Homework 3

Due: April 20, 11:59p ET

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

REQUIRED (40 points)

Problem 1 (8 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-identifiers 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 (4 points)

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

Problem 3 (8 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 on Instagram a selfie with a vaccine card. 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 (8 points)

For each of the cases, answer the following questions (based on lecture and course readings):

i) What are the potential legal issues
ii) Which area of law seems most relevant?
iii) In your view, is the action in question legal (Yes/No/Depends)
iv) What are the main reasons for your choices?

(a) Cornell sells student transcripts to a headhunting firm.

(b) Coursera sells performance information about you to a headhunting firm.

(c) Acxiom (a data broker) obtains divorce records from the New York State Courthouse

(d) A party guest posts photos of your house on social media.

(e) The IRS hands over President Biden's tax returns to you, because you asked.

Problem 5 (4 points)

(a) What does the 4th Amendment say? (In your own words)
(b) Briefly describe the original scope of the 4th Amendment.

(c) Define “reasonable expectation of privacy” and (ii) explain how it extends the original scope, per your answer in (b)

Problem 6 (6 points)

Pick two out of the three cases and answer the following questions based on readings, lecture, and group work. You may continue to work in your group but students should include answers in HW individually even if your answers are identical with groupmates.

i) Names of the people in your group

ii) Today, would the government need a warrant for … ? Briefly explain your answer, based on the five factor test. It’s ok to make factual assumptions (per discussion in class), as long as you say what they are.

(a) 10 years of chat records (e.g. on Whatsapp, iChat, FB)

(b) Facial recognition scanner

(c) Drone surveillance of an individual as they go about their day

Problem 7 (2 points)

(a) What data are protected under GDPR?

(b) Based on what we’ve covered in the technical parts of the course, why might this be a problem?

PICK YOUR OWN – should add up to 20 points

For each problem you pick, do the entire problem (i.e., you cannot choose-and-mix subproblems)

Problem CS1 (12 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.

```python
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 CS2 (8 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?

INFO1 (8 points)

(a) What does the (US) Privacy Act have in common with the EU General Data Protection Regulation (GDPR)?

(b) How does the GDPR define the consent requirement and how does this differ from the typical expectation embodied in the US consumer privacy domain?

(c) List and briefly describe 3 major differences between the GDPR and the CCPA.

(d) Provide one reason explaining how each legal framing (i.e. GDPR: Human Rights, CCPA: Consumer Privacy) can be limiting when it comes to protecting privacy.

INFO2 (7 points)

(a) Under the Fourth Amendment, describe what constitutes a search, seizure, and probable cause.
(b) Briefly summarize the landmark information privacy case *Katz v. United States*. Why did this case overturn *Olmstead v. United States*? (Hint: Physical intrusion/property)

(c) Pick another landmark case that deals with technology and explain how it affected the way legal scholars think about the Fourth Amendment.

INFO3 (2 points)

(a) In the US, The Privacy Act of 1974 was a major disappointment. Why?

(b) The Bill of Rights is severely limited in its protection of privacy. Explain.

INFO5 (3 points)

(a) Why is the US's approach to privacy regulation called sectoral?

(b) Name two sectoral laws and briefly say what they regulate

(c) Name one other part of the world that takes a different approach