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

Josep Domingo-Ferrer[1], Krishnamurty Muralidhar[2], David Sánchez[1]

[1] Universitat Rovira i Virgili, Tarragona, Catalonia

UNESCO Chair in Data Privacy

[2] University of Oklahoma, U.S.A.

{josep.domingo,david.sanchez}@urv.cat,krishm@ou.edu

Wiesbaden, September 2023

Introduction

- Dick et al. (2023)[1] have proposed a Confidence Ranked Reconstruction (CRR) attack to reconstruct Census data from publicly released aggregate statistics.
- CRR reconstructs record prototypes: records that appear in the original data with some multiplicity, but without finding out their multiplicity.

[1] 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.

Attack claims

- Dick et al. (2023) claim they can assign a confidence to the reconstructed prototype to potentially compromise the privacy of Census respondents.
- The current and previous Chief Scientists of the U.S. Census Bureau have endorded this claim[2].

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

Confidence Ranked Reconstruction (CRR)

1. CRR takes as input aggregate statistics released from a data set.
2. A nonconvex optimization algorithm is used to reconstruct the data set.
3. The optimization objective is to minimize the distance between original responses and the reconstructed data.
4. The process is iterated to generate multiple reconstructed data sets.
5. Records that appear most frequently in the reconstructions are identified and ranked.
6. The confidence rank of a record is proportional to its frequency of appearance.

Further analysis of CRR (I)

- CRR requires converting categorical data during the input stage and reconverting back to categorical data at the output stage.
- CRR was used by Dick et al. (2023) to reconstruct the data from two levels of geographies (tract and block) from the 2010 Decennial Census and American Community Survey.

Further analysis of CRR (II)

- Whereas reconstruction is supposed to generate one data set that closely resembles the original, CRR generates multiple reconstructions using randomized synthesis.
- This implies uncertainty and less confidence in the reconstructions.
- On a closer examination, we will see that CRR does not measure reconstruction confidence at all.

Further analysis of CRR (III)

- CRR counts the frequency of appearance of records in the multiple synthetic reconstructions.
- This is the same procedure suggested by Rubin (1993) in his seminal paper on synthetic data.

Reconstruction vs ranking synthetic records

- We will show why true reconstruction and Dick et al. (2023) ranking procedure are dramatically different.
- Consider tract-level Census 2010 data used in Dick et al. (2023).
- These data contain 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).

Limitations in the reconstruction of Age and Race

- Ages above 99 are provided only in buckets ($100-104, 105-109,$ and 110 and higher).
- "Two or more races" category is sub-divided into 57 combinations of the other six race categories.
- Only overall counts (and no breakdown by Sex or Age) are provided.
- Thus, there is no way to reconstruct Age > 99 years or sub-divisions of "Two or more races" (we exclude them from our discussion).

Exact reconstruction by differencing

- Differencing the results in PCT12A-N allows us to exactly reproduce the original source data

  for all Ages 99 and below, for both sexes, for the first six race categories, and both ethniticies[3].
- This is an exact reproduction, and frequencies in it are true frequencies[4].

[3] 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.

[4] Yet, this is no real disclosure, because tables were computed on protected microdata.

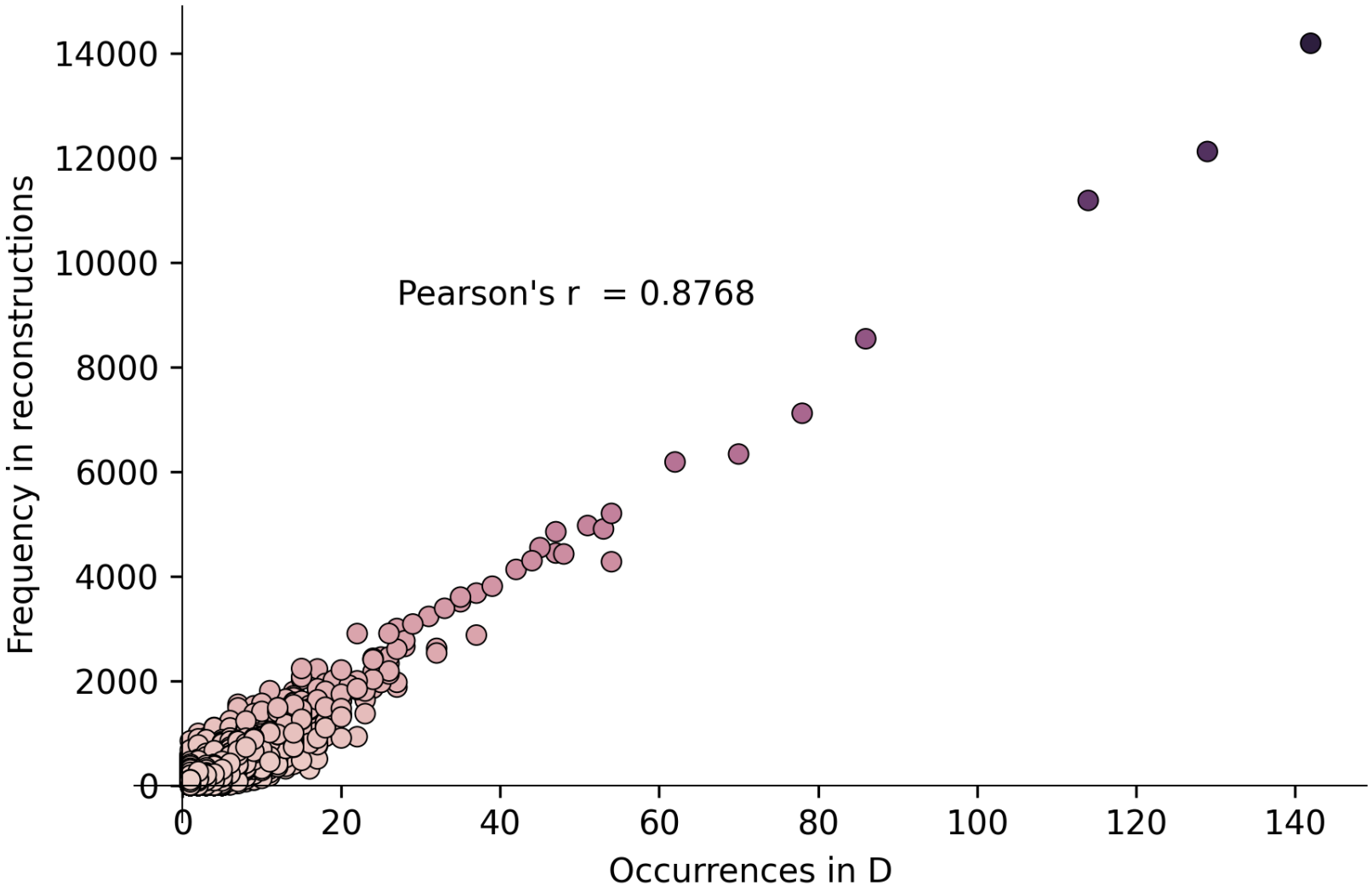
Dick et al.'s contraption

1. Convert categorical Race to a continuous attribute.
2. Feed data from tables P1, P6, P7, P9, P11, P12A-I, PCT12A-N to the optimization software to generate continuous synthetic data.
3. Reconvert continuous attributes to categorical as appropriate.
4. Repeat the two previous steps multiple times.
5. Record the frequencies of apperance of prototypes.

Highest ranks represent a privacy breach?

- In the above reconstruction procedure, the most common records in the original data will appear most frequently in the synthetic data and will get the highest ranks.
- Dick et al. (2023) consider a high rank to indicate a privacy breach (?).
- But... the most common original records are also the most protected (k-anonymity principle).

Correlation between CRR rank and frequency of appearance in original data



Records at risk not considered in the ranking

- Take the less common records, which are those at risk of re-identification.
- If they exist in the original data, they will appear in our simple differencing reconstruction.
- But... they may or may not appear in the synthetic data that are generated.
- Anyway, their frequency will be far less that of the most common records, and they will not be the highest ranked.

High rank does not imply high risk

- To claim that the most common records are more at risk compared to the less common or unique contradicts all SDC principles.
- 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.

High risk detection by low rank not guaranteed

- Many of the records that appear infrequently in the synthetic data never actually appear in the original data.
- 3.07% of the records in the original data never actually appear in the synthetic data.
- This is particularly relevant in sparse tables like the Census tables.
- Distinguishing between those records that appear infrequently in the original data, and those that never appear in the original data but appear infrequently in the synthetic data is an impossible task.

Conclusion

- CRR fails to reflect the actual privacy risk or the reidentifiability, unlike asserted by 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.

Thank you for your attention!
Vielen Dank für Ihre Aufmerksamkeit!
Gràcies per la vostra atenció!