Novemer, 19, 2017

Social Network Analytics, Empirical Exercise #5
Due on Tuesday, December 5, 2017 at 8:00am

Angel investors and startups on AngelList

Background on data: Many startups use AngelList as a platform for generating traction about their companies and securing early-stage funding from angel investors. The site's internal database contains a table that tracks how much traction, which is a measure of support for a startup's fundraising round, a startup receives over time. Both investors and startups can participate in fundraising campaigns. There is another table in the database that links each fundraising round to the investors and startups that participate in it. We use the two tables to build a network of users (investors and startups) and use ERGM models to make some predictions about the kinds of interactions that take place on the site. The network resulting from this data manipulation is provided on Canvas. Because ERGM (and Siena) models can be intensive to run, this manipulated file focuses on the fundraising rounds of 100 of the $\sim$ 5000 startups featured on the site. We are interested in the "has as a participant" relationship between startups in a funding round and the users and startups that participate in them. In this network, ties represent directed connections from actors (startups) to events (participants in a startup fundraising round). In focusing on the directed relation, we are not considering the relationships between round participants, as we did in the venture capital network exercise. Here, it is possible for startups to be connected with other startups, so we will not use bipartite methods on this network.

1. The file "startup_rounds_and_participants.csv" contains an edge list with the following columns:

   - id keys for each fundraising round
   - id keys for the startup being funded (these are startups only)
   - id keys for the participants in the round (these can be investors or startups, and the startup id key is consistent withe previous column)
   - the type of participant (indicator for whether participant is an investor or startup)
   - the date of the fundraising round
   - how much traction was gained during the fundraising round

   Use the network information provided and an ERGM model to predict whether reciprocation is more or less likely in this network. What if we only consider connections between startups in the model? Why does the second model appear different? Include an edges intercept in both models.

2. Use the network information and an ERGM model to predict whether startups are more likely to have larger fundraising rounds (i.e., the "has as a participant" relationship is more likely to exist) if they also participate in more funding rounds. You can run the model on the entire network or just relationships between startups. In either case, include an edges intercept as well as a mutuality covariate in the model.

3. Use the network information provided and an ERGM model to predict whether an investor is more or less likely to participate in a startup's round if it has participated in in one of its rounds in the previous year. Include an edges intercept as well as a mutuality covariate in the model.

4. Use the network and startup attributes and an ERGM model to predict whether a startup is more likely to have participants in its rounds that have gained similar levels of traction in the current year. Then, build a model to predcict whether a startup is more likely to have participants in its rounds that have gained similar levels of traction using a term for the current year as well as the previous year. Only consider interactions between startups in these models, and include an edges intercept as well as a mutuality covariate in both models.

5. Of the models run so far, which has the best fit to the actual network? Discuss the model fit using (1) the built-in mcmc diagnostics, (2) a simulation of the actual network using the results from the model, and (3) a goodness-of-fit analysis of the estimated model.

Extra challenge problem: RSiena provides a similar function to ERGM, but is more tailored for longitudinal data. Re-run the longitudinal models above (Questions 3, 4) using RSiena to determine if RSiena and ERGM produce similar results.