DS745 Network Analysis

Telling a Network Story: Global Online Course Social Interactions

Analysis using Exponential Random Graph Modelling (ERGM) and Network Visualization

Rick W. Lentz, MS, MS  
29 February, 2018 - [BEST VIEWED HERE](https://docs.google.com/document/d/1KSEmwLDosB1VRJZ_jyBS53zUQilKeCUAJVc2ojQQIKU/edit?usp=sharing)short line

# Dataset: Explanation

After reviewing the publically available datasets, I found the CMOOC dataset interesting since it captured social interactions between 3 teachers and 764 students of an distributed massive open online course offered in 2011. A link was added whenever party A referenced part B in their Tweet.

To perform the Exponential random graph model (ERGM) analysis, common elements of the object definition are extracted and added as node level attributes. Here is a table detailing the network’s vertex attributes.

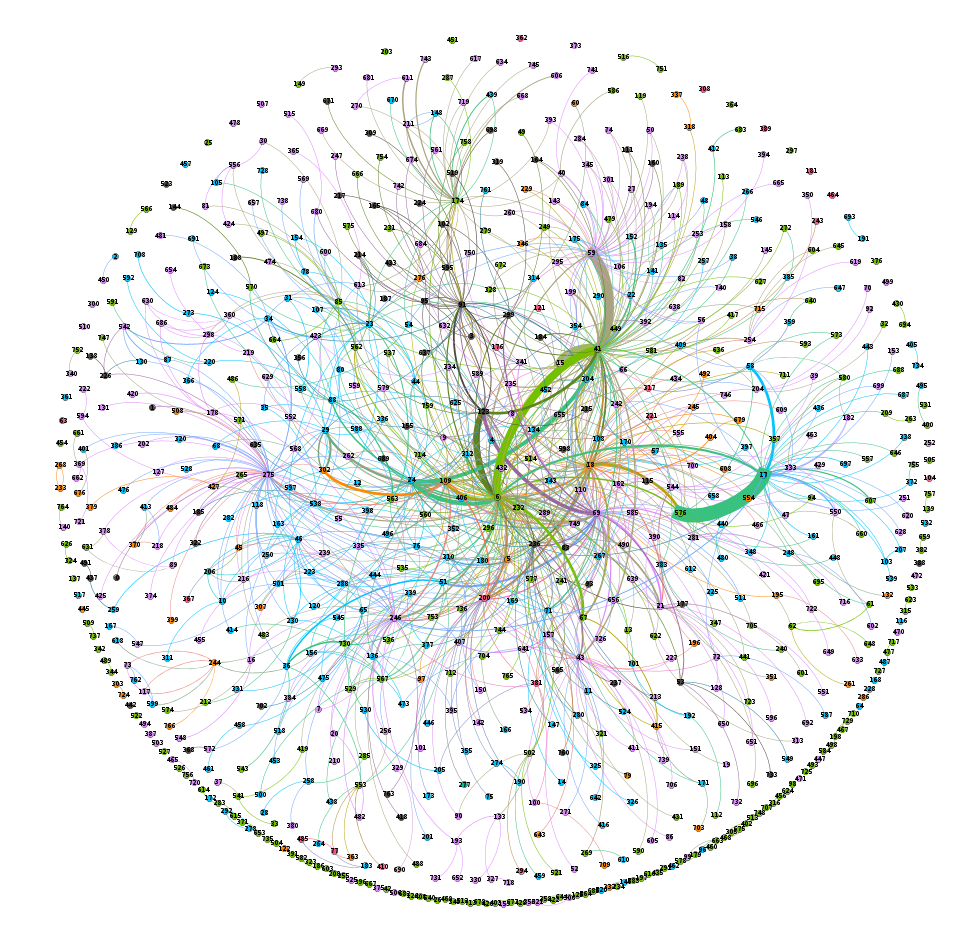
|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Example** | **Description** |
| id | 123 | Subject ID, range 1-767 |
| domain | Undergraduate, Community | Social cluster, group |
| gender | M, F, unk/other, Org | Reported gender or organization or unk |
| role | Instr, Student | Role in network |
| continent | Intl, Asia, Europe | Reported continent |
| socio\_tech | Social, Technical, Unk | Category Social or Techincal, or Unknown |
| work\_type | Mixed, Practice, Research | Employment role or Unknown |

# 

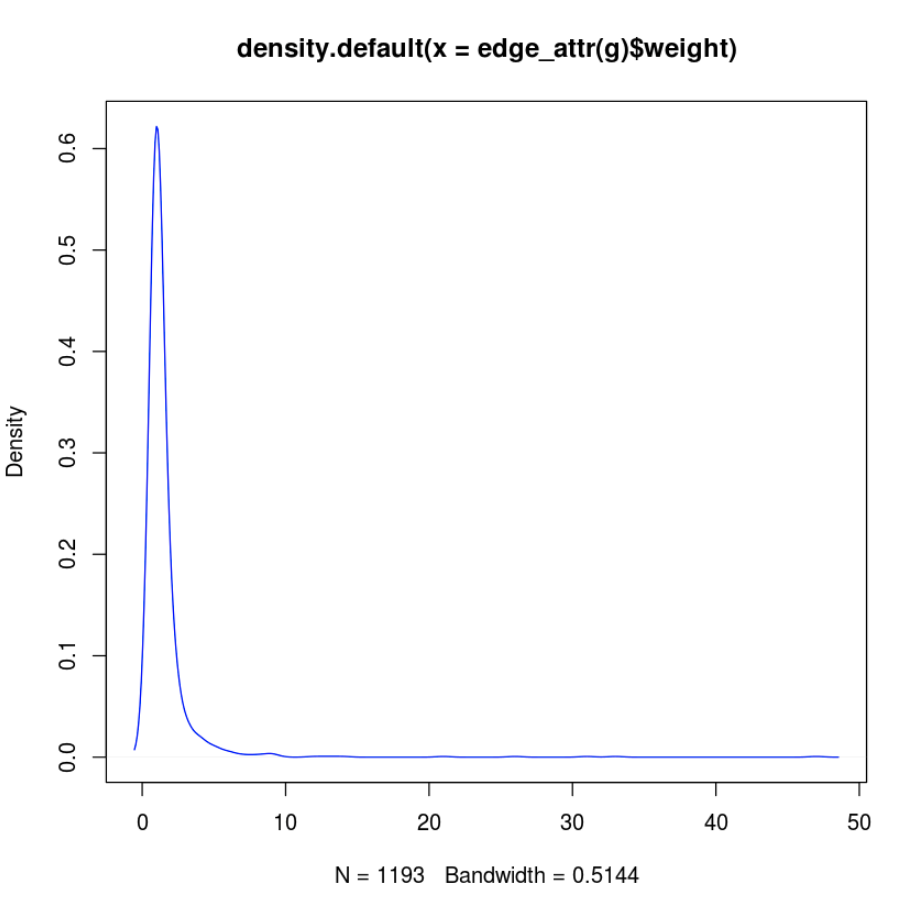
# Visualization and Network Summary

We start by looking at the overall color-coded structure of the entire network. Then analysis is shown for the network’s structural properties. These will help build up ideas about the features we may want to include in models created in later sections of this paper.

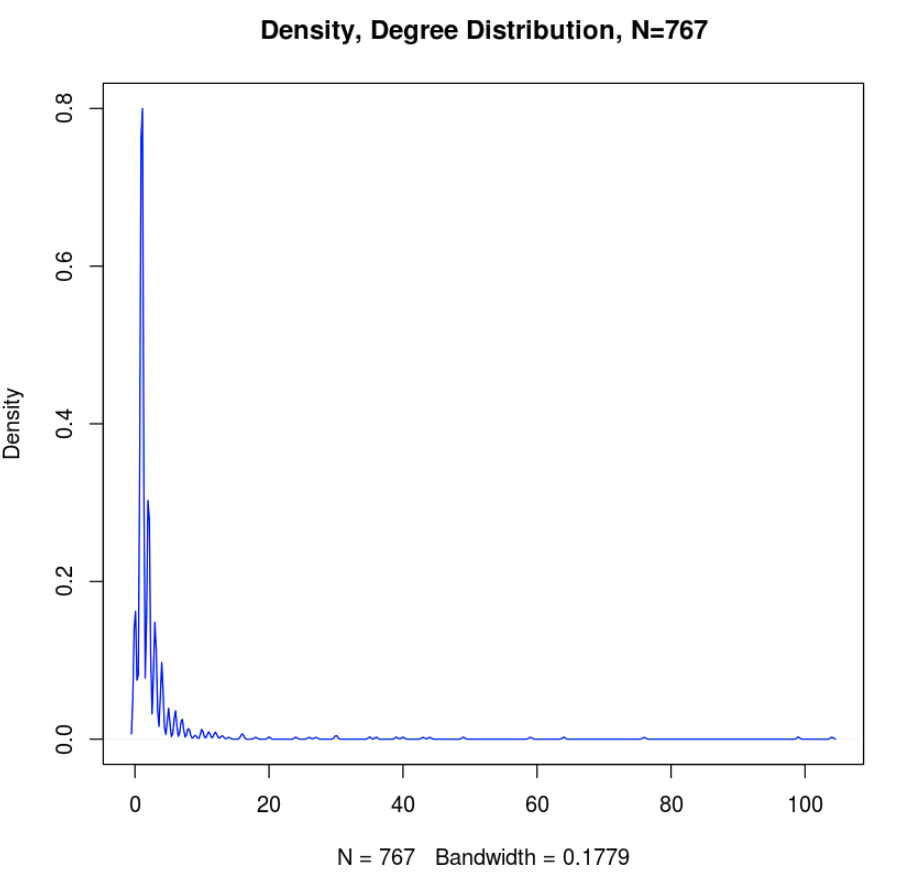
The network consists of 1193 interactions (edges) and has a small density value of 0.002.

The above Fruchterman Reingold visualization colors the links between persons according to geographical region. Other categories were used for this coloring, but less visual correlation was represented. We expect that regional association will be a large factor in our model.

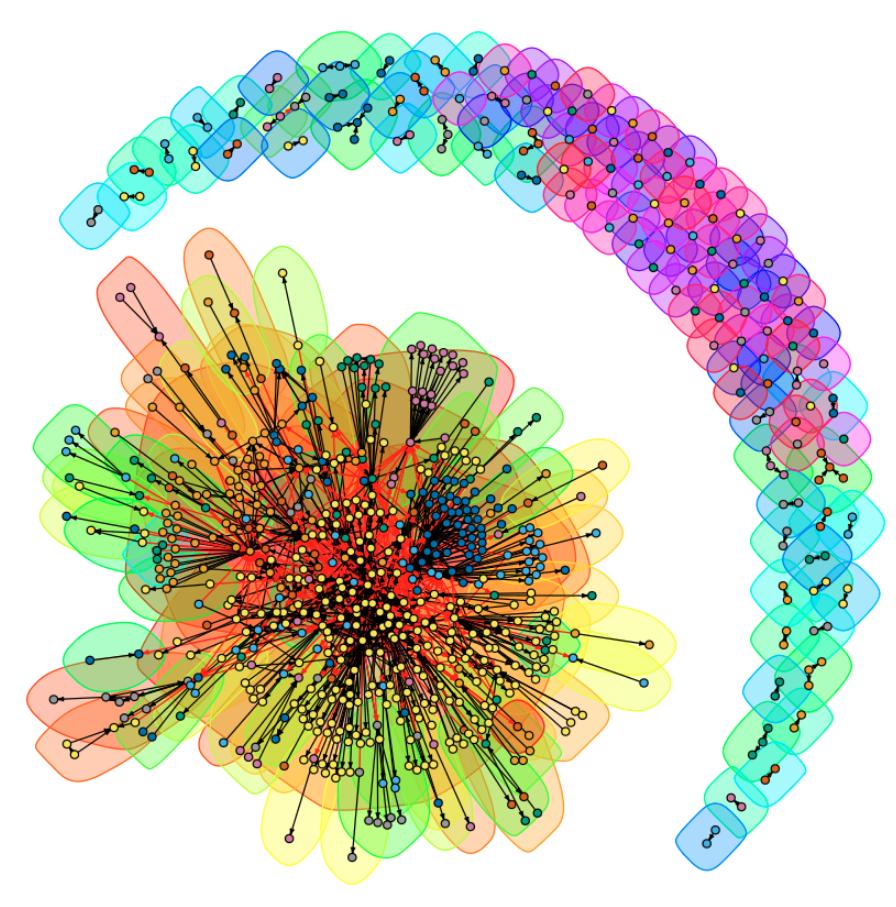
Next, we can look at the distribution of the edge weights (where parties mention each other in their messages). The the below plot, we can see that the distribution is skewed. Most of the values are less than five, and the maximum is 47.



Next, we can consider the degree distribution. This tells us the probability distribution (fraction of the nodes in the network with degree k) based on the number of connection the node has to other nodes. Most nodes have less than five connections, but some have over 100!



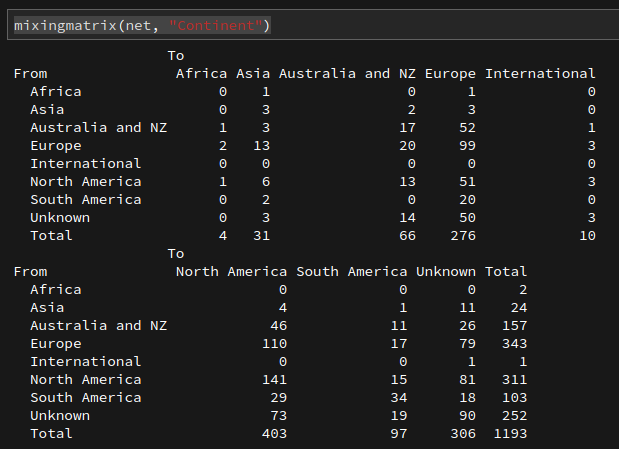
Community Detection

The below community detection algorithm reprentended the strongest of a variety of algorithms. ‘Walktrap’ was better than ‘fast greedy’, ‘edge betweenness’, ‘spinglass’, ‘leading eigenvector’, and ‘propagation’.

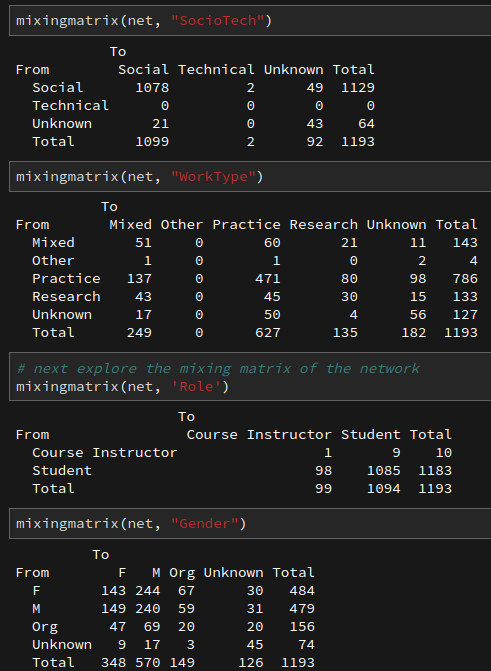
This provides an interesting perspective as it shows what one may intuit about the distributions of active, public communications in an online course. Here we see that many students mention just one or two others. But a few dominant nodes, seemingly sharing a common geography, have many social mentions.

We can build two hypothesis based on community clustering. One is that many students pair up in small groups. These small communities are physically separate from the main network (on the upper right arc above). Recalling that node color is based on geographic region, we concluded a second hypothesis, based on many clusters (e.g. the blue, purple, and yellow in the center) that have obvious groupings in the main, central network.

Finally, we use mixing matrices to view the actual numerical relationships between our attributes discussed above.

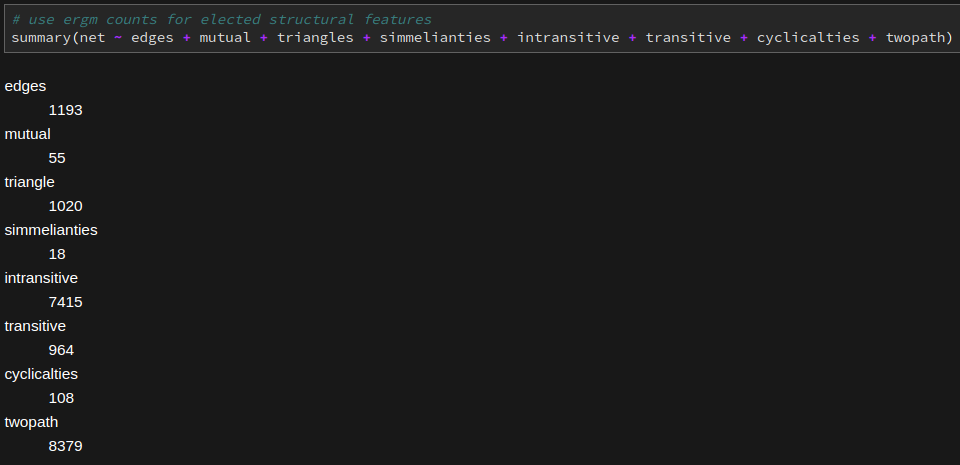


These relationships are seemingly not as clear cut as the images portray. Additionally, over 25% of the nodes are unknown. Next, we look at the next 4 attributes:



Under the role above, it stands out that there is little public instructor interaction, and little instructor mentioning of students. Student’s mainly mention students and about 8% of the time, instructors. In also looking at gender above, it appears as though male interaction is likely a factor we want want to consider.

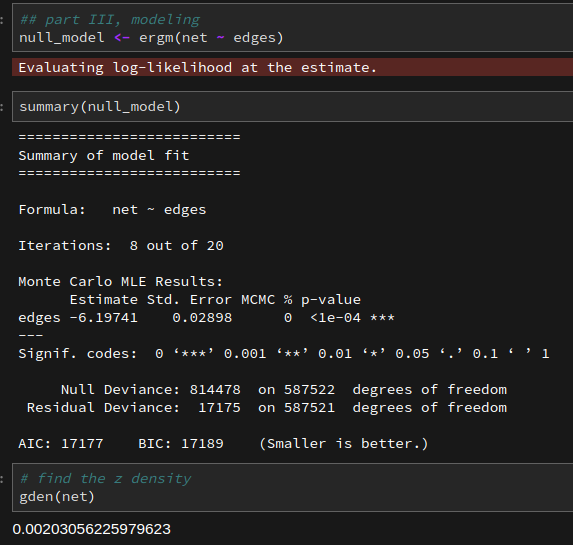
Finally, we need to consider factors associated with the network’s structural features. In the below summary table, we observe high two-path and intransitive properties.



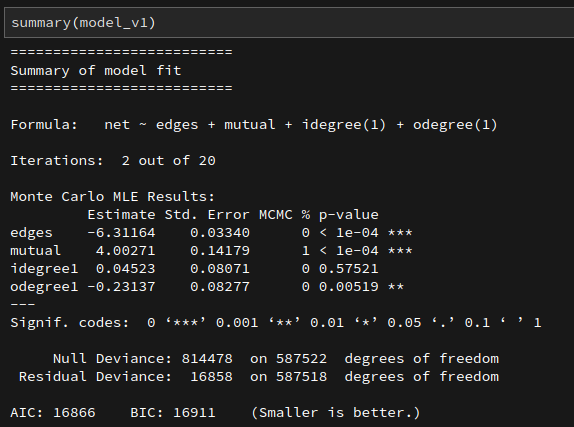
# Network Modeling

From the course notes we recall that goodness of fit is generated by the Monte Carlo Markov Chain algorithm in that it is used to approximate the Maximum Likelihood Estimation function.

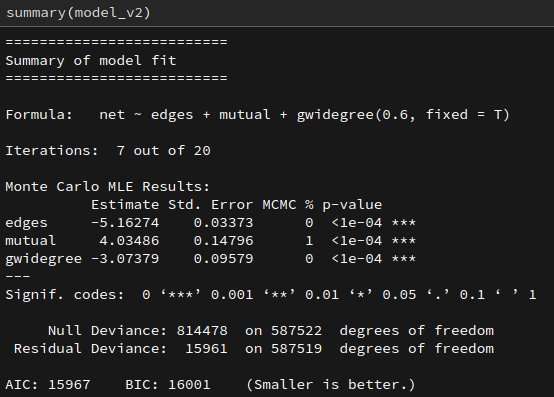
Network modeling with ERGM starts with a null model that is used as a benchmark for other models to assess goodness of fit. Typically the null model isn’t very good, since it only considers network density in relation to a randomly generated network. As we build in both network structural elements and certain edge and node attributes, the AIC coefficients tell us how the model is doing.



Here we see that the network density is less than 50% which is not unusual. We also see that the zDensity is 0.00203. This means that the probability of creating an additional edge by adding one more node is very low.

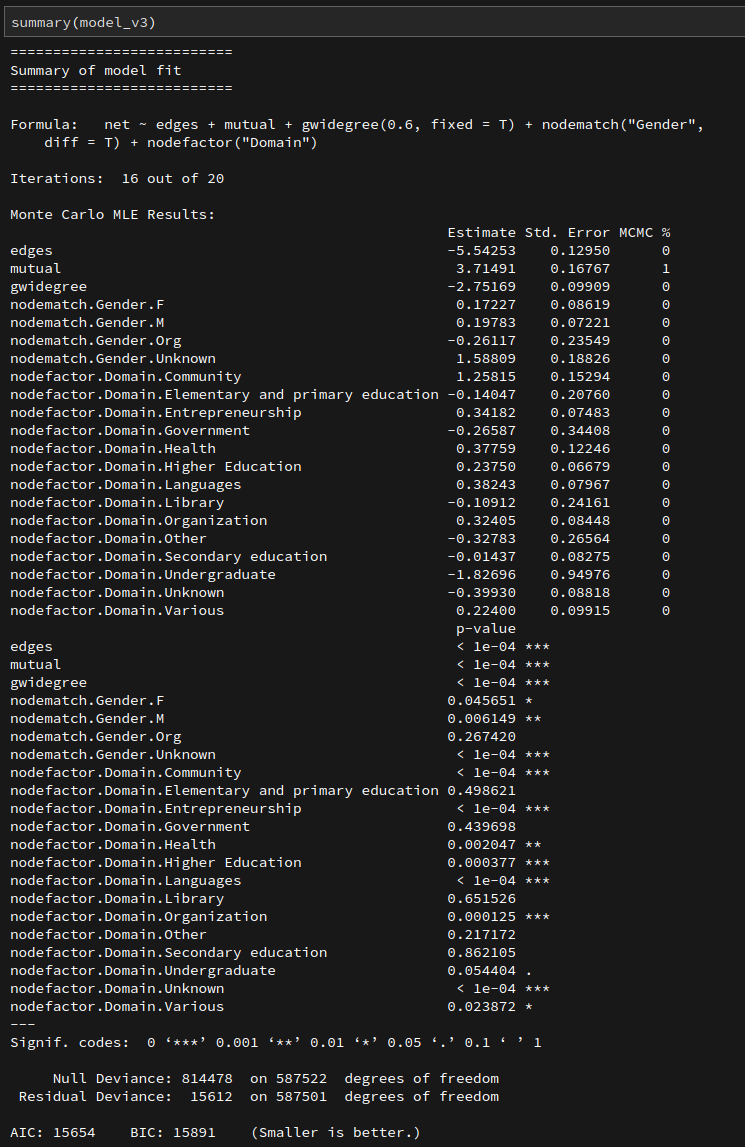


For my first of three non-baseline (null) models, I look at first order and mutual interactions. In the above summary, we can see that the mutual connections and out degree one, offer the strongest improvements over the null model (BIC improves from 17,189 to 16,911).



For the second non-baseline model, I keep mutual and add gwidegree to control for degree and closure. According to the above summary, this offers further improvement from BIC 16,911 to 15,994.

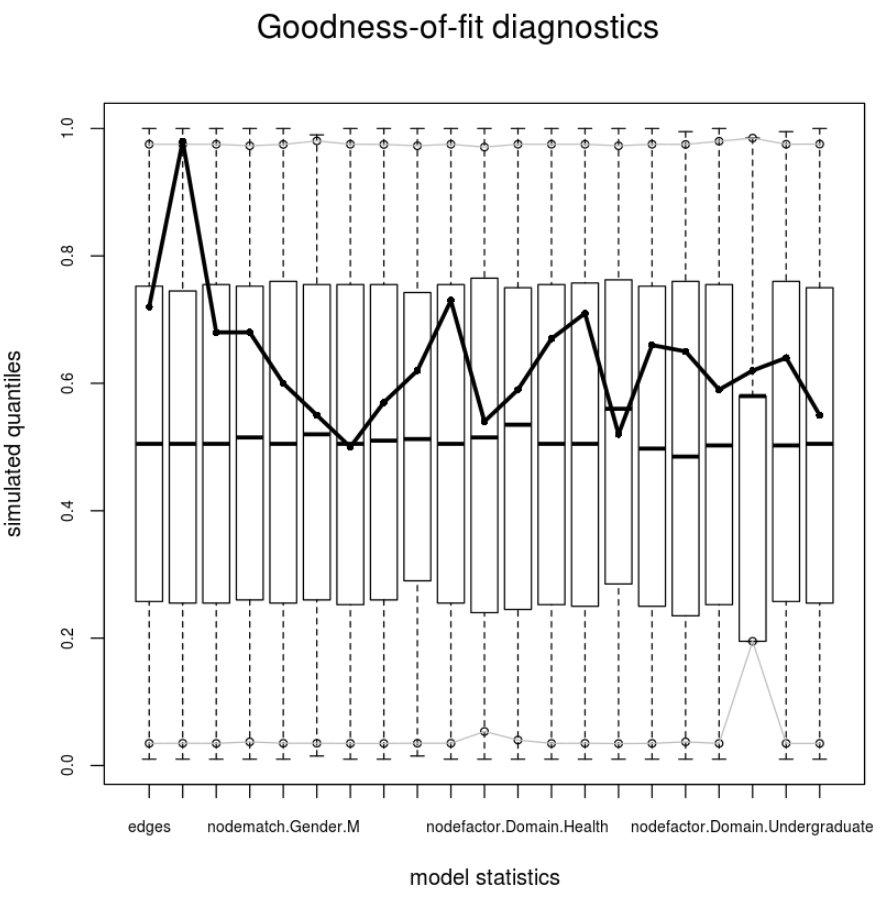
So, for the third model, what did we explore earlier that may help our model’s performance? Domain and gender both looked interesting. In the below summary, we can see how these new additions impacted our model:



The BIC took a nice improvement (from 15,994 to 15,891).

The next plots are helpful in reviewing the model’s goodness of fit, that is modelled versus observed performance. Here we run simulations and see that the model still struggles for some categories (while considering it it trying to capture edges, gender, and different domains):





Finally, we can observe the odds of the external and internal factors based on our initial thoughts and later findings:

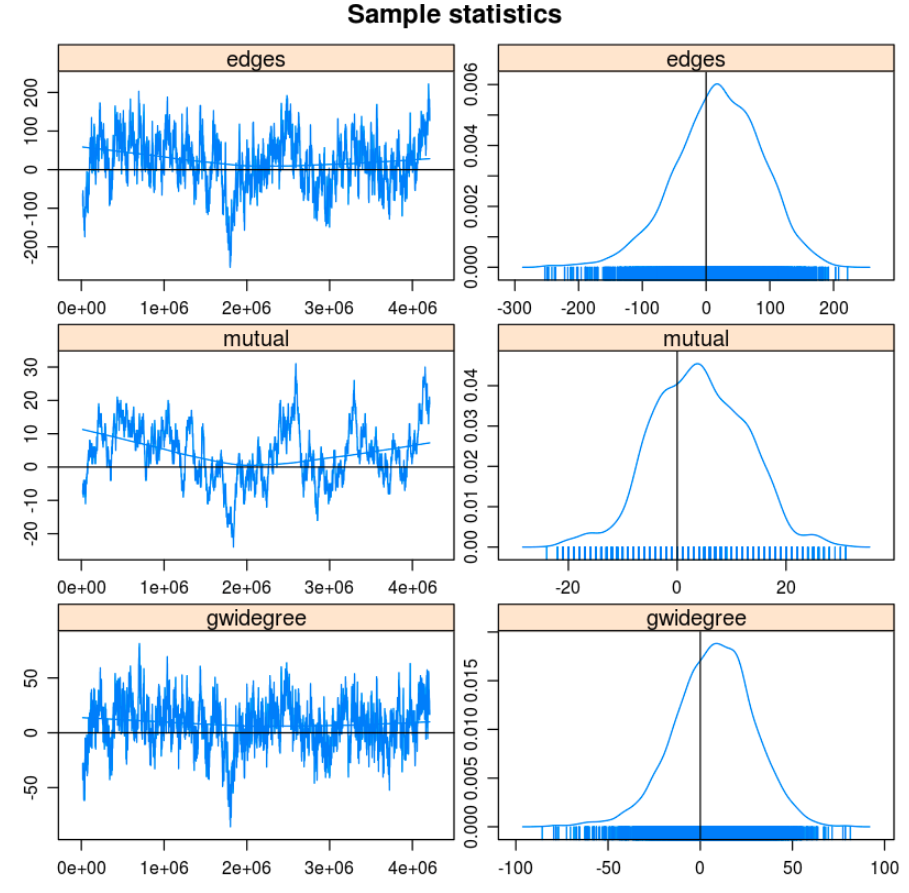


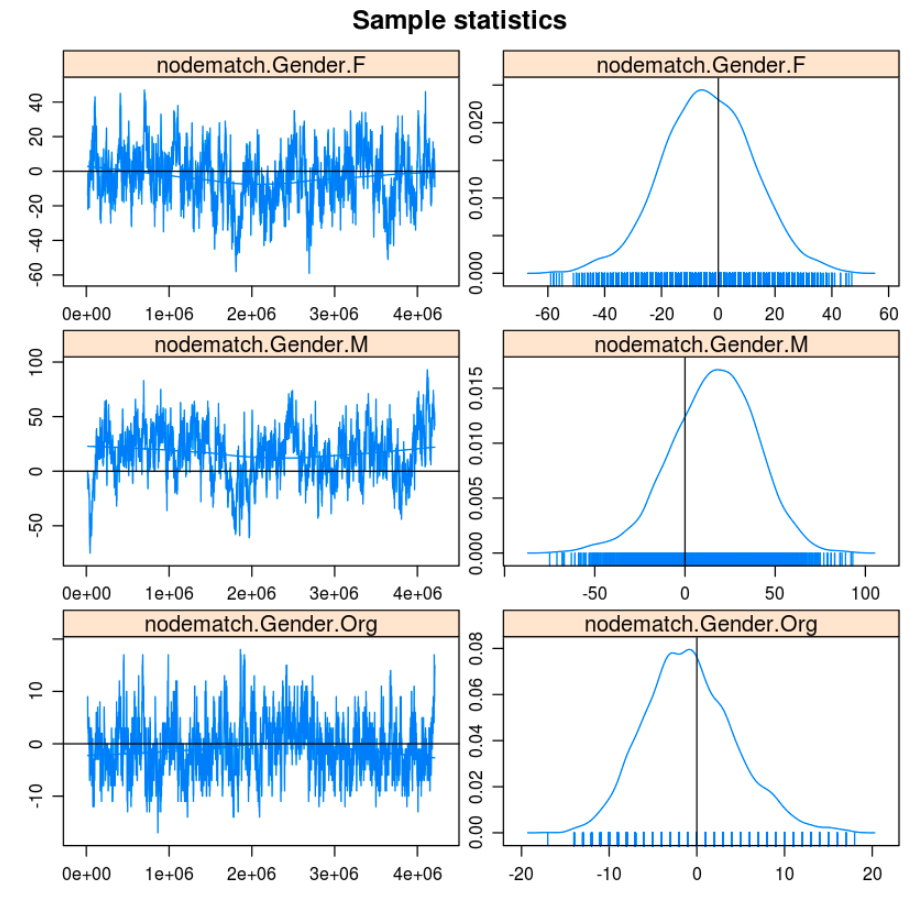
Based on the above, we can observe some homophily between males, higher education, institution accounts, community members, and entrepreneurs. These subgroups were factors of the information network’s flow and structural reciprocity.

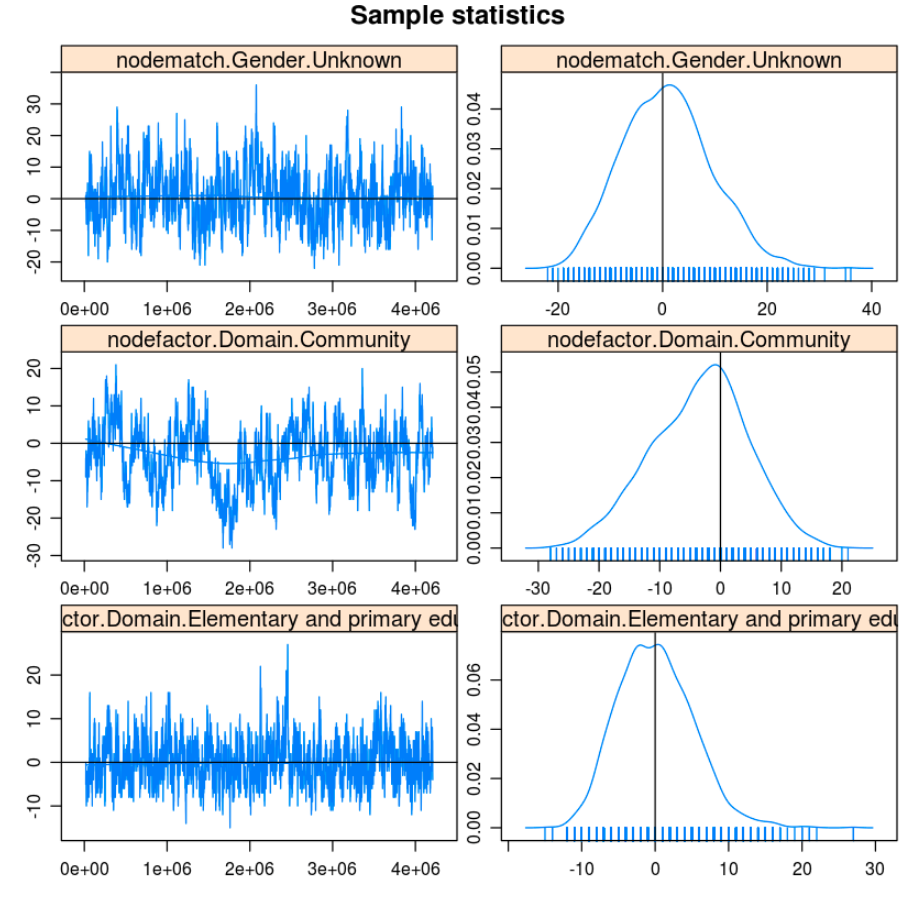
This analysis is publically available on my [GitHub](https://github.com/ricklentz) account under the creative commons license as it may be helpful for others studying social information exchanges or comparison between large online program behaviors.

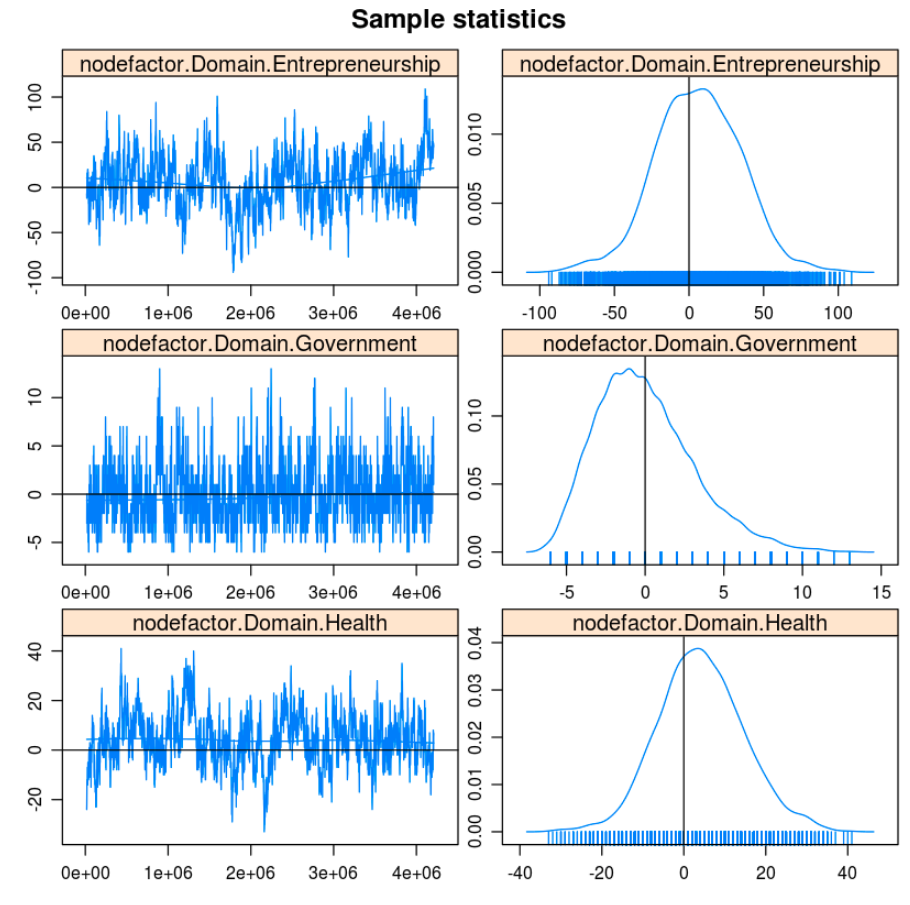
# 

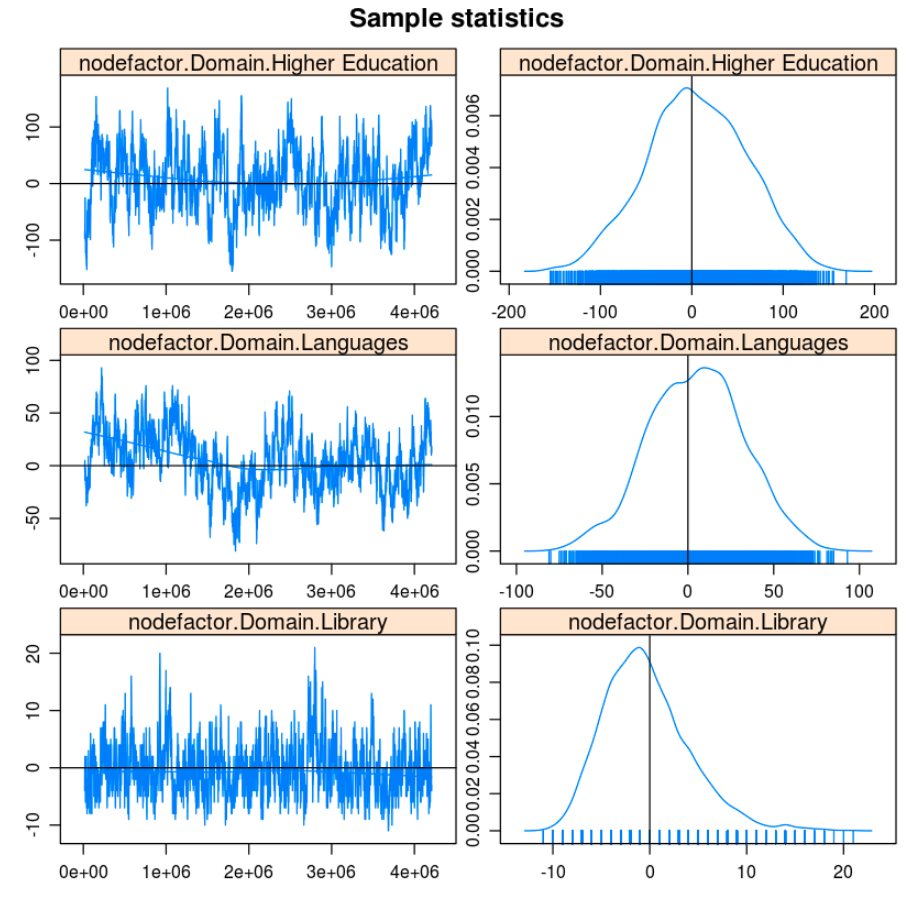
# Appendix A : Model III Sample Statistics Plots

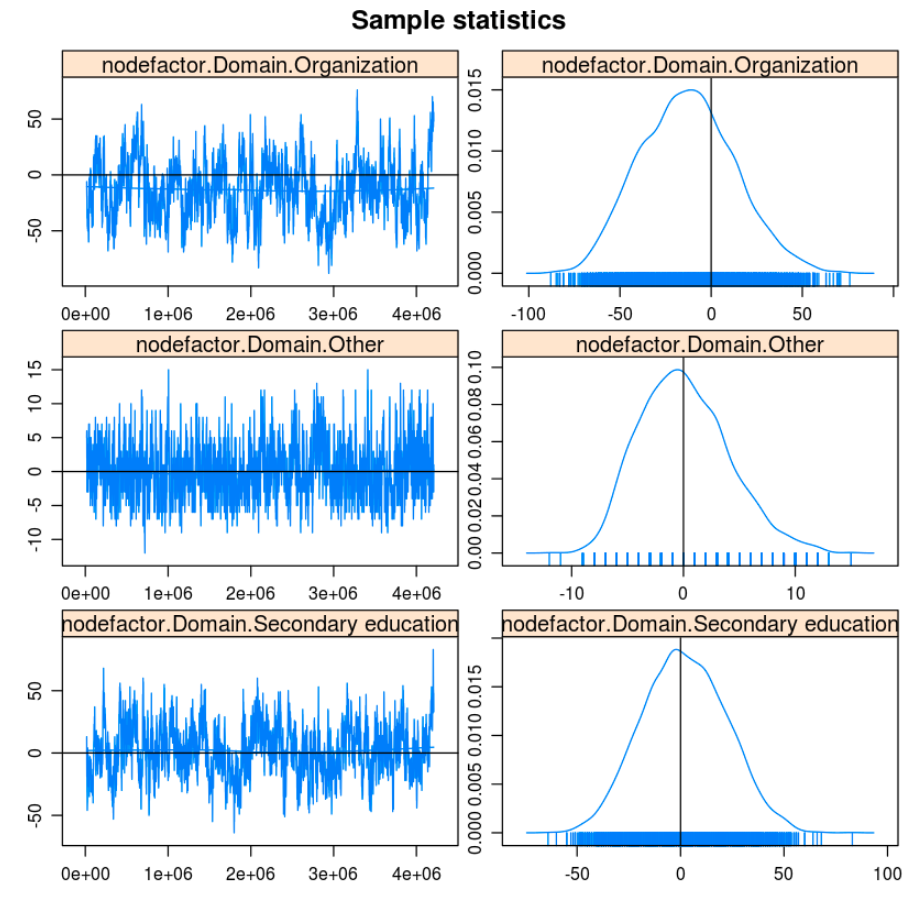


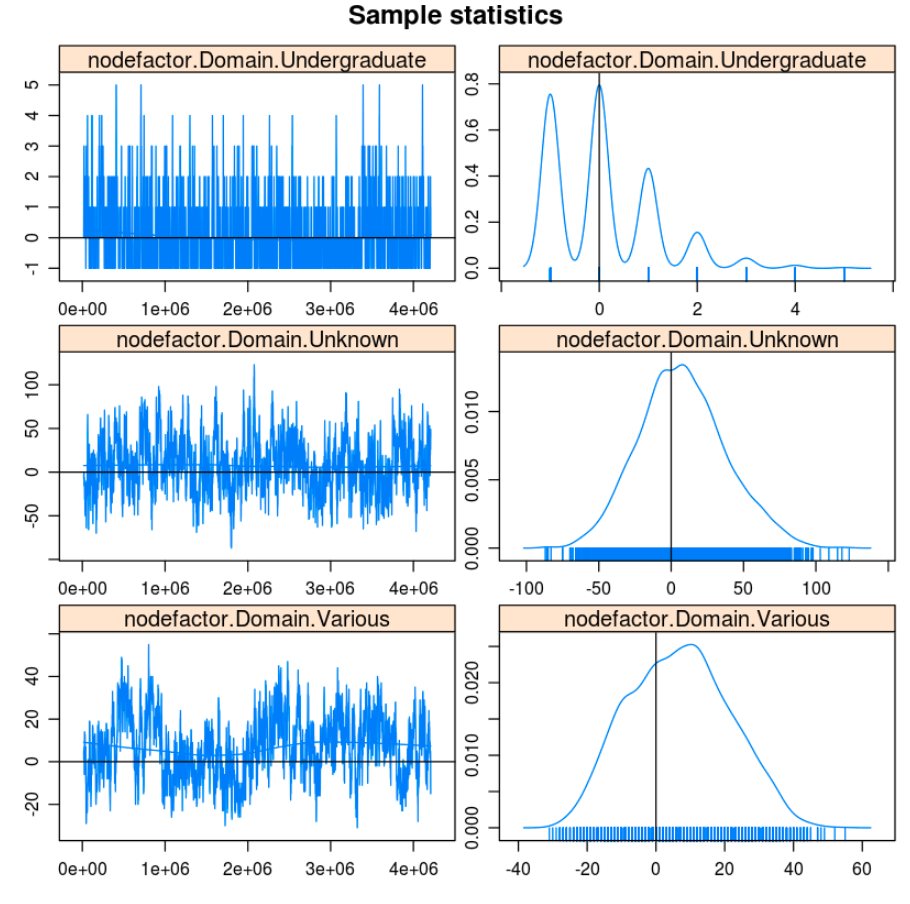






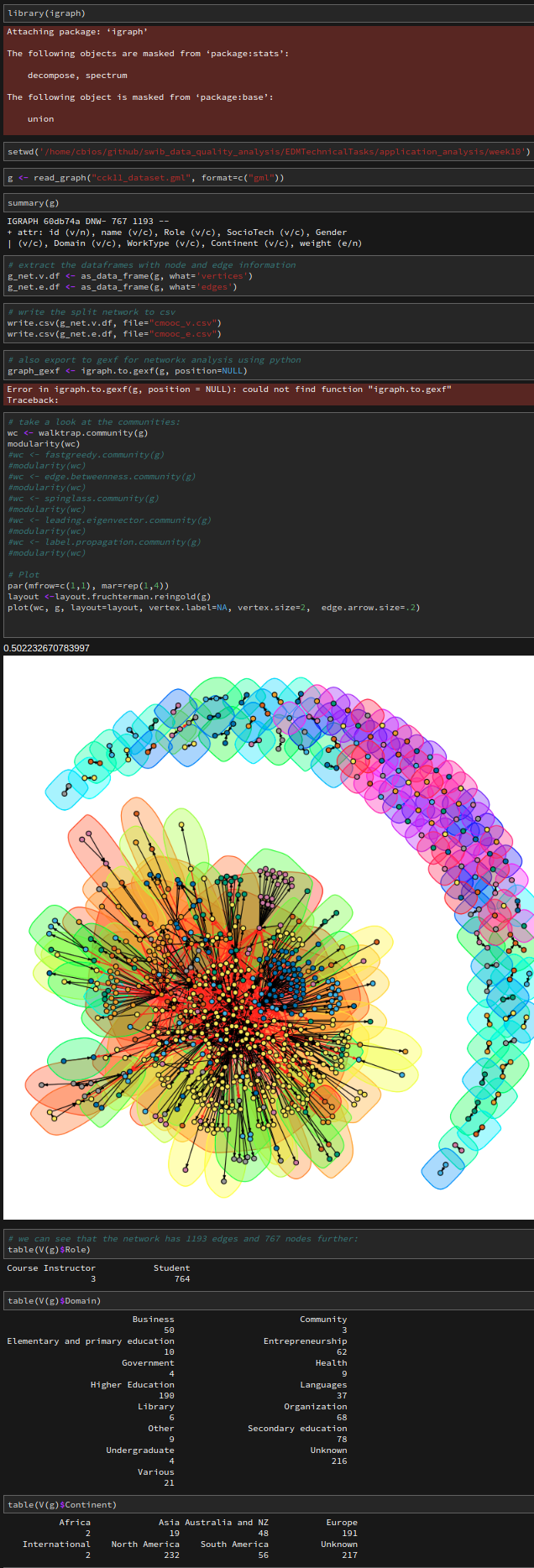


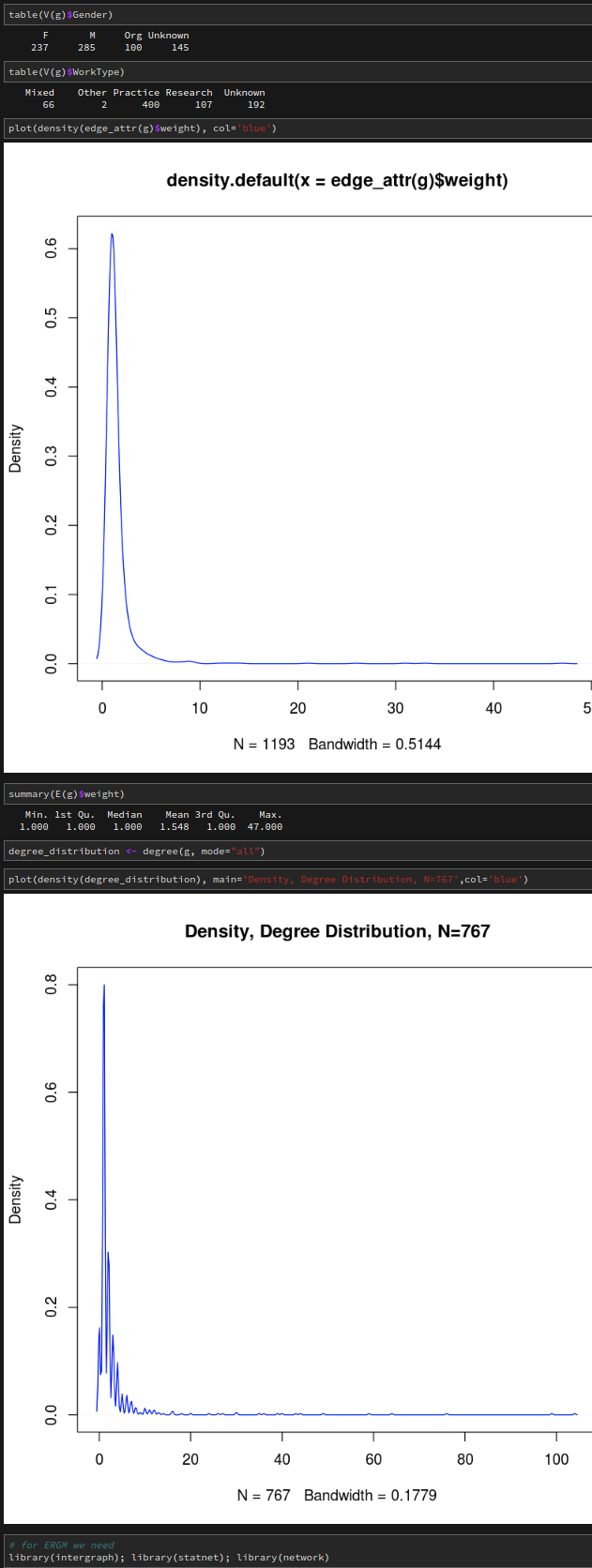


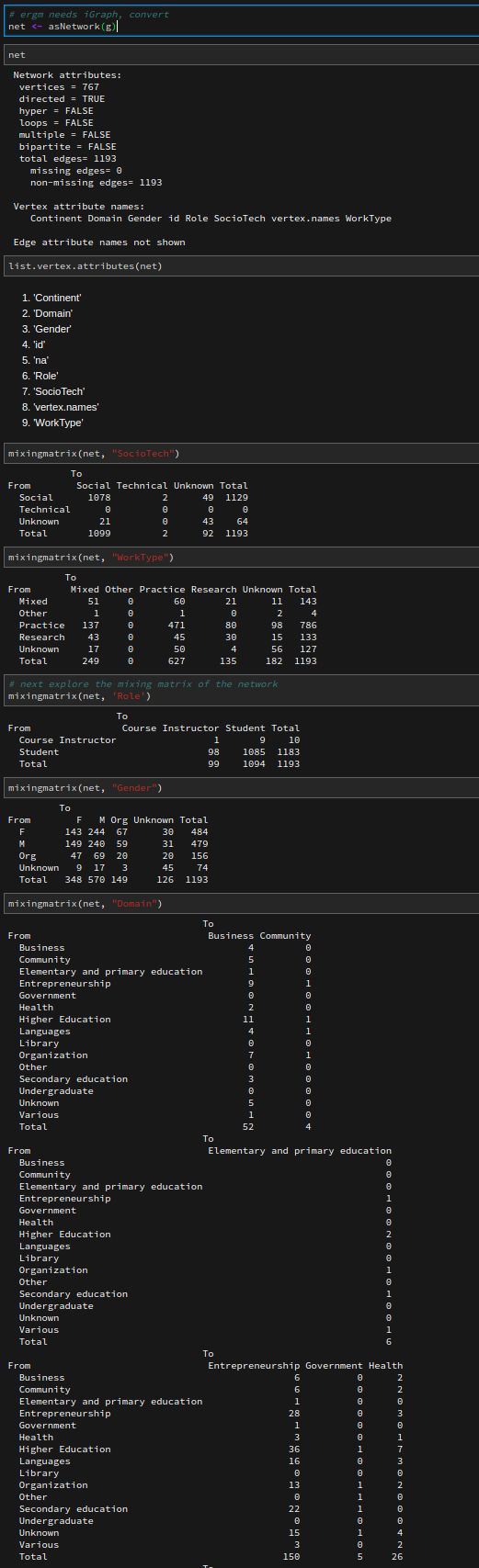


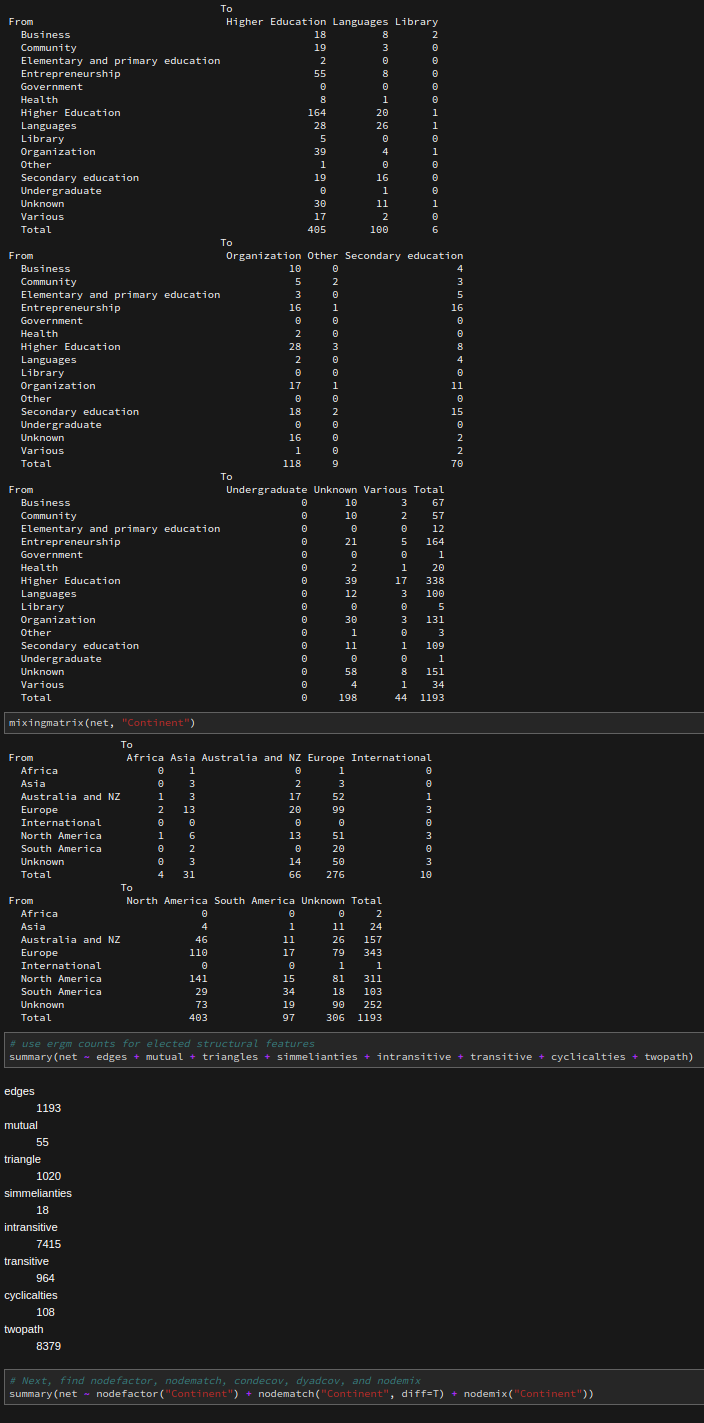
# 

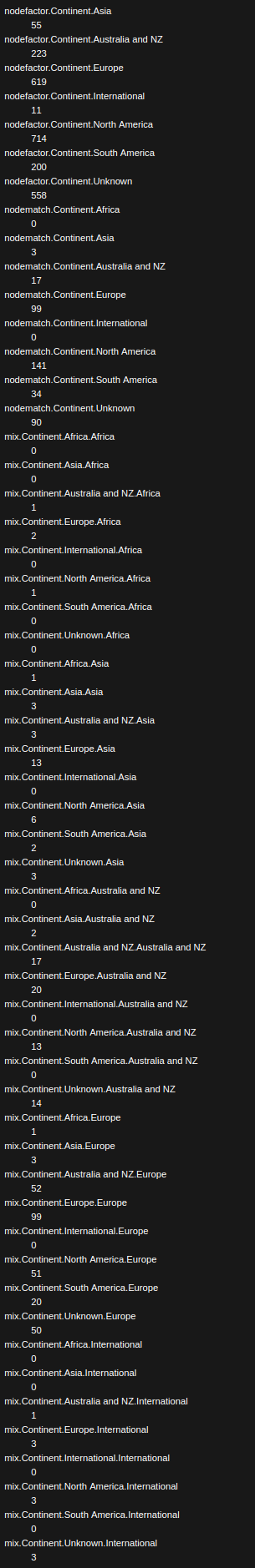
# R Code

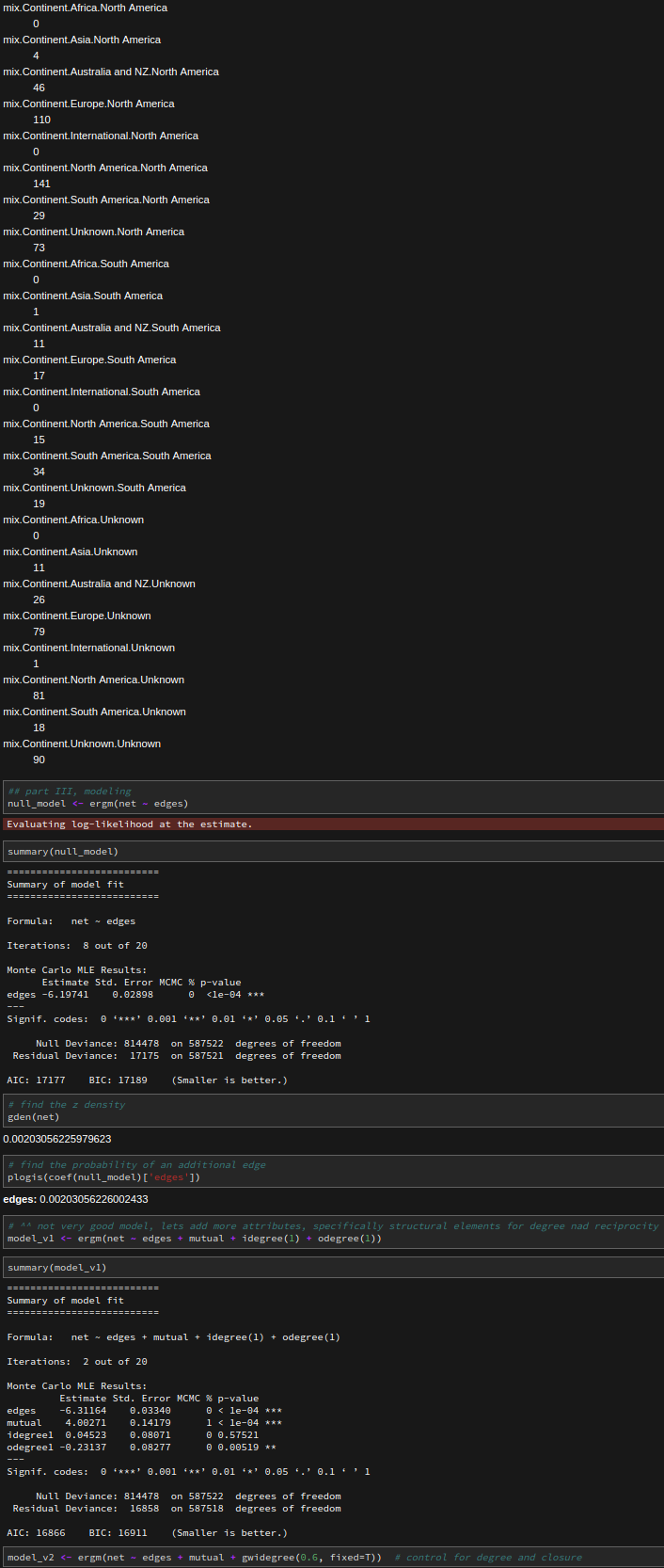


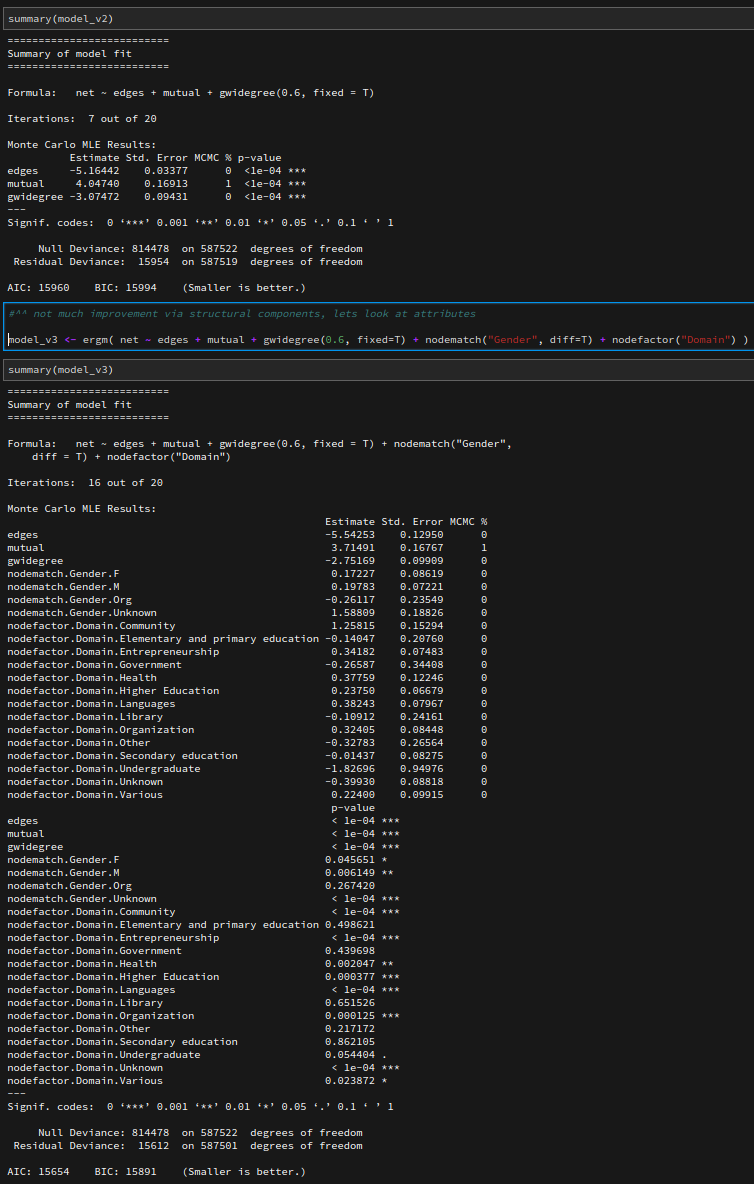


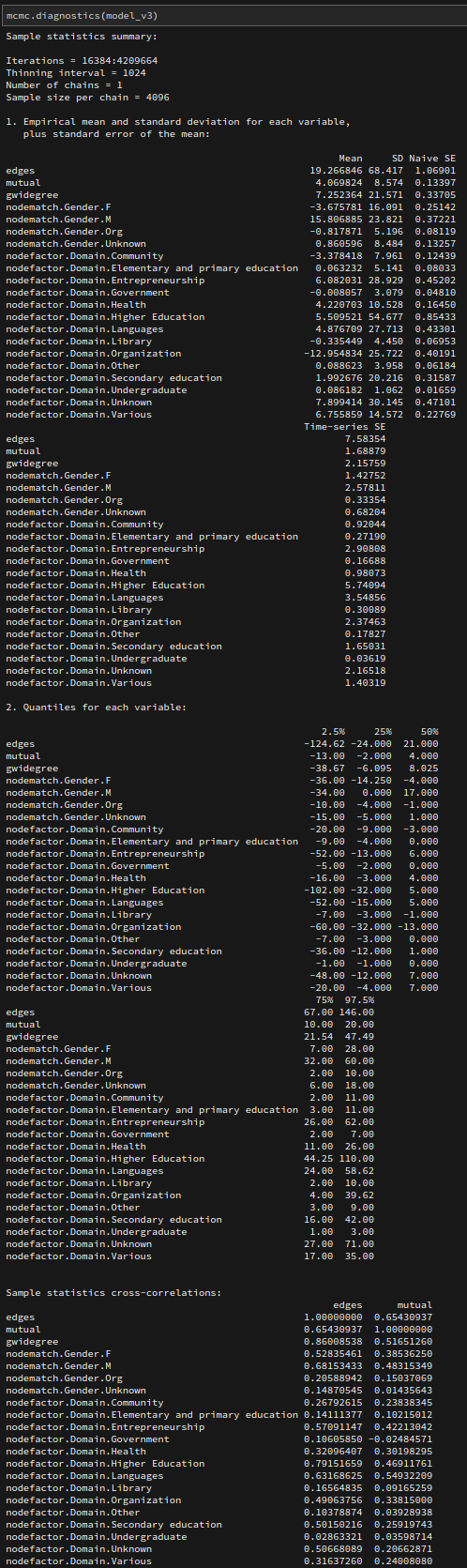


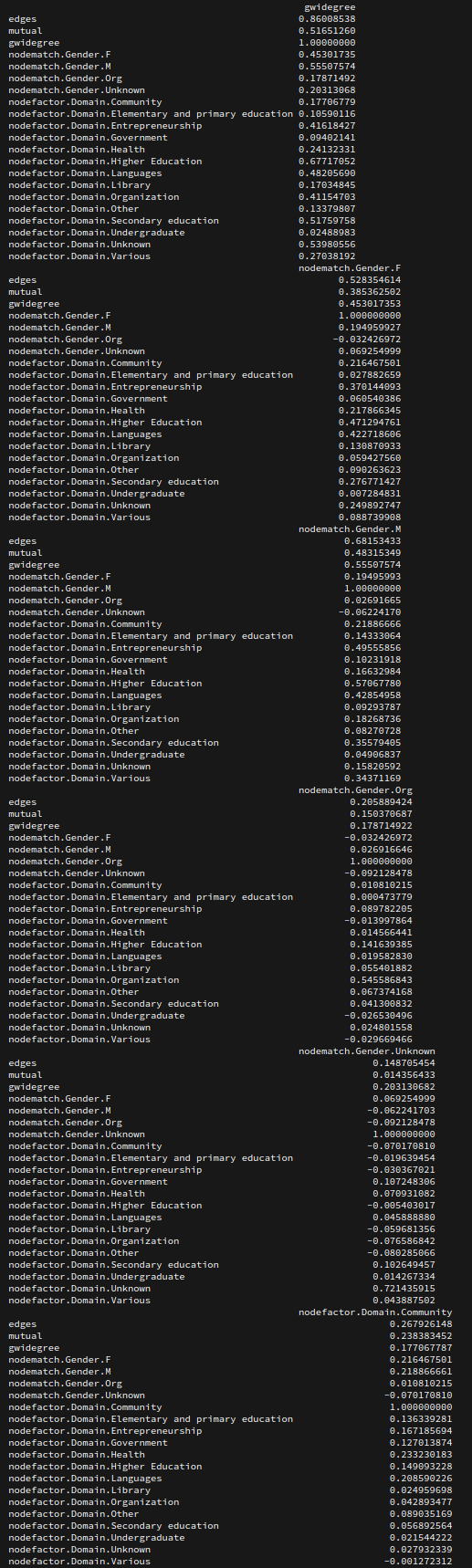


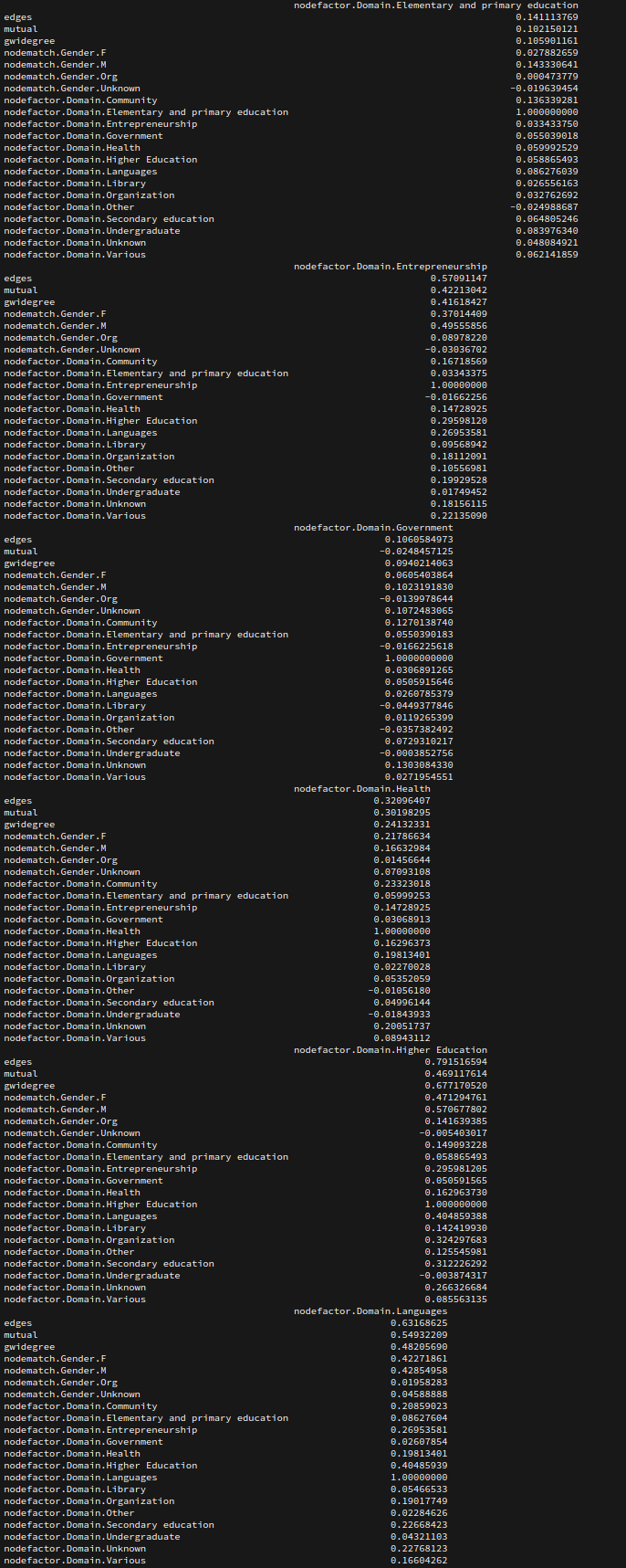


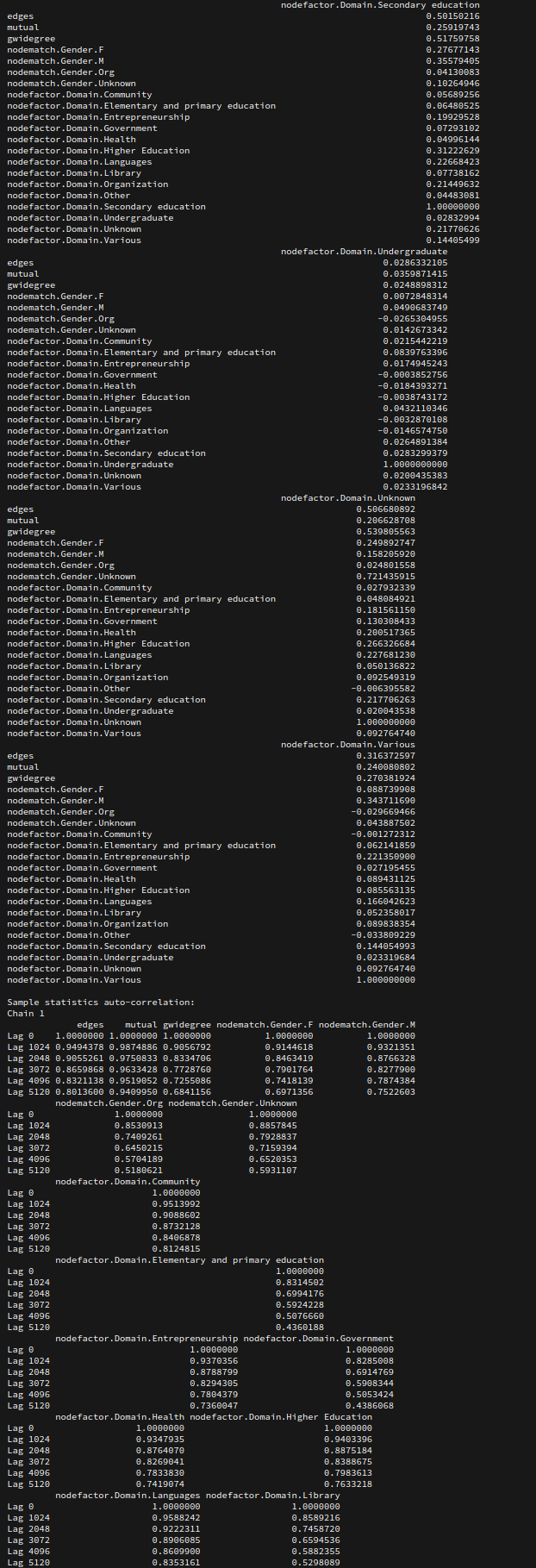


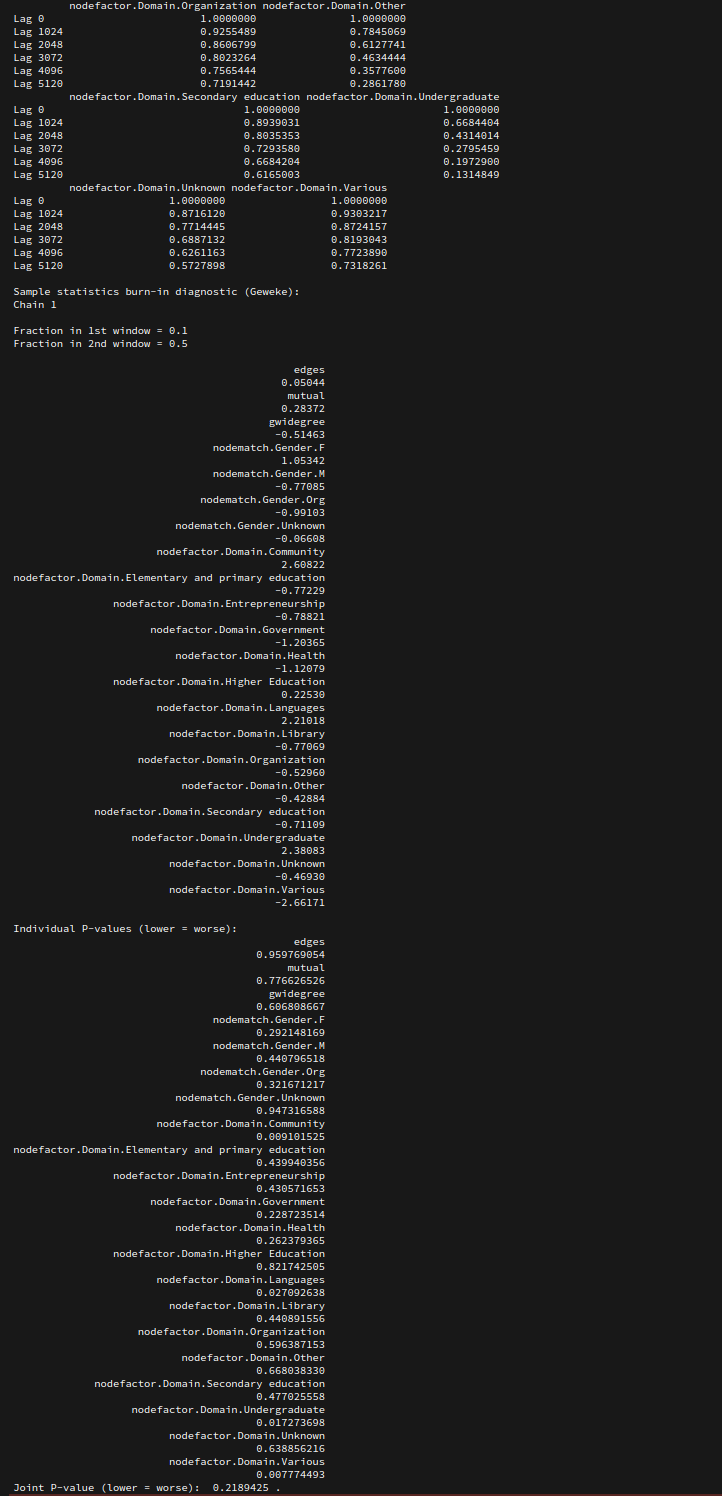




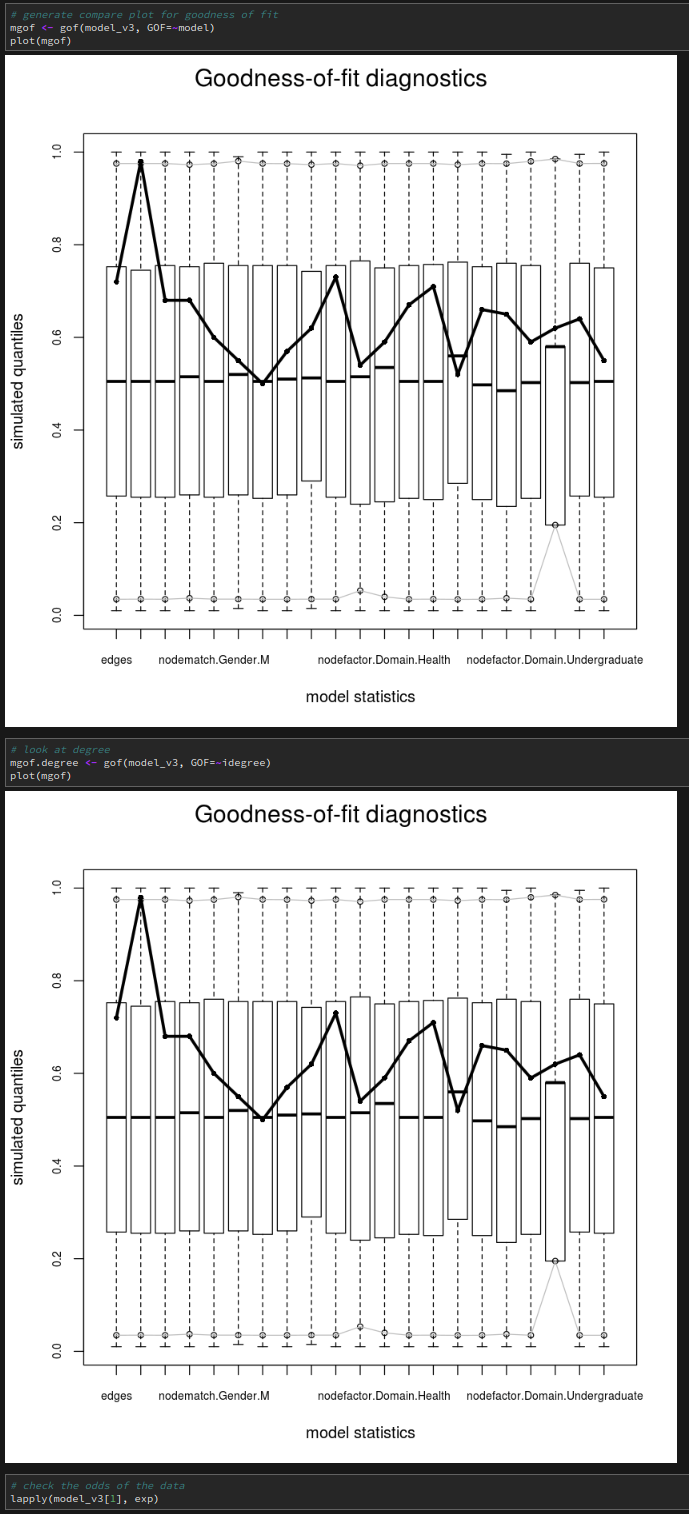


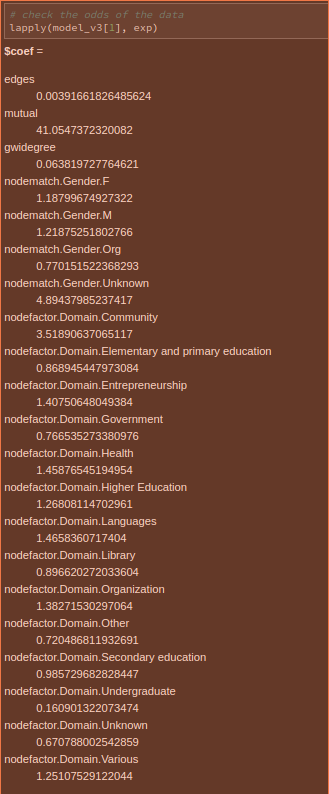






\*\*\*sample stats show in first appendix, removed for brevity \*\*\*





# References

CCK11 Dataset part of Gephi <https://gephi.org/users/download/>

Douglas A Luke, A User’s Guide to Network Analysis in R

J. K. Harris, An Introduction to Exponential Random Graph Modeling

ERGM tutorial <http://www.irrodl.org/index.php/irrodl/article/view/2170/3388>