Top-Down Recursive Spectral Clustering

For Heterogeneous and Multiscale Networks

Abstract: Real-world networks are often hierarchical, heterogeneous, and multi-scaled, while the idealized stochastic block models that are extensively studied in the literature tend to be over-simplified. In a line of work, we propose several top-down recursive partitioning algorithms which start with the entire network and divide the nodes into two communities by certain spectral clustering methods repeatedly, until a stopping rule indicates no further community structures. For these algorithms, the number of communities does not need to be known a priori or estimated consistently. On a broad class of hierarchical network models motivated by Clauset, Moore and Newman (2008), in which the communities are allowed to be heterogeneous and multi-scaled in terms of the size and link probabilities, our algorithms are proved to achieve the exact recovery for sparse networks with expected node degrees logarithmic in the network size, and are computationally more efficient than non-hierarchical spectral clustering algorithms. More interestingly, we identify regimes where no algorithm can recover all communities simultaneously while our algorithm can still recover the mega-communities (unions of communities defined by the hierarchy) consistently without recovering the finest structure. Our theoretical results are based on my newly developed two-to-infinity eigenspace perturbation theory for binary random matrices with independent or dependent entries.