Privacy Class Exercise

In this exercise, you will work in groups to disclosure-proof a dataset as much as possible while retaining as much of the value of the dataset as possible. There will be three steps to this exercise, which will be done in breakout groups with your groups. Before we start, you should have already designated one person in your group as the representative and provide me with an email address to send the data to. I have emailed the dataset as a CSV file as well as an accompanying short data dictionary. If you have not done this, please do this as soon as possible.

Step 1: Data exploration

You will be given a synthetic dataset containing information about people, including names, date of birth, phone numbers, and responses to survey questions. The dataset you receive will be a random half of an overall dataset. Other groups will have a different random half of the overall dataset. That is, you should expect about half of your dataset to overlap with any other group's dataset.

First, formulate a research question. Choose something simple enough to be done easily, but substantive enough to be interesting. Some example questions might be: "Is there an association between income and response to question A?" or "How does response to question B differ by age or race?" Using the data that you have, answer this question. You don't need to be overly thorough in this part, but try to do some basic steps such as making a quick graph or running a hypothesis test. Make sure you keep track of your research question, as you will need to try to answer this question in step 3.

Step 2: Disclosure-proofing your data

Your goal in this step is to create a dataset that removes all identifying information so that an individual cannot be re-identified. You will want to remove anything that can uniquely identify a person, such as social security number, but keep in mind that combinations of variables could also be used to identify individuals. The catch in this step is that you will need to make sure that the data is still useful and retains as much of its utility as possible. In the next step, another group will take your data and try to answer a research question. Your goal is to try to make sure they can answer that research question while making sure that they cannot re-identify anyone in the dataset.

You can do any number of operations to the data, including removing variables, transforming variables, adding noise, and so on. Remember, though, that any operation you do will remove some of the utility of the dataset. Keep track of the changes you make and provide a "data documentation" sheet with any important information about what was done to the dataset so that another group can use it for their analysis.

When you have created your new dataset with the updates/alterations, export it as a CSV file. Separately, export the IDs as a different CSV file. I will go around each of the breakout groups at this step and give you the email of the person you will need to send these files to.

Step 3: Swap data

You will then swap datasets with another group. Make sure you send the CSV of the disclosure-proofed dataset, the IDs, and the data documentation. Using your new dataset, try to answer the question you formulated in step 1. The data should come with information about what was done to it. Answer the following questions:

- Are you able to answer your question?
- Do you have doubts about the accuracy or generalizability of your results?
- What are steps that the data provider (in this case, the group that gave you this dataset) took that might have hindered your ability to answer your question?

Next, try to identify someone in the dataset you received. Recall that your dataset and the dataset you received should have about half of each overlap. Using the information in both datasets, try to see if you can recover the identify of a person in the dataset you received. When you are done, check the list of IDs that the group gave you to see if you were correct. You can also see if you can narrow it down to just a few possible IDs.

Class Activity Discussion

We will then reconvene as a group and discuss the results of this exercise. Think about the answers to these questions for this discussion:

- How did you balance the research and privacy needs for the dataset?
- What are some mistakes you might have made? Was the dataset secure enough? Do you think it kept its utility?
- What are steps you would take now if you could go back to better adjust the dataset?