Mauricio Sadinle (Carnegie Mellon University)

An Overview of Record Linkage

Record linkage techniques allow us to combine different sources of information from a common population in the absence of unique identifiers. Linking multiple files is an important task in a wide variety of applications, since it permits to gather information that would not be otherwise available, or that would be too expensive to collect. In practice, an additional complication appears when the datafiles to be linked contain duplicates.

In this talk we review the widely used Fellegi-Sunter framework for record linkage and its unsupervised implementations that aim to cluster record pairs into matches and non-matches. We also mention alternative supervised approaches to record linkage and duplicate detection that train classifiers on pairs of records with known matching status, and predict the matching status of the remaining record pairs. We explain the difficulties that these approaches pose in terms resolving non-transitive decisions and incorporating matching uncertainty into subsequent analyses of the linked files. We mention some recent advances that aim to resolve these problems and conclude with an overview of opportunities for future research.