Colloquium
Michigan State University
Department of Statistics and Probability

Jiashun Jin
Department of Statistics
Carnegie Mellon University

Fast Network Community Detection by SCORE

Tuesday, March 26, 2013
10:20 a.m. - 11:10 am
Refreshments 10:00 am
C405 Wells Hall

Abstract:

Consider a network where the nodes split into K different communities. The community labels for the nodes are unknown and it is of major interest to estimate them (i.e., community detection). *Degree Corrected Block Model* (DCBM) is a popular network model. How to detect communities with the DCBM is an interesting problem, where the main challenge lies in the degree heterogeneity.

We propose Spectral Clustering On Ratios-of-Eigenvectors (SCORE) as a new approach to community detection. Compared to classical spectral methods, the main innovation is to use the entry-wise ratios between the first leading eigenvector and each of the other leading eigenvectors. Let X be the adjacency matrix of the network. We first obtain the K leading eigenvectors, say, $\hat{\eta}_1, \ldots, \hat{\eta}_K$, and let $\hat{R}$ be the $n \times (K-1)$ matrix such that $\hat{R}(i,k) = \hat{\eta}_{k+1}(i)/\hat{\eta}_1(i)$, $1 \le i \le n$, $1 \le k \le K-1$. We then use $\hat{R}$ for clustering by applying the k-means method.

The central surprise is, the effect of degree heterogeneity is largely ancillary, and can be effectively removed by taking entry-wise ratios between $\hat{\eta}_{k+1}$ and $\hat{\eta}_1$, $1 \le k \le K-1$.

The method is successfully applied to the web blogs data and the karate club data, with error rates of 58/1222 and 1/34, respectively. These results are much more satisfactory than that by the classical spectral methods. Also, compared to modularity methods, SCORE is computationally much faster and has smaller error rates.

We develop a theoretic framework where we show that under mild conditions, the SCORE stably yields successful community detection. In the core of the analysis is the recent development on Random Matrix Theory (RMT), where the matrix-form Bernstein inequality is especially helpful.

To request an interpreter or other accommodations for people with disabilities, please call the Department of Statistics and Probability at 517-355-9589.