Assignment 1

2024-12-30

Nodal level analysis

By Marco Di Stasio, Alessandra Gandini, Gaudenzia Genoni, Yishak Tadele Nigatu

Introduction

This analysis focuses on the Borgatti_Scientists504 dataset, examining collaboration patterns through an undirected network where nodes represent individual scientists, and edges indicate collaborative relationships. Specifically, the structural positions of scientists from three fields — Management Sciences, Economics, and Behavioral Sciences — are compared to those from other disciplines to assess potential differences in their collaboration approaches. To achieve this, two measures of network centrality are calculated for evaluating positional importance, and their relationship with the disciplinary attribute is investigated.

Part 1

```r
#importing the libraries
library(sna)
library(igraph)
```

```r
#loading the dataset
load("./data/Borgatti_Scientists504.rda")
attributes <- Borgatti_Scientists504$Attributes
```

After importing the necessary libraries and loading the dataset, a specific cutoff (>2) is used to create a binary variable indicating whether a scientist has more than two collaborations (1) or not (0).

```r
social_net <- ifelse(Borgatti_Scientists504$Collaboration > 2, 1, 0)
```

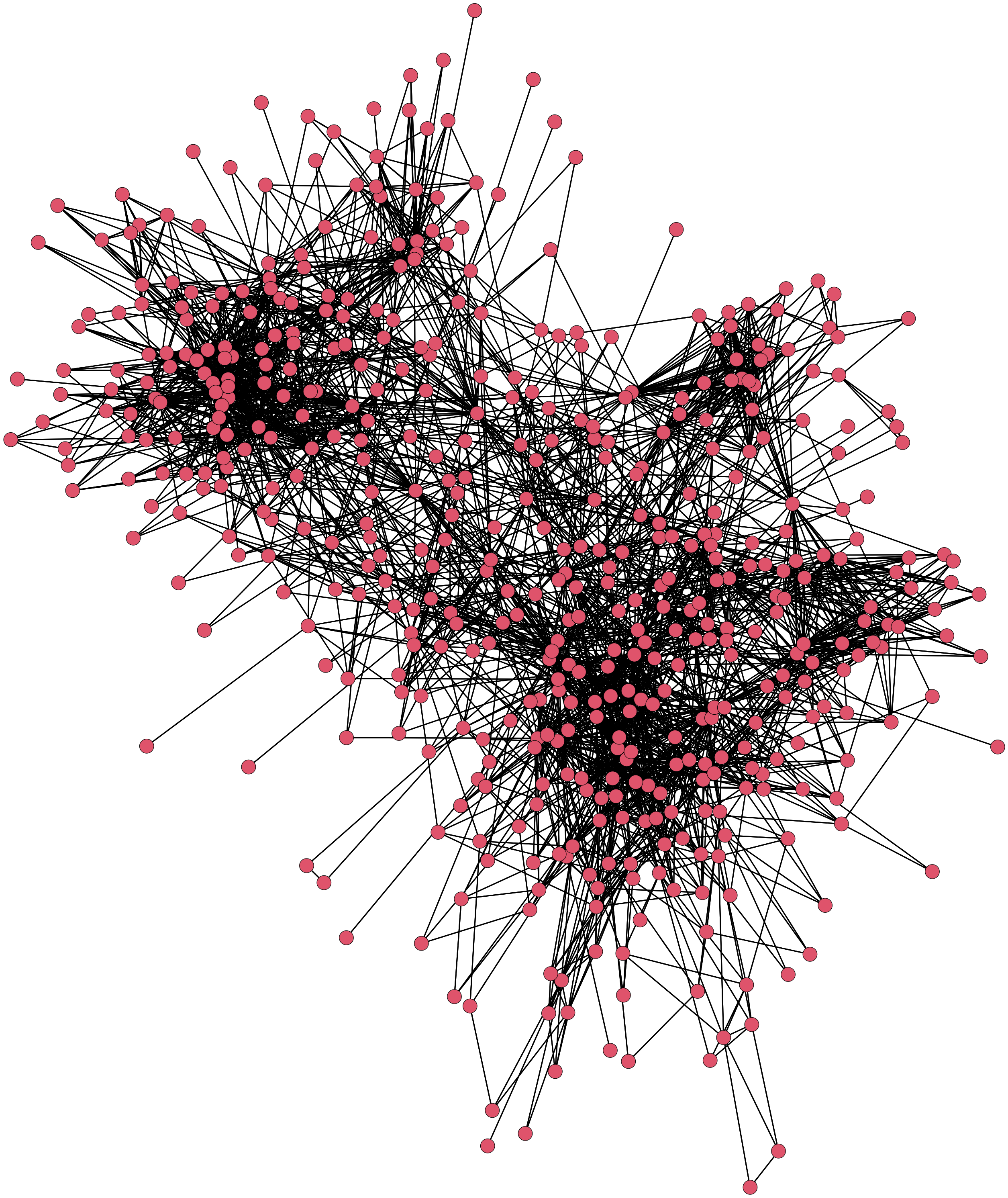
```r
#converting the adjacency matrix to an igraph object
g <- graph_from_adjacency_matrix(social_net, mode = "undirected", diag = FALSE)
```

Another important pre-processing step is creating a binary classification based on a specific nodal attribute. In this case, scientists belonging to departments 1, 2, or 5 (corresponding to Behavioral Sciences, Economics, and Management Sciences, respectively) were assigned a value of 1, while others (including NAs) were assigned a value of 0 in the new Type column.

```r
attributes$Type <- ifelse(attributes$DeptID %in% c(1, 2, 5), 1, 0)
```

The network is visualized as an initial exploratory step using gplot() from the sna package. The default graph layout is applied to visualize the network structure. The resulting graph suggests a polarization into two macro-groups; however, further analysis is required to draw more specific conclusions.

```r
gplot(social_net, gmode = "graph")
```

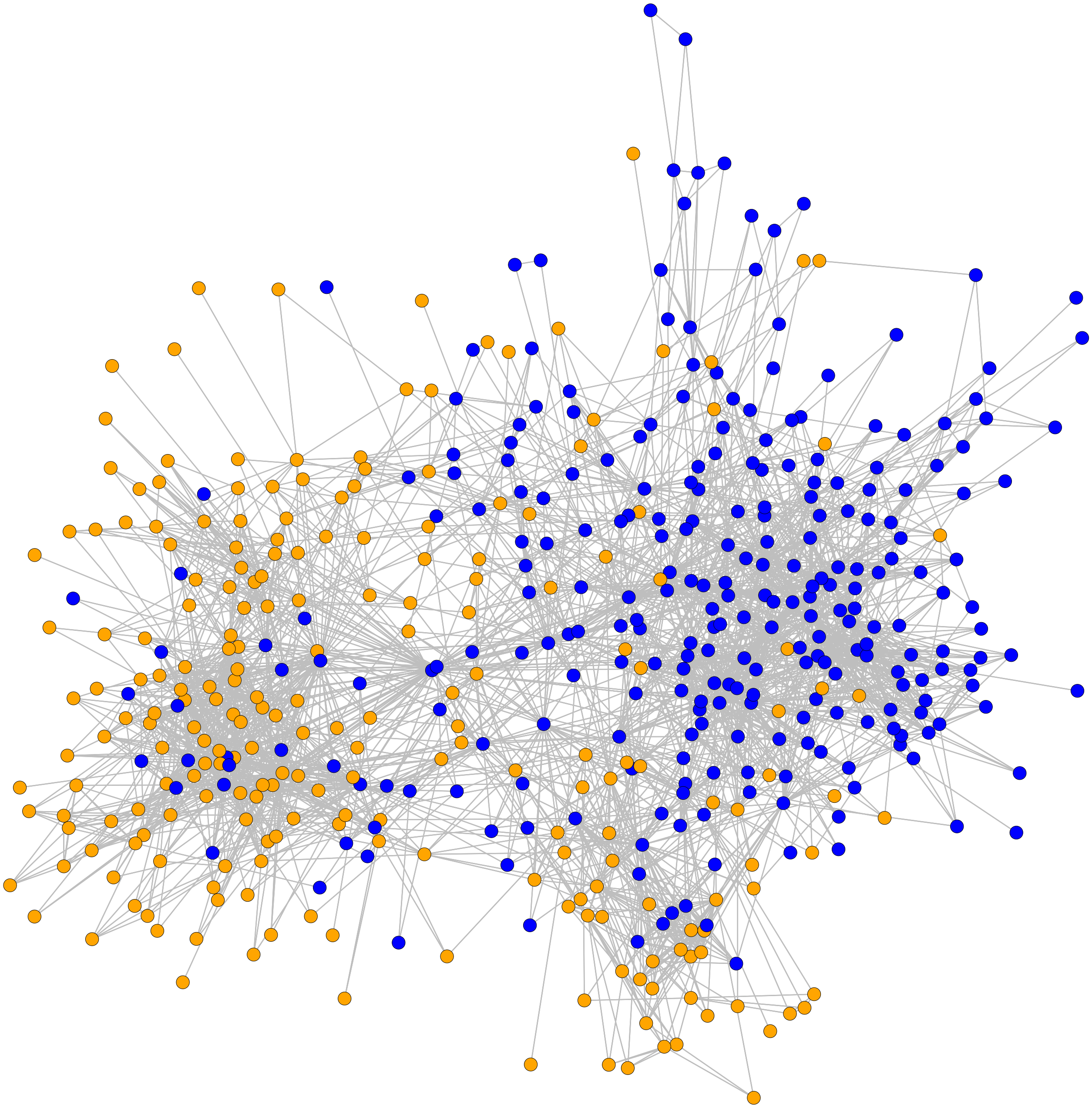


Part 2

A clearer visualization of the graph is achieved by color-coding the groups of scientists based on the previous categorization. Specifically, scientists from the departments of Behavioral Sciences, Economics, and Management Sciences are colored orange, while those from other departments are colored blue. The edges connecting the nodes are colored gray.

```
node_colors <- ifelse(attributes$Type == 1, "orange", "blue")
```

```
gplot(
  social_net,
  gmode = "graph",
  vertex.col = node_colors,
  edge.col = "gray",
  label.cex = 0.7
)
```



Interestingly, a clustering of the network can be observed, with orange nodes (representing the studied departments) collaborating more intensively among themselves and blue nodes preferentially connecting with their peers. This observation suggests, at least upon initial inspection, a degree of clustering among nodes based on departmental affiliation. Further analysis of the two measures of position, as detailed in the following parts, will provide deeper insights into collaboration patterns.

Part 3

In this analysis, the two chosen measures of position are degree centrality and closeness centrality, which will later be used in a correlation study.

- Degree centrality measures the number of direct connections a node has. In general, nodes with higher degrees are considered more prominent and are often viewed as more significant by other members of the network. When assuming that information flows along ties, degree centrality can be interpreted as an indicator of a node's level of exposure within the network (see Borgatti, Everett, Johnson, & Agneessens, 2022, p. 172).
- Closeness centrality is the sum of geodesic distances from a node to all other nodes in the network. In this case, the normalized version of closeness centrality is used, which involves scaling the values so that the maximum centrality is 1 and higher scores indicate greater centrality. In the context of information flow, closeness centrality can be interpreted as the minimum time required for something to reach a given node. A node with a high normalized closeness score is close to most other nodes, meaning that information originating from a random node can reach it more quickly. Additionally, since diffusion processes often introduce distortion, central nodes tend to receive more reliable information. Thus, a high normalized closeness centrality can be seen as an advantage, as it implies faster access to information with less distortion (see Borgatti, Everett, Johnson, & Agneessens, 2022, p. 179).

```r
# Degree centrality
degree_centrality <- degree(g, mode = "all")
```

```r
# Closeness centrality
closeness_centrality <- closeness(g, normalized = T)
```

A further step in the analysis is to fins the node(s) with the highest value for the respective measures of position.

```r
top_degree_index <- which.max(degree_centrality)

max_node_name <- V(g)$name[top_degree_index]

max_node_info <- subset(attributes, NodeName == max_node_name)

node_type <- max_node_info$Type
```

The node with the highest degree centrality, labeled N1514, has a degree centrality value of 70 (meaning it has 70 direct ties to other nodes). It belongs to one of the three departments examined (Behavioral Sciences, Economics, or Management Sciences).

```r
top_closeness_index  <- which.max(closeness_centrality)

max_node_name <- V(g)$name[top_closeness_index]

max_node_info <- subset(attributes, NodeName == max_node_name)

node_type <- max_node_info$Type
```

The node with the highest closeness centrality, labeled T0266, has a closeness centrality value of approximately 0.44. This value indicates that T0266 is relatively close to all other nodes: it is therefore well-positioned to disseminate information or resources across the network. It does not however belong to the three departments taken in consideration.

Part 4

A correlation analysis between degree centrality and closeness centrality within the network is conducted to determine whether nodes with more connections (higher degree) are also more strategically positioned to access other nodes (higher closeness).

```r
cor.test(degree_centrality, closeness_centrality)
```

```
##
##  Pearson's product-moment correlation
##
## data:  degree_centrality and closeness_centrality
## t = 25.191, df = 502, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7059481 0.7834298
## sample estimates:
##       cor
## 0.7472173
```

The analysis reveals a strong positive correlation between degree centrality and closeness centrality, with a correlation coefficient of 0.7472. This indicates that scientists with many direct connections (high degree centrality) are also well-positioned within the network to quickly access or disseminate information (high closeness centrality). The relationship is highly statistically significant, as evidenced by a very small p-value.

Part 5

The final analysis investigates how centrality metrics relate to the disciplinary attribute of scientists. By analyzing the correlations between degree centrality, closeness centrality, and departmental membership, the study explores whether scientists from the studied fields—Management, Economics, and Behavioral Sciences—differ in their positional characteristics compared to those from other disciplines.

```r
cor.test(degree_centrality, attributes$Type)
```

```
##
##  Pearson's product-moment correlation
##
## data:  degree_centrality and attributes$Type
## t = -2.64, df = 502, p-value = 0.008548
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.20229527 -0.02998615
## sample estimates:
##        cor
## -0.1170213
```

The correlation coefficient of -0.117 indicates a weak negative relationship between degree centrality and the attribute Type (type =1 for the three departments, type =0 the others).This means that individuals in the three specific departments tend to have slightly lower degree centrality compared to those in other departments. P-value is less than 0.05, so it's statistically significant (0.008548).

```r
cor.test(closeness_centrality, attributes$Type)
```

```
##
##  Pearson's product-moment correlation
##
## data:  closeness_centrality and attributes$Type
## t = -4.5818, df = 502, p-value = 5.823e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2827442 -0.1150215
## sample estimates:
##        cor
## -0.2003503
```

The correlation between closeness centrality and the departmental membership is -0.2004, indicating a weak negative relationship. It is statistically significant (p-value = 5.823e-06).

These results suggest that, compared to scientists from other disciplines, scientists from the three studied fields are slightly less central in terms of their direct connections (degree centrality) and their ability to efficiently reach other nodes (closeness centrality): they may collaborate more selectively or within smaller, denser clusters, reducing their overall centrality compared to peers in other disciplines. However, the correlation magnitudes are quite small, implying that disciplinary differences in centrality are modest and not a dominant feature of the network.

Conclusion

Iconclusion, the collaboration network of scientists from the Borgatti_Scientists504 dataset was analyzed to explore structural positions and their relationship with a disciplinary attribute. Degree centrality and closeness centrality were chosen as measures of position due to their ability to capture distinct yet complementary aspects of network prominence: direct connections and strategic accessibility.

The analysis of the whole network revealed a strong positive correlation (r=0.7472) between degree centrality and closeness centrality, indicating that scientists with many direct connections tend to occupy positions that allow efficient access to others in the network. Furthermore, when these centrality measures were correlated with the binary attribute of disciplinary membership, both revealed weak negative relationships. This hints at differences in collaboration patterns between the two groups, suggesting that scientists from the three studied fields collaborate slightly less widely and are marginally less accessible within the network compared to their colleagues in other disciplines.

Building on this exploratory work, future research could delve deeper into the network structure, such as examining the dynamics of individual departments or exploring additional attributes that may influence collaboration patterns.

Bibliography

Borgatti, S. P., Everett, M. G., Johnson, J. C., & Agneessens, F. (2022). Analyzing social networks using R. SAGE Publications.