Speaker: Robin Mitra (University College London)

Date: 13:15 on 02/11/2022 in 8 West 2.13

Title: Saturated count models for user-friendly synthesis of categorical data

Abstract:

Synthetic data methods are being increasingly used to protect data confidentiality. Large sparse categorical data sets pose some significant challenges for synthesis which makes many traditional methods unsuitable. We explore using saturated count models for synthesis. These are appealing as they allow large categorical data sets to be synthesized quickly and conveniently, as well as permitting risk and utility metrics to be satisfied a priori, that is, prior to synthetic data generation. Most well-known count models for synthesizing categorical data at the tabular level tend to utilise either Poisson or Poisson-mixture distributions. However, the latter are always over-dispersed, with a variance that is an increasing function of the mean. As a result, relatively more noise is applied to larger counts than smaller counts. But this is contrary to the objective of data synthesis, where larger counts are typically lower risk than smaller counts, and therefore require less perturbation. We thus additionally explore the benefits of using the discretized gamma family distribution (DGAF) for synthesis within the saturated model framework. The DGAF provides the synthesizer with control of the variance-mean relationship, allowing smaller counts to be over-dispersed and larger counts to be under-dispersed, which in turn produces synthetic data with greater utility. The benefits of the DGAF are illustrated empirically using a database which can be viewed as a good substitute to the English School Census.