Assignment $\mathcal{N}^{\underline{o}}$ 1

released: 03.10.2022 at 19:00 **due:** 19.10.2022 at 12:00

Task 1: Tie Dependence and CUGs **10 points**

Please use set.seed(1938) 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 `matrix.csv` in the folder `Assignment1.zip` contains the adjacency matrix of advice ties between 21 managers of a high-tech company. There is a tie from manager i to manager j if the managers seek another manager for advice.
The number of ties in the network is $m = 190$.
The mutual, asymmetric and null dyads are $M = 45$, $A = 100$, and $N = 65$.

(1) Compute the probability p of observing a tie if you choose one possible tie 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 `matrix.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)
cguInd <- 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 **10 points**

The folder `Assignment1.zip` contains information on trade between 18 companies.
The following information are provided:

- `trade1.csv`: valued adjacency matrix. The value in the cell x_{ij} is the amount of goods (in million CHF) sold by company i to company j in 2006
- `trade2.csv`: valued adjacency matrix. The value in the cell x_{ij} is the amount of goods (in million CHF) sold by company i to company j in 2007
- `attr.csv`: attributes of the companies:
  - *id*: the identifier of the company
  - *sector*: the sector in which the company operates (0 = public, 1 = private)
  - *size*: the number of employees of the company
  - *region*: the geographical region in which the company is located (1 = region A, 2 = region B, 3 = region C)

(1) Import the data. Use the QAP regression to test whether there is an association between the amount of goods traded in 2007 and that in 2006. Interpret the results.

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

Hp. 1: companies tend to sell goods to companies operating in the same sector
Hp. 2: companies in the public sector tend to sell more goods than companies in the private sector

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) Formulate two new hypotheses: one concerning the association between the network at time 2007 and the size of the companies, and the other the association between the network at time 2007 and the region in which the companies are located. The two hypotheses should require different operazionalitions of the explanatory variables (and not only differ in the use of a different attribute!)

(5) Test the hypotheses formulated in (4) by adding the corresponding variable in the MR-QAP specified in point (2). Comment on the result.

You are encouraged to work in groups of at most 4 people.
Please submit your solution (including R scripts!) using 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.