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

These attacks on statistical databases are no longer a theoretical danger.

BY SIMSON GARFINKEL, JOHN M. ABOWD, AND CHRISTIAN MARTINDALE

Understanding Database Reconstruction Attacks on Public Data

COMMUNICATIONS OF THE ACM | MARCH 2019 | VOL. 62 | NO. 3

Motivation of the authors

- "Published statistical tables are vulnerable to database reconstruction attacks (DRAs), in which the underlying microdata is recovered merely by finding a set of microdata that is consistent with the published statistical tabulations."
- "This article shows how such an attack can be addressed by adding noise to the published tabulations, so the reconstruction no longer results in the original data. This has implications for the 2020 census."

Example Data

- Hypothetical example developed by the authors to specifically highlight database reconstruction risk
- Disclosure limitation
  - **"Rule of three"**: All responses for fewer than three people must be suppressed
  - **Complementary suppression**: All responses that will violate "rule of three" by differencing or other computations must be suppressed

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

Garfinkel et al. (2019) Solution Procedure

SAT and SAT Solvers

The Boolean SAT problem was the first to be proven NP-complete.[9] This problem asks, for a given Boolean formula, whether replacing each variable with either true or false can make the formula evaluate to true. Modern SAT solvers work well and reasonably quickly in a variety of SAT problem instances and up to reasonably large instance sizes.

Many modern SAT solvers use a heuristic technique called CDCL (conflict-driven clause learning).[10] Briefly, a CDCL algorithm:

1. Assigns a value to a variable arbitrarily.
2. Uses this assignment to determine values for the other variables in the formula (a process known as unit propagation).
3. If a conflict is found, backtracks to the clause that made the conflict occur and undoes variable assignments made after that point.
4. Adds the negation of the conflict-causing clause as a new clause to the master formula and resumes from step 1.

This process is fast at solving SAT problems because adding conflicts as new clauses has the potential to avoid wasteful "repeated backtracks." Additionally, CDCL and its predecessor algorithm, DPLL (Davis–Putnam–Logemann–Loveland), are both provably complete algorithms: they will always return either a solution or "Unsatisfiable" if given enough time and memory. Another advantage is that CDCL solvers reuse past work when producing the universe of all possible solutions.

A wide variety of SAT solvers are available to the public for minimal or no cost. Although a SAT solver requires the user to translate the problem into Boolean formulae before use, programs such as Naoyuki Tamura's Sugar facilitate this process by translating user-input mathematical and English constraints into Boolean formulae automatically.

Sugar Input

Sugar input is given in a standard constraint satisfaction problem (CSP) file format. A constraint must be given on a single line of the file, but here we separate most constraints into multiple lines for readability. Constraint equations are separated by comments describing the statistics they encode.

Input for the model in this article is available at https://queue.acm.org/appendices/Garfinkel_SugarInput.txt.

- "Table 1 translates into 164 individual s-expressions extending over 457 lines. Sugar then translates this into a single Boolean formula consisting of 6,755 variables arranged in 252,575 clauses. This format is called the CNF (conjunctive normal form) because it consists of many clauses that are combined using the Boolean AND operation."
- "Brute Force" approach

Results of analysis: Table 4, Garfinkel et al. (2019)

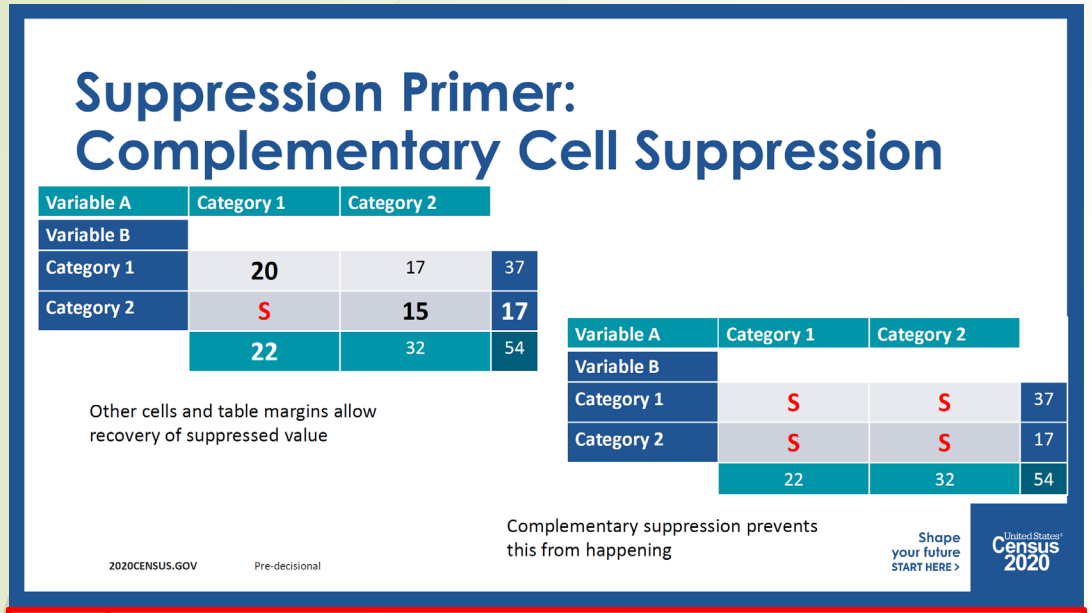
Age	Sex	Race	Marital Status
8	F	B	S
18	M	W	S
24	F	W	S
30	M	W	M
36	F	B	M
66	F	B	M
84	M	B	M

Intelligent analysis versus Brute Force

- No need for SAT Solver and Sugar
- Intelligent approach
  - (2A – 4A) discloses 4C completely
    - 2A, 4A, 4C discloses all Female Age
  - (2C – 4A) discloses 4B completely
    - 2B, 4B discloses 4D
  - (1A – 3B) discloses 3A
  - 4A also discloses Age for one Black Female
  - Median Age of Females discloses Age of two remaining Black Females
- Completes (Race, Sex, Age) characteristics of all individuals in the database. Only one combination of Age satisfies 3A, 3B. Reconstruction is complete!

Disclosure can be attributed mainly to the information in 4A

Suppression (Rule of three) and Complementary Suppression



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.

Determining the Privacy-loss Budget - Research into Alternatives to Differential Privacy

Michael Hawes (US Census Bureau)
Rolando A. Rodríguez (US Census Bureau)

Census Scientific Advisory Committee
May 25, 2021

United States™ Census Bureau

American Community Survey

Data Suppression

Improper implementation of disclosure limitation rules!

- Authors say: "statistic 4A is **an obvious candidate for suppression**" Suppressing 4A is not a choice. Complementary suppression requires 4A to be suppressed.

- "Rule of three" explicitly prohibits releasing Median Age for any group.
  - For odd size groups, it discloses the true age of a single record
  - For even size groups, it discloses the sum of true age of two records

- There is no reason to suppress responses to query 3A,

Once suppression and complementary suppression are properly applied

Statistic	Group	Age		
		Count	Median	Mean
1A	Total population	7	(D)	38
2A	Female	4	(D)	33.5
2B	Male	3	(D)	44
2C	Black	4	(D)	48.5
2D	White	3	(D)	24
3A	Single	3	(D)	16.7
3B	Married	4	(D)	54
4A	Black Female	(D)	(D)	(D)
4B	Black Male	(D)	(D)	(D)
4C	White Female	(D)	(D)	(D)
4D	White Male	(D)	(D)	(D)

- No unique reconstruction.
- Thousands of alternative solutions.
- Practically no additional information can be inferred.

Simplicity and versatility

- Suppression along with complementary suppression offers a simple, effective approach. For "Rule of three":
  - Population size (0 – 2) – No data release
  - Population size (3 – 5) – Population level data released
  - Population size (6, 7) – Population level and individual category level data released
  - Population size > 7 – Population level, individual category level, and multi-category (cross tabs) data released
    - The cross tabs will be a function of the number of categories
- Search Space = 2^{mn} (n = Group size, m = # of binary categories)

Further improper implementation of primary and complementary suppression

- The authors claim that suppressing 4A results in two feasible solutions. But only if the value of Median Age is available.
- The authors claim: "For example, dropping statistic 2A, 2B, 2C, or 2D still yields a single solution,"
  - There is no point in dropping just 2A and not 2B since you can infer 2B by differencing (1A – 2A) and vice versa.
- The authors claim: "Dropping 2A and 2B increases the solution universe to eight satisfying solutions. All of these solutions contain the reconstructed microdata records 8FBS, 36FBM, 66FBM, and 84MBM."
  - This is possible only because the authors assume responses to query 4A is released. Without 4A, *no information regarding the Sex attribute is released and it is impossible to infer the Sex attribute for any record*.

Main conclusion from this analysis

- We have shown that if disclosure rules had been properly applied, then there is no reconstruction.

- Even for this very small hypothetical example, carefully chosen by two senior Census researchers, properly implemented primary and complementary suppression makes it very difficult to reconstruct the data.

The *only* solution to the database reconstruction problem, according to Garfinkel et al (2019)

- "To protect the privacy of census respondents, the Census Bureau is developing a privacy-protection system based on differential privacy."
- "This article has explained the motivation for the decision to use differential privacy."
- "By using differential privacy, we can add the minimum amount of noise necessary to achieve the Census Bureau's privacy requirements."
- We investigate this claim

Laplace noise addition for this data set

- We limit our analysis to only three attributes (Age, Race, and Gender)
- There are two approaches to implementing Laplace noise addition
  - 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 ε/10.
  - To treat the entire data as a table consisting of Age, Race, and Gender.
    - Advantage: No splitting of ε
    - Disadvantage: Noise must be added to every cell in the table of (Age by Race by Gender) (125×2×2=500 cells) of which only seven have values
- We use the first approach.
  - The Census Bureau has announced that it is adopting a noise-injection mechanism based on differential privacy to provide privacy protection for the underlying microdata collected as part of the 2020 census. (Garfinkel et al)

Responses after Laplace noise addition

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

Summary of results

- With $\varepsilon = 1$, differentially private query responses are simply atrocious for both the Count and Mean Age queries.
- With $\varepsilon = 10$, the responses for the Count queries are better, but the responses for the Mean Age queries are still worthless.
  - 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.

Conclusions

- Even for a very small data set, even when a lot of information is released, simple disclosure prevention techniques, properly applied, are extremely effective at preventing database reconstruction.
- Garfinkel et al (2019) results do not justify differential privacy.
  - The database reconstruction was a direct result of improper application of the disclosure limitation rules
  - Even for this very small database, even with practically no privacy ($\varepsilon = 10$), the responses from the differentially private procedure are useless.
- Differential privacy is not always the solution.

Postscript

- In our opinion (**ε = 10**) represents ***"practically no privacy"***
- The US Census Bureau has chosen (**ε = 19.61**) for the 2020 US Census
  - "… the bureau settled on an epsilon of 19.61, significantly higher than where the dial was set in earlier versions …"
    - https://apnews.com/article/business-census-2020-55519b7534bd8d61028020d79854e909

You decide whether (ε = 19.61) provides any privacy!

Dziękuję bardzo