CS 5436 / INFO 5303 Spring 2022

Homework 2

Due: April 26, 11:59p ET

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

COMMON FOR CS AND INFO

Problem 1 (12 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 (6 points)

Several U.S. laws single out certain categories of information (e.g., driver licenses, health records, video rental records) for special privacy protection.

(a) What are the two main reasons for singling out these categories?

(b) Do these reasons still hold up in the face of modern de-anonymization and re-identification techniques?

(c) Is the IP address “personally identifiable information”? Give an argument why it is and why it is not.

Problem 3 (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 4 (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 (e.g., 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 (i.e., my Instagram feed) can tell that I was part of the dataset. Has differential privacy of the vaccination count been broken?

Problem 5 (4 points)

A company called Dataminr sells a product called First Alert to law enforcement and other government clients. This product monitors Twitter and generates alerts about possible criminal behavior. Law enforcement customers use First Alert for the “24/7 monitoring of crimes,” as well as to detect “future criminal action.” Alerts include information about potential demonstrations, including handles of users discussing plans for protests, whereabouts of protesters, locations of street blocking and sit-ins, etc.

Dataminr pays for fast, large-scale access to public tweets under Twitter’s Data Partner program. Twitter’s rules prohibit partners from using its data for “tracking, alerting or monitoring sensitive events,” specifically including protests and rallies. In practice, Twitter defines surveillance as the continuing monitoring of specific people and organizations. First Alert is based on public tweets only and doesn’t enable customers to search for or track specific accounts, among other limitations.

Many civil-liberties advocates define social-media surveillance broadly as systematic collection and analysis of information about people, groups and events targeted by authorities. Authorities respond that because the tweets are already public and no individual is under surveillance, there are no valid civil-liberties concerns about Dataminr’s product.

You can read this WSJ article for more information about the Dataminr controversy (you can use your free Cornell subscription).

Use the contextual integrity framework to analyze the flow of information from Twitter to law enforcement authorities via Dataminr.
Problem 6 (8 points)

(a) What does the 4th Amendment say?

(b) Briefly describe the original scope of the 4th Amendment.

(c) Define “reasonable expectation of privacy.”

(d) Briefly describe one of the examples presented in class where technology has challenged the traditional scope of the 4th Amendment and explain how the reasonable expectation test expands the scope of the 4th Amendment so that it covers it.

Problem 7 (8 points)

(a) Why is consent considered an important (perhaps essential) aspect of privacy regulation?

(b) Why is anonymity considered important in privacy regulation?

(c) Why is anonymity important in free societies? (Provide two reasons.)

(d) Briefly explain how the End Run argument exposes a problem:

(i) with treating anonymity as a privacy solution. (Distinguish the end run problem with anonymity with what you’ve learned in this course about challenges with achieving anonymity in the first place.) Hint: “Your name is noise!”
(ii) with treating consent as sufficient for protecting privacy.

Problem 8 (4 points)

(a) What data are protected under GDPR?

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

Problem 9 (6 points)

For the following cases, state which body of law is applicable. Then answer the question: Is this legal? (Yes/No/Depends) Provide a reason for your answer.

(a) Cornell sells student transcripts to a headhunting firm.
   i) Which source of law is applicable
   ii) Is this legal: yes/no/depends (provide a reason)

(b) Coursera sells performance information about you to a headhunting firm.
   i) Which source of law is applicable
   ii) Is this legal: yes/no/depends (provide a reason)

(c) Muslim Pro sells location data to 3rd parties, who sell to the U.S. military.
   i) Which source of law is applicable
   ii) Is this legal: yes/no/depends (provide a reason)
Problem CS1 (21 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 (9 points)

(a) What is longitudinal privacy?
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?

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?

INFO ONLY

INFO1 (8 points)

(a) Compare the GDPR and the CCPA by outlining 3 major differences between the two.

(b) 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 (13 points)

(a) Under the Fourth Amendment, describe what constitutes a search, seizure, and probable cause. [3]

(b) Briefly summarize the landmark information privacy case Katz v. United States. Why did this case overturn the Olmstead v. United States? (Hint: Physical intrusion/property) [4]
(c) Pick another landmark case that deals with technology and describe how it affected the way legal scholars think about the Fourth Amendment. [6]

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.

INFO4 (3 points)

Return to Question 7 and as

7(e) Referring to the Target case, explain how a company can make an “end run” around consent.

INFO5 (4 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