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

Title

Confidence-Ranked Reconstruction of Census Records Does Not Reflect Privacy Risks or Reidentifiability

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

Abstract

The possibility of mounting reconstruction attacks against census outputs has been given as an argument to protect those outputs using differential privacy. However, several authors (Ruggles and Van Riper, 2022; Muralidhar and Domingo-Ferrer, 2023) have shown that reconstruction is different from reidentification and that there are no privacy risks when many possible reconstructions exist. Dick et al. (2023) have recently presented a method to reconstruct record prototypes from aggregate query statistics of the US Decennial Census data, where we call prototype a record present with some multiplicity (i.e., number of repetitions) in the original data. The proposed method ranks the reconstructed prototypes by how frequently they appear in multiple reconstructions. The authors take this ranking as a confidence metric telling how likely it is for a reconstructed record to be actually present in the original data. They interpret this as a privacy threat and an indication of reidentifiability.

We show that their ranking does not properly consider the multiplicity with which a record appears in the original data and tends to award higher ranks to the most repeated records (which are intrinsically protected against re-identification by the repetition), whereas it misses outliers (which are those records that are really at risk of re-identification). Thus, their confidence-ranked reconstruction is ineffective at assessing the privacy risks, at detecting the most privacy-sensitive records (outliers), and at guiding re-identification attacks.

1. Introduction

Recently, Dick et al (2023) proposed a new reconstruction attack that allows them to reconstruct Census data using only publicly released aggregate statistics. The proposed procedure, which we refer to as Confidence Ranked Reconstruction (CRR), reconstructs *record prototypes*, that is, records that appear in the original data with some multiplicity, but it does not ascertain such multiplicity. Dick et al (2023) claim that they can assign a confidence to the reconstructed prototypes that allows the authors to potentially compromise the privacy of the Census respondents. Specifically, they claim that "Such a ranking could be used by an adversary to prioritize subsequent exploitation of private data—for example, for identity theft or to locate individuals of certain backgrounds." This conclusion has been readily endorsed by the current and previous Chief Scientists of the Census Bureau (Keller and Abowd 2023). Sánchez et al (2023) have pointed out that the claims of reidentification are unjustified. In this paper, we provide a detailed analysis in support of Sánchez et al (2023).

2. Confidence Ranked Reconstruction (CRR)

We now summarize CRR as described in Dick et al (2023). The input to CRR involves aggregate statistics released from a data set. This information is fed to a nonconvex optimization algorithm to reconstruct the database with the objective of minimizing the distance between original responses and the reconstructed data. The process is then repeated multiple times to generate multiple reconstructed databases. Records that appear most frequently in the (multiple) reconstructions are identified and ranked (with the record appearing most frequently being assigned the highest rank of 1). CRR also requires converting categorical data to continuous data during the input stage and reconverting back to categorial data at the output stage. CRR was used to reconstruct the data from two levels of geographies (tract and block) from the 2010 Decennial Census and American Community Survey. See Dick et al (2023) for specific results.

3. Further Analysis of CRR

One of the intriguing features of CRR is the need to generate multiple reconstructions to create the confidence ranking. Typically, the objective of a reconstruction mechanism is to generate a data set that closely (preferably exactly) resembles the original data set. In general, if the reconstruction mechanism results in many potential reconstructions, this implies uncertainty and hence less confidence in the reconstructions. By contrast, if the reconstruction mechanism results in a unique reconstruction, this implies that there exists only one possible data set that, when used as the input, can produce the output aggregate statistics. The fact that CRR requires multiple reconstructions seems to run against the confidence in the reconstruction process. On a closer examination, it becomes obvious that CRR does not measure reconstruction confidence at all.

Dick et al (2023) describe the nonconvex optimization as follows:

> The algorithms are closely related to recent methods for generating synthetic versions of sensitive datasets (5–7) while enforcing Differential Privacy (DP) (8–11). Crucially, these techniques are randomized, which allows us to repeat the reconstruction multiple times and get different results.

In other words, CRR is essentially the process of generating multiple synthetic data sets and analyzing the frequency of appearance of records in these multiple synthetic data sets.

This is *precisely* the same procedure suggested by Rubin (1993) in his seminal paper on synthetic data (albeit with a different statistical model for generating synthetic data). The key feature of Rubin's multiple imputation based synthetic data is that aggregating the results across the multiple data sets provides the user with the ability to perform valid statistical analyses. Hence, it is *necessary* that the multiple synthetic data sets closely resemble the true data set. Thus, a better title to the Dick et al (2023) paper would be, "Synthetic data is similar to the original data" which is stating the obvious.

4. Reconstruction versus Ranking Synthetic Records

In this section, we provide a brief analysis of why true reconstruction and Dick et al's (2023) ranking procedure are dramatically different. For the purposes of this illustration, we consider tract level Census data used in Dick et al (2023). At the tract level, the 2010 tabular data release contains tables (PCT12A-N) that provide COUNT(Age, Sex, Race) and COUNT(Age, Sex, Race, Ethnicity = Not Hispanic) for all values of Age (99 and below), both values of Sex, and seven categories of Race (White, Black, American Indian or Alaskan Native, Asian, Native Hawaiian or Pacific Islander, Other, and Two or more races).

Ages above 99 are provided only in age buckets (100 – 104, 105-109, and 110 and higher). In addition, the "Two or more races" category is further sub-divided into 57 combinations of the other six race categories. The breakdown of this category is only provided in Table P8 which is not one of the tables used by Dick et al (2023). As a result, there is no way to reconstruct Age > 99 years or sub-divisions of "Two or more races" since no queries can be generated to provide responses to these questions. For our purposes, we will limit our discussion to the combined "Two or more races" category. But in general, these two categories (people older than 99 or having two or more races) are relatively small.

Differencing the results in PCT12A-N appropriately gives us the ability to *exactly reproduce the original source data* for all Ages 99 and below, for both sexes, and the first six race categories, and both ethnicities (see Muralidhar (2022) for a detailed description). Since this is an *exact reproduction*, the frequency with which a particular combination of attribute values appears is its *true frequency*.

Now consider the Rube Goldberg style approach in Dick et al (2023). The categorical attribute Race is first converted to a continuous attribute (although it is not clear how this transformation is performed). The tabular data from tables P1, P6, P7, P9, P11, P12A-I, PCT12A-N are then fed to the optimization software to generate the (continuous) synthetic data. The output is then used to reconvert the continuous attributes to categorical attributes (where appropriate). The process is repeated multiple times. Finally, the frequency distribution of the appearance of the prototypes is recorded.

Given the requirements of generating synthetic data, it makes perfect sense that the most common records in the original data appear most frequently in the synthetic data. Dick et al (2023) consider this to be a serious breach of privacy. But it is not clear why. The most common records in the data are also the most protected since it is impossible to reidentify one individual from among a group of *k* individuals all of whom have *exactly the same attribute values*. Samarati (2001) calls this *k*-anonymity, and Gherke et al (2012) call it "privacy in a crowd".

By contrast, consider the less common records and assume that they are in Age 99 or less. If they exist in the original data, then they will always appear in the simple reconstruction procedure described above. They may or may not appear in the synthetic data that are generated. At the very least, the frequency with which they appear is far less than that of the most common records. Yet, it is precisely these records that are at risk of reidentification.

To claim that the most common records are more at risk compared to the less common (or unique) records is to contradict all basic principles of disclosure limitation. There is extensive literature showing that unique and less common records are far more at risk than the most common records. It is also basic common sense: the most identifiable person in a crowd is the one who is the most different from the rest of the crowd.

In summary, we would argue that the risk of reidentification is the opposite of what Dick et al (2023) claim. It is those records that appear infrequently that are at risk. But there is a problem with this approach as well. In many cases, when synthetic data is generated, many of the records that appear infrequently never actually appear in the original data. This is particularly true of sparse tables like the Census tables. Distinguishing between those records in the original data that appear infrequently and records that never appear in the original data but appear in the synthetic data infrequently is an impossible task.

5. Conclusion

We can conclude that CRR fails to reflect the actual privacy risk or the reidentifiability, unlike asserted in Dick et al (2023) and Keller and Abowd (2023). What is more, whenever there are multiple reconstructions compatible with a set of output aggregate statistics, there is no way of knowing which reconstruction is more likely to coincide with the original data.

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 Prizes to the first and third authors, and grant 2021SGR-00115). The first and third authors are 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

T. Dick, C. Dwork, M. Kearns, T. Liu, A. Roth, G. Vietri, and Z. S. Wu (2023) Confidence-ranked reconstruction of census microdata from published statistics. PNAS 120(18)e2218605120.

J. Gehrke, M. Hay, E. Lui, and R. Pass (2012) Crowd-blending privacy. In: Advances in Cryptology--CRYPTO 2012. Lecture Notes in Computer Science, Springer, pp. 479-496.

S. Keller and J. Abowd (2023) Database reconstruction does compromise confidentiality. PNAS 120(12)e2300976120.

K. Muralidhar (2022) A re-examination of the Census Bureau reconstruction and reidentification attack. In: Privacy in Statistical Databases (PSD 2022). Lecture Notes in Computer Science, Springer, pp. 312-323.

K. Muralidhar and J. Domingo-Ferrer (2023) Database reconstruction is not so easy and is different from reidentification. Journal of Official Statistics, to appear. https://arxiv.org/abs/2301.10213

D. B. Rubin (1993) Discussion: statistical disclosure limitation. Journal of Official Statistics 9(2):461-468.

S. Ruggles and D. Van Riper (2022) The role of chance in the Census Bureau database reconstruction. Population Research and Policy Review 41:781-788.

P. Samarati (2001) Protecting respondents identities in microdata release. IEEE Trans. on Knowl. And Data Eng. 13(6):1010-1027.

D. Sánchez, J. Domingo-Ferrer, and K. Muralidhar (2023) Confidence-ranked reconstruction of census records from aggregate statistics fails to capture privacy risks and reidentifiability. PNAS 120(18)e2303890120.