Differentially private data release by optimal compression
J. Köhler

Abstract

Differential privacy (DP) is the mathematical guarantee that the output of an algorithm operating on a data set has only a small dependency on any single individual in that set. A particular application of DP is the task of releasing privatized representations of data sets which should not be leaked to the public in their original form. Finding the right transformation of a data set, such that it is provably privatized but still preserves utility for the analysis task at hand, is a difficult task in theory and also practically unsolved for many applications. During this thesis we present an information theoretic approach to design DP data set release mechanisms, by reducing the problem to optimally compressing the data with respect to a measure of utility. As the optimal compression problem is inherently difficult to solve by itself, we analyze this approach for two linear instances of optimal compression for which an analytic solution exists and thus the analysis and sampling of privatized data sets is tractable. We further show in experiments that both methods cannot yield privacy/utility trade-offs that allow them to be used in practical tasks. We finish this research by proposing more complex approaches following this framework that are based on approximations and sampling methods and discuss their strengths and weaknesses.