Statistical Analysis of Networks
Statistics 218
Professor: Mark S. Handcock

Homework 1

Due Thursday, October 4, 2018

1) *Friendship:* The first question considers some descriptives for networks. The following list of 8 people were asked to name all of their friends from this list of 8. The results are shown below.

Respondent	Friend 1	Friend 2	Friend 3	Friend 4
Jacob	Michael	Emily		
Emily	Madison	Abigail	Jacob	
Michael	Joshua	Jacob	Matthew	Emily
Emma	Emily			
Joshua	Jacob	Michael		
Madison	Emily	Abigail		
Matthew	Jacob	Michael		
Abigail	Madison	Emily		

Figure 1: Adjacency matrix

a) Please answer this simple questions about the network:

(i) Is this a directed or undirected network?

(ii) How many possible ties are there in this network?

(iii) How many actual ties are there in this network?

(iv) What is the density (proportion) of ties in this network?

b) Draw the sociogram (plot) of this network of friends. You can do this with a software package or by hand.

c) Complete a simple EDA of the in-degree and out-degree distributions. Do they appear to be the same?

d) Construct a "mixing matrix" by gender. The entries in the mixing matrix are the counts of the number of edges between people with the row category to the people in the column category (See Figure 2 below).

2) *Degree distributions:* Degree distributions summarize the densities of ties of the population of nodes. In this question we explore the interactions between proteins of the yeast *S. cerevisiae*. The nodes are types of proteins in the yeast and a directed tie is said to exist if a protein binds to the target protein in a i"wet lab" experiment set up to test just this. Not all protein combinations are tested. Here we will consider a series of "mapping" experiments conducted in 2008 that covered approximately 20% of all yeast binary interactions (Yu et. al Science (2008))

a) Go to the home page of the "Yeast Interactome Project":
`http://interactome.dfci.harvard.edu/S_cerevisiae/`
From there download the interactions from CCSB-YI11. These comprise 1809 interactions among 1278 proteins. Construct a `network` object from this edge-list.

b) Construct the out-degree sequence for the network. Construct the in-degree sequence. Are the in- and out-degrees correlated? Use the sum of the in-degree and out-degree as an overall measure of the proteins activity (which we will refer to as its degree).

c) Fit degree distribution models using the `degreenet` package. Fit the classical discrete Pareto/Zipf law model discussed in the books:

$$P(K = k; \nu) = \frac{k^{\nu}}{\zeta(\nu)} \qquad \nu \geq 1, k = 1, 2, \ldots$$

Fit the Yule, Waring, Poisson, and Conway-Maxwell-Poisson models. Note: The corresponding functions are `ayulemle`, `adpmle`, `ayulemle`, `awarmle`, `apoimle`, `acmpmle`. See the help pages.

d) How can we compare the fits of the different models? Use the `lldpall()`,etc, functions to compute the corrected AIC and the BIC for the models. Summarize the fits of the models in a table. Which models fit best? Briefly comment on the models as a whole in terms of their fit.

3) *Components:* Continuing with the CCSB-YI11 network: Consider undirected the version of it where two proteins are linked if either of them binds with the other. This network has 1809 edges.

Gender	Male	Female	Total
Male			
Female			
Total			18

Figure 2: Outline of a "mixing matrix" by gender.

a) Compute the component distribution of the network. How large is the largest component? Does the network have a "giant" component? Note: Consider the `component.dist` function in the `sna` package.

b) Plot the subgraph comprised of the largest component.

c) Compute the (pairwise) matrix of geodesic distances between the proteins. Create a summary tabulation of the distances. What proportion of nodes-pairs are reachable (from each other)? What is the mean geodesic distance for reachable pairs? How many isolates are there in the network?