COS125 - Precept 8 (Ethics)

1 Confidentiality & Theories of Privacy

Samir and Samuel are married, which everyone knows. They tell their neighbors over dinner that they have a very normal sex life, and are intimate with each other about as much as the average person. One of the neighbors at that dinner tells Samir's boss at a restaurant what Samir has said. Another of the neighbors tells Samuel's boss at a government agency.

Has either of these neighbors violated confidentiality? Write 1-2 sentences explaining your answer.

Have they violated privacy? Use a theory of privacy you learned to write one sentence about the case of Samir and one about the case of Samuel, explaining your answers.

2 Database Search

Download `precept8.zip` from the precepts webpage, unzip and open the project folder. The folder contains two files with personal and healthcare information on (fictitious) residents of Massachusetts, `healthcare_data.csv` and `personal_data.csv`. The information is in CSV format, which Excel (on any platform), Numbers (on Mac) and LibreOffice Calc (on Linux) can all read.

Open `healthcare_data.csv` and sort the dataset by columns to count the number of matches for the following ages:

2	3	27	29	33	40	50	70	72	89	100

Now, sort again to count the number of matches for the following zip codes:

01060	01105	01604	01835	01841	01940	02138	02148	02149	02641	02726

3 De-Anonymization

We'll now see how a few (sometimes even one or two) seemingly innocuous pieces of information can be used to uniquely identify a person (thus de-anonymizing information from a public dataset).

For each of the combinations of feature values below, count the number of matching entries in `healthcare_data.csv`, a dataset with 1163 people. If there's only one, find (in `personal_data.csv`) and write their name.

- Age `2` and gender `M`:
- Age `27` and zip code `2184`:
- Age `29`, race `asian` and gender `F`:
- Age `36`, gender `F` and healthcare expenses above \$7,000,000:
- Age `40`, gender `M` and no recorded zip code:
- Race `native`:
- Age `72` and healthcare coverage below \$10,000:
- Age `100`:

Give an example of private information that can be revealed from healthcare coverage and/or healthcare expenses.

Imagine you are the employee of the Commonwealth of Massachussetts in charge of responsibly releasing the information in `healthcare_data.csv`. How would you modify the dataset to prevent de-anonymization? In particular, what is the effect of removing the `ZIP` feature/column?