ABSTRACT
In database management, record linkage aims to identify multiple records that correspond to the same individual. This task can be treated as a clustering problem, in which a latent entity is associated with one or more noisy database records. However, in contrast to traditional clustering applications, a large number of clusters with a few observations per cluster is expected in this context. In this work, we introduce a new class of prior distributions based on allelic partitions that is specially suited for the small cluster setting of record linkage. Our approach makes it straightforward to introduce prior information about the cluster size distribution at different scales, and naturally enforces sublinear growth of the maximum cluster size – known as the microclustering property. We evaluate the performance of our proposed class of priors using official statistics data sets and show that our models provide competitive results compared to state-of-the-art microclustering models in the record linkage literature.