Network Analysis: friendship relationships in two dutch classrooms

Davide Vandelli

febbraio 15, 2024

The following script is intended as ultimate evaluation for the Advanced Social Network Analysis, A.A. 2023-2024, taught by Filip Agneesens.

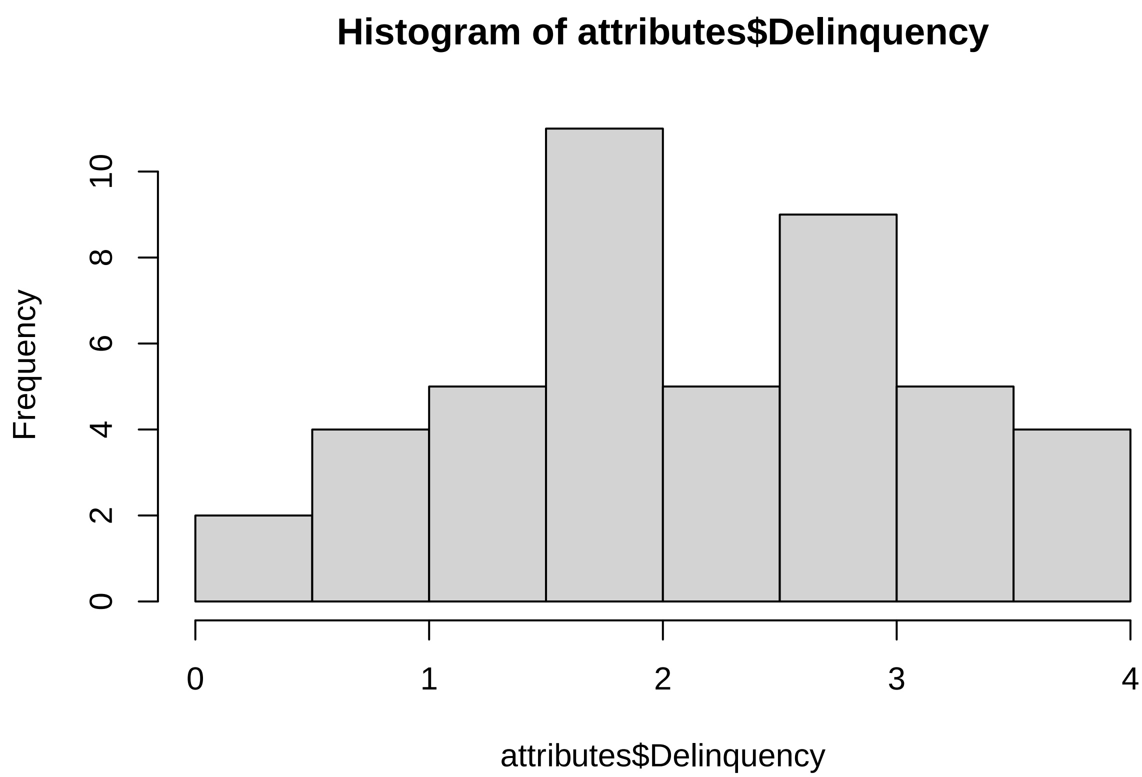
Retrieving data and libraries

```
library('xlsx')
library('GGally')
library('gridExtra')
library('grid')
library('network')
library('sna')
library('ggplot2')
library('igraph')
library('intergraph')
library('RSiena')
library('rgl')
friendship_1 <- read.xlsx("dataset1 exam.xlsx", sheetIndex = 1)
friendship_2 <- read.xlsx("dataset1 exam.xlsx", sheetIndex = 2)
attributes <- read.xlsx("dataset1 exam.xlsx", sheetIndex = 3)
ethnicity <- read.xlsx("dataset1 exam.xlsx", sheetIndex = 4)
```

```
table(attributes$Gender) #1 = Female, 2 = Male
```

```
##
##  1  2
## 25 20
```

```r
hist(attributes$Delinquency)
```




```r
any(is.na(friendship_1))
```

```
## [1] FALSE
```

```r
any(is.na(friendship_2))
```

```
## [1] FALSE
```

To prepare the data, I set the row names as the column "ID", then I label the gender with characters "female" and "male". Ultimately, for visualization purposes, I resize the observations in "delinquency" to slightly more than zero, otherwise the nodes won't appear. There is a total of 45 actors, with attributes *gender*, *delinquency*, *ethnicity* and the ties between them collected in two different times.

```r
ID_column <- friendship_1$ID
rownames(friendship_1) <- ID_column
friendship_1 <- friendship_1[, -1]

ID_column <- friendship_2$ID
rownames(friendship_2) <- ID_column
friendship_2 <- friendship_2[, -1]
```

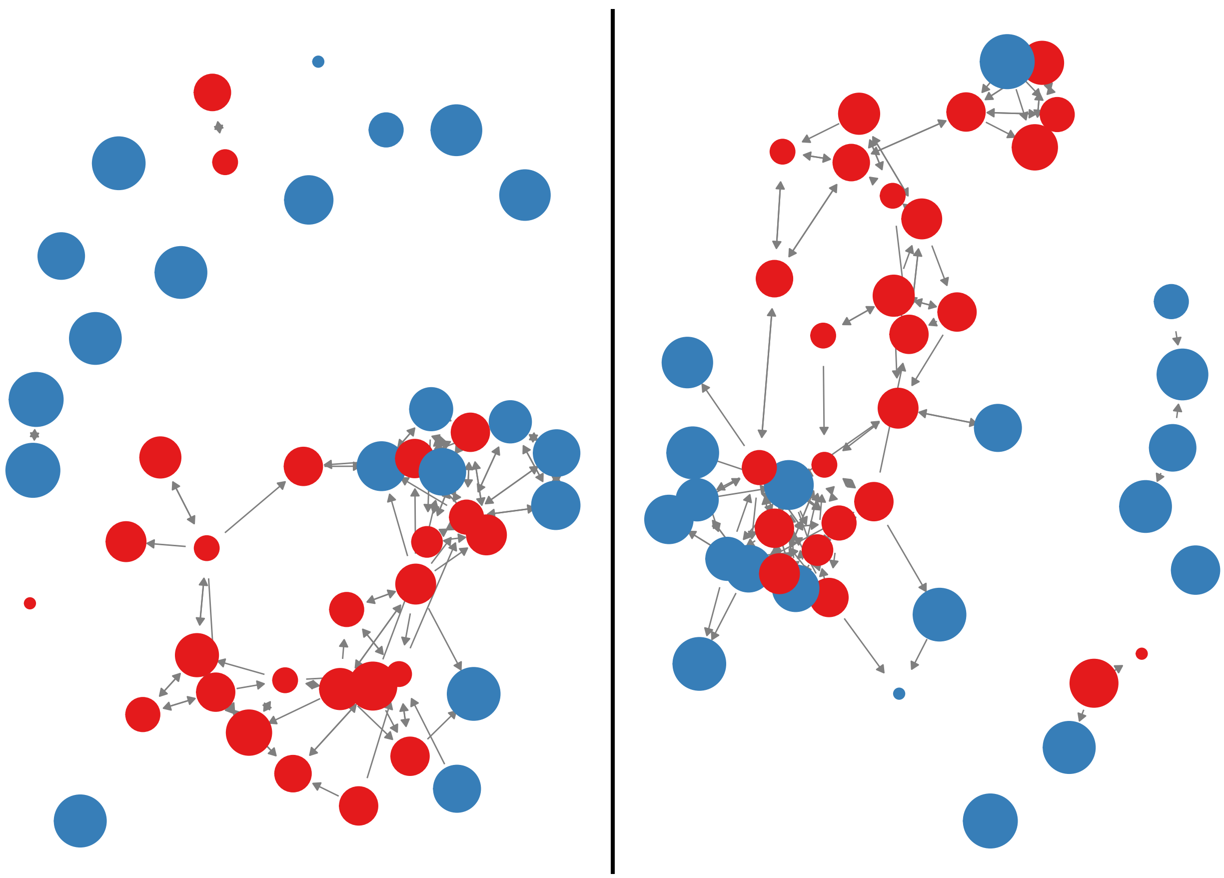
```r
ID_column <- ethnicity$ID
rownames(ethnicity) <- ID_column
ethnicity <- ethnicity[,-1]

gender_ch <- ifelse(attributes$Gender == 1, 'Female',
                    ifelse(attributes$Gender == 2, 'Male', NA))

attributes$Delinquency <- ifelse(attributes$Delinquency == 0, 0.1, attributes$Delinquency)
```

First network visualization

Exercise 1. Delinquency based on node size in a friendship network graph.

As a comparison, we may set aside time 1 and 2 friendship data to visualize delinquency, gender and directed ties. In red, females, in blue males.

On the left: time 1, on the right: time 2.



Reciprocity index

Exercise 2. Both time points 1 and 2. The measure for this network's reciprocity is *edgewise* or *tie reciprocity*, which is the proportion of *edges* which are reciprocated.

```r
sna::dyad.census(friendship_1)
```

```
##      Mut Asym Null
## [1,]  34   36  920
```

```r
sna::grecip(friendship_1,measure = c("edgewise"))
```

```
##       Mut
## 0.6538462
```

There are 34 mutual dyads ($i \leftrightarrow j$), 36 asymmetric dyads ($i \rightarrow j$) and 920 null dyads in the friendship network, resulting in a 0.654 reciprocity index - more than half of the non-null dyads are reciprocated. This represents a network where there are expectations for friendship, and many of them are reciprocated as it is closer to 1 (full mutuality in non-null dyads) than to 0.

```r
sna::dyad.census(friendship_2)
```

```
##      Mut Asym Null
## [1,]  43   56  891
```

```r
sna::grecip(friendship_2,measure = c("edgewise"))
```

```
##       Mut
## 0.6056338
```

There are 43 mutual dyads, 56 asymmetric dyads and 891 null dyads (no tie between i and j) in the friendship network. The reciprocity index is smaller than time 1, indicating a decrease in reciprocation, and possibly a centralization of the friendship network. Compared to the previous time, there are more mutual friendships, but there are also many more asymmetric dyads.

SIENA modelling

Exercise 3 To understand if there are "social forces" within our network that seem to affect how ties are made (or not made) we may implement a SIENA model using its related *r* package. In this scenario we have the possibility to run it as a longitudinal study, since we have two waves. The formula used to evaluate the network is a weighted sum of effects for the whole structure. For actor *i:* $f_i^{net}(x) = \sum_k \beta_k^{net} s_{ik}^{net}(x)$. The betas are the parameters and s are the structural effects. The structural effects are dependent on the network only, and in this scenario the following are used:

- network density (proportion of actual connections to possible connections),
- reciprocity (proportion of a mutual connection between nodes),
- transitive triplets (occurrence of indirect connections between nodes via a shared actor),
- number of three-cycles (triads of nodes in a network),
- squared in-degree popularity (sum of squares of inbound connections to a node),
- squared out-degree popularity (sum of squares of outbound connections from a node),

- squared out-degree activity (sum of squares of changes initiated by a node).

The dependent variable to predict are the friendships, or changes in network ties, between actors and waves. The **delinquency**, **gender** (both modelled as monadic covariate effects, which depend on the actor and alter, using covariate-alter popularity[1] and covariate-ego activity[2]) and the **ethnicity** (modelled as dyadic covariate effect) are added as independent variables.

In SIENA, the parameters for each of the covariates are estimated as the contributions to log-probabilities of increasing the dependent variable by 1 unit when the effect is increased by 1 unit (Ripley et. al, 2024), obtaining data from the ties in different moments of the network's structure and attributes.

Preparing the data is done as follows. First an array of 45 * 45 * 2 composed of the 2 friendship adjacency matrices. It is advised to standardize the numeric variables if their standard deviation is not within a range of 0.1 to 10 (*Ripley et al.*, 2024), but "Gender" and "Delinquency" both satisfy that condition as their standard deviations are $\sigma_G = 0.50$ and $\sigma_D = 0.97$.

```r
Friendship <- sienaDependent(array(c(as.matrix(friendship_1),
                                     as.matrix(friendship_2)),
                                   dim=c(45,45,2)))

Delinquency <- coCovar(attributes$Delinquency)
Gender <- coCovar(attributes$Gender)
Ethnicity <- coDyadCovar(as.matrix(ethnicity))
```

```r
df <- sienaDataCreate(Friendship, Delinquency, Gender, Ethnicity)

effects <- getEffects(df)

#fix(effects)
#effectsDocumentation()

#outdegree, reciprocity, transitive triplets, 3-cycles,
#squared indegree popularity, squared outdegree pop,squared Outdegree activity
effects <- includeEffects(effects, recip, density, transTrip, cycle3,
                          inPopSqrt, outPopSqrtm, outActSqrt)

#ethnicity
effects <- includeEffects(effects, X, interaction1='Ethnicity')

#gender (ego, alter, same)
effects <- includeEffects(effects, egoX, altX, sameX, interaction1='Gender')

#delinquency
effects <- includeEffects(effects, egoX, altX, sameX, interaction1='Delinquency')
print01Report(df, modelname='report_model')
```

No observations went missing from the first to the second wave.

The output for the primary statistics (saved in "report_model.txt") indicate a Jaccard index of 0.352. Since the Jaccard index represents the similarity index between the successive networks, it is a measure of stability. The average degrees increases in the second wave (from 2.311 to 3.156), and the Jaccard index is not low, which is equivalent to say that there should be a low turnover and we may consider the data an evolving network.

[1] Defined by the sum of the covariate over all actors to whom i has a tie, $s^{net}_{i74}(x) = \sum_j x_{ij} v_j$

[2] Defined by i's out-degree weighted by his covariate value.

```r
#effectsDocumentation(effects)
```

```r
SIENA.model <- sienaAlgorithmCreate(projname='SIENA_model', seed=25)
SIENA.results <- siena07(SIENA.model, data=df, effects = effects)
```

Results

Since there are only 2 waves, we obtain one rate parameter. The results are stored, in the short form, as an html file, and in their more verbose form in a .txt medium.

```r
xtable(SIENA.results, file="SIENA_results.htm", type='html')
sink("SIENA_results.txt")
summary(SIENA.results)
sink()
```

```r
SIENA.results
```

```
## Estimates, standard errors and convergence t-ratios
##
##                                        Estimate   Standard   Convergence
##                                                     Error      t-ratio
##
## Rate parameters:
##   0       Rate parameter                 7.7105  ( 1.3207   )
##
## Other parameters:
##    1. eval outdegree (density)          -1.2474  ( 0.5850   )    0.0164
##    2. eval reciprocity                   1.3899  ( 0.3077   )    0.0407
##    3. eval transitive triplets           0.6386  ( 0.1302   )    0.0435
##    4. eval 3-cycles                     -0.0982  ( 0.2271   )    0.0392
##    5. eval indegree - popularity (sqrt) -0.5037  ( 0.2042   )    0.0214
##    6. eval outdegree - activity (sqrt)  -0.1679  ( 0.1455   )    0.0361
##    7. eval Ethnicity                    -0.1233  ( 0.1862   )    0.0494
##    8. eval Delinquency alter            -0.1172  ( 0.1442   )    0.0541
##    9. eval Delinquency ego              -0.0001  ( 0.1463   )    0.0909
##   10. eval same Delinquency              0.0810  ( 0.3961   )   -0.0228
##   11. eval Gender alter                 -0.1926  ( 0.2732   )    0.0342
##   12. eval Gender ego                   -0.0462  ( 0.2789   )    0.0979
##   13. eval same Gender                   0.3545  ( 0.2005   )   -0.0486
##
## Overall maximum convergence ratio:    0.2202
##
##
## Total of 2741 iteration steps.
```

In this setting (and with seed = 25) the overall convergence ratio is 0.22, which means that the algorithm has converged, and it's suggesting that the model reached a relatively high level of convergence, as lower values (closer to 0) indicate a higher convergency; in the rSIENA manual it is advised to interpret results with a convergence ratio lower than 0.30 (*ibidem*). The rate parameter is 7.71 ± 1.32 . It is the estimated number of opportunities for change between the two observations per actor, so it is close to 8 or 7 opportunities of change of friendship per actor.

We may now discuss the parameters. We can test them, through dividing each estimate by its standard error hence obtaining the t-statistics.

Density: -2,13

Reciprocity: 4,53

Transitive triplets: 4,90

3-cycles: -0,43

Indegree pop. : -2,47

Outdegree pop. : -1,2

Outdegree activity: -1,15

Ethnicity: -0,66

Delinquency alter: -0,81

Del. ego: 0,00

Del. same: 0,2

Gender alter: -0,7

Gen. ego: -0,16

Gen. same: 1,77

The first 3 and in-degree popularity are bigger than 2 in absolute value, making them significant at 0.05 level, but the others do not.

"*As all other statistics are correlated with the density, the density is difficult to interpret by itself*" (*ibidem*).

These results indicate evidence for a tendency, within this network and networks similarly structured, for reciprocity and transitioning into a triplet, while also the indegree popularity diminishes. Close to significant levels is the gender being the same between nodes, indicating that there is a preference for same-gender friendships.

Less significant variables are still relevant to interpret as they indicate a lack of influence, and still serve a purpose for understanding "social forces".

"Gender alter" allows to estimate the likelihood of the alter's gender onto the ego, and viceversa does "Gender ego". In an identical logic, the delinquency is also modelled equivalently (and there's a more significant effect on the alter's delinquency rather than one's or the ego's).

3-cycles is the estimate of the generalized reciprocity, which is the contrary of a centralized or network-level hierarchical organization (usually represented as a loop between three nodes).

In-degree popularity utilizes the actor's in-degree ties count to estimate the likelihood friendship, and it is negatively associated, meaning that this sort of popularity doesn't make friends. On the other hand, to a weaker extent, out degree popularity and activity seem to not show a certain propensity with initiating connections with others. More specifically, out degree activity estimates the effect of present outgoing connections towards making new ones, while out degree popularity indicates the effect of the attractiveness of an actor.

Significance with Wald test. To test the significance of the parameters, a multiparameter Wald test is frequently applied to test the null hypothesis. The latter is defined as several parameters are 0. The requirements for this are overall maximum convergence ratio less than 0.25 (in our case it is equal to 0.1904, so close but not violating the constraint) and all t-ratios for convergence must be less than 0.1 in absolute value, which is also true.

```
## Tested effects:
##  Delinquency alter
##  Delinquency ego
##  same Delinquency
## chi-squared = 0.76, d.f. = 3;  p = 0.86.

## Tested effects:
##  Gender alter
##  Gender ego
##  same Gender
## chi-squared = 4.54, d.f. = 3;  p = 0.209.

## Tested effects:
##  Ethnicity
## chi-squared = 0.44, d.f. = 1; one-sided Z = -0.66; two-sided p = 0.508.

## Tested effects:
##  outdegree (density)
##  reciprocity
##  transitive triplets
##  3-cycles
##  indegree - popularity (sqrt)
##  outdegree - activity (sqrt)
## chi-squared = 194.58, d.f. = 6;  p < 0.001.
```

The parameters have been grouped based on their origin. When subgrouped together, having the same gender, and interactions with one's gender and an alter's gender are not significant enough to be evidence to disprove the null hypothesis, and the same can be applied to the delinquency and the interactions with one's delinquency and the alters'. Different is the result for Wald test when performed on the outdegree, the reciprocity, the transitive triplets, and in/out degree squared popularity, with a significantly bigger chi-square and smaller (and definitely significant) p-value with respect to the others.

`{r echo=FALSE Multipar.RSiena(SIENA.results, 7, 8, 9, 10, 11, 12, 13)`

When the constant covariates (dyadic included) are grouped for the Wald test, the test significance of their effects altogether increases, as the p-value is smaller, but not enough to be considered significant at any usual confidence level (p-value > 0.05). As an additional insight, we may test the hypothesis that the effects by parameters of "delinquency alter" and "gender alter" are the same as their "ego" counter part (same covariate, different node focus).

```r
testSame.RSiena(SIENA.results, c(8, 11), c(9, 12))
```

```
## Tested effects:
##  Delinquency alter  ==  Delinquency ego
##  Gender alter       ==  Gender ego
## chi-squared = 0.98, d.f. = 2;  p = 0.614.
```

It appears that they are not the same, hence their effects are different.

Additional comments The SIENA model is relatively complex to set up as a first-timer. Given this context, its potential is acknowledged, possibly with more theory on the meaning of its results.

Practicing these exercises was very helpful to understand more about the ASNA course, especially by enjoying rSiena manual, as this branch of statistical analysis is different from the typical analysis that the author practices.

References

- *Package 'RSIENA'* [digital manual], version 1.4.1., Jan, 2024. Accessed: Feb, 2024. [Online]. Available: https://www.stats.ox.ac.uk/~snijders/siena/RSiena_Manual.pdf

- Several materials (presentations, lecture scripts, etc.) from the course Advanced Social Network Analysis, A.A. 2023-2024, University of Trento, Sociology Department, taught by Filip Agneesens.