Nick Carney, Max Ivry
Professor Near
Dec 11 2023
Data Privacy
Project repo: https://github.com/NickCarney/3110-data-privacy-final-project

Caught Red Handed; an Educational Practice in Re-identification and Differentially Private Analysis of Los Angeles Crime Data (2010-2020)

Problem Statement:

The primary goal of this project is to demonstrate the susceptibility of Los Angeles crime Data to linking attacks, matching the names and information of criminals to their crime using the auxiliary dataset of the Condemned Inmate List of the California Department of Corrections and Rehabilitation. Additionally, a differentially private re-analysis of the dataset using the Laplace Distribution on several features shows a potential need for more deidentification measures. It is important to note that, while some re-identification was achieved, all information used in the project is publicly available, and this project only serves as an educational practice with no mal intent.

About the datasets:

The Los Angeles Crime Data (2010-2020) S is released by the Los Angeles Police department and is licensed under public Domain (CC0). There are 2117589 rows, each corresponding to one crime that contains 28 attributes. Some notable and potentially identifying attributes include crime location information (street, area name, district number etc.) date occurred, and date reported. The only mention of privacy in the dataset description is as follows: "Address fields are only provided to the nearest hundred block in order to maintain privacy". However, the dataset does not include the name of the offender and it is unclear what measure of privacy, other than address per crime, is maintained.

The auxiliary data A is the current Condemned Inmate List released by the California Department of Corrections and Rehabilitation, last updated December 6, 2023. There are 652 rows, which each correspond to one inmate with 8 attributes; attributes of importance include name (first and last), and reported date and offense date as quasi-identifiers. There is no statement on privacy in this data, so it is assumed all information is without noise.

Links to both datasets can be found in the References section.

Methodology:

The approach to re-identification implemented in the project was largely based on Chapter 1 of Programming Differential Privacy (Near et al.). First, the auxiliary data

was collected using the python package Beautiful Soup for web-scraping, appending the data to a plain text file "inmates.txt". Then, both datasets were cleaned so that the date reported of S could be compared with date reported of A. Iterating through A, if date reported of S matched date reported of A and date occurrence of A matched date sentenced of S, then the criminal information is appended to a list of potential matches. If the list is of size 1, then we have re-identified the criminal with the crime, displaying the name of criminal from A and type of crime from S. Through this approach, 9 people were matched to their crimes. However, many of the crimes did not seem worthy of capital punishment, which may suggest fault in our methodology, most likely in comparison between date occurrence of A and date sentenced of S..

Reattempting with a differentially private dataset:

In this section, noise was added to the identifying date columns to ensure epsilon differential privacy, and then the linking attack was re-attempted. The first approach is to generalize the day and year of each date. This was done by dividing the day and year of each rows identifying column by 10, casting to an integer, and then re-multiplying the day and year by 10. While this approach creates more matches, the dates are significantly changed, harming the accuracy but ensuring security.

The second approach uses the Laplace mechanism to add noise to each day in the identifying columns, using a total epsilon = 1 and sensitivity of 1 per call. This will ensure epsilon differential privacy via sequential composition, with epsilon = 2117589 for the total data set.

Results:

Results of linking were able to happen between the two datasets. This algorithm looked for ‘exact-matches’(links with a unique row matched) and ‘half-matches’ (links with two matching rows). With no differential privacy, it found 9 exact-matches and 10 half-matches out of the 131 potential condemned inmates whose crimes were reported between 2010 and 2019. These results could make one skeptical of their validity because many of the linked crimes are things such as identity theft and petty misdemeanors, and likely wouldn’t lead to being a condemned inmate. When adding differential privacy, the functions that added laplace noise to the day and the year values in each date led to finding 31 exact matches and 13 half matches (~33% of the data!). This could be happening due to a bit of luck caused by the number of total matches for each crime being decreased, increasing the number of exact-matches and half-matches. When adding differential privacy in the form of generalization, there were no exact-matches or half-matches found due to the day and year (individually) being too generic (only one digit) and leading to there being a lot of potential matches per inmate, but no exact or half-matches.

References:

Condemned Inmate List:
https://www.cdcr.ca.gov/capital-punishment/condemned-inmate-list-secure-request/

Los Angeles Crime Data 2010- 2020:
https://www.kaggle.com/datasets/sumaiaparveenshupti/los-angeles-crime-data-201020
20/data.

Near, Joseph P., and Chiké Abuah. Programming Differential Privacy. vol. 1, 2021.
programming-
dp.com.