Problem 1: Dataset De-anonymization.

(a) Define the following terms and give an example of each.
- Pseudonym
- Opaque identifier
- 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 the products that Republican voters are likely to buy. How would you go about this? Be specific about 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: Differential Privacy for Data Collection and Release.

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. 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.

Consider the use of randomized response as a data collection technique for a yes/no question (e.g. have you ever used an online dating application?) Each person will flip a coin that has probability $p$ of landing heads. If the coin lands heads, the person will respond truthfully, otherwise the person will respond “yes”.
(b) Given data collected using this protocol from \( n \) participants. If the total number of “yes” responses is \( k \leq n \), estimate the true number of “yes” participants. Your answer should be in terms of \( k, n, p \).

Next, we analyze the differential privacy properties of this protocol. Say the differential privacy goal is to obscure whether any participant (e.g. Alice) answered yes or no (the only two possible answers). One of the things we would want to show is:

\[
P[\text{Alice responds “yes” | Alice’s true answer is “yes”}] \leq e^{\epsilon} \times P[\text{Alice responds “yes” | Alice’s true answer is “no”}]
\]

(c) List the other three inequalities that we would want to show. Hint: they are variations of the inequality listed above.

(d) Solve for \( \epsilon \) in terms of \( p \). What values for \( p \) give the “most” and “least” privacy?

**Problem 3: Legal or no?**

For each of the following parts answer the following two questions:

i. Is this legal in the United States?

ii. To what domain of law must one refer to in order to answer the question? Why? (Answers may be brief - a few sentences at most)

Note: Grading will not depend on getting the answer in (i) correct, but in reasoning for (ii).

(a) FBI demands your web search history from Google.

(b) The IRS provides Donald Trumps 2017 tax records to the New York Times.

(c) Facebook obtains information about its users from Acxiom.

(d) Hackers breach Pokemon Go servers obtaining location and payment history of all players. (Talk about Niantic and Nintendo, not hackers.)

(e) Dictionary.com shares information about users to advertisers.

(f) Police place a GPS tracker on a suspects vehicle and monitor his travel for two weeks.

**Problem 4: Survey Methodology**

Describe at least 2 ways in which the theory of Contextual Integrity shaped the question design in Prof. Martin’s Location Privacy Survey.
Problem 5: How to Spy on Your Employees Without Becoming a Global Security Risk

In class, we talked about cases where a root certificate authority issued a certificate to its corporate clients granting them intermediate certificate authority. This allows the corporate client to self-sign certificates for any websites that they want to snoop on their employee’s traffic. With the self-signed certificate, they can intercept TLS/SSL traffic from their employees, decrypt, then forward along a separate TLS/SSL connection to the intended destination.

(a) Why is this solution not good? Does the corporate client receive power greater than what they require to snoop on employee TLS/SSL connections?

(b) Propose a different solution. Hint: How can the employer sign certificates that are validated by their employee devices without owning an intermediate certificate issued by a root certificate authority?