THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

**Summer 2021 Exam**

# MY461/MY561
## Social Network Analysis

Suitable for all candidates

**Instructions to candidates**

This paper contains five questions. Answer all five questions. All questions will be given equal weight (20%). Responses for each question should be a maximum of 500 words excluding tables and figures. Please include a bibliography with any cited sources (this does not count towards the word count).

The exam questions will be released on March 31, 2021. The exam is due on May 5, 2021 at 16:00.

Submission will be done through Moodle. Please submit your answers in a PDF file. You will be evaluated based on your responses to the 5 prompts. To help us determine where any errors were made, however, you must additionally submit an annotated R or Rmd file that presents the code used to arrive at your responses.

**Background information:**

For this exam, you will be analysing networks representing student exchanges between universities in Europe. These are part of the Erasmus programme (**E**u**R**opean Community **A**ction **S**cheme for the **M**obility of **U**niversity **S**tudents). Each year, thousands of European students leave their home institution to study at another "higher education institution" (HEIs, primarily another university) in another European country through this programme.

Data for this programme are available through the EU Open Data Portal, but here we will use elements from a version of these data that has already been cleaned and geocoded by Gadár et al (2020). While there are over 4000 HEIs involved in the Erasmus exchange programme, we will limit ourselves to consider only the subset of ~300 universities for which we have additional information on their academic impact, coming from the recent paper by Kosztyán et al (2021). This includes things like the Scimago and CWTS Leiden university rankings, both of which use information on journal publications to create a bibliometric measure of each university's academic impact (primarily using citation networks and centrality measures!).

From the larger set of HEIs involved in the Erasmus exchange, Kosztyán and colleagues include only those European universities that are also included in the Scimago and Leiden rankings.

We are providing you with six files:

The first file (HEI_metadata.csv) contains information on each of the "higher education institutions." The variables we will explicitly draw upon are as follows:

- Erasmus ID: the unique ID code for the university
- Region: based on the UN geoscheme, each university is located in Northern, Eastern, Southern, or Western Europe
- Latitude and Longitude: the coordinates (in decimal degrees) giving the exact location of the university
- 2013 Ranking: Scimago's rank of the university in 2013.
- Total academic staff (FTE): the total number of full-time employed staff at each university, which we can take as a rough proxy of the university's size
- POI_count: a count of the "points of interest" that are physically near each university's campus, such as restaurants, cinemas, hotels, and hospitals), which we can take as a measure of the amenities near the university, and perhaps as a rough proxy of urbanity.

The remaining files (main_el.csv, hum_el.csv, soc_el.csv, sci_el.csv, and eng_el.csv) are edge dataframes, where the first column represents the sending university, the second column represents the receiving university, and the third column represents the number of students who went from the first to study at the second. All of these files relate to student exchanges in the 2013-2014 academic year and are limited to only those universities that are included in the Kosztyán et al (2021) dataset. main_el includes a count of all student exchanges, while the others consider a subset of those, divided out by the academic subject the students were going to study.

- hum_el considers subjects in Humanities and Arts
- soc_el considers subjects in Social sciences, Business and Law
- sci_el considers subjects in Science, Mathematics and Computing
- eng_el considers subjects in Engineering, Manufacturing and Construction

So, the first row of main_el.csv shows that in the 2013-2014 academic year, two students from the University of Graz ("A GRAZ01") went to study at the University of Antwerp ("B ANTWERP01"). And similarly, the first row of hum_el.csv shows that in the 2013-2014 academic year, one student from the University of Graz went to study the Humanities and Arts at the University of Geneva.

Use these files to generate a number of networks, where nodes are the universities and edges are the number of students from University A who went to study at University B. Specifically, make the following six networks:

- **"main" network**: A weighted, directed network based on the main_el file, representing all student exchanges in 2013-2014 between the universities included in the the Kosztyán et al (2021) dataset.
- Four **"subject" networks**: one from Humanities and Arts (based on hum_el), one for Social sciences, Business and Law (based on soc_el), one for Science, Mathematics and Computing (based on sci_el), and one for Engineering, Manufacturing and Construction (based on eng_el).
- **"reduced" network**: The "main" network reduced to only include edges where at least 5 students went from University A to University B, and where any universities who as a result of this removal of edges are isolates are removed.

Citations:

Gadár, L., Kosztyán, Z. T., Telcs, A., & Abonyi, J. (2020). A multilayer and spatial description of the Erasmus mobility network. *Scientific Data*, *7*(1), 41.

Kosztyán, Z. T., Fehérvölgyi, B., Csizmadia, T., & Kerekes, K. (2021). Investigating collaborative and mobility networks: Reflections on the core missions of universities. *Scientometrics*.

With these networks in hand, answer the following questions.

1 Consider the overall metrics (density, average path length, transitivity, and reciprocity) for each of the four **subject networks** (1. humanities and arts, 2. social sciences, business and law, 3. science, mathematics and computing, 4. engineering, manufacturing and construction). How do the four networks compare to each other? Compare each subject network to a random network with the same number of nodes and edges, created with the Erdős–Rényi model. What do these comparisons tell you about the nature and structure of the relationships among the universities? In your answer, make sure to define each of the metrics and give an intuitive interpretation for them.

2 Which do you see as the most influential universities in the student exchange network? Identify two potential meanings of "influence," as proxied by different centrality measures. Justify your choice of each centrality measure. In that justification, present clear interpretations of what each centrality measure is capturing about the position of the countries. Calculate both of your chosen centrality measures on the **main network** and identify the university that has the highest value for each. Then, consider the LSE and the other London universities included in this network (City, Imperial, Kings, and Queen Mary). Plot histograms of your two chosen centrality measures, and overlay vertical lines demarcating the values for each of the London universities. How does the LSE compare to the other London universities? Make explicit reference in your responses to concepts related to node centrality covered in the course material. [Note that if you use edge weight for a centrality measure, you need to identify how the calculation interprets those weights (i.e., do higher values mean stronger connections or greater distance?); in those cases where the measure assumes that higher values mean greater distance, you should use 1/weight in the calculation.]

3 Visualize the degree distributions of the **four subject networks**. To facilitate comparison, overlay all four distributions on the same plot. You should also consider how to modify the plot (log axes, cumulative distribution, etc.) to make it more legible. Describe the distributions you observe and discuss the differences between the different subjects. Redo the analyses in problem 1 above but using the configuration model (instead of the Erdős–Rényi model). Discuss how the results differ from those in problem 1 and explain what causes the differences. What does the comparison with the configuration model tell us about the data?

4 How does geography influence student exchange relations? Calculate the probability of a student exchange tie within and between each region (using a blockmodel approach) for the **reduced network**. Evaluate the structural equivalency of the countries using the **main network**. Use the results of the structural equivalency calculation to divide the countries into six equivalency classes. Plot the **main network** with nodes positioned by their latitude/longitude (so that your network should look roughly like a map of Europe) and nodes coloured by their equivalency class. Make sure that your plots are legible and informative, with a legend. Interpret the six groups. Discuss what this implies about universities' student exchanges.

5 What helps predict whether students from one university go to study at another? Consider the results of the exponential random graph model (ERGM) run on the **reduced network** below:

```
==========================
Summary of model fit
==========================

Formula:   major1 ~ edges + nodeifactor("region", levels = c(1, 3, 4)) +
    nodeicov("rank_ranked") + nodeicov("staff_scaled") + nodeicov("POI_scaled") +
    mutual + gwesp(fixed = TRUE, decay = 0.8)

Iterations:  16 out of 20

Monte Carlo MLE Results:
                                  Estimate Std. Error MCMC % z value Pr(>|z|)
edges                            -4.7409930  0.0576203      0 -82.280   <1e-04 ***
nodeifactor.region.Eastern Europe  0.3329028  0.0578418      1   5.755   <1e-04 ***
nodeifactor.region.Southern Europe 0.3895257  0.0419288      0   9.290   <1e-04 ***
nodeifactor.region.Western Europe  0.0870115  0.0415069      0   2.096   0.0361 *
nodeicov.rank_ranked              -0.0016359  0.0002128      0  -7.688   <1e-04 ***
nodeicov.staff_scaled              0.5420552  0.0829627      1   6.534   <1e-04 ***
nodeicov.POI_scaled                0.2099600  0.0416916      0   5.036   <1e-04 ***
mutual                             2.5899157  0.0657558      0  39.387   <1e-04 ***
gwesp.fixed.0.8                    0.5478822  0.0204300      0  26.818   <1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 117796  on 84972  degrees of freedom
 Residual Deviance:  26264  on 84963  degrees of freedom

AIC: 26282    BIC: 26366    (Smaller is better.)
```

Here, “**rank_ranked**” is a re-ranking of the Scimago ranking of universities (the “2013 Ranking” variable) in the reduced network, such that the university with the highest global rank in the network has a value of 1, the second highest a value of 2, etc. “**staff_scaled**” rescales the original count of full-time academic staff at each university (the “Total academic staff (FTE)” variable) by dividing it by the maximum, so that the value is proportional to the university with the highest number of full-time staff. We can see this as a general measure of university size, with 1 being the maximum. “**POI_scaled**” is the count (with a maximum of 100) of the “points of interest” near the university (from the “POI_count” variable), rescaled so that the maximum is 1 (by dividing the original values by 100). Interpret each term in the ERGM (except edges) – how does each term influence whether or not a university sends students to another? Use odds ratios in your substantive interpretation of each term. In your discussion, make explicit reference to the concepts of triadic closure and reciprocity. Propose two additional terms to add to the model and provide a justification for them. What do you think they would capture that is currently missing from the ERGM?