CS 1657: Privacy in the Electronic Society

Project 3

Released: Monday, Apr 11

Due: Sunday, Apr 24, 11:59 PM

Motivation

In this course, we’ve discussed differential privacy and other techniques for obfuscating/anonymizing data releases. In this project, you will explore the use of such obfuscating and anonymizing techniques in practice.

Your submission should consist of the following components.

- Code that satisfies the C tasks below, in the language of your choice (check with your instructor if you’re not sure it will work for them)
- A writeup that completes the W tasks below **and** discusses your approach to the C tasks, mentioning specific lines, function names, etc. to help me understand your code. Each task should be discussed in your writeup.

All features, bugs, and other details regarding your code should be made clear in your writeup (i.e., do not submit a separate README or expect us to read every comment in your code). Each writing task should be clearly titled, and each code task should be clearly discussed in the writeup. In short, do not make us search for the components of your submission. Show off the hard work you did!

You may work individually or in a pair (a group of 2). If you work in a pair, make sure both members use the GitHub Classroom link to join the repository. (One member should create the team, then the second should join the existing team.)

In preparation for this project, you may want to read some additional references on differential privacy:

- Differential Privacy in Practice
- Video introduction to inverse transform sampling
- The inverse CDF method for sampling from a distribution
- Random sampling in numpy library for Python and Laplace distribution sampling method
    - For instance, you can do the following:

```
from numpy.random import default_rng
rng = default_rng()
sample = rng.laplace() # add params as needed
```

- commons library for sampling from Laplace distribution in Java (Source code)

Tasks

Task W0: Identify a publicly-available dataset containing both quasi-identifiers (e.g., location, age, gender) and potentially sensitive fields (e.g., preferences, medical conditions, interests). You may consider, for instance, the Netflix prize data, available at Academic Torrents or Kaggle. (These sites are also good sources to explore for an alternative dataset!)

To start your writeup, describe the source of your dataset and the content contained therein. You should also provide a link from which the dataset can be downloaded. **Please do not upload datasets to GitHub that are larger than 100 MB.**

Task W1: Describe how the release of this dataset in its current form reveals potentially sensitive information, and how this information could be linked back to the users involved through quasi-identifiers and an adversary's side knowledge.

Task C2: Transform the dataset so that it satisfies k-anonymity (or, at your discretion, something more sophisticated, such as l-diversity or t-closeness), for some value of k. As we discussed in lecture, this should involve clustering based on quasi-identifiers and generalizing those attributes to provide the desired privacy metric. Explain in full how you accomplished the desired property.

Task W3: Compare the original dataset to the output from Task C2. In what ways is less information revealed through this transformation? What information may still be revealed? Is the utility of this transformed dataset decreased relative to the original? Are there fields that certain attackers may know that you did not consider QIs? What would be the impact of this? Give examples as needed to clarify.

Task C4: Implement an algorithm to extract a particular insight from the data. For instance, you can compute something as simple as "the number of users who liked *Titanic*."

Task W5: Describe the algorithm implemented in Task C4 and explain why it does not satisfy differential privacy. In what way might the adversary differentiate between two neighboring datasets D and D' with high confidence, given an output from this algorithm?

Task C6: Implement a variant of your algorithm from Task C4 that yields the same insight in a way that satisfies ε-differential privacy, for some value of ε. As we discussed in lecture, one way to accomplish this is to add Laplacian noise to the value.

Keep in mind that this requires you to reason about the maximum impact a single user can have on the output! If a single user can change the output by up to A, you need to sample from a Laplacian distribution with a scale parameter of $A \times b$ to achieve $\frac{1}{b}$-differential privacy.

Task W7: Interpret the result and outputs of Task C6. In what way does this algorithm respect the privacy of the users represented in the dataset more than the algorithm used in Task C4? Has the utility been impacted, or is the insight equally valuable?

Dataset / Methods Choice

Up to 10 bonus points will be awarded for choosing an interesting and impressive dataset and/or methods for anonymizing. More points will be awarded for submissions that more thoughtful, interesting, creative, clever, etc.

Grading

Task	Points
Task W0	10 (bonus)
Task W1	10
Task C2	20
Task W3	15
Task C4	15
Task W5	10
Task C6	20
Task W7	10
Total	**Up to 110**

Note

As this course is an upper-level elective, you are being given a lot of freedom in terms of how you tackle this project. In exchange, you also have a lot of responsibility to demonstrate your hard work adequately to your TA and instructor. As such, there are tasks in this assignment that require you to discuss your code in detail. Your discussion should closely align to, and refer to, your specific implementation. Do not claim that your code does something that you know it does not (see the Academic Integrity Policy).

In this project, you may use any programming language and library functions that you prefer. Cite all sources. For instance, you may use a data parsing library to read in the dataset and convert it to a format that is easier to manage, or you may use a clustering library to determine which data points are most similar in the QI columns. I recommend against using existing libraries that are built to accomplish k-anonymity or differential privacy (unless your project has sufficient technical depth aside from these components). This is a 2-week project in an upper-level technical course; if you satisfy the requirements in only 4 lines of code, you're probably using a library function that allows you to skip the hardest (and most interesting!) parts of the assignment. If you need help deciding whether something should be permitted, contact your instructor.

Submission

Your writeup should be in PDF format. Submit it, and your code, by committing and pushing to your GitHub Classroom repository by the date and time listed above. Late submissions will not be accepted.