Utilizing ERGM for Evaluating a Communication Network

Michael Cho

Northwestern University

**Method**

The focus of this research is to test the following hypotheses regarding the communication network of 17 students who were designing military installations in a virtual simulation. The research question we have is do members have an independent tendency to retrieve information from other members or are there are dependent effects at work?

*Hypothesis 1: Members tend to allocate information to other members, from whom they retrieve information.*

*Hypothesis 2: Members tend to retrieve information from other members with high expertise.*

The primary method by which we can answer these questions to utilize an ERGM. First, we had to define the networks themselves. Here there are two primary components: an adjacency matrix of members retrieving information from other teammates regarding environmental quality (CRIeq), an adjacency matrix of members allocating information to other teammates regarding environmental quality (CAIeq), and a normalized expert score that is tied to the first network as vertex attributes.

Next, we had to define what sort of signatures might appear in this network to verify the hypotheses. One we can consider is just endogenous signatures such as the number of edges in the retrieval network (edges). Next, we can consider the number of reciprocal links in the network. One can request from another team member information about environmental quality; whereupon, at some later time that same team member asks the former for his or her input as well. This does not necessarily mean that information was allocated, just that there was a request made and it was reciprocal (mutual).

Perhaps, this system of communication is not on a dyadic basis and extends more between triads of team members. At the most optimal, we want these triads to be transitive and cyclical; therefore, we can count the number of links that belong to at least one triangle structure (gwesp) to verify this claim.

To fully verify the first hypothesis, we now bring in the CAIeq network as an edge covariate which is also defined among the same set of member vertices. Finally, for the second hypothesis, we can see the covariance between the in degree of nodes and the numerical attribute of their expert scores (nodeicov) as opposed to the covariance between the out degree of nodes and attributes of nodes (nodeocov). To further clarify, we would expect those with high expert scores to have a high indegree and perhaps those with low expert scores to have a high outdegree as they are continually seeking information from their peers whom they perceive as having more expertise. Once we define these network signatures, we can run the model and verify these hypotheses, get their log odds, as well as the significance of these results from their p-values defined at a 95% confidence level with a 0.05 alpha.

However, even if we are able to discern these effects as significant, we can’t fully be sure our model is correctly specified without first checking convergence of the MCMC chains and also comparing our observed network against simulated networks from which our model was derived. In regards to the later process, we can see how the observed network stacks up against the distribution of simulated networks by local substructures ( number of triangles) and also metrics that are currently not specified in our model (the distribution of indegree, outdegree, edge wise shared partners, and geodesic distances), the latter of which we can verify through a goodness of fit.

**Results**

First looking at the model results below, we can surmise that number of edges is very significant but lowers the log odds of a an edge forming in the information retrieval network (a team member seeking information from another). The retrieval network is not as dense as it could be. Second, retrieval communication doesn’t seem to occur through triads and reciprocity, which is quite strange given that edges of the allocation network do have a significant effect, but this relationship doesn’t seem to be reciprocal at all between dyads. However, we do have quite clear evidence of the second hypothesis as high expertise score individuals have a high indegree as lots of members are asking them for information about environmental quality.

==========================

Summary of model fit

==========================

Formula: CRIeq ~ edges + mutual + gwesp(0.2, fixed = T) + edgecov(CAIeq) +

nodeicov("EX") + nodeocov("EX")

Iterations: 2 out of 20

Monte Carlo MLE Results:

Estimate Std. Error MCMC % p-value

edges -7.0556 1.2101 0 < 1e-04 \*\*\*

mutual -1.4042 1.0065 0 0.16415

gwesp.fixed.0.2 0.5905 0.4356 0 0.17639

edgecov.CAIeq 2.1357 0.6870 0 0.00208 \*\*

nodeicov.EX 9.4695 2.3506 0 < 1e-04 \*\*\*

nodeocov.EX 1.6941 1.7444 0 0.33235

---

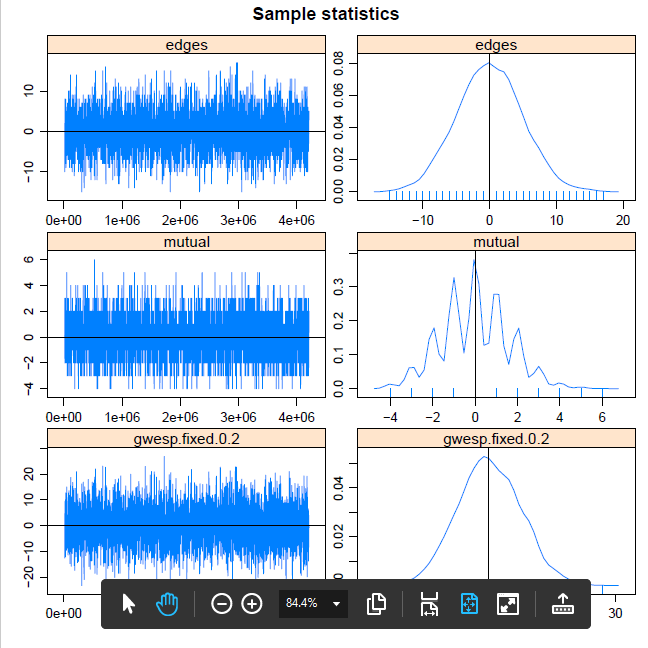
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

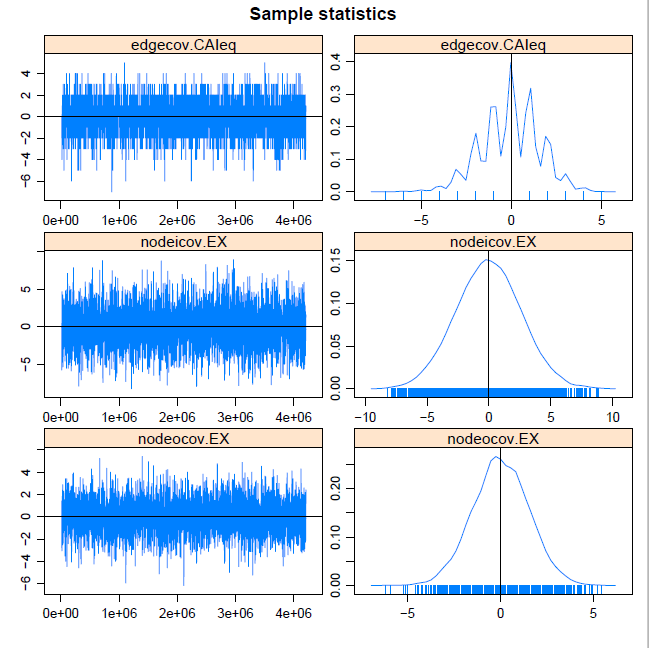
Null Deviance: 377.1 on 272 degrees of freedom

Residual Deviance: 125.2 on 266 degrees of freedom

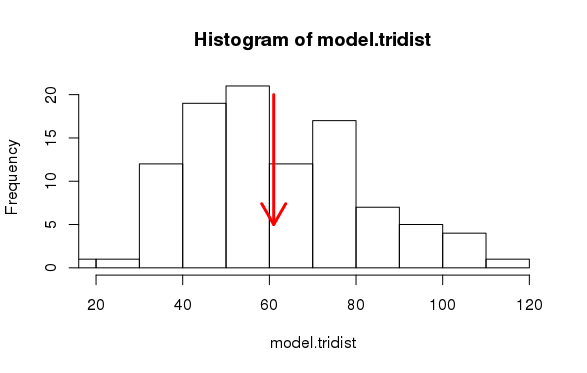
AIC: 137.2 BIC: 158.9 (Smaller is better.)

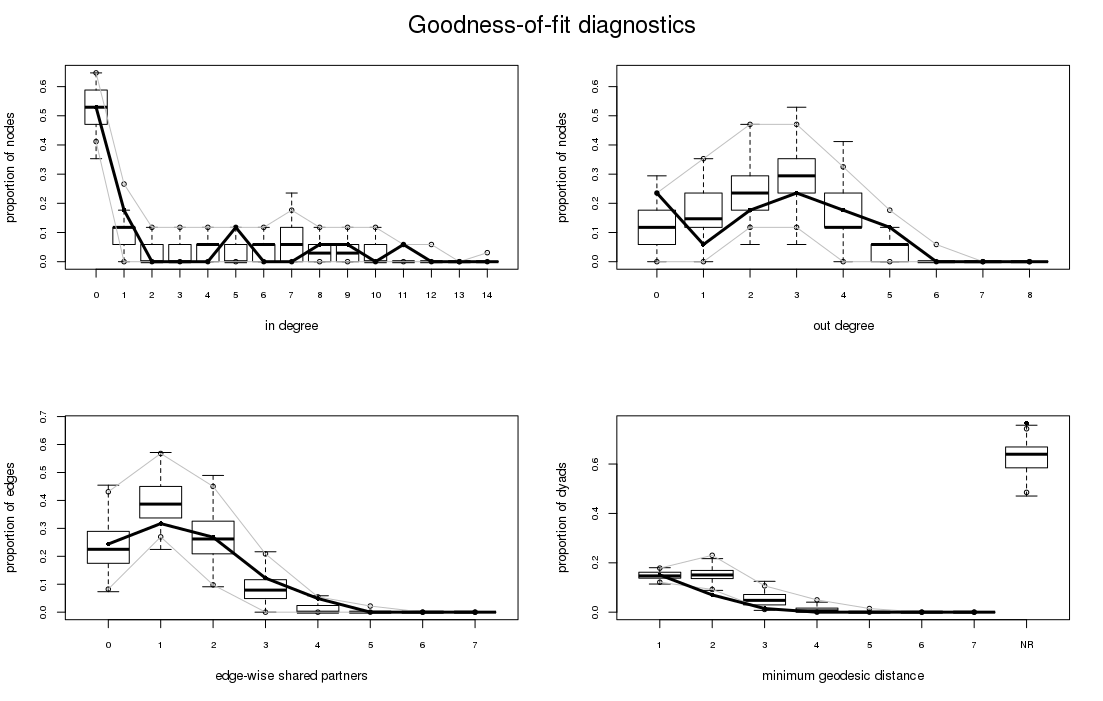
Moving onto the model fit itself, we first whether the MCMC chains converge. If these chains converge well, we would expect something resembling a normal plot on the right for each predictor variables centered around the normalized mean of 0. These visualizations verify the model converged well.





Finally, we assess the goodness of fit. We create the distribution of the number of triangles from the simulated network and see where the count from the observed network falls. If it falls toward the middle of the distribution, this just adds evidence of a good fit and indeed it is as the observed network has 61 triangles.



While the above is just a simple histogram, we can use the gof function to assess other network statistics that are in and not in the model. For these plots, the barplots representing the mean and quartile ranges of the simulations, the grey lines represent the 95% confidence intervals, and finally the black lines show the observed network. If the observed network falls cleanly between the confidence interval, we can be confident that model is well specified.

All of the measures in the gof function are outside of the model and the p-values fall into an acceptable range from 0.1 to 1. So after all these steps, we have a fully specified model.

Goodness-of-fit for in-degree

obs min mean max MC p-value

0 9 6 9.19 12 1.00

1 3 0 1.77 5 0.54

2 0 0 0.70 3 0.86

3 0 0 0.41 2 1.00

4 0 0 0.46 2 1.00

5 2 0 0.64 2 0.24

6 0 0 0.83 3 0.84

7 0 0 0.88 3 0.72

8 1 0 0.74 3 1.00

9 1 0 0.68 2 1.00

10 0 0 0.38 2 1.00

11 1 0 0.18 1 0.36

12 0 0 0.12 1 1.00

13 0 0 0.02 1 1.00

Goodness-of-fit for out-degree

obs min mean max MC p-value

0 4 0 1.84 5 0.22

1 1 0 2.72 7 0.44

2 3 1 4.26 9 0.74

3 4 1 4.64 10 0.90

4 3 0 2.82 7 1.00

5 2 0 0.66 5 0.30

6 0 0 0.06 1 1.00

Goodness-of-fit for edgewise shared partner

obs min mean max MC p-value

esp0 10 3 9.57 19 0.88

esp1 13 8 15.99 26 0.56

esp2 11 0 10.81 24 1.00

esp3 5 0 3.24 17 0.46

esp4 2 0 0.47 6 0.24

esp5 0 0 0.02 1 1.00

Goodness-of-fit for minimum geodesic distance

obs min mean max MC p-value

1 41 29 40.10 53 0.96

2 19 25 40.48 66 0.00

3 4 1 13.20 34 0.10

4 0 0 2.29 21 0.78

5 0 0 0.31 11 1.00

6 0 0 0.02 2 1.00

Inf 208 117 175.60 212 0.04

**Discussion**

The ERGM model we ran indicates that participants tended to retrieve information from those with high expertise. After running the model, we verified convergence and goodness of fit. The first hypothesis had no empirical support from the data, but also a cursory look at the network plots would verify that people aren’t reciprocating from retrieving to allocating information. Furthermore, the network signatures for the first hypothesis could be more well defined. Most triangles in the retrieval network didn’t have closure and transitivity, but it would have been interesting to include network statistics on these substructures to account for and verify the first hypothesis.