QBS 122 Spring 2021

Social Network Project

Due: June 4, 11:59pm

The over-arching goal of the project is to implement the techniques you've seen employed during lectures 2 through 5 of the social network module as well as some review of the material in Homework 7. It also involves identifying a network to use and transporting the data for that network into R.

After finding the network you want to analyze, please involve the methods discussed in class so as to demonstrate your mastery of them. This project will count for 15% of your course score, from which grades will be determined. You can make use of the scripts I've used in class and the analyses that I've done.

Please submit a written report the describes your network, the particular questions of interest to you, the methods you are using and why they are appropriate, the results and a short conclusion. The Appendix should include the script(s) you have used for performing your analyses.

Please state the source of your network and related data and acknowledge the persons responsible for creating the network. While you may discuss your project with others in the class (e.g., to seek general advice on how to perform a certain data operation or analysis) the project is to be your own work. If you do obtain a helpful idea from one of your classmates, it is important to acknowledge this. Carly, Josh, and I will continue our office hours through to June 15 so will be available the week that the project is due.

The deadline to submit your project will be 11:59pm on June 4. Projects are to be handed in electronically (via Canvas). You are welcome to hand your project in earlier than June 4.

Tasks and Specific Goals

1) Identify a network for which you are able to access relational and ideally attribute data as well.
2) Display your network graphically
3) Describe the structure of your network using summary measures of the network and of the actors' positions within the network using the SNA package or alternative.
4) **Primary goal:** State a research question involving at least one of:
   a) The structure of your network; are certain local configurations of ties between actors highly prevalent and thus appearing to be important constructs in the network? Test your hypothesis by modeling the network. Evaluate goodness of fit measures and also check MCMC diagnostics for your model. Be sure to interpret the key terms in the model.
   b) Whether or not the attributes of actors that are directly connected in your network are positively or negatively associated. Test your hypothesis by conducting a social influence analysis. Describe whether the social influence analysis is being estimated cross-sectionally or longitudinally, the estimation method used, and interpret the results.
   c) The comparison of multiple networks in relation to other (non-network) variables of interest. That is, use statistical models and network statistics of the network, or actors

within the network, to reduce each network to a vector of information that can be easily compared across networks and related to other variables measured on the actors.

You should do each of 1 to 3 and at least one of 4a, 4b and 4c (doing only one of 4a, 4b and 4c is sufficient if you do it well). If you only have relational network data for a single network and the available attribute information is non-mutable (i.e., are characteristics that do not change as opposed to being behavioral features), then your efforts in 4 will be limited to 4a. In that case, please perform a detailed analysis examining at least three network statistics (local configurations) and perform goodness of fit checks for each. Attempt to overcome lack-of-fit as needed and carefully monitor whether the model is fitting well (e.g., avoids degeneracy).

If you have mutable attribute data you can perform a combination of 4a and 4b such that a total of at least 3 analyses or statistical queries of the network are performed. You may focus on one of 4a and 4b but make sure you perform a detailed in-depth analysis.

If your network data includes multiple comparable networks with attribute data (e.g., of multiple regions, companies, hospitals) you may perform a combination of 4a, 4b and 4c such that a total of at least 3 analyses are performed. You may focus on one or two of 4a, 4b and 4c but make sure you perform an in-depth and comprehensive analysis involving a total of 3 statistical analyses, as above.

Sources of Networks

1) Your own network: You are welcome to search for a network off your own accord or use a network that you already have access to. Please ensure the network is suitable for the types of analyses performed in class.

2) Access a network at an online repository such as here: https://icon.colorado.edu/#!/networks

This site allows you to search within types of networks. I recommend focusing on social networks (contains many classic networks) and to prioritize networks with attribute information. For example,

- Zeggelink's Freshmen (1999) is a classic social network. Like many networks in this repository, it is a modestly-sized network
- Harry Potter Character Relations (2013) also contains multiple networks along with some attribute information.

3) The following is a website that Carly identified as a source to obtain publicly available network data sets: https://snap.stanford.edu/data/

The Networks with Ground Truth Communities contain network information and at least one piece of attribute information, the community membership of the actor. Unfortunately, community membership is not the type of variable that is suited for social influence analyses. However, these are great data sets for analyzing the structure of a network. As an illustration, to load the "email-Eu-core" network in R you can use code like:

```
netdata <- read.table('../Data/email-Eu-core.txt')
attrdata <- read.table('../Data/email-Eu-core-department-labels.txt')
```

As other examples, the Email network from a large European research institution and the Youtube network both contain relational and attribute information:

- Network of Emails from a large European research institution. Network: email-Eu-core.txt; Attributes: email-Eu-core-department-labels.txt
- Youtube social network and with ground truth communities used as attributes. Network: com-youtube.ungraph.txt; Attribute: com-youtube.all.cmty.txt

Advice for analyzing networks

Some networks are large, which can slow down computations. If you want to just analyze a subset of the network when applying procedures that are highly computer intensive, you can adapt the following code (which extracts a network with 500 actors). The code assumes that the network IDs (idfrom, idto) are numerically valued.

```
netdata <- read.table('../Data/network.txt')
names(netdata) <- c("idfrom","idto")
u=sort(unique(c(netdata$idfrom,netdata$idto)))
node501=u[501]
keep=(netdata$idfrom<node501 & netdata$idto<node501)
ndata=netdata[keep,]
attrdata <- read.table('../Data/attrdata.txt')
names(attrdata) <- c('id','cov1','cov2',...,'covk')
keep <- (id<node501)
attrdata <- attrdata[keep,]
```

Running your analysis on the reduced network can be helpful when perfecting model-specifications, even if ultimately you do end up modeling the full network.