Network Analysis

KU Leuven

Final Project

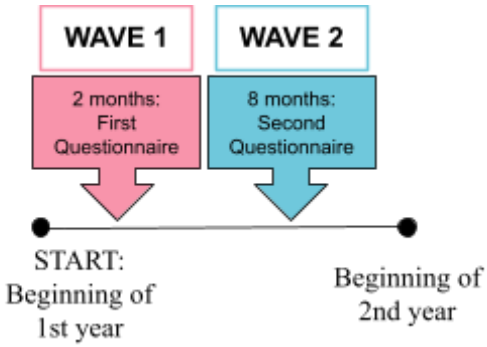
Gabriella Vinco (r0825871)

May 2022

2100

Introduction and Data Description

The data gathered for this report comes from RECENS (Research Center for Educational and Network Studies) where they collected data in a 4-wave longitudinal network survey about 43 Hungarian high-school classrooms spanning over the period of 2010–13. The data was gathered from 7 schools from the capital, one large town, and two smaller towns in Hungary. At the beginning of the study the students were around 14–15 year-old in 2010–11, and they were observed for the first 3 years of highschool. The data we are working with specifically only covers the first two waves which both took place within the first year of their highschool experience. The frequency in which the students were administered the questionnaires are displayed in the timeline below.



Within these questionnaires relationships were evaluated on a 5-point scale: friendship, liking, neutrality, dislike, and hate, where choices were mutually exclusive. Altogether, the dataset contains information on various social ties between students along 40 dimensions in each classroom, but we were assigned to work with the class code 2100. Wave 1 was collected in October 2010, only about a month after the start of the first high-school year, and wave 2 was 6 months later in April 2011. This is an interesting period to examine because we get to see the adaptation to a new environment, new classmates, and start to develop shared views about which behaviors or opinions are acceptable in the class and which are not.

1. Friendship Networks

Our selected classroom consisted of 23 students. Out of those 23 students only 3 were males and the other 20 were females. Had every student answered all the questions there was a possibility of 506 per each wave. For the first wave of data collection there was 8% of the data missing which would lead us to believe that two of the 23 students were absent or decided to not answer the questionnaire. In the second wave there was only 6% of the data missing indicating that there was one student absent or opted out. This is confirmed when looking at the data frames by the pattern in which the non answered questions were presented (whole rows incomplete/nearly incomplete). Considering that it's just one or two

nonparticipants per each wave that would typically be a pretty good turn out but the issue is we already have a smaller class to begin with.

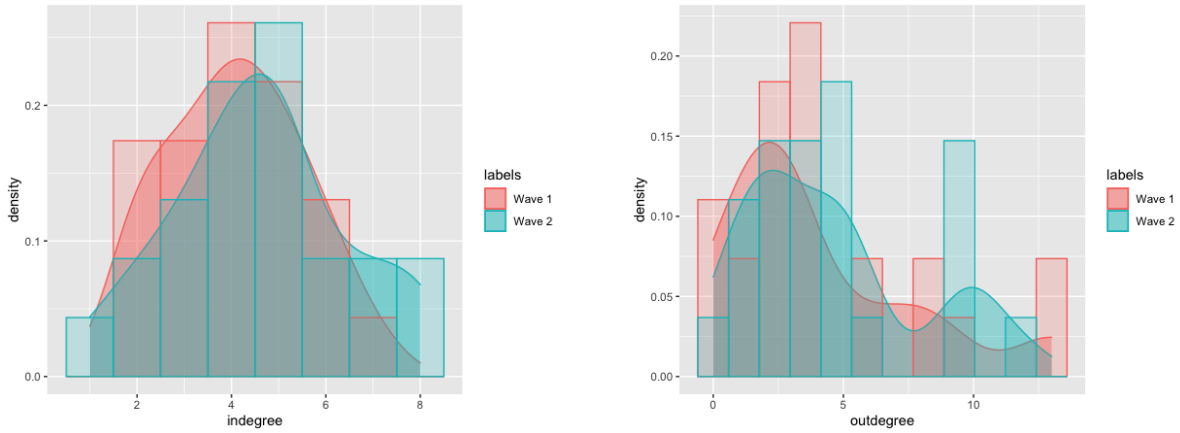


Figure 1.1: Histograms and Density Plots of Indegree and Outdegree for Both Waves

The first step that was taken after re-coding some of the variables was to check for density and reciprocities of the networks. In wave 1 the network density was 0.20 and in wave 2 the network density was 0.22. The reciprocities for wave 1 were 0.32 and for wave 2 were 0.42. The reciprocities needed to be readjusted to compensate for the large amount of zeros in the data. In this step we also examined the histograms and distributions of both the in and outdegrees of both waves as seen in the figure above. The indegree on the left shows slight improvements in student popularity especially with the tail on the right hand side of that plot. It's also similar in the outdegree plot on the right where we can see an increase in students that were more moderately and highly populated in the 2nd wave as compared to the first. In this exploration we also examined the structure of the friendship networks in each of the waves as well as included the other variables (sex and drinking). These network plots can be seen in the figure below.

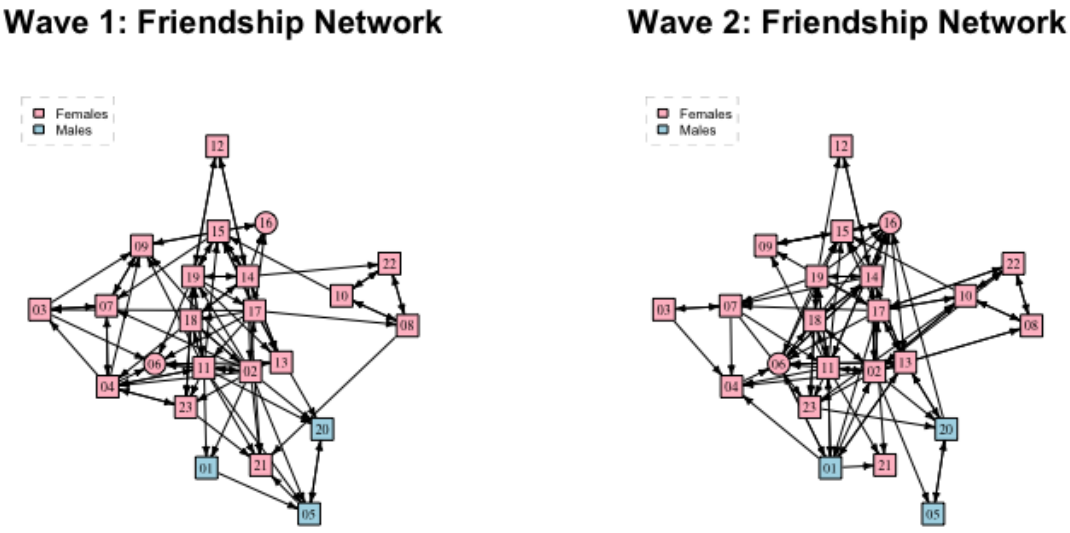


Figure 1.2: Friendship Networks with Color Indicating Sex and Shape Indicating Drinking

In these networks the colors indicate the sex of the students with conforming to gender stereotypes using pink for females and blue for males. Also the shape of the nodes indicate drinking (square) or non drinking (circle), to determine this we re-coded non-drinkers and non-reporters together to form a binary variable between drinking or not. From wave 1 to wave 2 you can notice some differences, for example student 16 has grown in popularity as well as student 06. This is particularly noticeable because they are the two students who responded with a 1 or 0 for drinking which would lead us to believe that they don't drink. The reciprocity increased quite a bit from wave 1 to wave 2 however these are still moderate to low values.

2. Relations Between Networks and Waves

Firstly we will examine the assortativity values and homophily which demonstrates the correlation and attachment between similar nodes. The first aspect we look at is the degree homophily which is -0.03 for wave 1 and -0.08 for wave 2. From this we would conclude that there was no degree homophily. This is not very surprising given the size of our dataset and its interconnectivity, typically it would affect larger less connected networks. In addition to this general evaluation we are going to observe assortativity values among different variables and between the different waves. Another way to gauge the similarity between the two waves is to investigate the Jaccard Index. For our 2 waves of the friendship variable the index was 0.432 which is interpreted as saying the two waves share 43.2%. This varies quite a bit from the results that we got from the Hamming distance matching coefficient with a value of 0.814. This matching coefficient is a ratio of similar items whereas the Jaccard Index determines similarity by 1 to 1 matching.

The first variable we are going to examine is the effect associated with the sex of the students and their friendships. Below in Table 1, we can see the average densities for friends based on the sex of the students. Due to the drastic difference in proportions between sexes in our assigned class we chose to use the normalized average density. However the reported ratios for wave 1 had quite a difference between them. Female students tended to select other female students 8.96 times more than they would select male students.Whereas male students were 132.25 times more likely to select another male student than female. This goes to show that in the first wave the 3 male students really stuck together. In the second wave things balance out a lot more, where the ratio for females selecting other females stays pretty constant at 8.17 times more likely. The ratio for males choosing males has decreased a lot being at only 10.94 times more likely to choose another male over a female. It shows that within that 6 month time period the males became a lot more comfortable becoming friends and developed better friendships with the females in that class.

Table 2.1: *Sex Based Densities*

Normalized Average Density: Wave 1

	Female	**Male**
Female	1.193	0.13
Male	0.019	2.46

Normalized Average Density: Wave 2

	Female	**Male**
Female	1.12	0.14
Male	0.14	1.49

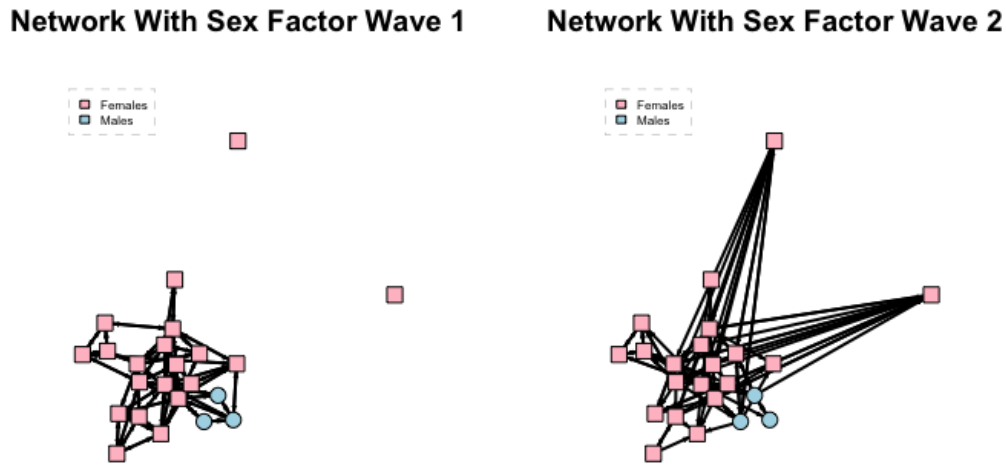


Figure 2.1: Friendship Network with Sex Factor

Table 2.2: *QAP for Friendship and Same Sex*

	Estimate	Exp(b)	Pr(>=|b|)
(intercept)	-2.1669	0.1145	0.00
x1	2.4646	11.759	0.00
x2	0.1838	1.2019	0.57

The next variable we look at is the effect associated with drinking and the relationship it has with student's friendships. Earlier we made the assumption to group the couple non responses (0) with the non-drinkers (1) to categorize into drinkers and non-drinkers. With the network graphs below we are better able to see the individual levels of drinking and their impact with relationships in the network. There is no real clear indication that the drink factor level has any effect on the friendship network because students of various drinking levels are still well connected. We will further investigate this later on in the modeling portion of the project.

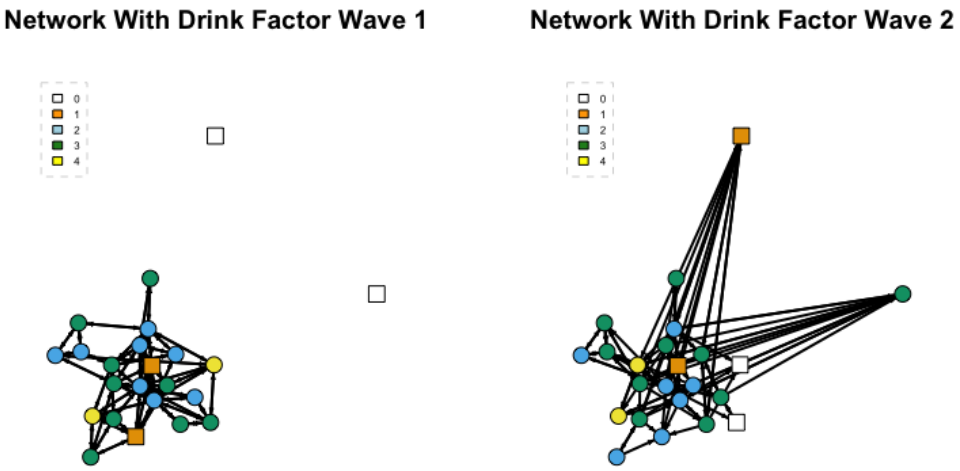


Figure 2.2: Friendship Network with Drinking Factor

3. Micro Patterns in Friendship Networks

We now focus our analysis on investigating the individual structure of each of the friendship networks which we do by using Exponential Random Graph Models (ERGM). This method proved to be the most robust and manageable method so that was why it was ultimately selected over other options. The downside of using the ERGM model is that it cannot handle any missing data. To accommodate this model, we re-coded the missing values to zeros. During the process of running the models we encountered a repeating error with the 'mutual' (reciprocity) effect when running the models. It's still unclear why this happened since we viewed earlier on that there was a moderate level of reciprocity in both waves, but hopefully this provides an explanation as to why it's not included in our models below. The first model displayed in Table 3.1 below is for the data in wave 1. While not all of the effects were significant to a 0.05 level, the last two, nodeicov("sex") and nodeocov("sex") were borderline and included because the close proximity.

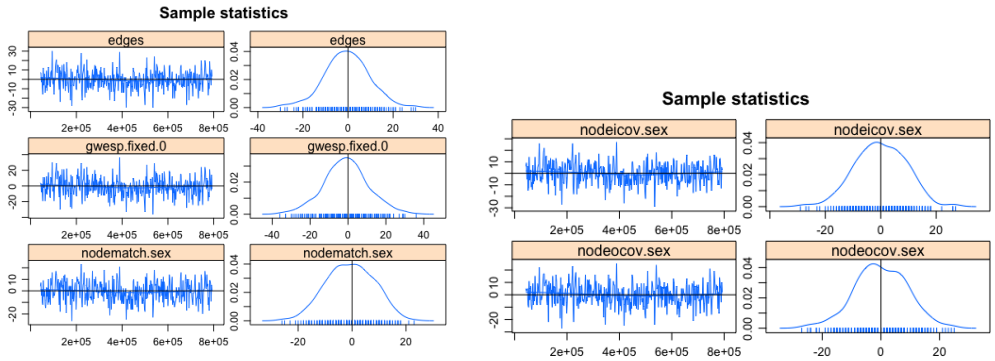


Figure 3.1: MCMC Plots - Wave 1

In the plots above we can see that there are no real trends or anything leading us to believe that the estimates did not converge. The plots all above display a roughly bell shaped curve and the trace plots show variation around 0. We can also see in the table below that the gwesp values are significant and also have a positive coefficient indicating that this model outperforms a randomized model with the other model constraints.

Table 3.1: Monte Carlo Maximum Likelihood Results Wave 1

ergm(formula = friend1 ~ edges + gwesp(0, fixed = T) + nodematch("sex") + nodeicov("sex") + nodeocov("sex") + nodeofactor("drink"))

	Estimate	Std. Error	MCMC %	Z value	Pr(>|z|)	
edges	-2.9663	0.4985	0	-5.950	< 1e-04	***
gwesp.fixed.0	1.5734	0.3447	0	4.565	< 1e-04	***
nodematch.sex	1.4204	0.3668	0	3.873	0.000108	***
nodeicov.sex	-0.751	0.3966	0	-1.894	0.058243	.
nodeocov.sex	-0.7406	0.3863	0	-1.917	0.055223	.
Signif. codes	‘***’ 0.001	‘**’ 0.01	‘*’ 0.05	‘.’ 0.1	‘ ’ 1	

We now move on to wave 2 where we can see once again the bell shaped curves and variability in the trace plot as seen below. We can also see below in Table 3.2 that there is not as strong gender homophily as there was in the previous wave. This makes sense with our previous findings given that there are only 3 males in the class and upon the second wave they started to befriend more girls rather than sticking together.

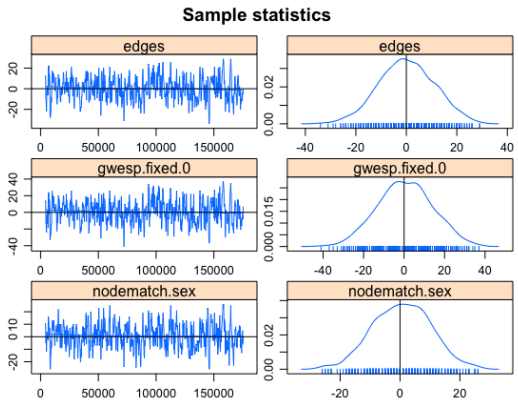


Figure 3.2: MCMC Plots - Wave 2

Table 3.2: Monte Carlo Maximum Likelihood Results Wave 2

ergm(formula = friend2 ~ edges + gwesp(0, fixed = T) + nodematch("sex") + nodeofactor("drink2"))						
	Estimate	**Std. Error**	**MCMC %**	**Z value**	**Pr(>|z|)**	
edges	-3.1924	0.3404	0	-9.378	< 1e-04	***
gwesp.fixed.0	1.1954	0.2624	0	4.556	< 1e-04	***
nodematch.sex	0.4029	0.2043	0	1.972	0.0486	*
Signif. codes	'***' 0.001	'**' 0.01	'*' 0.05	'.' 0.1	' ' 1	

While the drink factor was included in both models it proved to have no homophily. With the interpretation we had the effect of drinking or not drinking had no impact on the friendships. However we do see that gender plays an important role in friendships. Another thing to mention is that although borderline, the in and out stars included in the first wave model are not significant at a 0.05 level. This would conclude that the probability of ties is not significantly affected by the student's popularity or amount of friends they have. As for goodness of fit, the two plots below demonstrate that it falls within the acceptable boundaries and would lead us to believe that there are no issues.

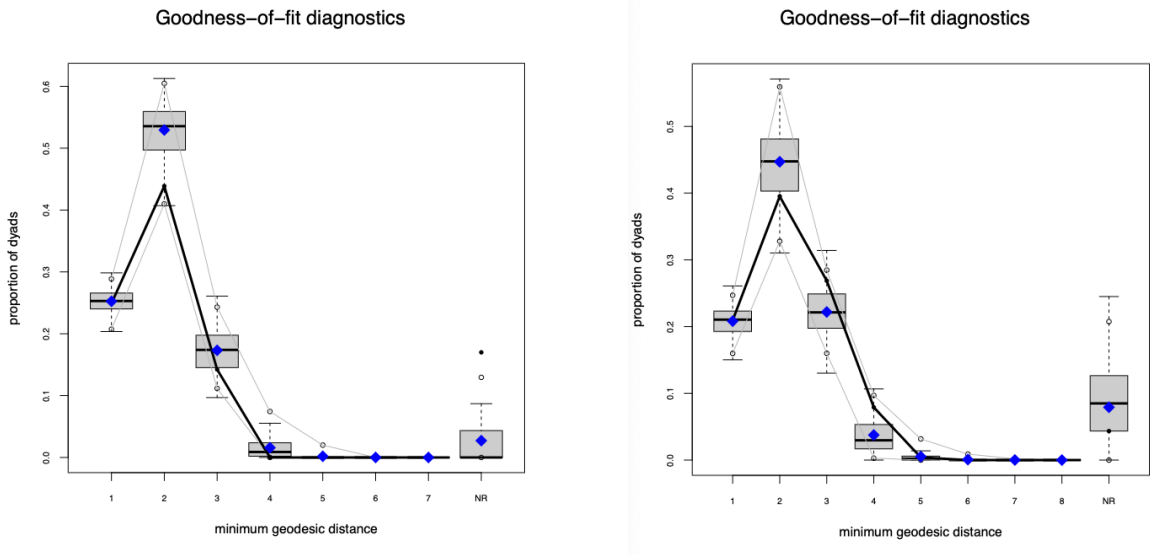


Figure 3.3: Goodness of Fit Plots for ERGM Models

4. Micro Mechanisms of the Friendship Network

The last part of the analysis is observing the micro mechanisms and their effect on the friendship network. We do this by using Stochastic Actor Oriented Models (SAOM) to assess our data. For the following statistics, missing values (if any) are not counted. From Table 4.1 we first take a look at the Average degree values. We can see a prominent decrease going from 6.6 in wave 1 to 4.93 in wave 2. In

context it makes sense since wave 1 was near the beginning of the school year at a new school so I am sure students were looking to make new friends and connections more than they typically would.

Table 4.1: Siena Function Results

Network density indicators:

Observation time	1	2
Density	0.300	0.224
Average degree	6.600	4.930
Number of ties	126	106
Missing fraction	0.170	0.065

The average degree is 5.765

Directed dyad Counts:

Observation	**total**	**mutual**	**asymm.**	**null**
1.	420	126	0	294
2.	456	60	84	312

Tie changes between subsequent observations

periods	**0 => 0**	**0 => 1**	**1 => 0**	**1 => 1**
1 => 2	244	28	66	53

For each of the Goodness of Fit graphs below the points all fall within the interval so there is no concern about fit. These plots indicate that our model is able to reproduce through simulation a product that is accurate to the observed original values. Table 4.2 below gives us a good evaluation of our model. The first thing observed is that none of the convergence t-ratios are greater than 0.1 indicating that convergence is adequate. We can also see that reciprocity is quite a strong factor which in the context of a relationship amongst friends is not a surprising result. Other than that it confirms what we saw in the previously ran models: there is gender homophily and there is no drink homophily.

Table 4.2: Siena Model Results

	Estimate	**Standard Error**	**Convergence t-ratio**
Rate	12.2779	(2.0917)	
Outdegree (density)	-0.0158	(0.8254)	0.0624
Reciprocity	1.8192	(0.4121)	0.0505
Transitive Triplets	0.4051	(0.1111)	0.0378
3-cycles	-0.5507	(0.2282)	0.0381
Indegree - popularity	-0.3685	(0.1686)	0.0581
sex.coCovar alter	0.2941	(0.4834)	0.0399
sex.coCovar ego	-0.1024	(0.3455)	0.0506
same sex.coCovar	0.1467	(0.3653)	0.0755
Overall maximum convergence ratio:	*0.0911*		

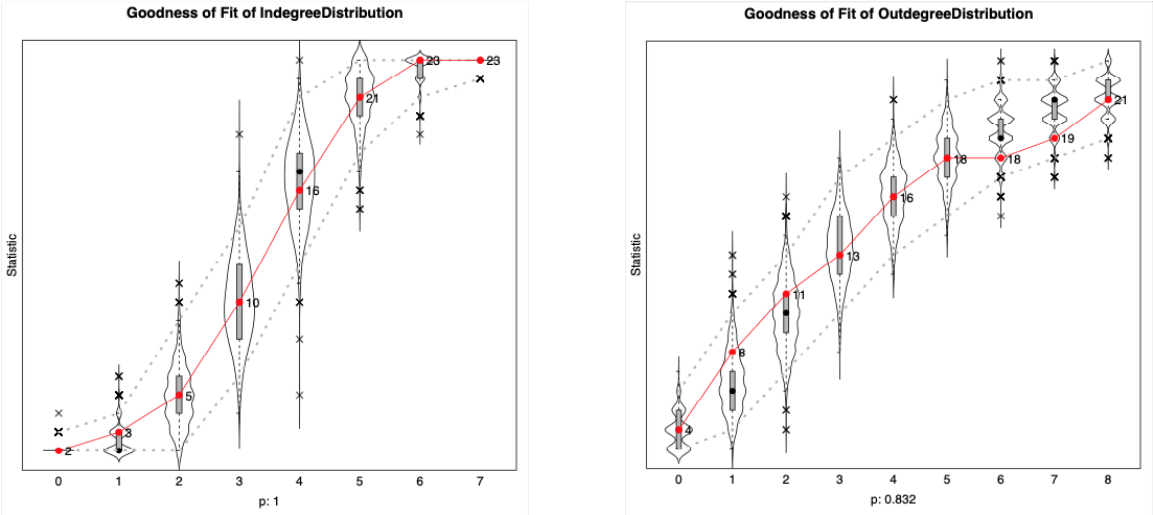


Figure 4.1: Goodness of Fit for SAOM - Indegree (left) and Outdegree (right)

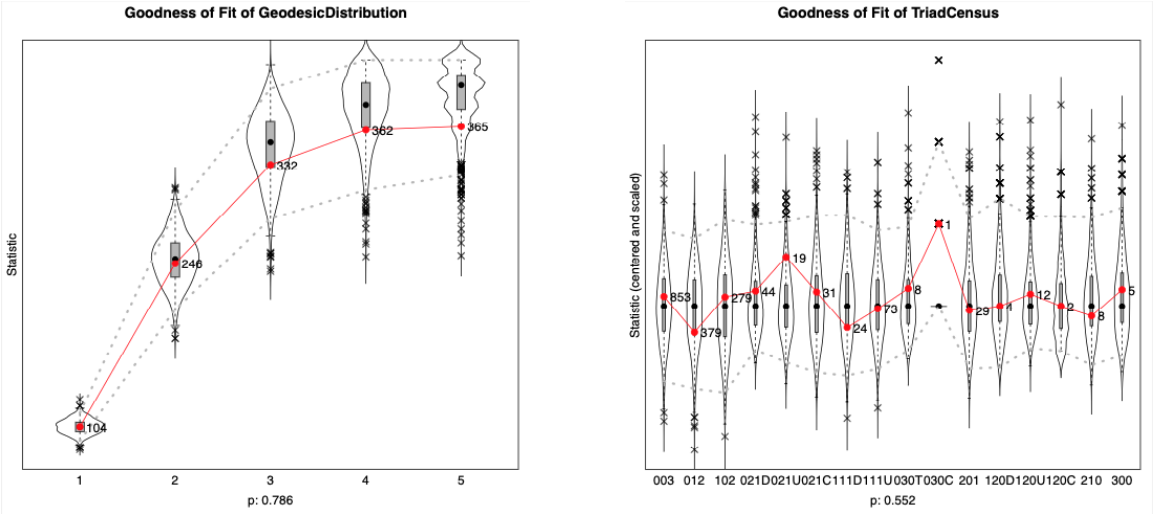


Figure 4.2: Goodness of Fit for SAOM - Geodesic (left) and TriadCensus (right)

5. Hypothesis for Future Tests

For this project we opted out of using the trust data that was provided with the original dataset. It would be interesting to see its interaction with the factors that we have observed in this project. I would hypothesize that it has quite a strong connection to the friendship network, since you wouldn't be friends with someone over a period of time if you didn't trust them. Another addition that would be interesting to see is the further 2 waves. For this project we only dealt with the first two waves which were 6 months apart in the first year. Continuing and seeing the longitudinal evolution of the friendship network would be quite interesting in understanding the social dynamic over the teenage age group. To go along with this

idea we could evaluate how the drinking factor is connected over a longer period of time. While in this project it wasn't very relevant, these were younger teenagers. So my hypothesis would be that drinking plays a bigger part of the friendship network as the group gets older. Last idea for the future would be to have a more balanced dataset in regards to sex of the students. Having such an imbalance of 3 males/20 females made it more difficult to really have a clear understanding on the interaction between the sexes. Had there been more males in the data, we wouldn't have seen such drastic numbers. So in the future it would be nice to view a more accurate overall view on the effect on friendship between the sexes.