Bayesian inference for linked and duplicated data

We review advances on general Bayesian methods for performing record linkage and deduplication across multiple databases and making inference using the resulting matched records. In particular, we cluster records to hypothesize latent entities using a formal hierarchical Bayesian model. Our novelty is that model comprises both the matching variables and the other features included for inferential aims. The pattern of links between records which are directly linked to latent true individuals is represented as a bipartite graph, in which records are directly linked to latent true individuals, and only indirectly linked to other records. Using such an approach, we can account for the matching uncertainty in inferential procedures, (e.g., multiple linear regression, capture-recapture) based on probabilistically linked data, and at the same time, we can generate a feed-back propagation of the information between the working statistical model and the record linkage process. We illustrate our methods on simulated and real data.