Integrated Session 3: Database Privacy

Due: Thursday, 28 October 2021 at class start

Total Homework/Assignment Points: 100

Your neighborhood association is conducting a questionnaire that asks its members their age, sex, and race. You participate truthfully, since you know that the association depends on the accuracy of the results, but you are also wary of your nosy neighbors, who are interested in juicy gossip about how old you are (because you look so good, naturally). When the survey is complete, the association publishes the results. You notice that they include not individual records (or 'microdata') but statistics taken over the neighborhood as a whole. It looks like this:[1]

statistic	group	count	median	mean
1A	total population	7	37	35
2A	female	4	39.5	41
2B	male	3	25	27
2C	black or African American	3	40	48
2D	white	4	26.5	25.25
3A	single adults	(D)	(D)	(D)
3B	married adults	4	38.5	45.25
4A	black or African American male	(D)	(D)	(D)
4B	black or African American female	(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)

(The entries labeled '(D)' indicate suppression. For these attributes the counts are 2 or below). *You are reassured: intuitively this database, reporting counts, means, and medians, and with so much suppressed, poses little risk to your privacy.*

Or does it? You decide to perform a technical audit to be sure, attacking the database's protections in order to reconstruct the microdata about yourself.

Setup

Download the file Garfinkel_SugarInput.txt, and open it in a plain text editor.[2] Search for the string 'Q1.' you will see a comment that looks like this (comment lines always begin with two semicolons):

```
;; ==========================
;; Q1. YOUR ANSWER HERE
;; ==========================
```

For each question in this assignment, search for the respective comment to find the place in which to write your response. Below the comment, and without any semicolons, you will write *constraints* that come from

[1] Adapted from Garfinkel, Abowd, and Martindale (2019).

[2] Use Notepad in Windows, TextEdit in plaintext mode in MacOS, and gedit/vim/emacs in other BSDs or Linux distributions. Do not open the file in a word processor like Word.

the statistics reported in the table. For example, there is a constraint that the neighborhood's mean age is 35: not every assignment of ages to you and your neighbors would have that mean—some are ruled out. That insight, combined over all the statistics, is the basis of the attack.

Representing Constraints

1. There is a constraint on the possible sexes in the data: all people are either FEMALE or MALE. The constraint on sex for person 1 is written like this:

   ```
   (int S1 0 1)
   ```

   This says that S1, the sex of person 1, is 0 (for female) or 1 (for male). Notice at the top of the file under 'DEFINITIONS' that 'female' and 'male' are both defined as numbers, and so are other words.[3] Type in the constraint that S1 is female or male (given above), then complete the remaining constraints for $S2, S3, \ldots, S7$, following the same model. Complete the exactly similar range constraints for Ages, Races, and Marital Statuses (the description for each of these is given in Garfinkel_SugarInput.txt, in the comment immediately below the Q1 comment).

2. In addition to specifying the range of values that a variable can take, constraints can also take the form of equations. For example

   ```
   (= (+ A1 A2) (* A1 A2))
   ```

   would represent the constraint that individuals 1 and 2 had ages whose sum and product were equal (that's not true; it's just an example). For Q2, write the constraint expressing that the mean age of the neighborhood is 35, using only =, +, and *.

3. Write a constraint representing that the female count is 4. *Hint: remember that* 'female' *is defined to be 0, while 'male' is defined to be 1, and think about sums.*

4. Write the constraint that the mean female age is 41. *Hint: Follow the model for question 2 above, but use the* FEMALE_AGE1, FEMALE_AGE2, ..., FEMALE_AGE4 *variables.*

5. Write a constraint on the range of possible values for number of people in the *suppressed* category SINGLE_ADULT_COUNT.

Reconstructing the microdata

1. First, save the file with your edits. Next copy the file you've just edited to our server using SCP, as follows, replacing 'username' with the username sent to you via email, and entering the password also sent by email when prompted. For MacOS and other BSDs, and for Linux distributions, do this (the '$' is the prompt, which might look different on your computer. You only type in the part after the prompt):

   ```
   $ scp /path/to/Garfinkel_SugarInput.txt username@198.199.111.52:db_reconstruction
   ```

   For Windows the command is very similar, see this short walk-through.

2. Next connect to the server over ssh, change to the db_reconstruction directory, and solve the *constraint satisfaction problem* you have defined:

   ```
   $ ssh username@198.199.111.52
   $ cd db_reconstruction
   $ ./reconstruct.sh
   ```

[3] Question 4 in 'Interpreting the Results' below asks how privacy relates to this constraint from a different perspective.

Did you succeed in recovering your microdata? Who else do you observe in the results?

Interpreting the results

1. List three questions that you have about how the attack in this assignment might apply to other databases.
2. Suppose you were able to find three possible assignments for your age and other data from the audit, instead of one. Describe why this might still represent a harm from nosy neighbors.
3. A different approach to mitigating database privacy risk is to leave the collected microdata unchanged, but restrict access to it, and to develop policies to govern its use. What do you see as the strengths and weaknesses of this approach, as compared to techniques for protecting privacy by transforming released data?
4. Suppose a neighborhood with a large number of gender-nonconforming people is given a questionnaire with an item related to sex, offering MALE and FEMALE responses, and that there is no item asking about gender. As a result, many people in that neighborhood leave the question blank, and write in an additional item. Is this an example of another kind of database privacy problem? Why might people do this? Why do databases *matter*?
5. In this assignment we were protecting our age from nosy neighbors. How would the situation change if the collection were performed on demographics by a national government? Discuss broadly who is vulnerable to demographics data collection, and why.

References

Garfinkel, Simson, John M. Abowd, and Christian Martindale (Feb. 21, 2019). "Understanding database reconstruction attacks on public data". In: *Communications of the ACM* 62.3, pp. 46–53. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3287287. URL: https://dl.acm.org/doi/10.1145/3287287 (visited on 03/25/2020).