CS 5436 / INFO 5303 Fall 2024
Homework 4

Due: December 5, 11:59p ET

This is an INDIVIDUAL assignment.
You may discuss, but each student must submit their own work.

REQUIRED (50 points)

Problem 1 (9 points)

(a) Define the following terms and give an example of each.

Identifier

Pseudonym

Quasi-identifier

What are the differences and similarities between the terms? Why are quasi-identiers important in anonymization research?

(b) You are given an anonymized loyalty-card database of a major national retailer. For each customer, it contains that customer's ZIP code, gender, date of birth, and the list of his or her purchases over the past year. Your objective is to identify products that Republican voters are likely to buy. How would you go about this? Be specific about the use of auxiliary information.

(c) A case based on Latanya Sweeney's de-anonymization research went to the courts. In Southern Illinoisan v. Illinois Department of Public Health, the court rejected the claim, arguing:

The court posited that it was not reasonable to believe that someone with less knowledge, education, and experience in this area would be as successful as Dr. Sweeney in using the information provided to arrive at the same results Dr. Sweeney reached. Are there two people in the entire state of Illinois who could replicate Dr. Sweeney's results with the same limited data or are there two thousand? Are there

zero or are there a million? These questions are significant because without some sense of the magnitude of the alleged threat of which the defendants complain, it is very difficult for this court to determine whether the data in question reasonably tends to lead to the identity of specific persons.

Discuss the court's reasoning.

Problem 2 (2 points)

What are the key characteristics of database owners that, in Paul Ohm's view, should be subject to new privacy regulations? Why?

Problem 3 (6 points)

Differential privacy guarantees that what can be learned about a data subject from a differentially private data release is close to what could have been learned if the analysis had been performed without that individual's data. Generally, this is achieved by adding random noise to function outputs in order to obscure the true value of each individual data point. The amount of noise needed to provide the same level of privacy increases for function outputs with large sensitivity.

(a) In technical terms, make precise the underlined concepts in the paragraph above.

(b) Differential privacy is sometimes interpreted as hiding the presence of an individual in a dataset (eg, the list of all NY residents who received a Covid vaccine). Suppose I publish a selfie with a vaccine card on Instagram. Now any adversary with access to this auxiliary information (ie, my Instagram feed) can tell that I was part of the dataset. Has differential privacy of the vaccination count been broken?

Problem 4 (9 points)

(a) What is longitudinal privacy?

(b) Every time a client submitted a randomized response, epsilon has to be "charged" against the privacy budget. How does RAPPOR deal with the problem of a client submitting multiple reports based on the same underlying value, quickly exhausting the privacy budget?

(c) Suppose Google wants to use RAPPOR to find 1,000 most common URLs that cause the Chrome browser to crash. Approximately how many RAPPOR reports does Google need to collect?

Problem 5 (3 points)

A recent paper on data extraction from large language models shows that simply prompting these models with short texts can reveal potentially sensitive information. Explore this idea using Google Colab and come up with 5 different prompts that produce interesting, i.e. privacy sensitive, results. What were your criteria for choosing the prompts that worked for you?

Problem 6 (9 points)

In the federated averaging algorithm, which is the core of federated learning, users' data stay on their devices; only model updates are shared with the central aggregator.

(a) What is secure aggregation?

(b) Why confidentiality of users' data cannot be achieved without secure aggregation?

(c) Developers of federated learning frameworks assert that they protect privacy because "your data stays in your device". Is this claim sufficient or necessary to establish that federated learning achieves privacy? Why or why not?

Problem 7 (12 points)

To complete this portion of the homework, please cut-and-paste the write-up from the November 26, group activity.

– HERE –

Problem 8 (Bonus 2 points)
Copy from and link to the Privacy Design in-class activity.

(a) Summarize your group's ideation discussion. Include an overview of the picked technology, and data. What is your privacy goal, and what privacy strategies did you consider?

(b) Provide a link and screenshot of your user flow, like a service blueprint or flowchart. You can collaborate with your group to complete it if needed.

PICK YOUR OWN – should add up to 10 points
For each problem you pick, do the entire problem (i.e., you cannot choose-and-mix subproblems)

Problem CS1 (10 points)

In this exercise, you will experiment with local differential privacy and try to see for yourself how many respondents it requires to produce accurate estimates. For simplicity, we will focus on yes/no questions (e.g. did you ever cheat on a homework assignment?)

(a) Implement a Python routine that emulates an individual user. This routine should take two inputs: (1) the bit indicating the user's true answer, (2) epsilon. The output is the user's randomized response. To compute the response, you will need an implementation of Laplace noise; feel free to use numpy.random.laplace.

```
def local_dp(answer: bool, epsilon: float) -> bool:
        # your code goes here
```

Submit your implementation of the above function signature, it will be graded as part of this homework.

(b) Suppose for 5% of Cornell students the true answer is "yes" (the bit is 1). Assume N=40 students in a class. Use your randomized response routine to emulate individual students' answers and compute an estimate of the number of students in a class who cheated on an assignment.

What value of epsilon did you choose and why?

What estimate did you produce (give a single number)?

Repeat the experiment using N=23,000 (total Cornell enrollment). What is your estimate now?

Problem INFO1 (10 points)

(a) Why are many privacy settings in apps so difficult to use? Describe the two key reasons discussed in the Privacy Design lecture. (2 points)

(b) Companies have leveraged the insights of a particular discipline (field of study) to navigate around meaningful user consent. What discipline is that? (1 point)

(c) What are these methods influencing user choice against their interests commonly called? (1 point)

(d) Provide an example of a common privacy setting and calculate the total number of configuration options a user might encounter when given complete control. Using the CI tuples, create a table that details these information flows. (3 points).

(e) What is the main goal of Privacy by Design? (1 point)

(f) What privacy-by-design strategies were discussed in class? Which is your favorite? Explain. (2 points)