MY461
SOCIAL NETWORK ANALYSIS

EXAM PROJECT - MAY 2021

Candidate Number: 17483

Question 1

METRICS OF THE OBSERVED SUBJECT NETWORKS

Network	Density	Average Path Length (L)	Transitivity	Reciprocity
Humanities	0.092	2.1874	0.2918	0.5066
Social Science	0.0972	2.1432	0.3073	0.5656
Science	0.0455	2.5869	0.1753	0.2613
Engineering	0.0458	2.593	0.2895	0.449

Table 1: Metrics of the Observed Subject networks

Density indicates the proportion of connections that exist in a network out of all possible connections (*Equation 1*).

$$\text{Density for directed network} = \frac{number\ of\ edges}{number\ of\ dyads}$$

Equation 1

Out of the four networks, the Social Science has the highest density of 0.0972 *(table 1)*, indicating that the probability of a university sending one or more Social Science students to another University is 9.7%, considering as possible connections only links between universities that are part of the Social Science network. The Humanities has a similar density, whereas for Science and Engineering are almost half. A higher density doesn't necessarily mean more links, because it's proportional to the size of the network, which varies (*table 2*).

METRICS OF THE SIZE OF THE OBSERVED SUBJECT NETWORKS

Network	Number of Vertices	Number of Edges	Density
Humanities	270	6684	0.0920
Social Science	287	7978	0.0972
Science	290	**3812** ↑	**0.0455** ↓
Engineering	266	**3225** ↓	**0.0458** ↑

Table 2: Metrics of the size of the observed subject networks. Shows that the four subjects have different network size and this doesn't necessarily reflect the density as seen in the Science and Engineering network measures.

The **Average Path Length (L)**, is the average of all geodesic paths, considering both directions. Humanities and Social Science are better connected, since their L is lower than the other two. L=2.1874 for Humanities, means that on average a student to be send from one university to another, needs to be sent to 1.1874 (2.1874-1) in between, if this was allowed from the Erasmus Program.

Transitivity measures network clustering and here directionality is ignored. The Social Science network has the highest transitivity of 0.3073 meaning that there is a 30.73% chance that two universities that have exchanged students with a shared univeristy have also exchanged students among themselves. We can also derive that an exchange relationship with a shared university affects more the Social Science universities than the Science.

Reciprocity measures the frequency of two universities exchanging students in both directions when they already exchange in at least one. A reciprocity of 0.5656 for the Social Science means that if university A sends one or more Social Science students to university B, there is 56.56% chance that B also sends one or more students to A. Science has much fewer mutual connections.

From reciprocity and transitivity we can infer that Science students/universities don't manage to form as good relationships as Social Science do, in order to influence mutual connections or connections with shared universities.

None of these measures consider *weight*, the number of exchanged students.

OBSERVED vs ERDOS-RENYI SUBJECT NETWORK METRICS

Network	Density	Average Path Length	Transitivity	Reciprocity
Observed Humanities	0.0920	2.1874	0.2918	0.5066
Random Humanities	**0.0920**	**1.9999**	**0.1755**	**0.0952**
Observed Social Sciences	0.0972	2.1432	0.3073	0.5656
Random Social Sciences	**0.0972**	**1.9657**	**0.1851**	**0.1015**
Observed Sciences	0.0455	2.5869	0.1753	0.2613
Random Sciences	**0.0455**	**2.4865**	**0.0912**	**0.0477**
Observed Engineering	0.0458	2.5930	0.2895	0.4490
Random Engineering	**0.0458**	**2.5132**	**0.0914**	**0.0434**

Table 3: Comparison of the overall metrics for the Observed and Erdos-Renyi generated networks, for all 4 subjects.

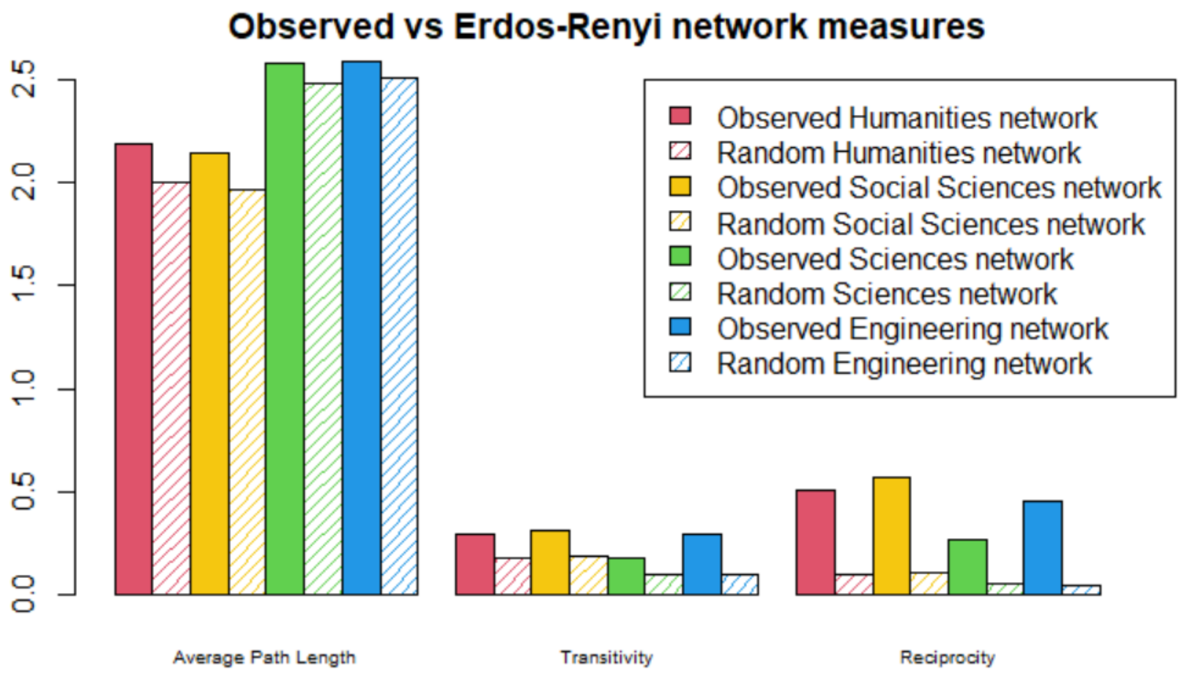


Figure 1: Visual comparison of the overall metrics for the Observed and Erdos-Renyi generated networks, for all 4 subjects. Density is excluded as it doesn't vary between the observed and randomly generated networks.

The measures of the randomly generated Erdos-Renyi graphs, can be used as the null expectation, if there were no social dynamics to make the clustering or reciprocity more likely. The density is the same, as it was the given parameter to the model. All other measures are lower in the Erdos-Renyi, implying social dynamics. The Engineering network has the biggest difference between the two graphs, suggesting that social structures have a stronger effect *(Figure 1)*. For example, administrative teams of Engineering students could be actively seeking new connections taking advantage of existing connections of their neighbours (transitivity) and that students form meaningful relationships when in exchange, influencing their classmates in the receiving university to participate in the Erasmus and visit them back (reciprocity). Additionally, the randomness of the Erdos-Renyi model is a way for "long-ties" to be generated and hence results in the lower L, since the model assumes equal/similar degree, which is not the case here.

Question 2

One meaning of influence is the number of students that a university receives (weighted indegree), since the university gets the opportunity to expose them to its culture, research approaches, students and professors. When the exchange students go back to their main university, they will sequentially influence their peers and their wider environment accordingly. Similarly, the number of universities from which the university receives students (non-weighted edge indegree) has similar but not same meaning of influence. Hosting students from a variety of backgrounds implies that the multicultural environment created in the university can add value to all students. The Universities of Granada and Bologna *(tables 4&5)* appear as the most influential regarding indegree. Valencia has the 3rd position regarding student indegree but doesn't show as high diversity when looking at the university indegree.

The high indegree of both universities and students can be an indication against the diversity-bandwidth trade-off (Aral & Van Alstyne, 2011) and hence an indication of strong connections. The bandwidth could be proxied by the student indegree and the diversity by the university indegree, slightly deviating from the literature as it is defined by tie range.

TOP STUDENT IN-DEGREE UNIVERSITIES

University <chr>	Student Indegree <chr>
Universidad de Granada	1123
University of Bologna	1053
University of Valencia	1032
Complutense University	1026
Charles University, Prague	841
University of Vienna	697

Table 5: Top 5 universities with highest student indegree. It was calculated based in the weighted edges.

TOP UNIVERSITY IN-DEGREE UNIVERSITIES

University <chr>	University Indegree <chr>
Universidad de Granada	183
University of Bologna	173
Complutense University	171
Charles University, Prague	170
Lund University	156
University of Barcelona	155

Table 4: Top 5 universities with highest university indegree. It was calculated ignoring the weighted edges.

Influence can also account for connections beyond local, by considering the importance of the neighbours, defined by the proportion of outgoing students to the particular university compared to the total outgoing students. The centrality measure of **PageRank** is a proxy of it and reveal Granada and Bologna universities as the most influential again *(table 6)*.

PageRank centrality can be considered a proxy of university **social capital**. Here, a high social capital can mean a university having important incoming student exchanges. The connections that are developed through the exchanges become part of a university's "wealth", in the sense that students get attached with the university, people, the area and processes and this facilitates further exchanges but also other activities between the universities and stronger relationships with the market that students will later be part of.

TOP PAGERANK UNIVERSITIES

Uni <chr>	PageRank <chr>
Universidad de Granada	0.0162
University of Bologna	0.0157
Complutense University	0.015
University of Valencia	0.0139
Charles University, Prague	0.0122
University of Copenhagen	0.0119

Table 6: Top 5 universities with highest university PageRank value.

London Universities

For both indegree and PageRank, LSE has the lowest value and hence the least influence among London universities *(figures 2&3)*. The number of incoming students for LSE is less than 10, whereas the rest of the universities receive more than 100. City university has a significant advancement in both. Since the indegree is in absolute terms, the size of the university can impact the value. Assume that the university size can be proxied by the number of full-time academics *(table 7)*, we see that LSE has the smallest size and City the highest, meaning that LSE doesn't have as many resources to accommodate numerous students. The PageRank ranking of the London universities reveal the importance of their connections and LSE doesn't seem to have many.

It's interesting to note that Queen Mary university has higher incoming universities than Kings, but when looking at PageRank Kings appear to have a higher influence and hence we can assume that even though its connections are not as many, they are more important.

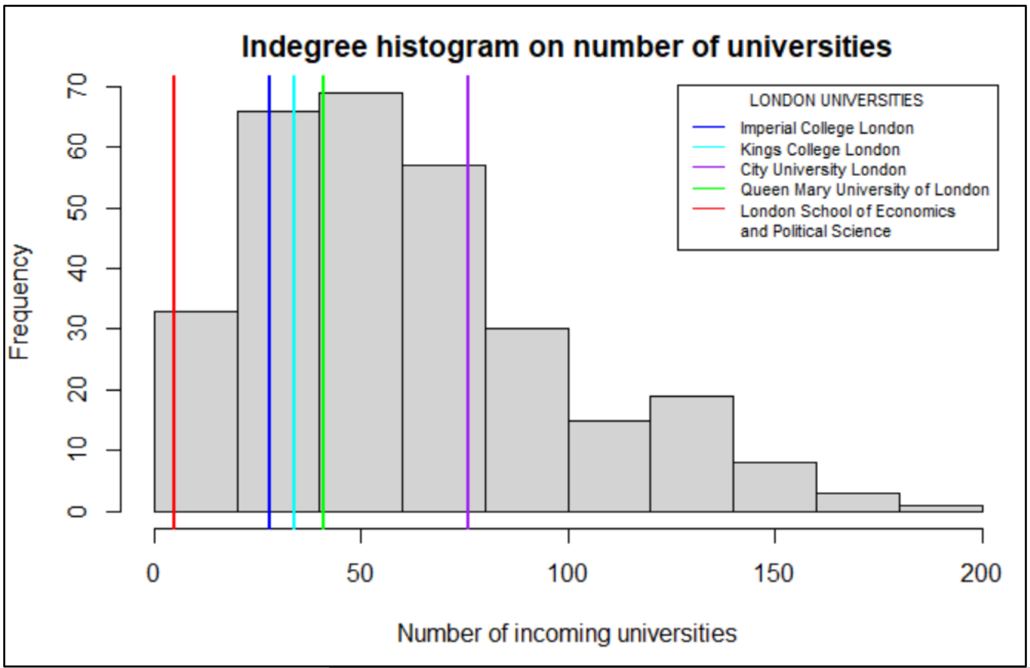


Figure 2: University indegree position of the London universities.

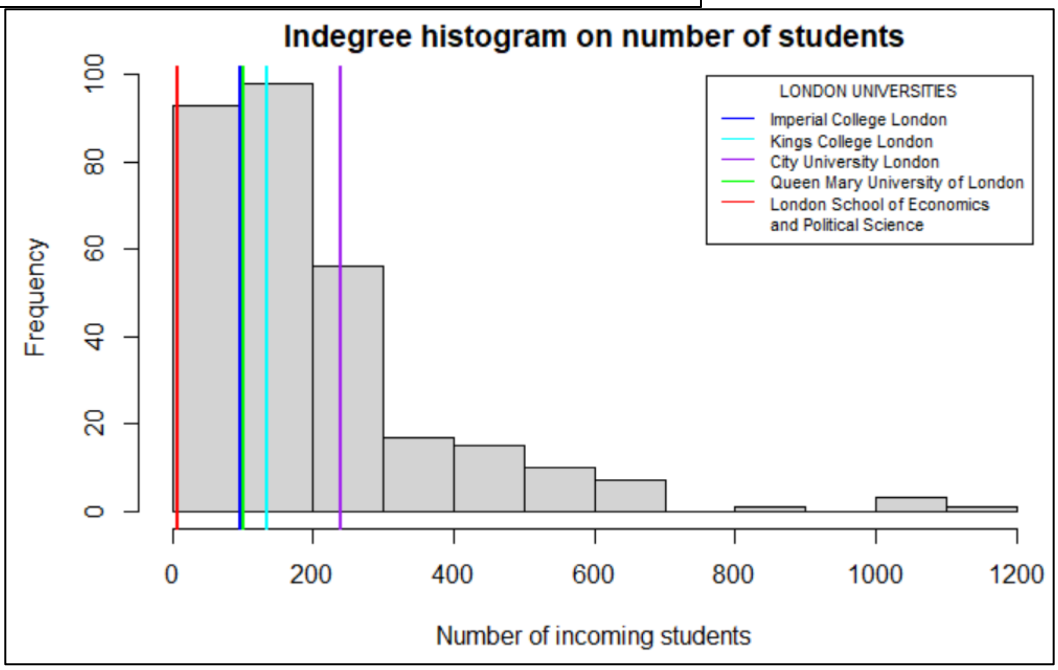


Figure 3: Student indegree position of the London universities.

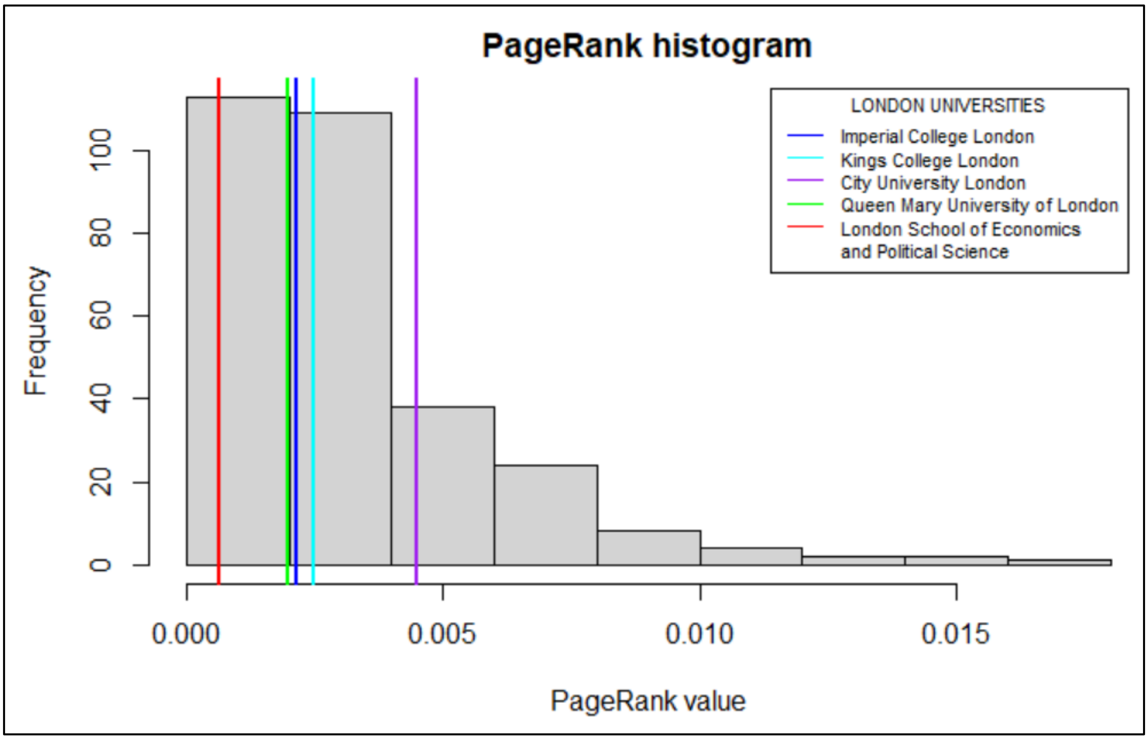


Figure 4: PageRank position of the London universities

PROXY OF LONDON UNIVERSITIES SIZE

University	Total academic staff (FTE)
City University London	5380
Imperial College London	3720
King's College London	3460
Queen Mary University of London	1760
London School of Economics and Political Science	1050

Table 7: Number of total full-time academic staff for each London university to be used as a proxy of their size.

Question 3

The indegree distributions are **heavy-tailed** *(figures 5&6)*, because the probability of observing extreme values is non-zero, as the median line reveals on the left side of the plot. Humanities and Social Science appear with a heavier tail, making the student exchnages less robust to potential failure of one of the high indegree universities to host students.

The proportion of universities with 1 indegree in the Engineering network is almost double of the other networks. Hence, higher indegrees have a lower proportion for Engineering universities compared to other networks. Figure 7 zooms in and we see that Science also has higher proportion in lower indegrees than the other two, indicating a **weakly scale-free** property for Engineering and Science. Figure 10 makes this obvious with a straignt line fitting the tail, opposite to the Social Science and Humanities line.

We can derive that Science but mostly Engineering universities tend to accept students from few universities, whereas Hummanities and Social Science tend to have a variety on the universities they accept students from. For example, this could mean for the Engineering network that strong relationships are created between students or that the administartive team has difficulties in creating new paths for its students. However, we shouldn’t ignore that we are **comparing networks of different size** *(table 8)*. This is clearer when plotting the indegree distribution on the actual number of universities that students are sent from *(Figure 8)*, showing that a similar number of universities (450-500) from Engineering and Humanities neworks receive students from only one university (degree=1).

METRICS ON THE SIZE OF THE NETWORKS

	Indegree <dbl>	Max indegree <dbl>	Number of vertices <dbl>
Humanities	6684	110	270
Social Science	7978	117	287
Science	3812	80	290
Engineering	3225	79	266

Table 8: Metrics about the size of the network.

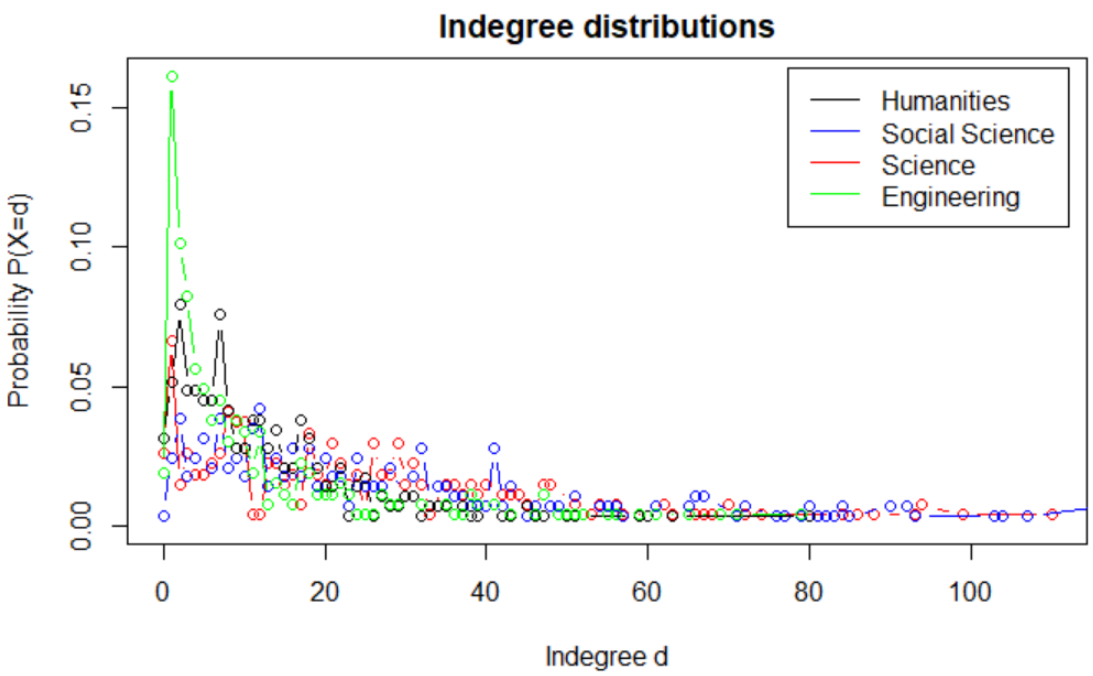


Figure 5: Indegree distributions of the subject networks.

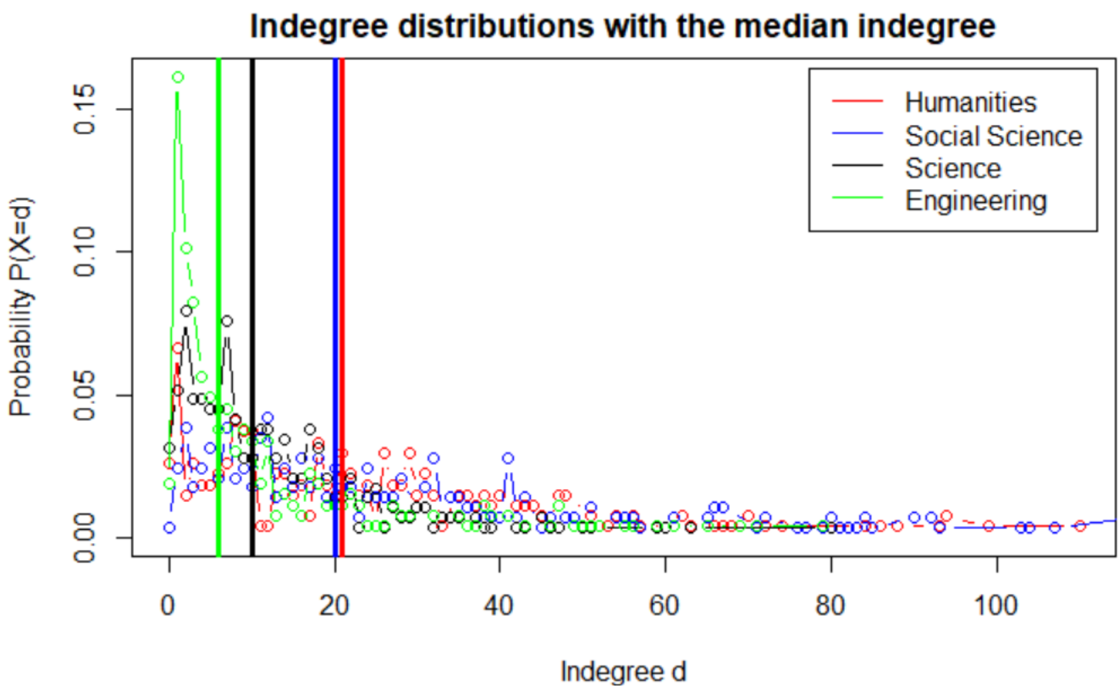


Figure 6: Indegree distributions of the subject networks with a vertical line indicating their median. The non-zero probability of extreme values can skew the mean, so this is the reason for plotting the median (and not the mean) in these plots.

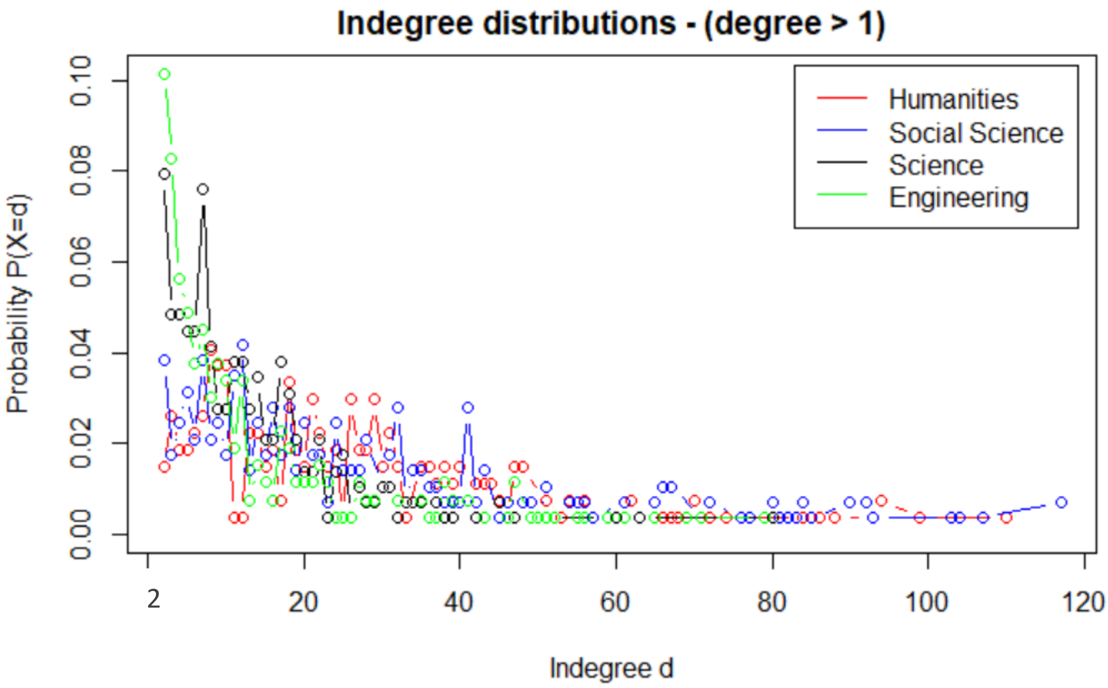


Figure 7: Zoomed- in Indegree distributions of the subject networks by looking for degrees higher than 1.

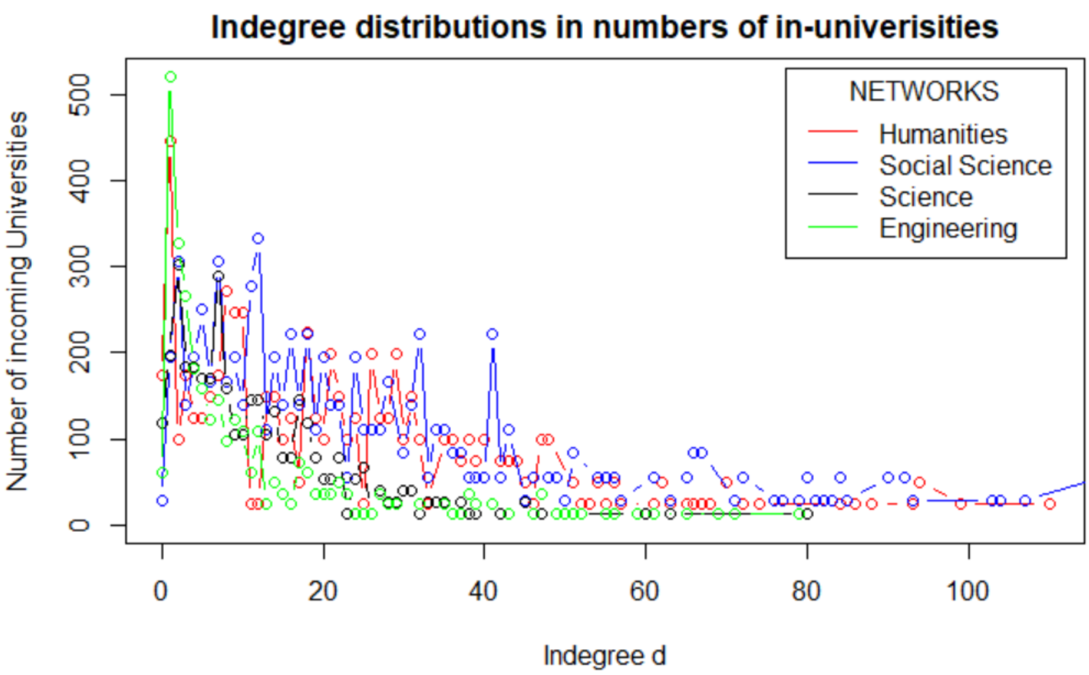


Figure 8: Indegree distributions of the subject networks, considering the number of universities that students are sent from.

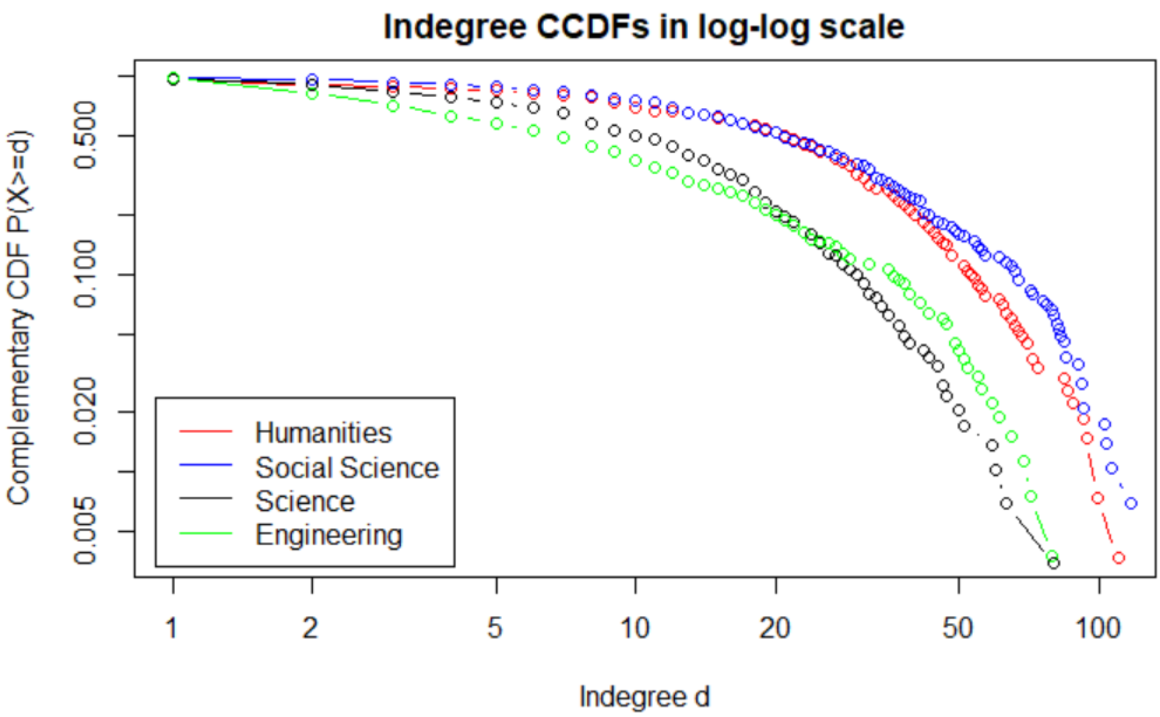


Figure 9: Plot of the complementary cumulative distribution function, namely P(X≥x). Plots the probability of observing that indegree or larger in the data.

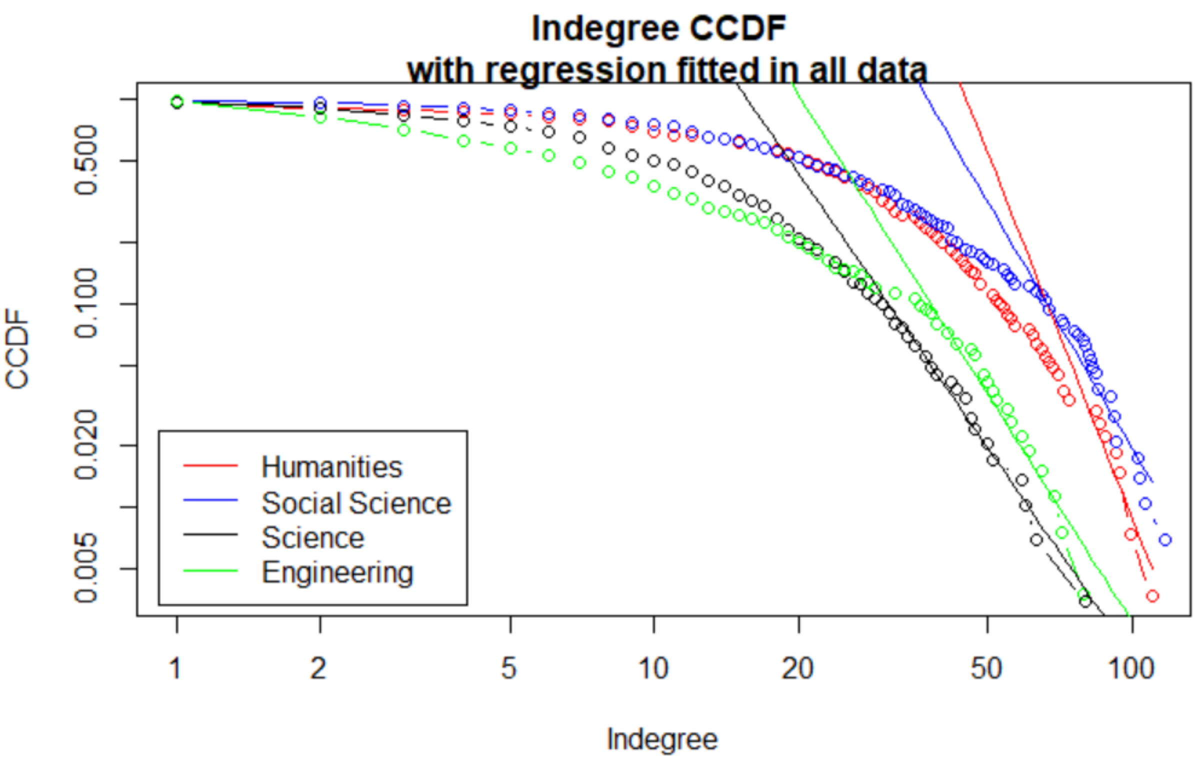


Figure 10: Indegree complementary cumulative distribution functions for all subject networks with a straight line fitting their tail. It facilitates the presence or not of the power law.

OBSERVED vs CONFIGURATION SUBJECT NETWORK METRICS

Network	Density	Average Path Length (L)	Transitivity	Reciprocity
Observed Humanities	0.0920	2.1874	0.2918	0.5066
Configuration Humanities	**0.092**	**2.2083**	**0.3032**	**0.1488**
Observed Social Sciences	0.0972	2.1432	0.3073	0.5656
Configuration Social Sciences	**0.0972**	**2.164**	**0.3101**	**0.1638**
Observed Sciences	0.0455	2.5869	0.1753	0.2613
Configuration Sciences	**0.0455**	**2.591**	**0.1724**	**0.0695**
Observed Engineering	0.0458	2.5930	0.2895	0.4490
Configuration Engineering	**0.0458**	**2.6134**	**0.2752**	**0.1483**

Table 9: Comparison of the overall metrics for the Observed and Configuration generated networks, for all 4 subjects

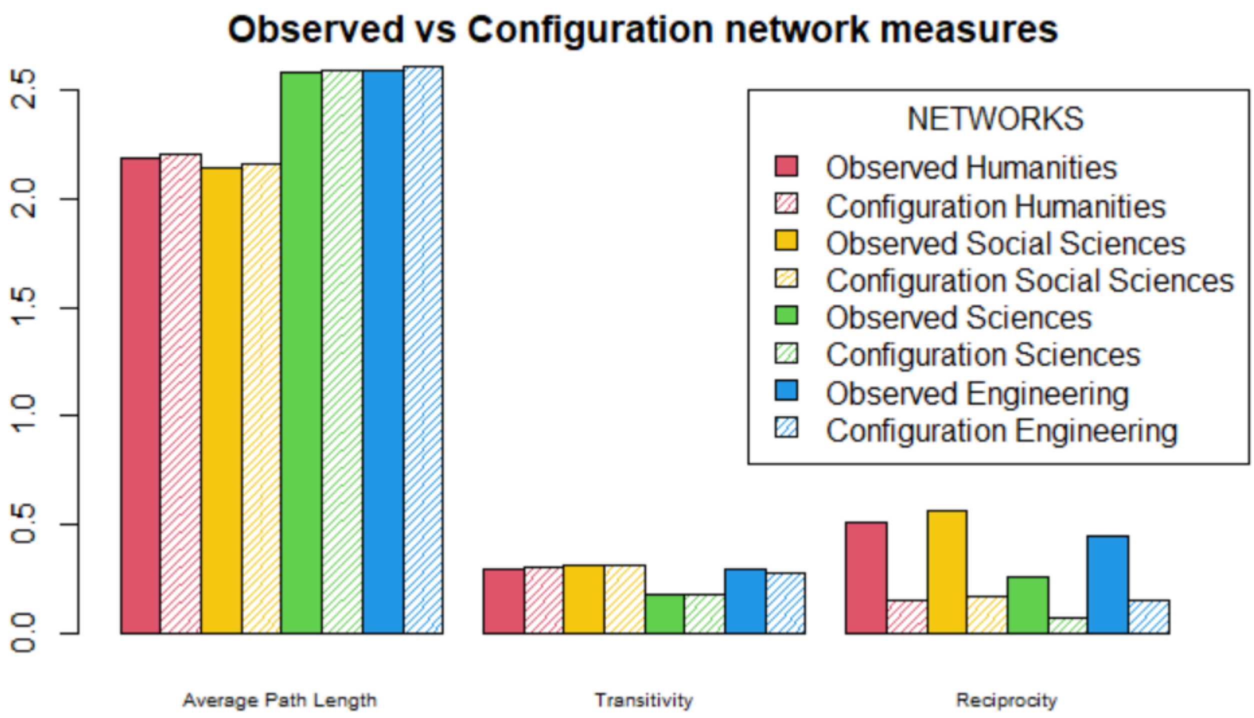


Figure 11: Visual comparison of the overall metrics for the Observed and Configuration generated networks, for all 4 subjects. Density is excluded as it doesn't vary between the observed and randomly generated networks.

Average Path Length (L) and Transitivity have been approached almost perfectly from the configuration models *(table 9 & figure 11)*, because the real degree distribution is a known parameter to the configuration model, opposite to the Erdos-Renyi model (Newman, 2010).

L is only a bit higher in the configuration graphs, probably because the method used to generate them allows for loops and multiple edges, which doesn't align with the reality. Multiple edges and loops are ignored when calculating L, so the configuration graph will have less edges to examine than the observed one.

If social forces were present for exchanges to be influenced by shared neighbours, we would expect that the observed network has significantly higher transitivity than the configuration. This is not the case here, so we can assume that the observed transitivity is the random we would observe given the degree distribution. So, a university exchanging students with many other universities (high degree) could be because there are efforts from the team to have a variety and maybe get a better place in the ranking and it is not necessarily because of using references from existing collaborators.

On the contrary, the configuration models significantly underestimate Reciprocity. It is one more confirmation that social structural forces are in place for universities to mutually exchange students, possibly because of strong friendships that students create when they visit a country and they want their friends to visit them back.

Question 4

From the blockmodel *(figure 12)* we see that Southern European universities have the highest probability of 8.73% in exchanging five or more students with each other. The least within-region collaboration is observed in Northern European universities with a probability of 1.41%. Worth noting is the collaboration between Northern and Western universities, which is mostly based on Westerns sending students to Northerns (6.59%) with no indication for mutual correspondence (1.17%). A more mutual exchange between regions is seen between Southern and Western universities (5.05% and 5.29%).

Eastern Europe is underrepresented since it has only 25 universities in the reduced network, when the other regions have over 70 *(table 10)*. However, it is the second most frequent region to send students with a probability of 19.37% and the third region to receive students. From the total measures we can see that Northern universities have the least probability to send students (8.06%), Southern the highest (24.29%) and the same hierarchy is for receiving students.

BLOCKMODEL ON REGION

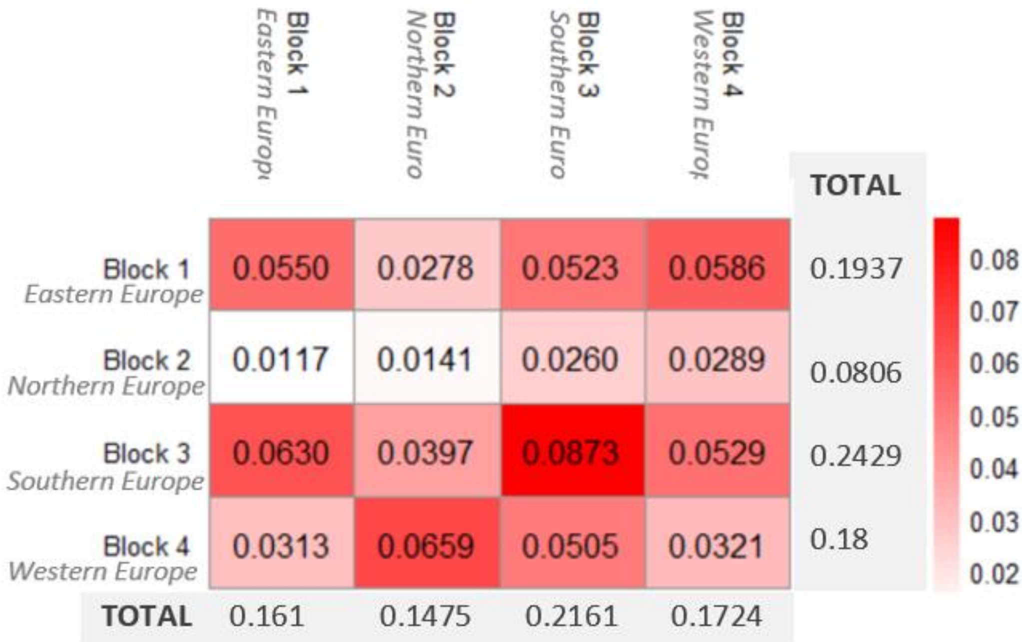
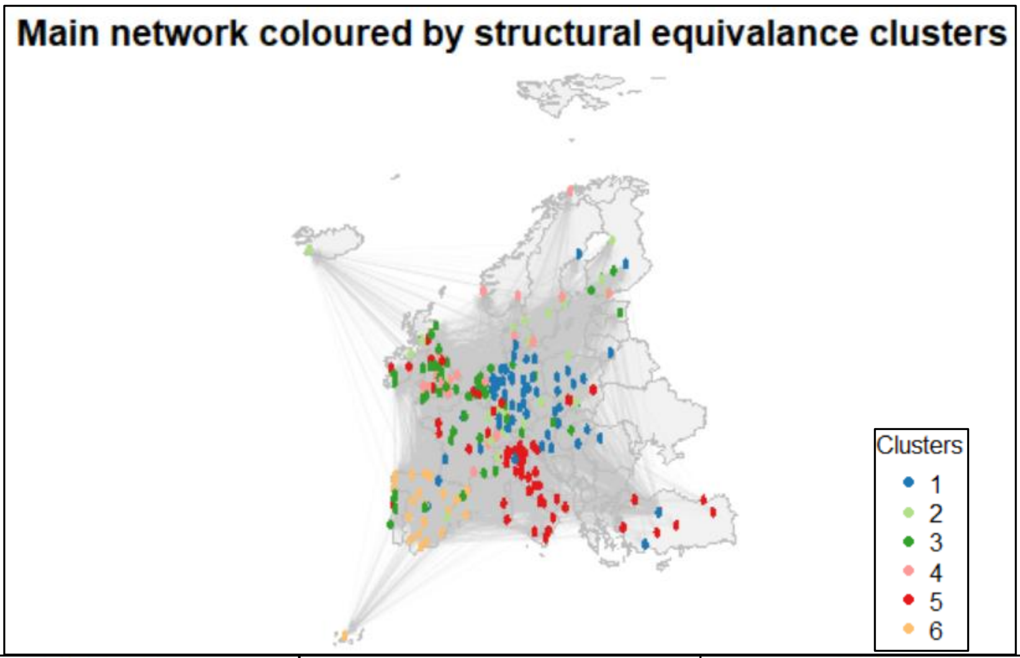


Figure 12: Stochastic block model results presenting the probability of a student exchange tie within and between each region for the reduced network.

Figure 13: Main plot showing universities on a map, coloured by the cluster they belong to, based on their structural equivalence.



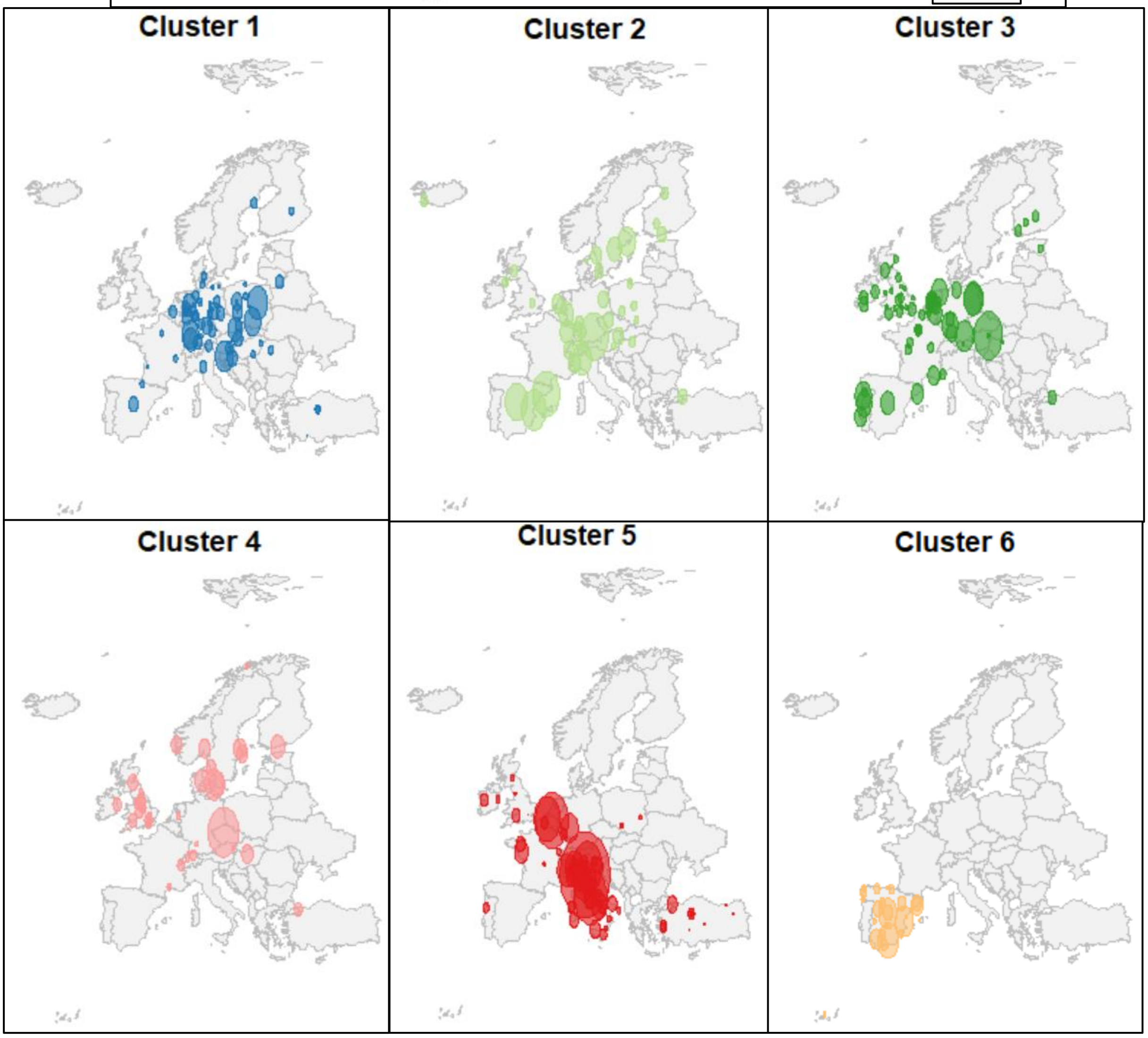


Figure 14: Six plots, subsets of the main network by the universities of each cluster, with size relative to the student degree of each university.

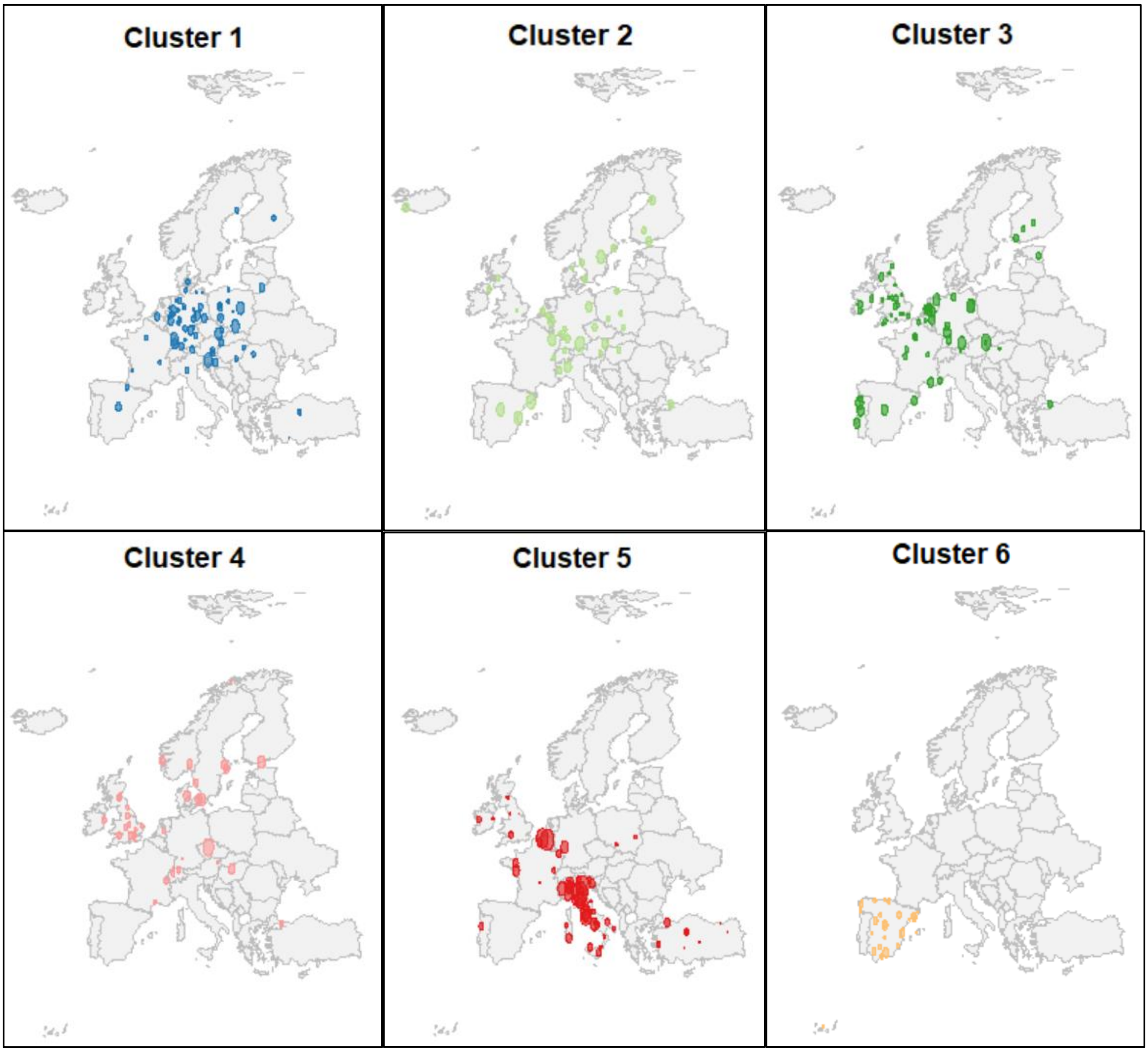


Figure 15: Six plots, subsets of the main network by the universities of each cluster, with size relative to the university degree.

NUMBER OF UNIVERSITIES IN REGIONS

Region	Number of Universities
Eastern Europe	25
Northern Europe	72
Southern Europe	94
Western Europe	101

Table 10: Number of Universities that belong to each region.

Universities were divided into clusters *(figures 13-15)* based on their structural equivalence assessed using Pearson's correlation both on incoming and outgoing students. Structural equivalence looks at universities' immediate connections, revealing sets of universities with a shared position in the network. For example, the perfect structural equivalence would be in universities that form exchange ties to the same universities with the same direction and number of students (weight). Therefore, for universities that are within the same cluster we can infer that reasons for that are similar exposure to information regarding the ranking of universities, the interest in the area of the university to visit, the similar culture they share that influences their decision etc.

Cluster 6 includes only universities in Spain and is the only cluster concentrated in one region and one country. We could infer that Spanish universities and students influence greatly each other or that policies on the national level allow for exchanges in particular universities only. **Cluster 1** has universities mostly in Western and Eastern Europe but also very few in South and North. From the latter ones that are further away and with a fairly low student-degree we can infer that students from the West and East universities when seeking for an exchange in a different region, they have very few choices, probably because they are aware of only these universities. **Cluster 2** spreads across Europe with the biggest presence in the South and West. It also appears in all the limits of Europe, like Iceland, Turkey, and Finland. **Cluster 3** includes mostly UK and West-East universities, but also with presence in Portugal and the countries of Cluster 2. **Cluster 4** includes mostly UK and Scandinavian universities with a strong presence in a Czech university. The vast majority of **Cluster 5** is in Italy and West Europe.

Comparing figure 14 and 15 we see that structures are more equivalent in terms of the number of universities rather than the number of students, since the size of the vertices is more homogenous in figure 15.

Question 5

ODDS RATIOS AND CIs OF ERGM TERMS

	Lower CI	Odds Ratios	Upper CI
nodeifactor.region.Easter Europe	1.2455	1.3950	1.5625
nodeifactor.region.Southern Europe	1.3598	1.4763	1.6027
nodeifactor.region.Western Europe	1.0057	1.0909	1.1834
nodeicov.rank_ranked	0.9979	0.9984	0.9988
nodeicov.staff_scaled	1.4615	1.7195	2.0232
nodeicov.POI_scaled	1.1368	1.2336	1.3387
mutual	11.7169	13.3286	15.1621
gwesp.fixed.0.8	1.6617	1.7296	1.8002

Table 11: Odd Ratios, Lower and Upper limits of their 95% Confidence Intervals for the estimated coefficients of the ergm terms. All values are rounded to 4 decimal places. None of the Confidence Intervals contains 1, which means that we are confident that the Odds Ratios correctly represent the direction of the influence of each term.

All terms are significant at the 5% level, looking at their p-values and CIs. Interpretations of the terms below assume that all other factors are controlled.

The reference category for the **nodeifactor("region",levels=c(1,3,4))** is *Northern Europe.* 1.395 Odds Ratio for **nodeifactor.region.Eastern_Europe** means that a university is 1.395 times more likely to receive five or more students if it is located in Eastern Europe than in Northern. *nodeifactor.region* terms Odds Ratios are greater than 1, hence Northern Europe has the lowest probability to receive students.

Based on **nodeicov.rank_ranked** for each increase of 1 unit in the rank_ranked, the likelihood of forming an incoming tie will be 0.0016 times lower (1-0.9984).

For **nodeicov.staff_scaled**, since the variable staff_scaled ranges between 0 and 1 we say that if a university moves from the state of having 0 full-time academic staff to having the most out of all universities, then its likelihood of forming an incoming tie will increase 1.7195 times.

Based on **nodeicov.POI_scaled**, since the variable POI_scaled is scaled ranging between 0 and 1, it means that if a university with 0 points of interests gets 100 points, then its likelihood of forming an incoming tie will increase 1.2336 times. Similarly, if its points of interest increase by 1 then its likelihood of forming an incoming tie will increase 1.002336 times {[(1.2336–1)/100]+1}.

mutual and *gwesp* are structural terms. **mutual** influences the reciprocity of the network. Its Odds Ratio is 13.3286, meaning that if university A sends students to university B, then the likelihood of B sending students to A will increase 13.3286 times. Social structural forces can be assumed, like when A sends students to B it means that communication channels have already been opened, making it much simpler for B to also send students to A.

gwesp captures the transitivity of a network, hence the tendency of ties to close triangles. It influences the propensity of a university to exchange students with another either way, given that they both exchange students with shared universities. Its Odds Ratio is 1.7296, meaning that for a 1 unit increase in gwesp.fixed.0.8, the likelihood of forming a tie that closes a triad increases 1.72956 times. gwesp.fixed.0.8 steadily decreases with a rate of 0.8 as a pair of universities has more existing shared universities *(equation 2)*.

$$w = e^{\alpha} \sum^{n-2} \{1 - (1 - e^{-\alpha})^{i}\} p_i$$

Equation 2: Where w is the GWESP statistic, α is the GWESP decay parameter, pi is the number of actor pairs who have exactly i shared (edgewise) partners, n is the number of nodes in the network (Hunter 2007).

NODEMATCH("Country_Code") can be an additional term, as it captures the effect of a student going to the same country, since a disassortativity mixing of -0.1148 was calculated. A possible behavioural rule is that students usually decide to join Erasmus to gain experiences and go in different environments.

Another term is **DGWESP()** to supplement gwesp, since it additionally accounts for directionality. The type of triangles that will count towards this statistic will be only pointing into 1 university directly or indirectly. This is based on the behavioural rule assumption that the most attractive universities in terms of ranking, location etc. will be the goal of every student, but when there are no vacancies, they select the next attractive university.

References

1. Aral, S., & Van Alstyne, M. (2011). The diversity-bandwidth trade-off. American Journal of Sociology, 117(1), 90-171.
2. Hunter D. R. (2007). Curved Exponential Family Models for Social Networks. *Social networks*, *29*(2), 216–230. https://doi.org/10.1016/j.socnet.2006.08.005
3. Kosztyán, Z. T., Fehérvölgyi, B., Csizmadia, T., & Kerekes, K. (2021). Investigating collaborative and mobility networks: Reflections on the core missions of universities. Scientometrics.
4. Newman, M. E. J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23), 8577–8582.
5. Newman, M. E. J. (2010), Networks: an introduction , Oxford University Press , Oxford; New York .
6. Niemeijer, R. (1973). Some Applications of the Notion of Density, in J. Boissevain and J.D. Mitchell (eds) Network Analysis. The Hague: Mouton.