Topics in Data Science, Data Protection — Spring 2020
CS 4364 — L. Longpré

Class notes 3/5/2020

Exercise

Suppose the table below reflects integer data that has been rounded to the closest multiple of 10, and if the data ends in 5, it is rounded up. (For example, 65 is founded up to 70.) Find two possible solutions. (I believe there are no other other solutions different from permutations of the two solutions.)

20	20	50
20	20	50
50	50	100

DNA databases

Follow the link on our course website to the recent news on how access to a DNA database helped police charged someone for murder. Read the article (5 minutes) and discuss with your neighbor(s) the following quesitons:

1. Was it for police to lie in this case in order to solve the crime?
2. Was there a leak of privacy when police's investigation accessed the DNA database?

Balancing research with privacy and protection

Read introduction/abstract of the two Genome/DNA related papers on the course website.

Differential privacy

Purpose: (paradox?) Learning nothing about an individual while learning useful information about a population.
Other purpose: address the possibility of "side information" like re-identification attacks.
Example: From medical database, learn that smoking causes cancer. Health insurance increases rates of smokers.

Query auditing not feasible in general.
Example: Suppose people's height considered highly sensitive. Side information: a friend told you that he is two inches shorter than the average Texan.
Formalization of "access to a statistical database should not enable one to learn anything about an individual that could not be learned without access" is impossible.
Aspects about impossibility result: applies whether your friend is in the database
Differential privacy: database disclosure reveals essentially the same information about an individual whether or not the individual is part of the database.
Definition to classify algorithms that release statistics about data. The goal is for an observer seeing the algorithm's output not being able to tell if a particular individual's information was used by the algorithm. In other words, the result of the algorithm should not be differentiable if an individual is included or not.
Definition:
A randomized function K gives ε-differential privacy if for all data sets D_1 and D_2 differing on at most one element,

$$Pr[K(D_1) \in S] \leq exp(\varepsilon) \cdot Pr[K(D_2) \in S]$$