ANALYSIS OF DIFFERENTIALLY PRIVATE NEW YORK CITY TAXI DATASET GENERATED USING PRIVBAYES

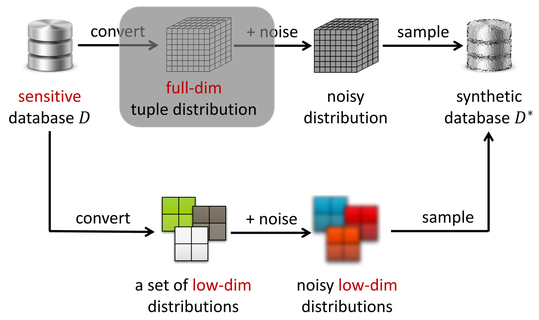
Yanping Zhang, MS[1]; Lai Wei, MS[2]; Naman Jain, MS[2]
[1]ECE Department, Duke University [2]CS Department, Duke University

Motivation

The NYC & Limousine Commission releases the collected yellow taxi trip records on its website every month which can be used to conduct various analysis and research. However, the trip records include sensitive attributes such as pick-up and drop-off dates/times, trip distances, passenger counts etc. Directly releasing this dataset may leak the privacy information for individuals. Therefore it is important to find a differentially private algorithm to release this dataset.

Problem Statement

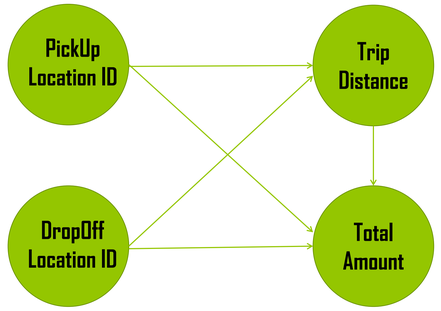
The NYC Taxi dataset is high-dimensional, which contains 17 attributes and six million records. When releasing high-dimensional dataset, we confront two challenges, including Output Scalability and Signal-to-Noise Ratio. So, we use PrivBayes to release the NYC Taxi dataset. PrivBayes approximates the high-dimensional distribution of the original data with a data-dependent set of well-chosen low-dimensional distributions, which effectively address the above two challenges.



Dimension Reduction in PrivBayes Algorithm

Methods and Materials

STEP 1: Construct a Bayesian network $\mathcal{N}$
The illustration below is a simplified instance of Bayesian Network constructed using four out of the total 17 attributes.
(Differentially private technique is used)



$$\Pr[*] \approx \Pr[Pickup\ LocationID] \cdot \Pr[Dropoff\ LocationID] \cdot \Pr[Trip\ Distance|Pickup\ LocationID, Dropoff\ LocationID] \Pr[Total\ Fare\ Amount|Pickup\ LocationID, Dropoff\ LocationID, TripDistance] \cdot$$

STEP 2: Compute noisy conditional distributions
- Materialize the joint distribution of Attribute-parent Pairs $\Pr[\mathbf{X}_i, \prod_i]$
- Inject Laplace noise into $\Pr[\mathbf{X}_i, \prod_i]$ to obtain a noisy distribution $Pr^*[\mathbf{X}_i, \prod_i]$
- Derive a noisy conditional distribution $Pr^*[\mathbf{X}_i | \prod_i]$

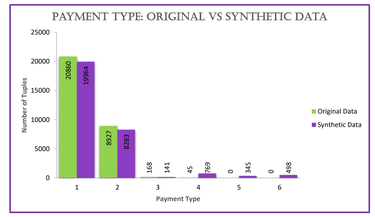
(inject noise – Laplace mechanism)

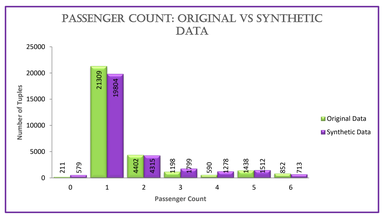
STEP 3: Generate synthetic data
- Use the Bayesian network(step2) and noisy conditional distributions(step3) to derive an approximate distribution of the tuples in D
- Sample tuples form the approximate distribution to generate a synthetic dataset

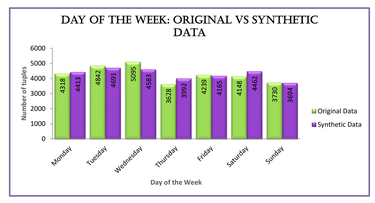
(post-processing: no privacy issues)

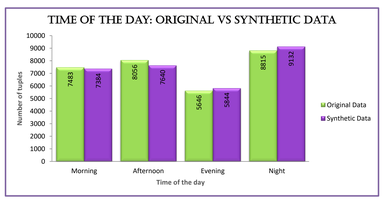
Results

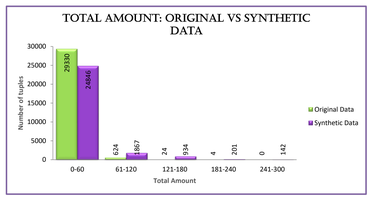
As per the problem definition, we did release the NYC yellow taxi dataset (for January 2018). As mentioned, we think that the pickup/dropoff time for any customer should not be known directly as it is sensitive data. So we substituted those columns for days of the week and time of the day columns. For the released dataset, we used epsilon = 0.8. We used epsilon = {0.2, 0.4, 0.6, 0.8} to calculate 1-way and 2-way marginals. Following are the graphs that show the frequency for 5 attributes that hold significant importance. We achieve great accuracy for most of these attributes.

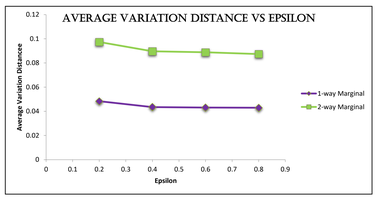












Future Work

One of the possible future improvements for us is to customize the bayesian network. We can delete the edges that are not important to our queries. We can then compare the accuracy using a set of range queries and select the best Bayesian network that is suitable for us. We could also help do some analysis based on the generated synthetic data in Todd W. Schneider site which is a popular and interesting data analysis website.

Reference

JUN ZHANG, GRAHAM CORMODE, CECILIA M. PROCOPIUC, DIVESH SRIVASTAVA, XIAOKUI XIAO. 2017. PrivBayes: Private Data Release via Bayesian networks (Both paper and Slides)