Department of Decision Sciences

Statistics Seminar

Methods for Quantifying Conflict Casualties in Syria

Rebecca Steorts
Duke University

Thursday, 23rd May 2016
12:30pm Room 3-E4-SR03 Via Rontgen 1 Milano

Abstract

Information about social entities is often spread across multiple large databases, each degraded by noise, and without unique identifiers shared across databases. Record linkage—reconstructing the actual entities and their attributes—is essential to using big data and is challenging not only for inference but also for computation. In this talk, I motivate record linkage by the current conflict in Syria. It has been tremendously well documented, however, we still do not know how many people have been killed from conflict-related violence. We describe a novel approach towards estimating death counts in Syria and challenges that are unique to this database. We first introduce a novel approach to record linkage using a new Bayesian nonparametric property (BNP)—microclustering, and then a model that assumes this property. Many BNP methods assume the size of each latent cluster grows linearly with the number of total data points (records). However, this is not the case for such tasks as record linkage, and specifically with the Syrian conflict. Micro clustering, assumes instead a sublinear growth, or rather that the the size of each latent cluster is grows negligibly compared to the total number of data points. Advantages of our construction are that our model quantifies the uncertainty in the inference and propagates any linkage uncertainty into subsequent analyses. We then introduce computational speed-ups to avoid all-to-all record comparisons based upon locality-sensitive hashing from the computer science literature. Finally, we speak to the success and challenges of solving a problem that is at the forefront of national headlines and news.

Department of Decision Sciences

Via Röntgen 1 - 20136Milano

Tel. 02 5836.5632
Fax 02 5836.5630