January 16, 2014
Rebecca Nugent, Carnegie-Mellon University
Sidney Smith Hall, Room 1074
3:30pm *Refreshments served at 3:10pm

Solving the Identity Crisis: Large-Scale Clustering with Distributions of Distances with Applications in Record Linkage

"Will the real Rebecca Nugent please stand up?" Deduplication is the process of linking records corresponding to unique entities within a single database, where each unique entity may be duplicated multiple times. Most often these records are largely text, but can also contain continuous or categorical fields (e.g. last name vs age). We frame deduplication as a clustering problem, where each observed record belongs to a cluster corresponding to some latent unique entity in the underlying population. In large-scale deduplication problems (e.g. over 300 million records in the US Census), calculating and analyzing the necessary pairwise comparisons is computationally infeasible. Instead, we adopt a divide-and-conquer strategy and re-frame our problem as clustering records given several estimates of their similarity (or inter-record distance). More specifically, we define the distance between record-pairs to be a monotonically decreasing transformation of their pairwise matching probabilities, which are obtained via a supervised learning approach (given some set of match/non-match labels). We build an ensemble of classifiers and estimate a distribution of pairwise matching probabilities for each pair of records. We then cluster the records by mapping features of these distributions to the best approximation of the true distance between records. With respect to the choice of clustering method, hierarchical clustering can be applied to resolve pairwise transitivity violations, but most deduplication problems have too many records to reasonably use apply it here. If time permits, we discuss an additional blocking scheme, sequential in nature, to help reduce the number of comparisons needed. In general, these clustering approaches can be used when we have very large datasets and/or unavailable or uncertain distances between observations. We show results from the identification of unique inventors in the United States Patent and Trademark Office patent-inventor database.