Network Modeling HS 2023
C. Stadtfeld, A. Espinosa-Rada, A. Uzaheta

Assignment № 2

released: 20.11.2023 at 19:00 **due:** 7.12.2023 at 12:00

Task 1: Network evaluation function 4 points

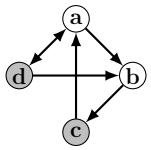
Consider a SAOM which objective function is specified by: the following statistics

$$f(i, x, \beta) = \beta_1 s_{1i}(x) + \beta_2 s_{2i}(x) + \beta_3 s_{3i}(x) + \beta_4 s_{4i}(x, v)\,,$$

with $s_{1i}(x)$ the out-degree (density), $s_{2i}(x)$ the reciprocity, $s_{3i}(x)$ the transitive reciprocated triplets (the reciprocated tie is the tie $i \leftrightarrow j$) and $s_{4i}(x, v)$ the same covariate effects. Where x_{ij} denotes the presence or absence of a tie between actors i and j, and v_i denotes a covariate value for actor i.

(1) Give the mathematical formula for each effect. You can use the RSiena manual to look for the formulas.

(2) Given the current state of the network, with the colour of the nodes representing a binary attribute taking categories 1 (white) and 2 (gray),



and given that $\beta_{1i} = -1.5$, $\beta_{2i} = 2$, $\beta_{3i} = 1$ and $\beta_{4i} = 1.5$, what is the probability that in the next mini-step:

i. actor ⓒ adds a tie to ⓑ?

ii. actor ⓑ adds a tie to ⓐ?

iii. actor ⓐ deletes the tie to ⓑ?

iv. actor ⓓ does not change anything?

Task 2: Simulations from SAOM **8 points**

The file `simSAOMs.R` contains the code to simulate the network evolution between two observations from a SAOM with an evaluation function specified by outdegree, reciprocity and transitive triplets effects statistics. It also includes the code to produce violin plots for the triad census counts.

(1) Implement the missing code so that the function `simulation` can be used to simulate the network evolution. Document the code. The algorithm is described in the file *Simulating from SAOM* available in the Lecture notes and additional material section on Moodle. Unconditional simulation is used.

Hint: a useful function for the implementation is `sample`
Tip: If you want to implement an efficient code for the simulation, you can use change statistics (How much would the statistic change if the tie is toggled?). In this way, creating the network with the toggled tie is unnecessary for computing the effect statistics. For transitive triplets, it is useful to consider that a tie from a sender can play one of two roles in this network structure.

(2) Consider the two adjacency matrices in the files `net1.csv` and `net2.csv`. They are observations of two networks collected on a set of 22 actors at time t_1 and t_2, respectively. Estimate the parameters of the SAOM with outdegree, reciprocity and transitive tiplets statistics using the function `siena07`.

(3) Conditioning on the first observation, generate 1,000 simulations of the network evolution using the function `simulation` developed in (1) and setting the parameters with the results of the model estimated in (2). Compute the triad census counts for each simulated network. Save the results in an `R` object[1], named *triadCensus*, in which rows are the index of a simulated network and columns are the type of triads.

Hint: use the function `triad.census()` from the `sna` package to compute the triad census counts.

(4) Use the simulated values of the triad census counts to evaluate the model's goodness of fit. The second part of the code was written to this aim, complete the missing pieces of code to produce the violin plots. Additionally, write the code to compute the Mahalanobis distance and the p-value used in `RSiena` to assess the fit of the model with respect to the triad census auxiliary statistic. Remember to drop statistics with variance of 0 for the plot and Mahalanobis distance computation, report which statistics suffer this issue. The code should compute the following quantities:

[1] You can use an object of class matrix/arrays or data frame

i. standardize the simulated network statistics, i.e., centered and scaled values of each type of triad given in the *triadCensus* object. Named the resulting object as *triadCensusStd*.
(The centered and scaled values are computed as $x_{\text{std}} = (x - \bar{x})/\sigma_x$ with $\bar{x}$ the average and σ_x the standard deviation of the simulated distribution).

ii. the variance-covariance matrix of the standarized simulated network statistics $\hat{P}$ and its generalized inverse.
Hint: useful functions are `cov()` and `MASS::ginv()`

iii. standardize the observed values of the triad census counts in the second observation with $\bar{x}$ and σ_x as in i.

iv. compute the Mahalanobis distance for each simulated and the observed network using the standardized values (computed on i. and iii.).
(The Mahalanobis distance is computed as $x_{\text{std}}^T \hat{P}^{-1} x_{\text{std}}$)

v. compute the percentage of simulated networks with Mahalanobis distance equal or greater than the observed network Mahalanobis distance.

Run the complete code to obtain the violin plots and the test on the Mahalanobis distance. Would you think that the model has a good fit based on the triad census auxiliary statistics and the p-value compute in (4)? Justify your answer.

Hint: a useful function to apply the same function to the rows or columns of a data frame(array) is `apply`

(Please do not modify existing code even though more efficient solutions can be implemented)

Task 3: Estimation and interpretation of SAOMs **8 points**

The folder `Glasgow.zip` contains data collected by Michell and West (1996) under the "Teenage Friends and Lifestyle Study" [2]. The dataset was collected on a cohort of 160 students followed over two years starting in February 1995, when the pupils were aged 13, and ending in January 1997. The friendship network of the pupils was observed at three-time points. Pupils were asked to name up to six friends and provide information on their socio-demographic characteristics along with the use of substances, such as tobacco and alcohol consumption. In the following, we analyse the data of the 129 pupils who were present at all three-time points.

The folder contains the following files

- `f1, f2, f3.csv`: adjacency matrices of the friendship networks
- `demographic.csv`: data frame containing information on gender (1 boy, 2 girl) and age
- `logdistance.csv`: logarithm of the distance (in kilometers) between the houses of the pupils
- `alcohol.csv`: alcohol consumption coded as 1 (non), 2 (once or twice a year), 3 (once a month), 4 (once a week) and 5 (more than once a week);

(1) Let us start by considering the friendship network as the only dependent variable

(1.1) Compute the Jaccard index to evaluate if the data contains enough information to investigate the evolution of the friendship network. Comment on the results.

(1.2) Specify a reasonable model to test the following hypotheses on the friendship evolution:

i. Students tend to be friends with popular pupils
ii. Students tend to be friends with pupils with similar alcohol consumption to their own
iii. Students tend to be friends with students that live in the same neighborhood (living nearby)

Do not forget to control for the basic endogenous and exogenous variables.
Hint: take a look at the practicals on SAOM and on the introductory paper on SAOMs for inspiration

[2] Data description and download from `http://www.stats.ox.ac.uk/~snijders/siena/siena.html`

(1.3) Estimate the model, check its convergence and fit, and comment on its parameters.

(1.4) Are the hypotheses i.-iii. supported by the data? Argue for your answer.

(2) We now investigate the co-evolution of friendship (network dependent variable) and alcohol consumption (behavioral dependent variable).

(2.1) Use the model specification developed in (1) for the selection part of the model. Specify a reasonable model for the influence part to test the following hypotheses:

iv. Popular students tend to increase or maintain their level of alcohol consumption

v. Students tend to adjust their alcohol consumption to that of their friends

(2.2) Estimate the model, check its convergence and fit, and comment on its parameters.

(2.3) Are the hypothesis iv.-v. supported by the data? How the conclusions about the test on the hypothesis i.-iii. differ with respect to the findings in (1). Argue for your answers.

(2.4) Given the model estimated in point (2.2.), do we have evidence for selection processes only, influece processes only, or both selection and influence processes? Argue for your answer.

(3) Discuss how the model in (2) could be improved (by adding new effects) so that geodesic distances and degree distributions might be better represented. Provide theoretical justification for the new effects you propose to add in the model specification.

You are encouraged to work in groups of 3 or 4 people. It is a requirement for the submission to belong to a group in Moodle.
Please submit your solution in a ***PDF****. It should contain all the plots, results, comments, and answers to the tasks in the assignment. The PDF should be named Assignment02_GroupXX.pdf; for example, for group 9, the file name is Assignment02_Group09.pdf (Groups with numbers 1 to 9 pad a zero on the left, 01 to 09).*
As companion files to the PDF submission, please submit one or more `R` *scripts (or an* `Rmd` *or* `qmd` *file) with the code used to generate the results presented in the PDF.*
Remember to put your names on the PDF and `R` *scripts you submit to Moodle. Only one member of the group should submit the solution. Do not forget to report the names of all the group members in the documents you submit, PDF and* `R` *scripts.*