UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Database reconstruction is very difficult in practice.

Krishnamurty Muralidhar (University of Oklahoma)
krishm@ou.edu

Abstract

The U.S. Census Bureau has motivated the use of differential privacy to protect the outputs of the 2020 Decennial Census by highlighting the dangers of reconstruction attacks (see Garfinkel, Abowd and Martindale (2019) "Understanding database reconstruction attacks on public data", Communications of the ACM, 62(3):46-53). We examine in detail the running example in that paper and we conclude it reveals quite the opposite: database reconstruction appears to be very difficult even for very small databases if classical statistical disclosure control techniques are properly applied (e.g. complementary cell suppression). In contrast, the use of differential privacy entails a very large utility loss even when the parameter epsilon is chosen to be as large as 10 (in which case practically no privacy is achieved).

Database reconstruction is very difficult in practice!

Krish Muralidhar[*] and Josep Domingo-Ferrer[**]

[*] University of Oklahoma, USA, krishm@ou.edu
[**] Universitat Rovira i Virgili, Catalonia, josep.domingo@urv.cat

Abstract: The U.S. Census Bureau has motivated the use of differential privacy to protect the outputs of the 2020 Decennial Census by highlighting the dangers of reconstruction attacks. We examine in detail the running example in that paper, and we conclude it reveals quite the opposite: database reconstruction appears to be very difficult even for very small databases if classical statistical disclosure control techniques are properly applied (e.g. complementary cell suppression). In contrast, the use of differential privacy entails a very large utility loss even when the parameter epsilon is chosen to be as large as 10 (in which case practically no privacy is achieved).

1 An Illustration of Database Reconstruction

In Garfinkel, Abowd and Martindale (2019), senior methodologists from the U.S Census Bureau highlight the dangers of reconstruction attacks. A closer examination of their paper reveals quite the opposite: *it shows database reconstruction is very difficult even for very small databases*.

In their paper, the authors use an example data set consisting of a total of **seven** individuals. They claim that this is a realistic example since “The 2010 U.S. Census contained 1,539,183 census blocks in the 50 states and the District of Columbia with between one and seven residents.” For every individual, the example data set reports: (1) Race – Black/African American (B) or White (W), (2) Gender – Female (F) or Male (M), (3) Marital Status – Married (M) or Single (S), and (4) Age (a numerical integer between 1 and 125). The information in Table 1 is released, which can be construed as the output of statistical queries on the original data set.

The authors also note the following: “Notice that a substantial amount of information in Table 1 has been suppressed—marked with a (D). In this case, the statistical agency’s disclosure-avoidance rules prohibit it from publishing statistics based on one or two people. This suppression rule is sometimes called ‘the rule of three,’ because cells in the report sourced from fewer than three people are suppressed. In addition, complementary suppression has been applied to prevent subtraction attacks on the small cells.” (page 48, Garfinkel et al., 2019).

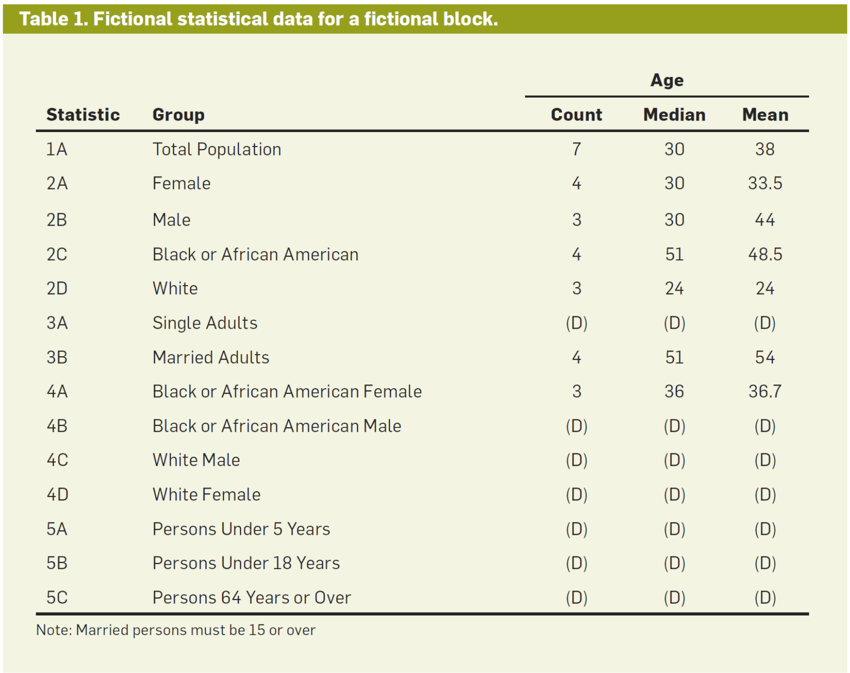


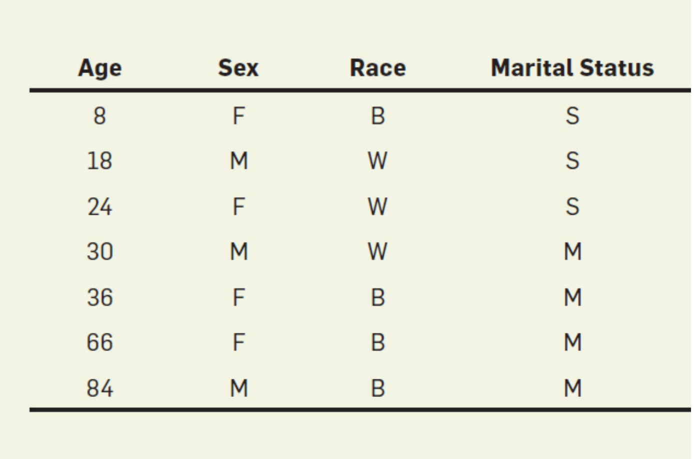
Table 1. Fictional statistical data for a fictional block.

			Age	
Statistic	Group	Count	Median	Mean
1A	Total Population	7	30	38
2A	Female	4	30	33.5
2B	Male	3	30	44
2C	Black or African American	4	51	48.5
2D	White	3	24	24
3A	Single Adults	(D)	(D)	(D)
3B	Married Adults	4	51	54
4A	Black or African American Female	3	36	36.7
4B	Black or African American Male	(D)	(D)	(D)
4C	White Male	(D)	(D)	(D)
4D	White Female	(D)	(D)	(D)
5A	Persons Under 5 Years	(D)	(D)	(D)
5B	Persons Under 18 Years	(D)	(D)	(D)
5C	Persons 64 Years or Over	(D)	(D)	(D)

Note: Married persons must be 15 or over

Table 1. Statistics released on the example data set (Table 1of Garfinkel et al (2019))

Using the information in Table 1 and a very sophisticated SAT Solver, the authors go on to show that the seven individuals in the original data set can be reconstructed. Table 2 shows the reconstructed data set.



Age	Sex	Race	Marital Status
8	F	B	S
18	M	W	S
24	F	W	S
30	M	W	M
36	F	B	M
66	F	B	M
84	M	B	M

Table 2. Disclosed information from the analysis of the released data (Table 4 of Garfinkel et al (2019))

Simple reconstruction is possible only because the disclosure prevention requirements were not properly implemented. Garfinkel et al (2019) claim "In Table 1, statistic 4A is an obvious candidate for suppression—especially given that statistics 4B, 4C, and

4D have already been suppressed to avoid an inappropriate statistical disclosure." The authors do not seem to realize that *this was not a choice*. We offer the following definition of primary and complementary suppression from a Census Bureau source (U.S. Census Bureau 2020, page 5):

> Primary and secondary/complementary suppression. Primary suppression protects against identity/attribute disclosure by replacing cells or records with a marker that identifies they have been suppressed or show as "No Data" (Antal et al., 2017). Secondary suppression involves suppressing additional nonflagged cells so that suppressed values cannot be derived through inferential disclosure. Alternatively, all problematic variables or entire flagged groups or geographies could be suppressed from dissemination (UNECE-CES, 2015).

With respect to the application of suppression rules for aggregate data, we find in another Census Bureau source (U. S. Census Bureau 2016, page 8):

> When there is a mean or aggregate in a table for a given geographic area that is suppressed by this rule, complementary suppression must be performed on other means or aggregates to show that area so that the suppressed mean or aggregate cannot be derived via subtraction.

In their example, Garfinkel et al (2019) have a suppression "Rule of three" and complementary suppression. In terms of primary suppression, "Rule of three" explicitly prohibits the release of information concerning two or fewer individuals. This implies that the "Rule of three" prohibits release the median age of any group regardless of size since the release of the median discloses the age of a single individual (group size is odd) or two individuals (group size is even). This directly contravenes the "Rule of three" requirements.

For complementary suppression to be properly applied in this example, response to query 4A *must be suppressed*. Since no complementary suppression has been applied, we can difference: (a) all (four) African Americans and (three) African American Females to disclose the values for the (single) African American Male, and (b) All (four) Females and (three) African American Females to disclose the values for the (single) White Female. Further, the release of the median age of (Black Females, White, and Female) results in the complete disclosure of the entire data set. Garfinkel et al (2019) also make the same error in other illustrations. For instance, they claim that "All of these solutions contain the reconstructed microdata records 8FBS, 36FBM, 66FBM, and 84MBM. This means that even if statistics 2A and 2B are suppressed, we can still infer that these four microdata records must be present." If 2A and 2B are suppressed, this inference is possible *if and only if* response from 4A is released. But *there is no point in suppressing (2A, 2B) if 4A is released*. When 4A is also suppressed (which is required by complementary suppression), reconstruction is impossible.

All the database reconstruction illustrated in Garfinkel et al (2019) is a direct result of not applying primary and complementary suppression properly. When the rules are properly applied (suppression of the median and complementary suppression of query 4A in Table 1), even for this toy data set, there are thousands of alternative solutions, making database reconstruction impossible.

2 Applying Differential Privacy to this Data Set

Garfinkel et al (2019) claim that the only way to prevent database reconstruction is to use noise injection based on differential privacy. In this section, we evaluate this claim. We applied Laplace noise addition to protect the data using two privacy levels ($\varepsilon = 1, 10$). Note that the second privacy level entails very weak protection for this small data set. To make the discussion easier, we limit our analysis to only three attributes (Age, Race, and Gender). There are two approaches to implementing Laplace noise addition:

(1) To treat each query as an independent query in a total of 10 queries (5 count queries and 5 mean queries) with ε being split for each query as $\varepsilon/10$.
(2) To treat the entire data as a table consisting of Age, Race, and Gender. The advantage of this approach is that the value of ε does not have to split among the different queries. The disadvantage is that a complete table of (Age by Race by Gender) would consist of a total of $(125 \times 2 \times 2 = 500)$ cells of which only seven cells have a non-zero value. To satisfy differential privacy, it would be necessary to add noise to every cell in the entire table (since there are no structural zeros), which would result in noise overwhelming the true values.

We chose the first approach. Table 4 contains a summary of the implementation parameters.

Overall ε	1		10	
ε per query	0.1		1.0	
	Count	Age	Count	Age
Global Sensitivity	1	124	1	124
Laplace Shape Parameter	10	1240	1	124
Noise Variance	200	3075200	2	30752

Table 4. Parameters for Laplace noise addition

We also implemented some common-sense output requirements: (a) all count values are set to zero when they are negative; (b) all count values are rounded to the closest integer; and (c) the mean age is limited to be between 1 and 125. Table 5 gives one realization from applying Laplace noise to the responses.

Description	Statistic	True Values		$\varepsilon = 1$		$\varepsilon = 10$	
		Count	Mean Age	Count	Mean Age	Count	Mean Age
Total Population	1A	7	38.0	0	1	4	32.1
Female	2A	4	33.5	4	125	4	54.9
Male	2B	3	44.0	0	1	1	36.4
Black or African American	2C	4	48.5	0	125	3	110.2
White	2D	3	24.0	18	125	4	51.4

Table 5. Output statistics after implementing Laplace noise addition

These results should not surprise anyone. With $\varepsilon = 1$, differentially private values are simply atrocious for both the Count and Mean Age queries. For $\varepsilon = 10$, the figures for the Count queries are better, but the figures for the Mean Age queries are still worthless. This was to be expected considering that the global sensitivity for the Age variable is so large that the noise dominates the true value. These results are only one realization. If we repeat the simulation several times, the conclusions stay the same.

3 Summary and Conclusions

In summary, if the article by Garfinkel et al. (2019) proves anything, it proves that *even* for a very small data set, *even* when a lot of information is released, if *simple* disclosure prevention techniques are *properly* applied, *uniquely reconstructing the data set is very difficult*. It is important to remember that this was a *hypothetical* scenario created by two senior scientists from the Census Bureau plus an academic. If this is the scariest scenario that they can come up with, then we have little to worry from database reconstruction.

To use database reconstruction attacks to justify the use of differential privacy is doubly worse. *Even* for this very small database, *even* with practically no privacy (ε = 10), the performance of a differentially private procedure is terrible. Common sense and simple disclosure prevention (properly applied minimum cell count requirement with complementary suppression) is completely adequate to prevent reconstruction in this case. A slightly higher minimum cell count (with complementary suppression) would make the reconstruction even more difficult. We do not mean to imply that other procedures will not be necessary in other scenarios. But we certainly do mean to imply that *differential privacy is not the only solution*.

References

Antal, L., T. Enderle, and S. Giessing (2017) Harmonised Protection of Census Data in the ESS: Statistical disclosure control methods for harmonised protection of census data, Eurostat Centre of Excellence on Statistical Disclosure Control, The Hague.

Garfinkel, S., J.M. Abowd, and C. Martindale "Understanding Database Reconstruction Attacks on Public Data," *Communications of the ACM*, 62(3), 2019, 46-53.

United Nations Economic Commission for Europe - Conference of European Statisticians - UNECE-CES (2015) Recommendations for the 2020 Censuses of Population and Housing, United Nations Publications, New York, NY.

U.S. Census Bureau (2016) American Community Survey: Data Suppression. Sep. 27. https://www2.census.gov/programs-surveys/acs/tech_docs/data_suppression/ACSO_Data_Suppression.pdf

U.S. Census Bureau (2020) Disclosure Avoidance and the Census, Select Topics in International Censuses, October. https://www.census.gov/content/dam/Census/library/working-papers/2020/demo/disclosure_avoidance_and_the_census_brief.pdf

Acknowledgments. Support from the European Commission (projects H2020-871042 "SoBigData++" and H2020-101006879 "MobiDataLab"), the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer) and the Spanish Government (project RTI2018-095094-B-C21 "Consent") is gratefully acknowledged. The second author is with the UNESCO Chair in Data Privacy, but the views in this paper are the authors' own and not necessarily shared by UNESCO.