Database Reconstruction is Very Difficult in Practice

Krish Muralidhar, University of Oklahoma (krishm@ou.edu)

Abowd (Senior Scientist at Census) and Garfinkel (Senior Computer Scientist at Census) wrote a paper ("Understanding Database Reconstruction Attacks on Public Data," *Communications of the ACM*, 62(3), 2019, 46-53) highlighting the dangers of reconstruction attacks. A closer examination of the 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. The authors 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 each individual we have: (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 following information is released.

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

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."

Using this information and a very sophisticated SAT Solver, they go on to show that we can reconstruct the individuals in the database. I reconstructed the database using logic and arithmetic (see end of this note for details).

Individual	Age	Race	Gender	Marital Status
1	84	B	M	M
2	66	B	F	M
3	36	B	F	M
4	30	W	M	M
5	24	W	F	S
6	18	W	M	S
7	8	B	F	S

This is a very poor example for many reasons. First, even though the authors claim that complementary suppression has been applied, from the example it is obvious that the only restriction that has been applied is the "rule of three" without any complementary suppression. Because no complementary suppression has been applied, we can difference all (four) African Americans and (three) African American Females to disclose the values for the (single) African American Male.

If you applied complementary suppression to this data, then responses to queries 3B and 4A will also be suppressed. You can attempt to recreate the database with the remaining information. You still do not need SAT Solver to solve this problem, just simple logic/arithmetic. Without responses to queries 3B and 4A, you cannot reconstruct the database uniquely. There are only four different combinations of Race and Gender of which you can eliminate two (see end of note for the combinations of Race and Gender). Now you have two possibilities for which you solve for Age. No reconstruction on Marital Status is possible (other than Minimum Marriage Age).

Age	Race	Gender	Marital Status	Individual	Age	Race	Gender	Marital Status
84	B	M	--	1	90	B	M	--
66	B	F	--	2	72	B	F	--
36	B	F	--	3	36	W	F	--
30	W	M	--	4	30	B	M	--
24	W	F	--	5	24	W	F	--
18	W	M	--	6	12	W	M	S
8	B	F	S	7	2	B	F	S

We could argue that there is partial reconstruction (missing the Marital Status). There is only one individual who has the same Age, Race, and Gender in both options (Individual 5). Four other individuals have the same Race and Gender but different age (Individuals 1, 2, 6, and 7). The remaining individuals have a different Race or Gender. Overall, for such a small data set, there is

considerable uncertainty about the reconstruction. But even this partial reconstruction is possible *only because of the median age*.

Releasing both the mean and median (particularly median) age for this small data set is highly disclosive. *The authors are fully aware of this*, observing that query 2B (information regarding the three Males) reduces the number of possible age combinations from 317,750 to only 30. In conjunction with the other information, this quickly reduces to only 2. If you eliminate the information on median age, there are thousands of possible combinations of age*, making accurate reconstruction impossible even for this very small data set.*

Applying Differential Privacy to this Data Set

An alternative to implementing simple disclosure limitation techniques is to use a technique based on differential privacy to protect this data. In this section, I applied Laplace noise addition to protect the data using two privacy levels ($\varepsilon = 1, 10$). Note that the second specification is very weak privacy for this small data set. To make the discussion easier, I limit my 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 with 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), resulting in noise overwhelming the true values.

I chose the first approach. Here is 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

I also implemented some commonsense 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) Mean age is limited to be between 1 and 125. Here is 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

I am sure that these results are not surprising to any of us. With $\varepsilon = 1$, the results are simply atrocious for both the Count and Age queries. For $\varepsilon = 10$, the results for the Count queries are better, but the results for the Mean Age queries are still worthless. This is not surprising at all 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. We could do simulations to replicate the results, but it is not going to change our conclusions.

Summary

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, *reconstructing the data set is practically impossible*. It is important to remember that this was a *hypothetical* scenario created by two senior scientists from the Census Bureau. 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 ($\varepsilon = 10$), the performance of a differentially private procedure is terrible even for this simplistic example. Commonsense (not to release both mean and median of age) and simple disclosure prevention (properly applied complementary suppression) are completely adequate to prevent reconstruction in this case. I do not mean to imply that other procedures are unnecessary in any scenario. But I certainly do mean to imply that *differential privacy is not the only solution*.

The Abowd paper says (above the title on the first page): "**These attacks on statistical databases are no longer a theoretical danger**." Freudian slip?

Simple Approach for Reconstructing the Original Data

(1) Knowing that there four Black individuals (2C) and three of them are Females (4A), we identify there is only one Black Male, age 84.
(2) Knowing there are four females (2A) and three of them are Black (4A), we identify that there is only one White Female, age 24.
(3) Knowing that the median age of males is 30 (2B) and the one Black Male is of age 84, we know that one of the While Males must be age 30 and the other White Male is age 18.
(4) Knowing that the median age of Black Females is 36 (4A), we know that one of the Black Females must be age 36. Since the mean age is not 36, we also know that one of the Black Females must be age less than 36 and the other must be age greater than 36.
(5) Knowing that the median age of Black is 51 (2C), one Black Male age 84, and one Black Female age 36, one of the remaining Black Females must be age 66.
(6) Knowing that the average age of Black Females is 36.7 (4A), one Black Female is age 36 and one Black Female is age 66, we know that the remaining Black Female is age 8.
(7) Knowing that the average age of Married Adults is 54 and median is 51 (3B), we compute that the Married Adults are Black Male age 84, Black Female age 66, Black Female age 36, and White Male age 30.

Possible values for Race and Gender Combinations

	Option 1		Option 2		Option 3		Option 4	
	Male	Female	Male	Female	Male	Female	Male	Female
Black	0	4	1	3	2	2	3	1
White	3	0	2	1	1	2	0	3