CS 1699: Privacy in the Electronic Society

Project 3

Released: Tuesday, April 12

Due: Sunday, April 22, 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.

- A writeup that completes the W tasks below.
- Code that satisfies the C tasks below, in the language of your choice. Note that you **are allowed** to use external libraries in this project.

All features, bugs, and other details with your code should be made clear in your writeup (i.e., do not submit a separate README or expect your TA 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 your TA search for the components of your submission. Show off the hard work you did!

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 Box 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 advanced, 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. You are allowed to use published code such as arx for this, as long as you cite your sources.

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? Give examples as needed to clarify.

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

Task W5: Discuss why the insight from Task C4 does not satisfy differential privacy. In what way might the adversary differentiate between two neighboring datasets D and D' with high confidence?

Task C6: Implement a variant of your program from Task C4 so that it gives the 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*b* to achieve *1/b*-differential privacy.

Task W7: Interpret the result of Task C6. In what way does this method protect the users represented in the dataset more than the method used in Task C4?

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**

Submission

Upload your code and writeup to the Box folder created for you with the name `cs1699-p3-abc123`, where `abc123` is your Pitt username.