Assignment 3: Data Privacy

In this assignment we will work on how to identify individuals from improperly anonymized datasets, and how to properly anonymize data using the Safe Harbour method. Note that the datasets provided are synthetic data.

Tasks / Learning Goals

- Explore how individuals can be identified from improperly anonymized datasets using small amount of external data, and simple matching procedures
- Gain experience with properly anonymizing datasets using the Safe Harbour method

Due Date

11:59 pm Monday, May 8th, submitted on TritonED.

Submitting Assignments

You will submit a Jupyter notebook file (.ipynb) to TritonED. Make sure that the file you submit has the following filename (filled in with your student ID number):

'A3_A#########.ipynb'

Grading Rubric

There are 2 parts to this assignment, with the following point values:

Part 1: Identifying Individuals	5 points
Part 2: Anonymizing Data	5 points

Detailed Instructions

This is a brief outline of the tasks in the assignment. The full, detailed instructions are included in the Jupyter notebook that you will use for this assignment. Note that all the data in this assignment is synthetic.

Part 1: Identifying Data

In this part of the assignment we will explore how to identify individuals from improperly de-identified datasets.

Data:

- A (badly) de-identified data set of users for some company.
    - anon_user_dat.json
- Publicly scrapable information on people who work in relevant companies
    - employee_info.json

Tasks:

- Using e-mails that are left in the dataset, we will try to infer information from these about where people work.
- We could use this information to scrape information of people who work at those companies (for the assignment, this information is given to you in 'employee_info')
- We can then match this employee information with data in our original dataset to de-identify people.

Part 2: Anonymizing Data

After we have explored how sloppily anonymized datasets can be de-anonymized, we will now explore how to properly anonymize datasets using the Safe Harbour method. The relevant details for applying Safe Harbour are covered in the lecture slides and/or in the assignment notebook, but if you want/need more information, a full overview is available here:
https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/

Data:

- The original user data, including identifiable information
    - user_dat.json

Tasks:

- Drop columns of data that contain identifiable information
- Drop any individuals in the data set that do not meet Safe Harbour inclusion criteria
- Recode zip codes, following Safe Harbour standards