Faculty of Science
Master of Statistics and Data Science
Academic year: 2021 - 2022

Network Analysis [G0W14a]

Report

by

Khachatur Papikyan r0825613

Network Description: main characteristics of the friendship network at the two observations, relations between gender, drinking, and friendship in wave 1 and the changes by wave 2

The network to be analyzed consists of 26 students of the classroom 2200 from the RECENS high-school study dataset. Of the 26 members of the network, 7 are boys and 19 are girls. The friendship network is recoded into a binary 0-1 network, where the 1 values represent the presence of ties between actors and the 0 values signify their absence. It is evaluated at 2 time points. It is worth mentioning that there were 2 students absent while evaluating the network at the second time point. The 1-4 scale drinking behavior of the students is also taken into account both at the first and the second time points. In terms of density, the network at first time point has scored 0.17 and at the second time point 0.18, which is interesting since at first glance the network might be considered quite stable as there haven't been observed big changes in the densities between two time points. Unlike the simple density, the number of reciprocated ties tells that indeed there has been quite some evolution in the network since from time point 1 to time point 2 the percentage of mutual ties has increased from 37.8% to 53.6%. Quite an important amount of changes can be observed from the histograms of indegree and outdegree distributions over time, which also signifies that the network has been evolving between the two time points. In particular, table 1 shows that in terms of indegree distribution the network at week 1 reminds more of a random network, which is also logical since in week 1 students are just getting to know each other and seeking for good peer matches and thus they may be trying to make more connections that will later on be filtered out according to certain personal preferences. So in this sense it is expected that there will be changes in the popularity of students meaning that there will be more students with lower level of popularity and less students with higher level of popularity. The results of this filtering out process are depicted in indegree distribution at week 2, where it is shown that indeed a large portion of students has now only indegree of 2 signifying that the students have probably found their own people and made their own small clusters while communicating to the other clusters and to the rest of the network through a small number of popular students with high indegrees. In a sense this may remind of the so-called "Matthew effect", thus the "rich get richer" phenomenon may be suspected. Through the inspection of the histograms for outdegree distribution the "small world" phenomenon may be hypothesized, which is yet to be addressed in order to be further validated.

Through a brief visual inspection of Figure 1 it can be stated that indeed quite some evolutions have been happening throughout time with the network, even though no significant changes were observed while comparing the values of network densities. The network has become considerably denser in the upper right area where the highest concentration of drinkers is depicted, which may be a sign of preferential attachment in terms of drinking behavior and thus,

drinking homophily. The transitive triangle of girls in the upper left corner of the friendship network from week 1 has dissolved over time with significant changes in drinking behaviors of 2 out of 3 girls from this triangle moving from the categories of moderate drinkers to the category of heavy drinkers connecting to the upper right cluster of heavy drinkers through an almost non-drinking boy. But it has to be noted that in the view of the absence of the third girl at the evaluation at week 2, this interpretation is hard to claim valid.

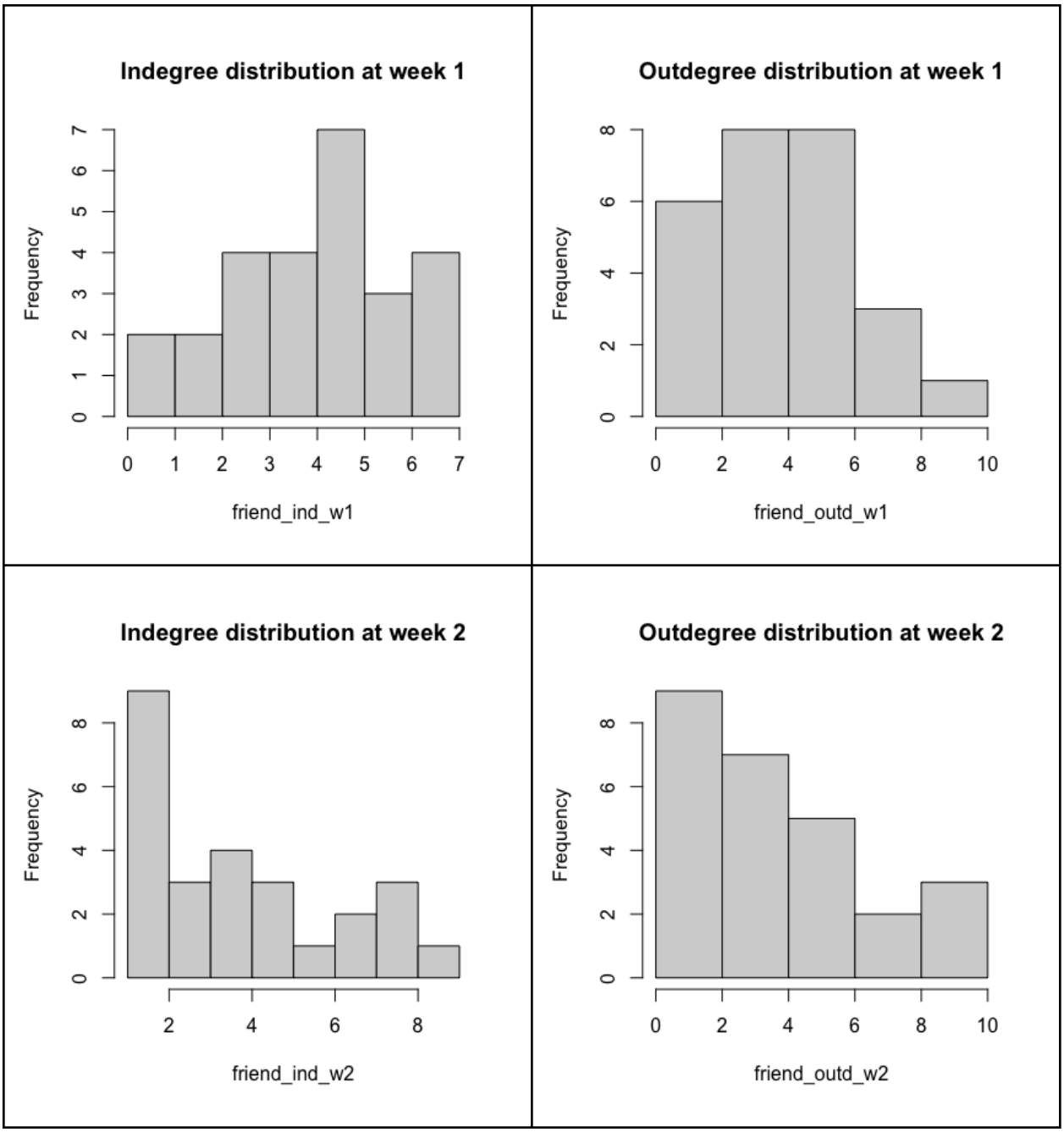


Table 1: Indegree and Outdegree distributions at weeks 1 and 2

The density of the network seems to have considerably decreased in the lower half of the network dominated by girls with certain changes of drinking behavior in the increasing direction. It is hard to visually hypothesize about the presence of gender homophily since it does not really seem to be present graphically, consequently this will later be addressed more formally in order to get a much clearer insight.

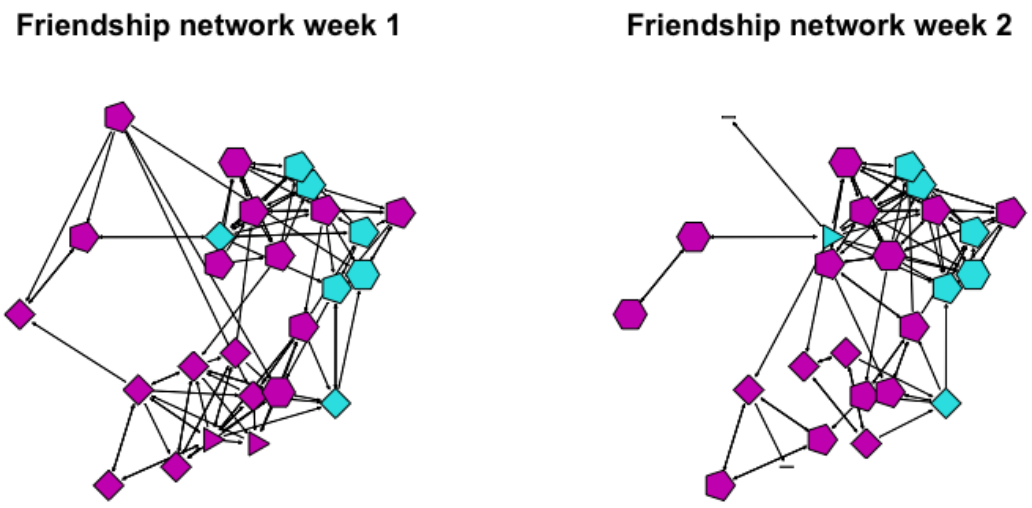


Figure 1: Friendship networks at weeks 1 and 2 (sna package).
Blue nodes: boys, Pink nodes: girls. Node shapes represent the drinking behaviors, the higher the number of vertices, the more the student drinks on 1-4 scale. Triangle: drinking behavior 1, Square: drinking behavior 2, Pentagon: drinking behavior 3, Hexagon: drinking behavior 4. The nodes shown with "–" represent the absent students at week 2.

	Gender homophily	Drinking homophily
Week 1	0.2743724	0.0801256
Week 2	0.1571906	0.1551292

Table 2: Gender and Drinking homophily

Table 2 shows that there does not seem to be strong homophily neither in terms of gender, nor in terms of drinking behavior. The former was stated through graphical inspection as well. It is worth mentioning though, that the coefficients for both gender and drinking homophily have been considerably changed between the two time points. The preferential attachment has

decreased in terms of gender and increased in terms of drinking habits. The latter was also depicted graphically, when stating that the network has become denser in the upper right corner of the graph for week 2 in Figure 1 with a relatively higher concentration of more intensive drinkers. It is also worth mentioning that for the first week there was a stronger preferential attachment in terms of gender rather than drinking behavior, whereas for the second week it is already harder to really distinguish which form of attachment is dominating since the scores for both gender and drinking homophily are very similar with the former having a slight dominance.
Interesting details had popped out when investigating and comparing the friendship selection tables by gender for weeks 1 and 2. In particular, for both of the weeks boy-boy connections were the most popular types of ties due to having the highest densities in the corresponding subgraphs. This was followed by girl-girl connections. For both of the weeks girls were slightly more inclined towards sending ties to boys than the boys. It is also observed that the density of the girl-girl subgraph has decreased while the one for the boy-boy subgraph has increased, which may also be related to the drinking habits of the boys since they might have had more opportunities of spending time together because of that and thus, make more connections with one another.

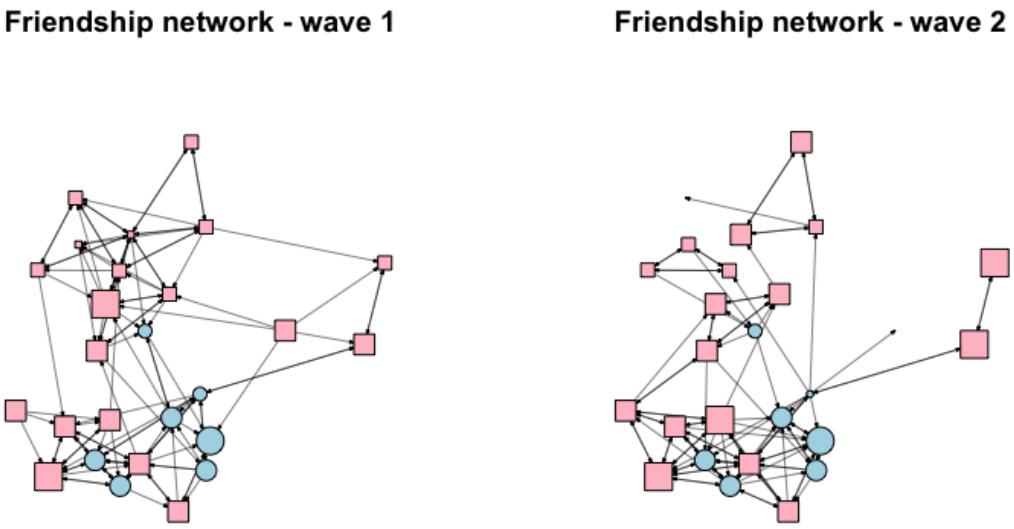


Figure 2: Friendship networks at weeks 1 and 2 (igraph package).
Blue circles: boys, Pink squares: girls. Node sizes represent the drinking behaviors, the bigger the node is, the more the student drinks on 1-4 scale. The 2 empty nodes represent the 2 absent girls at week 2.

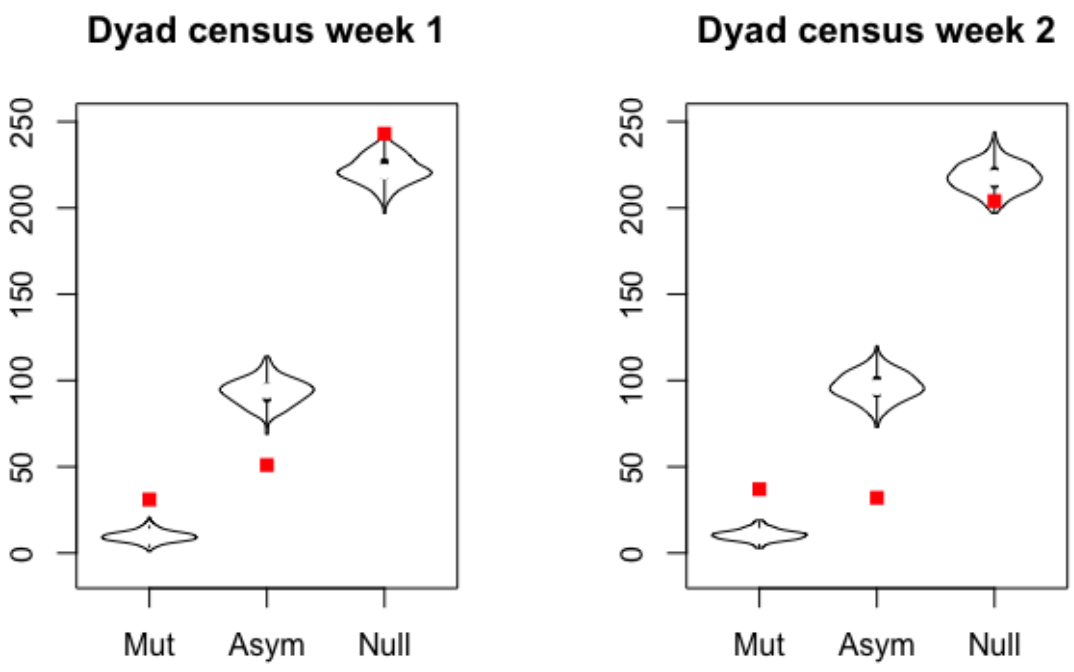


Figure 3: Comparison of dyads of observed vs 200 randomly generated networks at weeks 1 and 2.

Figure 3 shows that the mechanism of forming dyads in the observed networks is significantly different from that of the simulated 200 random networks of the same size and density as the observed network. In particular, the observed network contains significantly more mutual ties (which have been slightly increasing between the 2 time points) than the random networks and significantly less asymmetric ties (which have been slightly decreasing between the 2 time points). Even though the number of no ties of the observed network is not significantly different from that of the random networks, nevertheless, it has been decreasing over time meaning that the network has become somewhat denser.

Figures 4 and 5 show that some of the selected triad types are significantly different in their frequencies from what would be expected by random chance. In particular, the number of simple transitive triads and cycles in the observed network is significantly lower than in simulated networks. On the other hand, there are significantly more triads in the observed network providing more sophisticated types of transitivity and cyclicity such as the ones encoded 210 and 300. Moreover, the number of triads of type 300 has considerably increased over time. This may also be related to the fact that the network has become denser in the upper right part of Figure 1 or in the lower left part of Figure 2. In simple words this may mean that for some reason there is a higher tendency of reciprocation of ties in that region of the network. One may hypothesize that this tendency is related to the fact that the above mentioned is the drinking region.

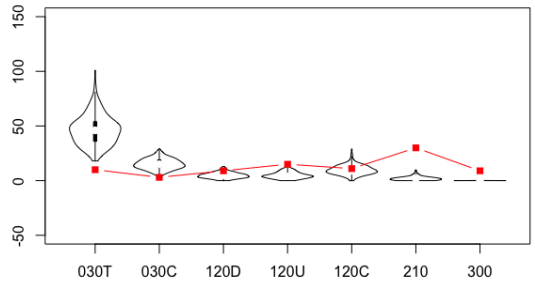


Figure 4: Comparison of triads of observed vs 200 randomly generated networks at week 1

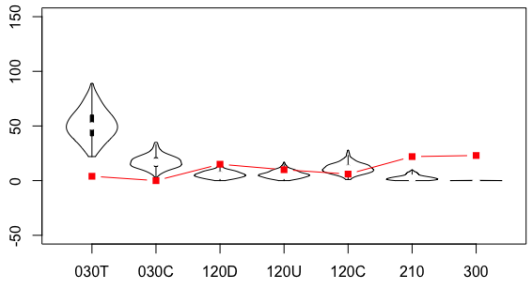


Figure 5: Comparison of triads of observed vs 200 randomly generated networks at week 2

In terms of Hamming distance, only 12.3% of the ties are different between the 2 networks. Analogously, from the point of similarity/stability of the network, the Simple matching coefficient is 0.87. Thus, it seems that it is a quite stable friendship network. However, for sparse networks one will always get high similarity (or low distance), because most of the ties are absent. In this regard, Jaccard index would be more useful. The value of Jaccard index is 0.452, which means that only a bit more than 45% of the ties that existed in at least one observation were stable. So actually, it turns out that the friendship network has been quite dynamically evolving.

It would also be interesting to check whether there can be found certain groups whose members are more similar in terms of who they are connected to. In the view of current assignment, indirect and direct blockmodeling, partitioning based on k-cores and fast-greedy community detection algorithms were used for the network from week 1. Figures 6 and 7 show the cluster dendrogram based on which the number of clusters for indirect blockmodeling were decided, the sociomatrix of the original week 1 network along with the indirect blockmodeling solution with 3 clusters, as well as the solution from the direct approach to blockmodeling (where also 3 clusters were considered by expecting complete or null blocks). It is worth mentioning that the direct approach would technically yield much lower errors if regular blocks were also considered, but that would come at the expense of less interpretability since it was simply making one big regular block while leaving the other blocks to consist of single nodes no matter the number of prespecified expected clusters. What can be observed from the blockmodeling solutions is that there are 2 groups the students are more inclined towards communication within and a third group where the communication takes a more occasional character. The 2 denser groups very rarely communicate between each other, but both of them occasionally communicate with members of the third group where the ties were not so dense.

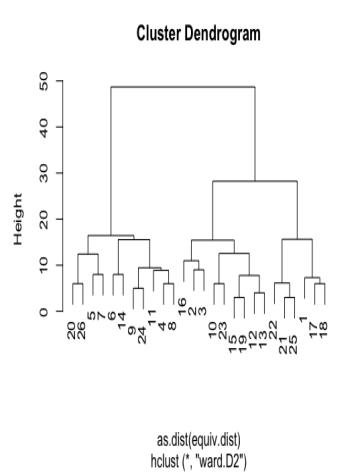


Figure 6: Cluster dendrogram for indirect blockmodeling

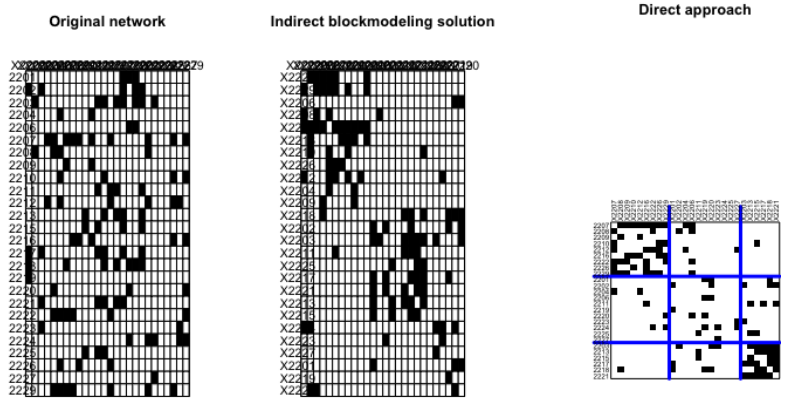


Figure 7: Original network in matrix form along with indirect and direct blockmodeling solutions

Figures 8, 9, 10 and 11 show the clusterings based on different criteria and algorithms for week 1 network. The solutions from the indirect blockmodeling approach and from fast-greedy algorithm are quite similar to each other. This gives a bit more confidence in stating that initially the network could be characterized by 3 clusters.

All in all, almost 53% of the triads in the connected week 1 network are closed triads, which means that it is a relatively well-clustered network, because if there were no clustering (e.g. a chain), then the diameter of this network would be 25 and not 4. So it does seem like a "small world". After generating 200 random networks of similar size and density with that of week 1, the comparison of the transitivity scores of the observed network (0.53) with that of the random networks (0.25) shows that the level of clustering is much higher in the observed network. The diameter of the observed network (4) is relatively longer compared with that of the random networks (3.4). So it would be concluded that it is a highly clustered network, but it is hard to claim that there are different communities that are connected with just a few connections between them (i.e. though there might be articulation points, there are many paths going through them, so it is not only just 1 path connecting different parts of the observed network). So by one criterion it is a "small world", but by the other it is not really.

Concerning the week 2 network, even though there have been observed quite some changes in terms of the closed to open triads ratio (0.55) or the diameter of the network (5), but the fundamental conclusions that were drawn about week 1 network in terms of the network being small world or not have remained the same. In terms of maximal cliques, the number of 3, 4 and 5-cliques has decreased over time, whereas the number of 2 and 6-cliques has increased almost twice and 3 times respectively. Overall, for both of the networks the smallest cliques were the 2-cliques and the largest cliques were the 6-cliques. Interestingly enough, the number of the smallest as well as the largest cliques have considerably increased over time leading to the hypothesis of some students forming groups around a certain phenomenon while leaving the

others that are not so interested in that phenomenon for whatever the reason form their own, much smaller groups by possibly directly communicating with a much limited amount of people (e.g. having only 1 friend). This may, in fact, be considered a very natural selection process since every additional tie is costly and it is natural to communicate to the rest of the network (including the denser part) through some central nodes instead of directly communicating to every single other node. In other words, knowing the right people is to be preferred over spending time and other resources on knowing and befriending with everyone.

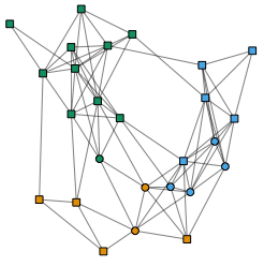


Figure 8: Clusters based on indirect blockmodeling. Circles: boys, Squares: girls

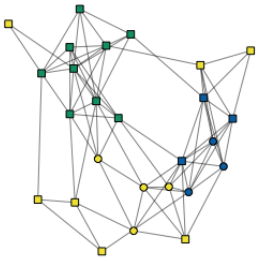


Figure 9: Clusters based on direct blockmodeling. Circles: boys, Squares: girls

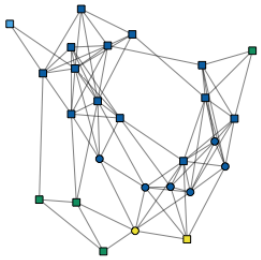


Figure 10: Clusters based on k-cores. Circles: boys, Squares: girls

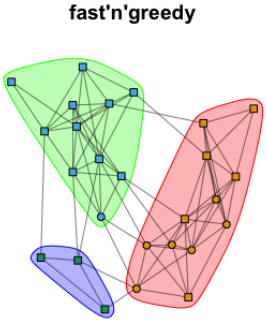


Figure 11: Communities based on fast'n'greedy algorithm. Circles: boys, Squares: girls

Important micro patterns for separately explaining the structures of the two friendship networks (ERGMs)

The comparison of the 2 networks separately through ERGMs yields quite interesting results in terms of the underlying mechanisms generating the networks. In particular, for both of the weeks the students were not inclined towards making no matter what connections, i.e. it comes not as a surprise that every additional tie would be costly and so even the initial connections of week 1 would not come completely at random. The significant and negative coefficients of the parameter estimates for the edges are stating this phenomenon. The probabilities of a tie being present are 9.6% and 3.1% for weeks 1 and 2 respectively. If the tie is present, then in both of the networks there is a high tendency of the tie being reciprocated with probabilities 87% and 96% correspondingly. Moreover, when adding a tie, with 79.3% and 75% probabilities it contributes to a formation of transitive triads. Gender homophily also seems to be playing a significant positive role in both of the networks in a sense that if a tie is created than with 62% and 60% probabilities it is going to be between actors of the same gender correspondingly in weeks 1 and 2. A more interesting picture arises in terms of the drinking behaviors of the students. As a remark it has to be mentioned that for modeling purposes the drinking behaviors of the 2 absent students from week 2 were considered as unchanged and thus, for these 2 students the drinking behavior values were borrowed from week 1. Despite its insignificance in week 1, in week 2 the drinking behavior does come into play. That said, if a tie is created, then it is significantly increasing the chances of the tie to be formed between students having similar drinking habits supported by a probabilistic base of 60.4%. Moreover, the results suggest that the higher value of drinking behavior contributes to the popularity of the student. On the other hand, the significant negative coefficient of the estimate for the outdegree covariate of drinking suggests that the higher value of drinking behavior contributes to the activity of the student in a negative direction. In other words, the more the students drink, the more popular they may be and the more ties they may receive (64% probability for this type of ties), whereas in terms of sending ties it may be less likely to see from their side (39% probability for this type of ties). Overall, the rise of significance of the drinking behavior over time is not so surprising because it had already been hypothesized in the network description part while comparing the corresponding assortativity scores.

The MCMC diagnostics for convergence and the models' GOF checks have turned to be successfully made.

Week 1

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

Monte Carlo Maximum Likelihood Results:

                Estimate Std. Error MCMC % z value Pr(>|z|)
edges          -2.24542    0.81962      0  -2.740  0.00615 **
mutual          1.88631    0.34990      0   5.391  < 1e-04 ***
gwesp.fixed.0   1.34351    0.26004      0   5.167  < 1e-04 ***
twopath        -0.10868    0.04836      0  -2.247  0.02461 *
nodematch.sex   0.48578    0.17456      0   2.783  0.00539 **
nodeicov.sex   -0.33008    0.24244      0  -1.361  0.17337
nodeocov.sex   -0.13158    0.25097      0  -0.524  0.60008
nodematch.drink  0.20560    0.16711      0   1.230  0.21858
nodeicov.drink   0.04476    0.13708      0   0.327  0.74403
nodeocov.drink  -0.09460    0.13385      0  -0.707  0.47972
---
```

Week 2

```
ergm(formula = friend2 ~ edges + mutual + gwesp(0, fixed = T) +
    nodematch("sex") + nodeicov("sex") + nodeocov("sex") + nodematch("drink") +
    nodeicov("drink") + nodeocov("drink"))

Monte Carlo Maximum Likelihood Results:

                Estimate Std. Error MCMC % z value Pr(>|z|)
edges          -3.4339     0.5779      0  -5.942  < 1e-04 ***
mutual          3.1305     0.4415      0   7.091  < 1e-04 ***
gwesp.fixed.0   1.1030     0.2606      0   4.232  < 1e-04 ***
nodematch.sex   0.4259     0.1771      0   2.405  0.01618 *
nodeicov.sex   -0.4632     0.2920      0  -1.586  0.11276
nodeocov.sex   -0.2730     0.2926      0  -0.933  0.35065
nodematch.drink  0.4243     0.1546      0   2.745  0.00604 **
nodeicov.drink   0.5782     0.1891      0   3.058  0.00223 **
nodeocov.drink  -0.4379     0.1897      0  -2.309  0.02096 *
---
```

Table 3: ERGMs for weeks 1 and 2

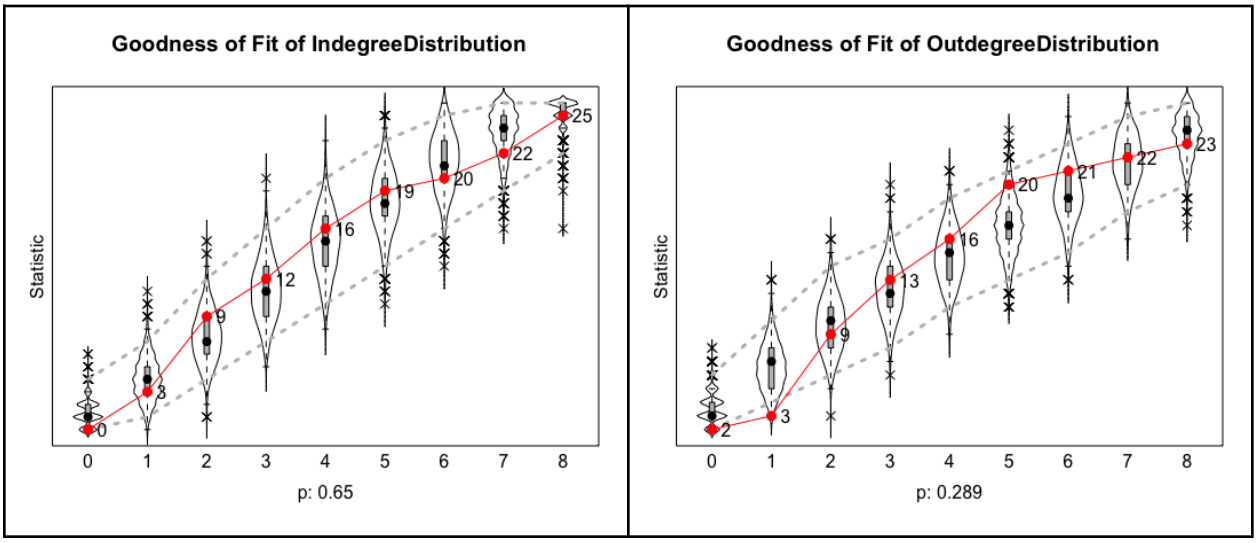
Micro mechanisms shaping the evolution of the friendship network (SAOMs)

In terms of evaluation of the evolution of the friendship network in the current classroom, SAOMs come handy. Here, instead of explaining the structure of the two friendship networks separately, the overall evolution is explained. Some more network descriptives along with information about dependent network, constant actor covariate gender and dependent behavioral variable drinking behavior can be found here. In particular, one might find it useful to add that there were 449 no ties that remained no ties, 42 no ties that became ties, 38 ties that dissolved and 66 ties that stayed untouched from time 1 to time 2. As the Jaccard distance is greater than 0.3, there were no problems with convergence envisaged. The basic rate parameter (6.9886) of the dependent network, which tells how many opportunities of doing something (i.e. of creating a tie, dissolving a tie or doing nothing) each actor got between the 2 time measurements, shows that the network has been facing quite some dynamic changes. The edge parameter outdegree, which also represents the density, will always be negative and does not need to be interpreted, it is a baseline.

The first thing checked after the estimation process is that the maximum convergence ratio is lower than 0.25 meaning that the model has converged and the Siena run was successful. If this value were lower than 0.35, then one might say that the model had almost converged, but then it would perhaps be worth rerunning by using the current results as starting values. The second things are convergence t-ratios, and a good convergence is achieved in this sense as well since all of them are below 0.1 in absolute values. If any of those t-ratios were higher than 0.1 in absolute value, then the model would not be considered as well-converged. Almost okay value could be 0.15, but then rerunning would again be recommended. According to the results of estimation that can be found here, there is a high tendency for reciprocal ties within the evolution of the network. There is also a significant tendency of forming transitive triplets when passing from network 1 to network 2. In the meanwhile, there is a significant tendency of not forming

3-cycles, so the actors are more inclined to have transitive triplets rather than to form 3-cycles, which might be rather strange in social relations. In terms of indegree-popularity it can be stated that those who have high indegree are not so popular (negative parameter value), consequently if one has a lot of friends then there is a tendency that the ties to that person will dissolve, so people will not form new ties to that person (it is something opposite to rich get richer). So, since the indegree-popularity parameter speaks about the popularity of the alters according to their indegrees, if the alter has a high degree, the ego will not connect to him (because popularity is a property of alter and activity is a property of ego). It has to be noted though, that based on the estimation results the indegree-popularity is not significant. Regarding the ego-alter controlled gender homophily effects, from the estimation results it follows that gender homophily does not really play a significant role for shaping the evolution of the network. Something similar can be stated about drinking behavior as well. These results, in fact, seem to be more coherent with the hypotheses drawn from the descriptives of the network.

The GOF plots from Table 4 suggest that the model has been quite a good fit for the evolution dynamics of the network (with the exception of certain configurations of triads). Figure 12 shows the observed week 2 network along with one of the simulated networks based on the estimated model. Certainly the simulated network is not a perfect replica of the observed one, but overall it is not too far from the original as well. Otherwise, it would be too good to be true!



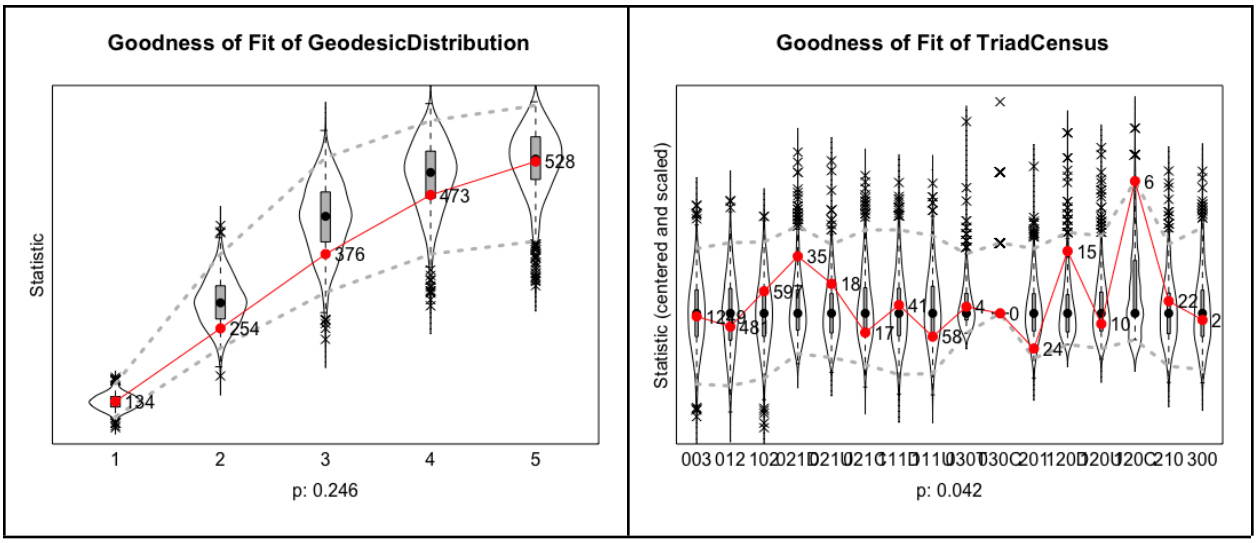


Table 4: GOF plots

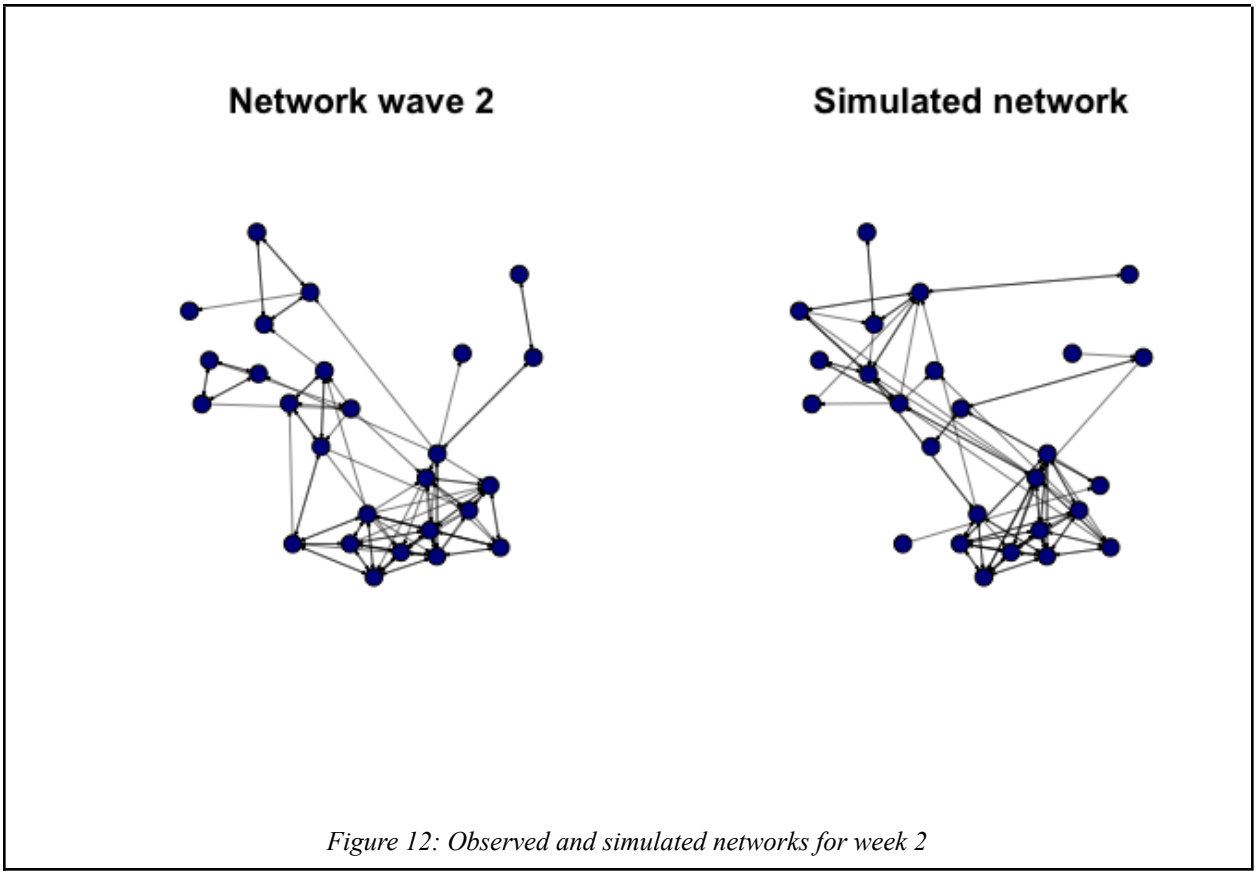


Figure 12: Observed and simulated networks for week 2

Hypotheses about other micro mechanisms that could potentially be tested with the discussed methods

Many other micro mechanisms can be sitting at the heart of the processes generating the observed networks separately or encompassing the temporal evolution of the classroom network as a whole. E.g. one might be interested in diving deeper into hypothesizing about purely structural effects based on internal processes of the network. One such hypothesis could be drawn about the "alternating" star effects stating "Are more ties assumed to matter less?", i.e. checking whether the effect is lowering down with higher degrees. The task of checking this hypothesis can, in fact, be easily accomplished with ERGMs once the corresponding network has been put into undirected form.

Another person might find it useful to further investigate the threetrails and to check a hypothesis whether certain trail types significantly contribute to the formation of the network at one of the distinct time points. In ERGMs this can be done by adding a $threetrail$ term to the model. For a directed network, this term adds four statistics (or some subset of these four specified by the levels argument), one for each of the four distinct types of directed three-paths. If the nodes of the path are written from left to right such that the middle edge points to the right (R), then the four types are RRR, RRL, LRR, and LRL. That is, an RRR 3-trail is of the form i → j → k → l, and RRL 3-trail is of the form i → j → k ← l, etc. There is no requirement that the nodes be distinct in a directed 3-trail. However, the three edges must all be distinct. Thus, a mutual tie i ↔ j does not count as a 3-trail of the form i → j → i ← j; however, in the subnetwork i ↔ j → k, there are two directed 3-trails, one LRR (k ← j → i ← j) and one RRR (j → i → j ← k).

Even though in terms of triadic closure transitivity has been proven to be dominating over cyclical closure, a higher order cyclicity hypothesis check such as the significance of k-cycles with k ranging from 4 to N (i.e. network size) might also contribute to the fit of the model.

Plenty of other assumptions can be made and hypotheses can be checked about other and perhaps more complicated aspects of social networks such as dependence assumptions about edge-triangles, bow-ties, reciprocal path dependence and so on not only with the functions already available in the above-discussed packages, but also with custom made ones.