Privacy Aware Computations
May 5, 2021

Differential Privacy Temporal Map Challenge: Sprint 3

Title: Histogram Sparsity Reduction Alongside Histogram Combination

Main Idea:

We plan to build histograms from the generalized permutation of columns and augment this main histogram with information from the "publicly available data" provided as training data and secondary histograms to generate synthetic data.

Pre-processing:

Publicly available data:
We create histograms from the training ground truth data identical to ones we create from the private dataset. The goal is to to sample from these bins corresponding to bin counts its private histogram counter-part.
We had planned to use other information to adjust the samples taken above, however, due to lack of time, those plans have been abandoned.

Private data:
We group the data by taxi ID, then create a histogram from the bellow queries:

1: Is the average trip in a day made by this taxi: 5 or less trips, 5-11trips, 11-17 trips, 17 or more
2: What is the standard deviation of trips made daily from tthe above average? bins curated coresponding to the average
3: A special query designed to classify driver type:
  a) Is 50% of the trips made from the same pick up zone?
  b) Is 30% of the trips or more made from the same pick up zone?
  c) Are there 2 pickup zones with 25% of the trips each?

Some bins are not included to improve specificity, such as the bin corresponding to 3a-true and 3b-false because this conndition is impossible.

(source code: function create_taxi_stats(rows, taxi_histogram))

Privatization and Privacy Proof:

We use the Gaussian method to privatize our histogram. Each driver can only fall into one bin because of the disjoint nature of the queries, which yields a sensitivity of precisely one.

Post Processing:

We match the bins from the public histogram to the ones from the private one. For each bin, we sample as many driver weekly trip data from the public histogram as the count in the private one. To avoid picking the same driver twice, we take on a strategy where with a probablity of 80% (dropout rate in the code) we re-sample in the case of a duplicate, and include the duplicate as another entry with probabilty of 20%.

References:

1. G. Cormode, T. Kulkarni, and D. Srivastava, “Marginal Release Under Local Differential Privacy,” *Proceedings of the 2018 International Conference on Management of Data*, 2018.