not entitled to use a different definition of disclosure risk in assessing the appropriateness of the tabular data that was released. They are certainly entitled to use *attribute or value disclosure* as the criterion to decide whether a new disclosure protection procedure must be implemented.

What CBRT is not entitled to do is to compute the risk using one definition and use it to imply another. This is precisely what CBRT has done. They have computed attribute disclosure risk for (Race, Ethnicity) variables and implied that it is identity disclosure risk. In common usage, the term "reidentification" means "identify (someone or something) again" (Merriam-Webster.com). Without *identity* there is no "reidentification" and *value disclosure does not require identification*.

In almost every presentation made by CBRT, the term used to describe the disclosure is "reidentification" implicitly implying identity disclosure. There is at least one case where the claim is explicit. In his declaration to the Court in the State of Alabama v. The Department of Commerce (Case 3:21-cv-00211-RAH-KFP), Abowd (2021a, page 8) states:

> The Census Bureau's own internal analysis, for example, confirmed that a modern database reconstruction-abetted re-identification attack can reliably match a large number of 2010 census responses to the names of those respondents—a vulnerability that exposed information of at least 52 million Americans and potentially up to 179 million Americans.

There can be no more explicit acknowledgement that CBRT consider their results to be disclosure of identity than the statement "match a large number of 2010 census responses to the names of those respondents". That this statement was made in a court declaration only lends further credibility that CBRT misrepresented "confirmed reidentification". This may also help explain how "the risk of re-identification is small" (Abowd 2018) became "reconstructed microdata permit between 52 and 179 million correct re-identifications from the 2010 Census." (Abowd 2021)

This is not just a question of semantics. Whether it is identity or value disclosure might have legal consequences as well. The relevant sections of Title 13 read as: "the Secretary may furnish copies of tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent," (Section 8(b)) and shall not "Make any publication whereby the data furnished by any particular establishment or individual under this title can be identified." (Section 9(2)) If all the CBRT has to offer is disclosure of value (of race and ethnicity) without identification then, since the disclosure is only inferential, the release of such data does not violate the confidentiality protection mandate. Whether there is a case to be made if there is only one individual in a (Block, Sex, Age Group) is certainly open for interpretation and for the courts to decide.

If CBRT truly believes that the disclosure risk associated with the 2010 Census tabular data release is unacceptable, it behooves them to make a case based on the *honest* presentation of the available evidence. Their current efforts do not fit that bill.

Acknowledgments

Partial support to this work has been received from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042 ("SoBigData++"), the MCIN / AEI / 10.13039 / 501100011033 / FEDER, UE (project PID2021-123637NB-I00, "CURLING"), and the Government of Catalonia (ICREA Acadèmia Prize to the second author, and grant 2021SGR-00115). The second author is with the UNESCO Chair in Data Privacy, but the views in this paper are the authors' own and are not necessarily shared by UNESCO.

References

Abowd, J.M. 2018. How modern disclosure avoidance methods could change the way statistical agencies operate. Federal Economic Statistics Advisory Committee (December 14, 2018).

Abowd, J.M. 2018b. The US Census Bureau Adopts Differential Privacy. KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining July 2018 Pages 2867 https://doi.org/10.1145/3219819.3226070

Abowd, J. M. 2018c. "Staring down the database reconstruction theorem" in Joint Statistical Meeting, (Vancouver, BC, Canada, 2018).

https://www.census.gov/content/dam/Census/newsroom/press-kits/2018/jsm/jsm-presentation-database-reconstruction.pdf (accessed 22 June 2023).

Abowd, J.M. 2021. Declaration of John M. Abowd. Case no. 3:21-CV-211-RAH-ECM-KCN, U. S. District Court for the Middle District of Alabama. Apr. 13.

Chen, G., Keller-McNulty, S. 1998. Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* 14(1), 79-95.

Gehrke, J., Hay, M., Lui, E., Pass, R. 2012. Crowd-Blending Privacy. *Proceedings of Advances in Cryptology - CRYPTO 2012*, 479-496.

Hawes, M. 2022. Reconstruction and Reidentification of the Demographic and Housing Characteristics File (DHC). *Presentation to the Census Scientific Advisory Committee* (Washington DC, USA, 29-30 September 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).

Lambert, D. 1993. Measures of disclosure risk and harm. *Journal of Official Statistics* 9 (313-319).

McKenna, L. 2019. "U.S. Census Bureau Reidentification Techniques." Research and Methodology Directorate, U.S. Census Bureau. April. https://www2.census.gov/adrm/CED/Papers/CY19/2019-04-Reidentification%20studies-20210331FinRed.pdf (accessed May 8, 2023).

Muralidhar, K. 2022. A re-examination of the Census Bureau reconstruction and reidentification attack. In: *Privacy in Statistical Database – PSD 2022*, Springer, pp. 312-323.

Ramachandran, A., Singh, L., Porter, E., Nagle, F. 2012. "Exploring Re-Identification Risks in Public Domains." In 10th Annual International Conference on Privacy Security and Trust, IEEE, Danvers MA, 2012. DOI: 10.1109/pst.2012.6297917

Ruggles, S. and Van Riper, D. 2021. The role of chance in the Census Bureau database reconstruction experiment. *Population Research and Policy Review*, early access since Aug. 22. https://doi.org/10.1007/s11113-021-09674-3

Ruggles, S., Fitch, C., Magnuson, D., & Schroeder, J. 2018. Differential privacy and census data: Implications for social and economic research. *AEA Papers and Proceedings*, 109, 403–408.

Samarati, P., Sweeney, L. (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression." https://www.semanticscholar.org/paper/Protecting-privacy-when-disclosing-information%3A-and-Samarati-Sweeney/7df12c498fecedac4ab6034d3a8032a6d1366ca6 (accessed 25 June 2023).