Survey of Differentially Private Accuracy Improving Techniques for Publishing Histograms and Synthetic Data

Jenni Reuben[1] and Boel Nelson[2]

[1] Karlstad University, Karlstad, Sweden
jenni.reuben@kau.se
[2] Chalmers University of Technology, Gothenburg, Sweden
boeln@chalmers.se

Poster Abstract

Drawing insights from data sets provide enormous social value. However, privacy violations are major impediments to access these data sources. For example, messages in social platforms is a valuable source for a social science researcher who wants to understand how individuals fleeing from wars organize themselves but access to such information jeopardize the privacy of those individuals. Differential Privacy (DP) is a privacy model that provide formal guarantees to the individuals that their participation in the data set is 'nearly' hidden. Differential privacy definition states that the results of an analysis on a data set remain 'essentially' the same whether or not an individual participates/does not participates in the data set. To this end, some differentially private analyses add noise to the output for obfuscating the contribution of any individual. The magnitude of the added noise is inversely proportional to the privacy guarantee.

In this work we focus on the analyses that produce histograms or synthetic data. Histograms and synthetic data are interesting to study because they provide a sanitized form of the original data for which access is restricted. There is growing interest in the scientific community to demonstrate different ways to improve the accuracy of differentially private histograms or synthetic data. However, there are a few work that systematize the knowledge gained by those scientific investigations. Thus our aim is to analyze and structure the state-of-art techniques that improve the accuracy of histograms or synthetic data published under differential privacy. We used the systematic literature review as the research method to summarize the state-of-the-art. Our preliminary result that categorize the state-of-the art is illustrated in Figure 1.

Fig. 1. Three approaches to improve the accuracy of differentially private histograms or synthetic data