Assignment № 1

released: 02.03.2026 at 20:00 **due:** 25.03.2026 at 12:00

Data Description: Czech School Data

The folder `Lintner` contains data on friendships from a sample of 276 Czech Grade 6 students. We provide a subset of eight classrooms from the twelve that were collected, along with the supporting student-level demographic, and literacy data. Gathered longitudinally at the beginning and end of the 2021/2022 school year, the data provide a relational insight into the nature and evolution of early adolescents' friendships. The following information are provided:

- `XX_WY.csv`: a data frame containting the friendship network at wave `Y` for classroom `XX`. For example, `10_W1.csv` contains the friendship network of classroom 10 at wave 1.
- `attr.csv`: a table containing individual attributes of all students in the sample.
  - *studentid*: the identifier of the student
  - *classroomid*: the identifier of the classroom
  - *gender*: self-reported gender (coded as "female" and "male")
  - *HISEI*: a measure of socio-economic status. The scale has a minimum of 16 and maximum of 90.
  - *literacy_end*: the literacy score at the end of the data collection.

Note on data cleaning! An important part of Social Network Analysis when working with empirical data is data processing and data cleaning. The data we have provided contains missing data (NA) in both the networks and the attributes. For this assignment, we ask that you clean your data appropriately. For task 1, it is sufficent to assume that missing values are 0. In tasks 2 and 3, we ask that when you handle missing data, you do so by removing the affected data from the dataset rather than imputing the missing data. As an example, if node i has missing data on only one variable which you require for a given analysis, you must remove it from all the required variables. In networks, this is done by removing the associated row and column. For monadic attributes, this is done by removing the associated row. Which parts of the data require cleaning depends on the analysis you are carrying out (i.e, if you are not using a certain variable in a model, missing data in that variable is not relevant). Please also remember to convert individual attributes as necessary in order to be able to include them in your analyses.

Task 1: Tie Dependence and CUGs **7 points**

Please use set.seed(161) and 5,000 permutations to solve the assignment

Let (Ω, P) be a probability space, Ω the set of events and $P : \Omega \to [0, 1]$ the function associating to each event in Ω its probability. Two events $A, B \in \Omega$ are statistically independent if

$$P(A \cap B) = P(A) \cdot P(B) \ .$$

If $\Omega = \{X_{ij}, i, j \in N\}$ is the set of all the possible ties in a binary network, ties are independent if

$$P(X_{ij} = a \ \cap \ X_{hk} = b) = P(X_{ij} = a) \cdot P(X_{hk} = b) \quad \forall \ i, j, h, k \in N, \quad a, b \in \{0, 1\}$$

The file `10_W1.csv` in the folder contains the adjacency matrix of friendship ties between 22 pupils at the first wave of observation. There is a tie from pupil i to pupil j if pupil i considers pupil j to be a friend.
The number of ties in the network is $m = 134$.
The mutual, asymmetric and null dyads are $M = 42$, $A = 50$, and $N = 139$.

(1) Compute the probability p of observing a tie if you choose one possible pair of nodes at random.

(2) Compute the probabilities p_M, p_A, and p_N of observing a mutual, asymmetric and null dyad.

(3) What would be the value of the probabilities in (2) if we assume tie independence and that the probability of observing a tie takes the value p computed in (1)?

(4) Given the values obtained in (2) and (3), would it be reasonable to assume tie independence when performing further analysis on the observed network? Justify your answer.

(5) Consider a different model that assumes all nodes have an outdegree of nine ($d = 9$). Compute the probability of a mutual tie assuming tie independence. Compare this probability with the probabilities computed in (2) and (3). Would it be reasonable to assume this model?

(6) Load the packages `sna` and `network` in R. Import the file `10_W1.csv` as a matrix object. The R function `cug.test` perform a CUG test in R. Run the following commands in R:

```
cguRec <- cug.test(obsMat, grecip, cmode = "edges", reps=5000)
cugInd <- cug.test(obsMat, centralization, cmode="edges", FUN.arg=list(FUN=degree,
cmode="indegree"), reps=5000)
cugTrans <- cug.test(obsMat, gtrans, cmode = "dyad.census", reps=5000)
```

(6.1) State the hypotheses and the conditional features of the tests
Use the help function ?cug.test to understand the command lines

(6.2) Interpret the results

Task 2: MR-QAP regression **9 points**

(1) Import the data for classroom 10. Build a QAP to test if friendship nominations in wave 2 are associated with friendship nominations in wave 1.

(2) Add to the model in (1) variables to test the following hypotheses simultaneously:

   i. Students with high ending literacy scores are less likely to receive friendship nominations.

   ii. A friendship nomination is more likely between a pair of studens of the same gender.

   iii. Students with a high HISEI value are more likely to send friendship nominations than those with lower scores.

   Argue for the definition of the variables. When several operationalizations are possible choose one of them.

(3) Estimate the model specified in (2). Interpret the coefficients of the model and determine whether the data support the hypotheses listed in (2).

(4) Repeat the analysis in (3) for all remaining classrooms.

(5) Without running any further statistical analyses, comment on the results across classrooms and what can be taken away from them as a whole. Are they generalizable? What would differences suggest about our hypotheses? You may also use visual aids (e.g., scatterplots, caterpillar plots, or other summary visualizations) to help interpret and communicate the results.

Task 3: Network Autocorrelation Model **9 points**

Now we want to explore the Network Autocorrelation between ending literacy score and friendship in the second wave. For each classroom, run a network autocorrelation model, including the student's gender and HISEI scores as exogenous variables.

(1) For classroom 10, run a Network Autocorrelation model, treating the ending literacy score as the dependent variable and the friendship in wave 2 as the network variable.

(2) Add gender and HISEI scores to the model as exogenous variables.

(3) Estimate the model specified in (2). Interpret the coefficients of the model and determine whether the data support the autocorrelation hypothesis.

(4) Reflect on how the results from the Network Autocorrelation model compare with those of the QAP specified earlier, especially in relation to the variables that appear in both models.

(5) Repeat the analysis in (3) for all remaining classrooms.

(6) Without running any further statistical analyses, comment on the results across classrooms and what can be taken away from them as a whole. Are they generalizable? What would differences suggest about our hypothesis? You may also use visual aids (e.g., scatterplots, caterpillar plots, or other summary visualizations) to help interpret and communicate the results.

Submission instructions: You are encouraged to work in groups of 4 people. It is a **requirement** for the submission to belong to a group.

Please ensure that:

- The submission includes a **single PDF** file that contains all essential information, including code, results, plots, and written explanations, as this PDF will be the primary document for grading. We suggest using Rmarkdown or Quarto to create the PDF document as these tools simplify the process to create a single document.
- The pdf is named *Assignment01_GroupXX.pdf*; for example, for group 9, the file name is *Assignment01_Group09.pdf* (Groups with numbers 1 to 9 pad a zero on the left, 01 to 09).
- Any accompanying R scripts (`.R`, `.Rmd`, or `.qmd`) should be zipped and included with the submission; these files will be referenced only if additional verification of computations is needed.
- Only one member of the group submits the solution.
- The names of **all** group members are reported in the documents you submit (PDF and R scripts).

Thank you for following these guidelines to ensure a smooth grading process!