CS211- Data Privacy : Final Project

Nikhil (nchoppa), James (jcastner)

Adding Differential Privacy to a Credit Card Clients Database from Taiwan in 2005

Goals

This project had a few simple goals:

- Identify a dataset with potential personally identifiable information (PII).
- Determine which statistics from this dataset would be useful.
- Create a system to generate differentially private statistics from this dataset (using epsilon as a reasonable privacy budget).
- Automatically generate a final PDF report when finished.

Dataset

The dataset we used for this project comes from Kaggle. The dataset contains information on default payments, demographic factors, credit data, payment history, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. While the data itself does not contain simple PII such as names, dates of birth, or SSN, it does contain information such as:

- Customer Age
- Customer Gender
- Education Level
- Marital Status
- Limit Balance
- Bill Amount
- Pay Amount
- **All things that could be used to re-identify individuals!**

Statistics

We picked a few basic statistics that we thought an analyst might be interested in. They are:

- Counts of most common education level
- Average credit card clients age
- Average limit balance of a credit card clients accounts
- Average monthly bill amount
- Average monthly payment amount
- Average monthly bill amount for credit clients in higher education (EDUCATION = 1 or 2)
- Contrast average monthly payments between sexes
- Counts of an individual’s marital status in comparison to if they default

Privacy Strategy

For all bill average statistics, we used the *Sparse Vector Technique* to determine an upper clipping parameter. We then generated a noisy sums and counts using the Laplace mechanism (and a portion of the total epsilon alloted for this query). Then we divided the noisy sum by the noisy count to get a differentially post-processing private average.

Report Generation

Finally, we calculated basic averages, error percentages and conditional averages for average queries and basic counts and conditional counts for the count queries. This data is then generated automatically into a .tex file and compiled into a .pdf file.

Mail to:

nchoppa@uvm.edu

jcastner@uvm.edu

Thank you!!!
...