STATISTICS SEMINAR

Shu Zhou

Master's Defense

Monday, July 20, 2015

Dickens Hall 207, 2:00 pm

Exploring Network Models Under Sampling

Networks are defined as sets of items and their connections. Interconnected items are represented by mathematical abstractions called vertices (or nodes), and the links connecting pairs of vertices are known as edges. Networks are easily seen in everyday life: a network of friends, the Internet, company or organization networks, metabolic networks, food webs, citation networks. With the growing popularity of social networks, it has also become an attractive way to reach target populations. The increase of available data and the need to analyze it has resulted in the proliferation of models for network data. However, for networks with billions of nodes and edges, computation and inference might not be achieved within a reasonable amount of time or budget. A sampling approach seems a natural choice, but traditional models assume that we can have access to the entire network. Moreover, when data is only available for a sampled sub-network conclusions tend to be extrapolated to the whole network/population without regard to sampling error.

The statistical problem this report addresses is the issue of how to sample a sub-network and then draw conclusions about the whole network. Are some sampling techniques better than others? Are there more efficient ways to estimate parameters of interest? In which way can we measure how effectively my method is reproducing the original network? W e explore these questions with a simulation study on Mesa High School students' friendship network. First, to assess the characteristics of the whole network, we applied the traditional exponential random graph model (ERGM) and a stochastic blockmodel to the complete population of 205 students. Then, we drew simple random samples and stratified samples of 41 students out of the 205, applied the traditional ERGM and the stochastic blockmodel again, and defined a way to generalized the sample findings to the population friendship network of 205 students. Finally, we used the degree distribution and other network statistics to compare the true friendship network with the projected one.

We achieved the following important results, 1) as expected stratified sampling outper- forms simple random sampling when selecting nodes; 2) ERGM without restrictions offers a poor estimate for most of the tested parameters; and 3) the Bayesian stochastic blockmodel estimation using a stratified sample of nodes achieves the best results.