Social Network Analysis

Iulian-Valeriu Cioată [1, 3] and Franca Maria Kiefinger [2, 3]

[1] Facultatea de Economie și Administrarea Afacerilor, Universitatea Alexandru Ioan Cuza, Iași, România
[2] Philosophische Fakultät, Universität Passau, Passau, Deutschland
[3] Faculté d'économie, gestion et administration économique et sociale, Université de Bordeaux, Pessac, France

iulian-valeriu.cioata@etu.u-bordeaux.fr
franca-maria.kiefinger@etu.u-bordeaux.fr

Abstract. This paper analyses the effects of the policy intervention on a network cluster, consisting of individual agents (scientists), and the changes between two periods (before and after treatment), using appropriate network statistics and visualizations. The paper aims to understand the gender differences observed in the cluster and the implications of male-to-female imbalances, throughout the cluster's evolution.

Keywords: NetworkX, Gender Analysis, Louvain community detection.

Table of contents

1 Introduction

The source code as well as all the other materials are available on: https://github.com/iuliancioata/Social-Network-Analysis.

This paper analyses the effects of the policy intervention on a network cluster, consisting of individual agents (scientists). The cluster has been observed in two consecutive periods, before and after a policy intervention. In the first period, none of the individuals were treated, while in the second period, all individuals were treated. All bilateral scientific collaborations are undirected, meaning that each dyad appears twice, once in each direction.

The data includes two csv files, 'nodes_cluster_G.csv' and 'edges_cluster_G.csv'. The "nodes table" contains time-invariant information on individual agents, while the "edges table" contains time-varying information on dyads. These tables can be connected via individual identifiers (id) and are presenting as follows:

Columns	Description
id <int64>	Each scientist is represented by a unique identifier.
coord <int64>	There is only 1 coordinator of the cluster, the other nodes take value 0.
core <int64>	If that person has been identified as a 'core' member of the cluster, it takes value 1, otherwise 0. This variable does not have any network relevance.
gender <int64>	It takes value 0 for male scientists and 1 for female scientists.

Table 1.1. Nodes of the network cluster

Columns	Description
source <int64>	Individual identifier of the first node of the dyad.
target <int64>	Individual identifier of the second node of the dyad.
pre_link <int64>	Active scientific collaboration in the dyad before treatment.
post_link <int64>	Active scientific collaboration in the dyad after treatment.

Table 1.2. Edges of the network cluster

2 Basic graph information with NetworkX

The French formatting for numbers is used in this paper, represented by a comma for decimal separator and a space for thousands separator, example: 1 234,567.

At first sight, graph G has **381 nodes** or vertices, connected by **2098 edges**, representing both time periods. The graph is **disconnected**, which means it contains separate components that have no connection between them [1], and **undirected**, in which the edges do not have any direction associated [1], as stated in introduction. The **graph's density** approximated at 0,029 indicates a low global level of connectivity between all nodes, respectively that only a small proportion of the total possible edges in the network exists. [2]

Plotting the **degree distribution** of graph G in Fig. 2.1, a rapid drop off can be observed, from left to right, meaning that most nodes have low degree, and only few nodes have very high degree [3]. The **average degree** of 11,013 might suggest a medium level of connectivity between its nodes.

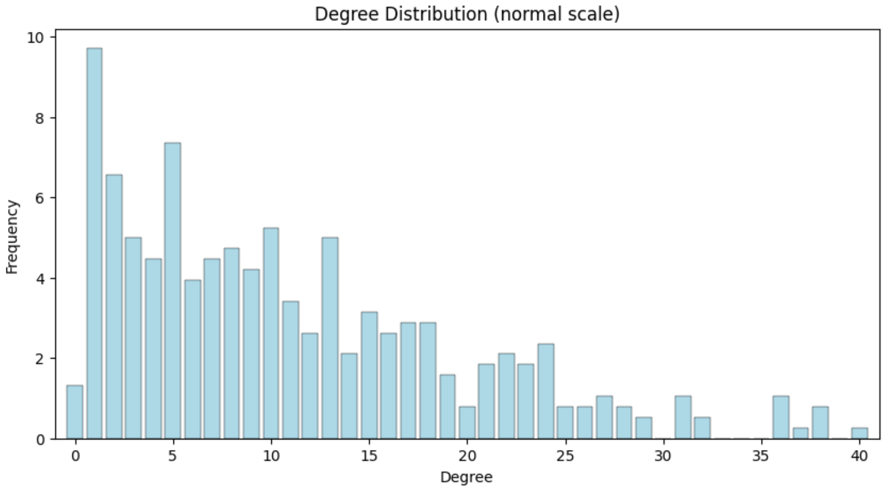

Fig. 2.1. Degree distribution (normal scale)

Adding the degree distribution on a log-log scale, in Fig. 2.2, and comparing it to a theoretical **exponential distribution**, it concludes that the degree of scientists' cluster follows an **heterogenous degree distribution,** meaning that the degrees of the nodes are **unevenly distributed**, with some nodes having a very large number of connections (high degree) and others having very few connections (low degree) [1]. Comparing the cluster with a theoretical **Erdős-Rényi** random network, it suggests that it represents a **social network** with **random topology** with **rather uniform hierarchy**, not high enough to form evident sub-clusters between nodes, but noticeable.

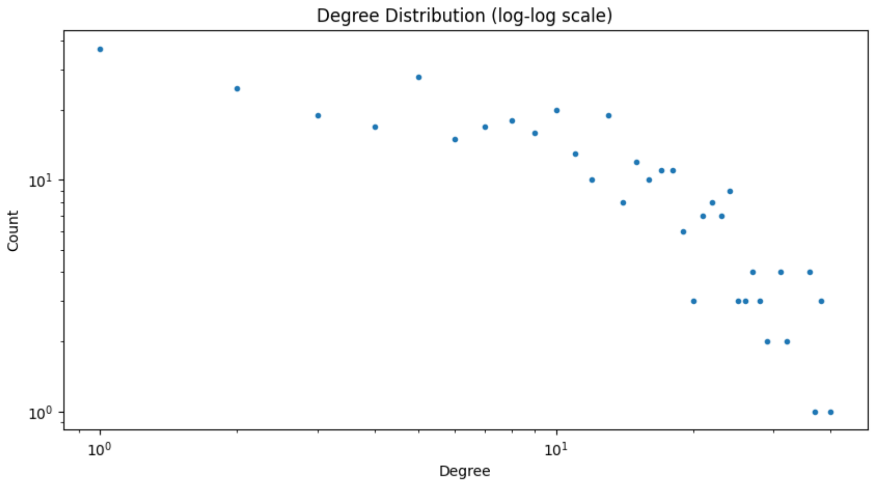

Fig. 2.2. Degree distribution (log-log scale)

The hierarchy is given by **nodes' centrality**, meaning that the few high degree nodes are more central and influential in the cluster than low degree nodes, but the **information flow** is still moderately quick and efficient, with few apparent bottlenecks [1], seen in Fig 2.3.

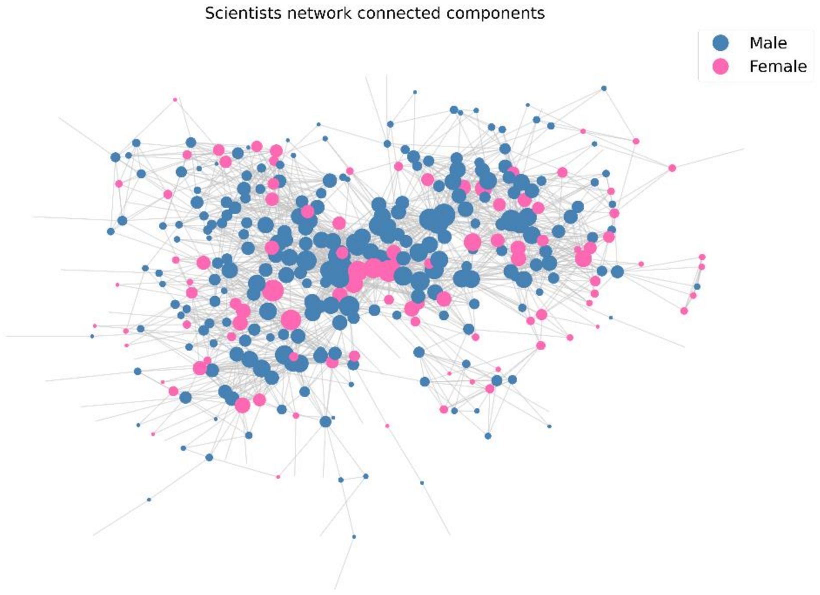

Fig. 2.3. Scientists network connected components

At first glance, the illustration of the scientists' cluster, visible in Fig. 2.3, shows a rather great share of males, compared to females, with 11 isolated nodes, connected by 3 edges, from which 6 male and 5 female scientists. Also, plotting only the connected components reveals an average shortest path length of 3,814, that suggests that the nodes in the network are relatively closely connected, indicating that information or influence can spread relatively quickly through the network.[1]

This is confirmed by the histogram of **local cluster coefficient distribution**, in Fig. 2.4, creating approximately 3 levels of hierarchy, around 0, between 0,3 and 0,6 and around 1. Also, the **average clustering coefficient** of 0,467 represents a moderate tendency of nodes to form clusters or groups with near neighbours [4], with a significant **local density** of connections in the neighbourhood of each node. Additionally, the **transitivity**, or global clustering coefficient of 0,41, neither highly clustered nor highly random, smaller in size than the local average, suggests that nodes in the network tend to form **triangles** [5], where if two nodes are connected to a common neighbour, they are also likely to be connected to each other.

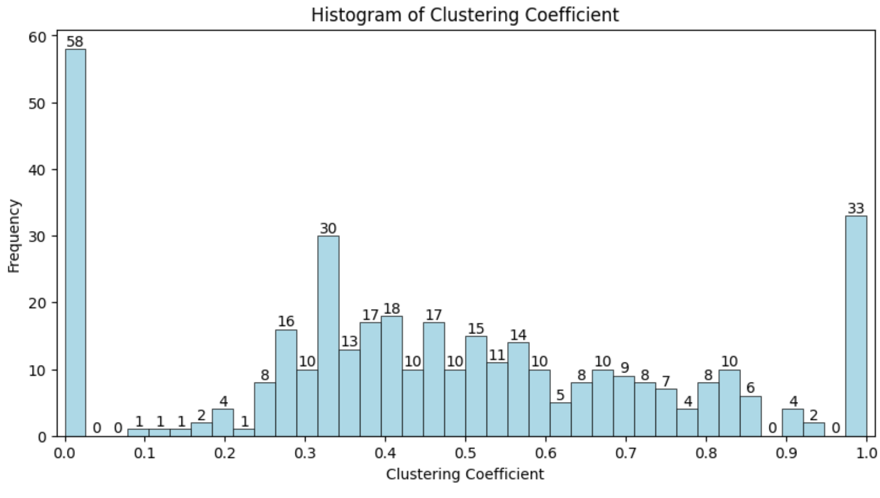

Fig. 2.4. Histogram of clustering coefficient

3 Time analysis based on gender

Dividing the analysis between gender: male and female; and between periods: before and after treatment, we can observe the evolution of the network structure.

Firstly, the graph of male nodes, before the treatment (M) can be seen in Fig. 2.1. It is a **disconnected** graph with **271 nodes** and **746 edges.**

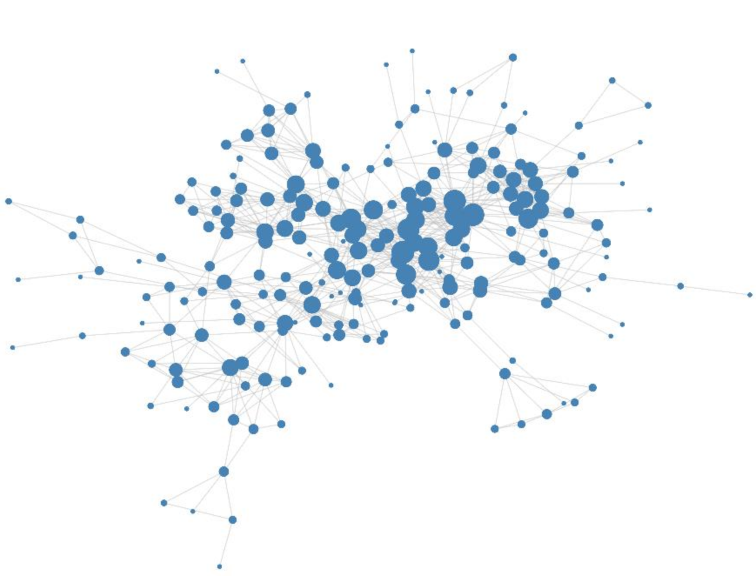

Fig. 3.1. Male Graph before treatment

Its aspect is very similar to **Fig. 2.3,** due to males' greater share in graph composition. Males' graph metrics present as following:

- M**oderate average degree** of 5,506, which suggests that individuals in the network are connected to each other, but not extremely densely.[3] On average, each male scientist has about 5,5 male connections.
- **Relatively low average clustering coefficient** of 0,35, which suggests that the males' network is relatively sparse and may not have many tight-knit subgroups. The neighbours of each male node are connected to each other, to some extent.[4]
- **Moderate transitivity** of 0,442, greater than the average local clustering coefficient, meaning that the network tends to form small groups or clusters of interconnected individuals, in moderation, being quite a few triangles in the network compared to triplets.[5]
- **Relatively low average closeness centrality** of 0,143, meaning that on average, nodes can quite quickly reach other nodes in the network, being moderately connected.[6]
- **Very low average eigenvector centrality** of 0,029, concluding no evidence of an overall influential male node. On average, the nodes in the males' network are not highly influential to each other or connected to other highly influential nodes.[7]
- **Very low average betweenness centrality** of 0,007, concluding no evidence of a node that could act as a bridge or connector between different parts of the network.[8]

Almost no change has been produced after the policy intervention, as the male graph lost only 1 edge, now having **745 edges.** The graph presents with the following statistics: **similar average degree** of 5,498; **increasing average clustering coefficient** of 0,409; **slightly increasing transitivity** of 0,46; **similar average closeness centrality** of 0,141; **similar average eigenvector centrality** of 0,032 and **identical average betweenness centrality** of 0,007. The only noticeable change that can be conducted, for the period after the treatment, could be that the local clustering coefficient went from relatively low to moderate, meaning that the neighbours of each male node became moderately well connected to each other.[4]

Secondly, the graph of female nodes, before the treatment (F) can be seen in Fig. 2.2. It is a **disconnected** graph with 110 nodes and 76 edges.

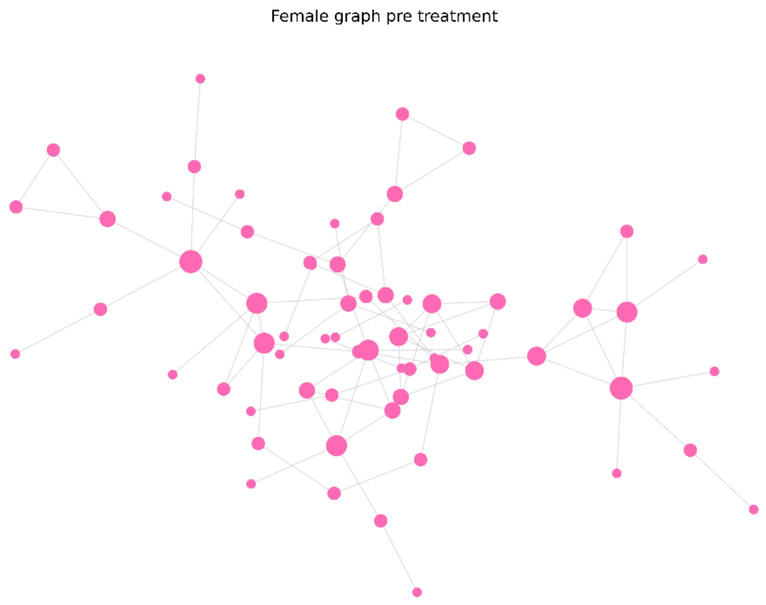

Fig. 3.2. Female graph before treatment

Its aspect is very different to Fig. 1.3, due to females' smaller share and in graph composition and very low connectivity. Females' graph metrics present as following:

- **Very low average degree** of 1,382, which suggests that agents in the network are weakly connected to each other.[3] On average, each female scientist has only about 1,4 female connections before the treatment.
- **Low average clustering coefficient** of 0,166 indicates that the network is sparse, or less dense.[4] The number of clusters or subgroups is low, and the neighbours of each female node are poorly connected to each other.
- **Relatively low to medium transitivity** of 0.387 indicates a less dense network, with moderate likelihood that two female nodes that are connected to a third node are also connected to each other.[5]
- **Low average closeness centrality** of 0.027 indicates that on average, the nodes are relatively sparse, and information may not flow easily through the network.[6]
- **Very low average eigenvector centrality** of 0.031 justifying that there is no overall influential woman in the network, indicating that nodes are relatively unimportant one to each other.[7]
- **Very low average betweenness centrality** of 0.002 indicating that the nodes are relatively at a long distance, and that the nodes may not be key connectors between different parts of the network.[8]

On the other hand, there are visible changes in the females' graph, reaching to **107 edges** after the treatment, as seen in Fig 2.3.

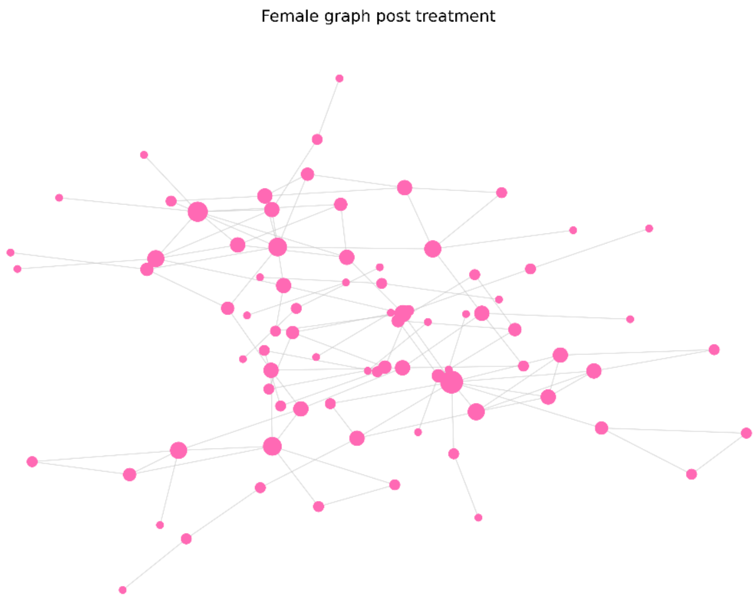

Fig. 3.3. Female graph after treatment

Females' graph metrics after the treatment have generally increased, but not drastically. The **major increase** is seen in the **average degree** of 1,945, from very low to low, which means the women are more connected after the treatment, but still almost **2,8 times less connected** than males. Also, it can be observed an **increased average clustering coefficient** of 0,28, indicating the rising tendency of cluster creation; an **increased average closeness centrality** of 0,042, denoting that information can flow more easily through the network after treatment; a **slightly increased average eigenvector centrality** of 0,035 and **slightly increased average betweenness centrality** of 0.003; and **identical transitivity** of 0.378. In conclusion, the females' graph has increased its density by strengthening the connections between nodes, but this rising is not crossing the lower to moderate limit.

Fig 3.4 observes the coordinator's cluster, before the policy intervention. The male node with id 7440, has a **degree** (or number of neighbours) of 20 and a **relatively high clustering coefficient** of 0.315, expected from the coordinator role to create a cluster around it. Although the node does not have the highest degree nor the highest clustering coefficient of the network (captured in both time periods).

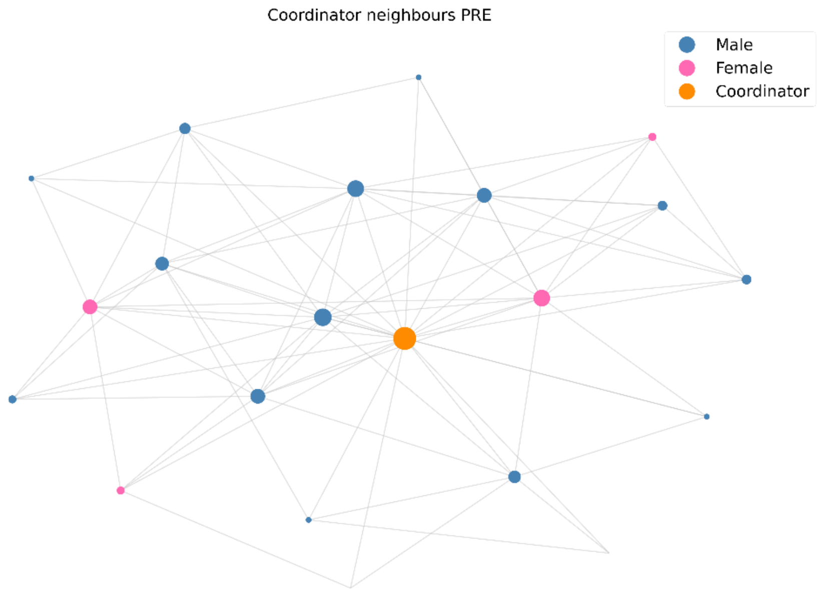

Fig. 3.4. Coordinator's cluster before treatment

The coordinator centrality measures consist of:

- **Relatively high closeness centrality** of 0.254, indicating that the coordinator is central in its cluster and is well connected to the other nodes, having a certain degree of influence on them.[6]
- **Moderate eigenvector centrality** of 0.144, which indicates that the coordinator is moderately influential in the network and is also connected to some other highly connected nodes.[7]
- **Moderate betweenness centrality** of 0.016 which suggests that the coordinator plays a moderate role in controlling the flow of information through the network, acting as a bridge between nodes.[8]

The gender distribution in coordinator's cluster, before the treatment is represented by 17 males (81%) and 4 females (19%). Knowing that it is a social network, this signifies a **certain underrepresentation** of woman in the coordinator cluster, before the treatment.

Furthermore, Fig. 3.5. depicts the coordinator's cluster evolution, after the treatment. In this period, the **degree** of the male coordinator increases to 30 (1,5 times more than previous period) and the **local clustering coefficient** slightly increases to 0,356, meaning that the cluster is denser, with somewhat stronger connections.

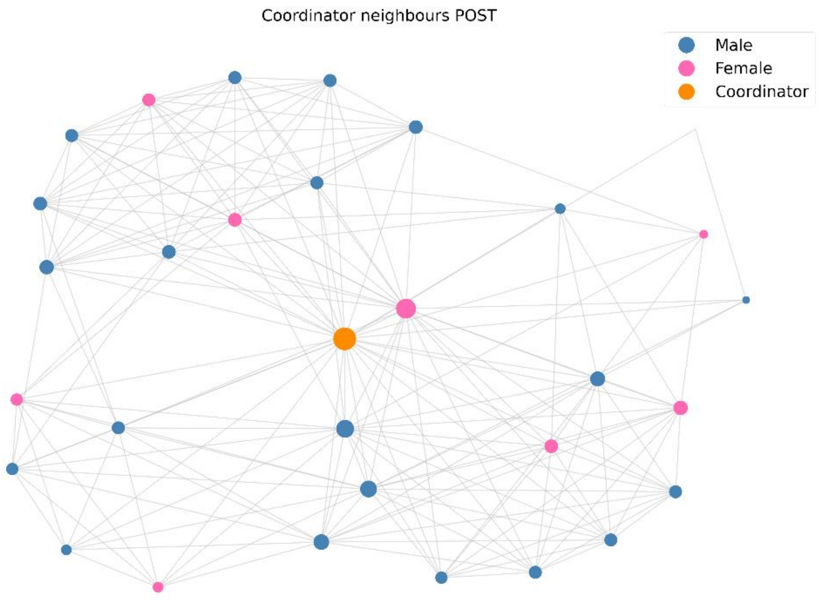

Fig. 3.5. Coordinator's cluster after treatment

After the policy intervention, the coordinator's **closeness centrality** and **betweenness centrality slightly increased** to 0,276, respectively to 0,018; the major increase could be seen in the connection of the coordinator with other influential nodes, which correlates with the **eigenvector centrality** of 0,264 (1,8 times more than previous period).

Also, the gender distribution has changed after the policy, summing 22 males (71%) and 9 females (29%), which represents an improvement of male-to-female ratio, but the gender imbalance persists.

4 Gender analysis for both periods

Taking into consideration both periods, it becomes more obvious the exponential distribution of the social network, meaning that nodes with more connections tend to increase their degree rapidly over time, while nodes with few connections tend to increase their degree to a lesser extent.[1] This instance can be observed in the males' graph and females' graph when both periods of time are analysed.

Males' and females' averages, in both periods, are described in Table 4.1:

Metric	Male	Female
Average degree	8,568	2,636
Average clustering coefficient	0,415	0,307
Transitivity	0,407	0,339
Average closeness centrality	0,231	0,097
Average eigenvector centrality	0,035	0,040
Average betweenness centrality	0,009	0,018

Table 4.1. Male-Female averages

The most significant disparity occurs between the average degree of men and women, as **male nodes have 3,25 times more connections** within the graph **than female nodes**, taking both periods into consideration. Also, on average, male nodes' neighbours tend to be more connected to each other, have greater likelihood that two of their neighbours are also connected to each other and are closer to many other nodes. While female nodes, on average, have more influential neighbours (due to proximity to male nodes) and tend to lie on more shortest paths between other nodes in the graph.

The only significant difference is not only in the same of average degree, but in the degree distribution as well. Figure 4.1. analyses this difference, by comparing the male with the female degree distribution, in both times.

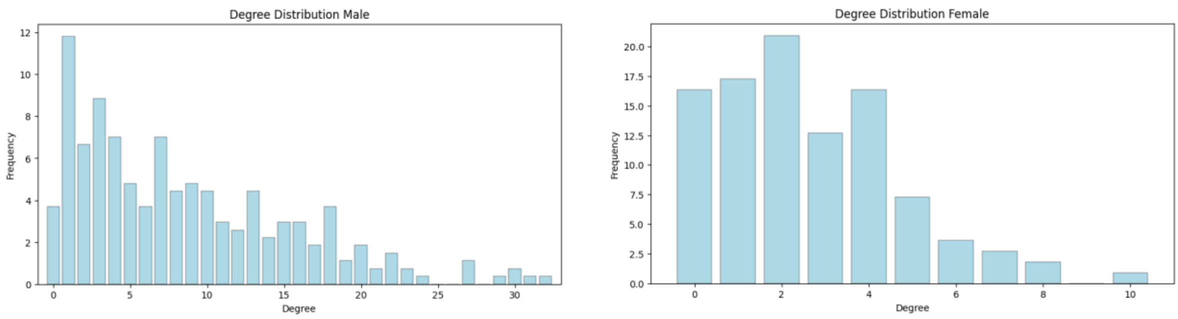

Fig. 4.1. Male-Female degree distribution

Both distributions can be considered **exponential**, as the male degree distribution looks alike the overall degree distribution, given the larger number of male nodes (271) of the total (381), but the female degree distribution is **more homogenous**, meaning that the nodes' connections are more evenly distributed. Also, the gender distribution for both periods is similar to the second one, with a ratio of **2,46 men for 1 woman**.

Within the coordinator's cluster, captured in both periods, a disproportion of 30 male nodes (76,9%) to 9 female nodes (23,1%) can be observed, or a **3,33:1 male-to-female ratio**, meaning that women are less representative in the coordinator cluster, compared to the whole network.

5 Community detection

For a better understanding of the network structure, Fig. 5.1 identifies groups or clusters of nodes that are densely connected within themselves and relatively sparsely connected to nodes in other groups, based on **Louvain** community detection algorithm, which maximizes modularity in a greedy manner [9]; as well as **Girvan-Newman**, for hierarchical structure clarifications. [10]

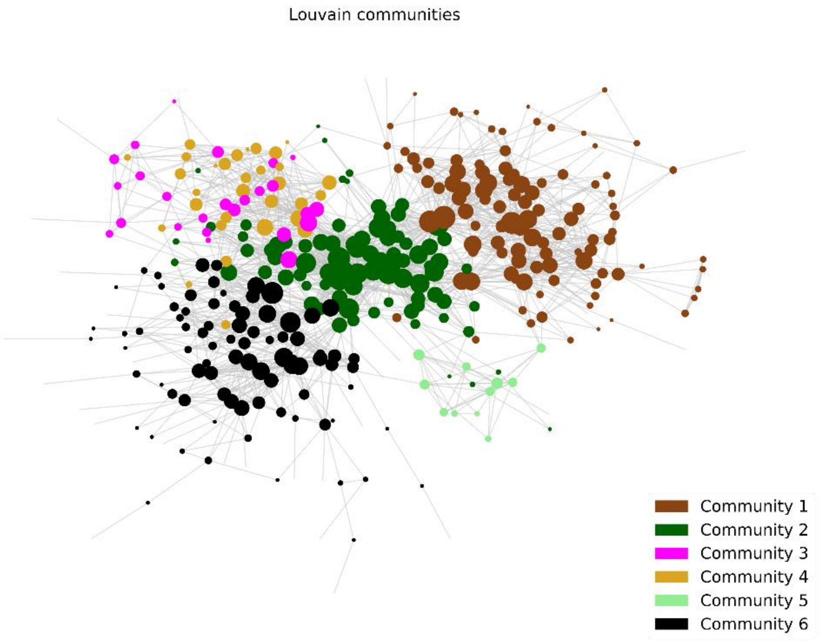

Fig. 5.1. Louvain communities

Computing the **modularity** value of 0,611, it indicates that the network has a relatively high degree modularity, suggesting that the nodes in the network tend to form tight-knit communities.[11] Considering the moderately hierarchical structure of the cluster, when applying Girvan-Newman community detection algorithm, only 2 communities are revealed with a low modularity of 0,031, similar to Louvain Community 5 and the rest of the graph. In conclusion, the cluster has a relatively low hierarchical structure, with communities that are rather overlapping on borders or do not have strict limits.

The **cluster coordinator** is part of the second Louvain community, with 35 (out of 38) of its neighbours in the same community, concluding that the coordinator's group is part of the most central group, Community 2.

Dividing each community based on nodes gender attribute, can give a clearer picture of the gender distribution in the groups, as seen in Fig. 5.2, Fig. 5.3, and Fig. 5.4.

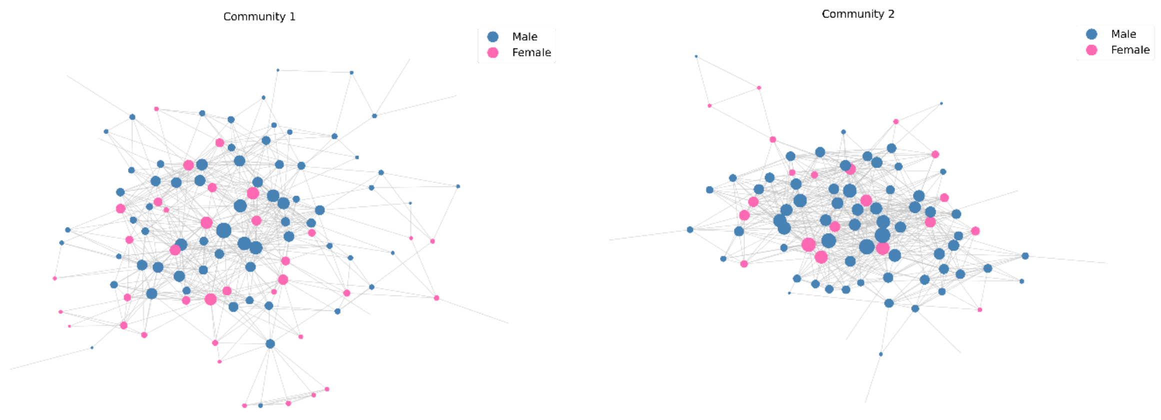

Fig. 5.2. Louvain Communities 1 & 2

Community 1 is characterized by **111 nodes** and **552 edges**, with a **low density** of 0,09, being the **largest community**. It has the **most balanced gender distribution out of all communities**, with 71 males (64.55%)

and 39 females (35.45%), or 1,82:1 male-to-female ratio. It also has the minimum transitivity of 0,424, meaning that, globally in the cluster, nodes with common neighbours don't necessary connect themselves.

Community 2 has **89 nodes** and **624 edges**, with **rather low density** of 0,159. It has a gender distribution of 65 males (73.86%) and 23 females (26.14%) and also the **maximum degree** of 14,022, out of all communities.

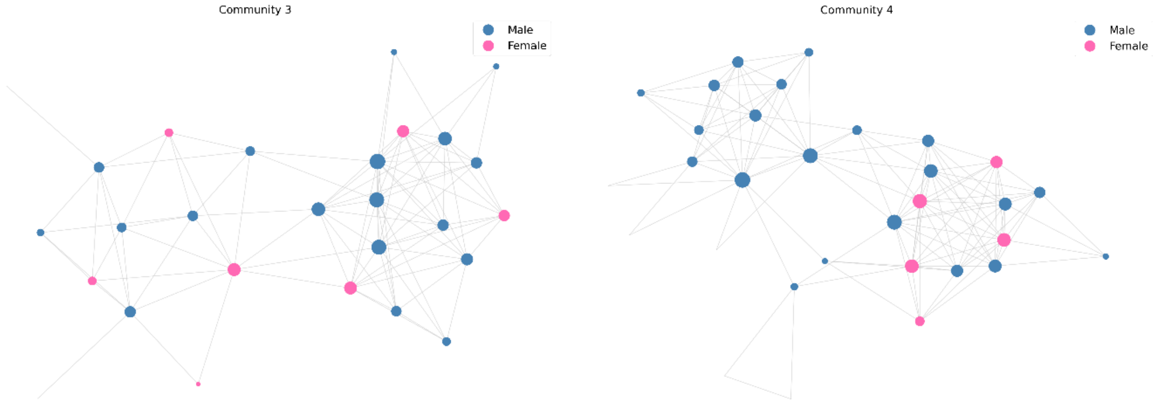

Fig. 5.3. Louvain Communities 3 & 4

As seen in Fig. 5.1, communities 3 and 4 are intersecting each other, both forming somewhat bottlenecks. Community 3 is represented by **27 nodes** with **97 edges** and a **moderate density** of 0,276, while Community 4 is represented by **31 nodes** with **126 edges** and a **similar moderate density** of 0,271.

Community 3 has 20 males (74,07%) and 7 females (25,93%). On the other hand, Community 4 is highly populated by 25 males (80,65%) and 6 females (19,35%), or a 4,16:1 male-to-female ratio, being the most imbalanced community in term of gender distribution. On the left side of the group, a region with only male connections can be observed. Also, it has the **maximum local clustering coefficient** of 0,765, meaning that nodes tend to create clusters around them and their neighbours.

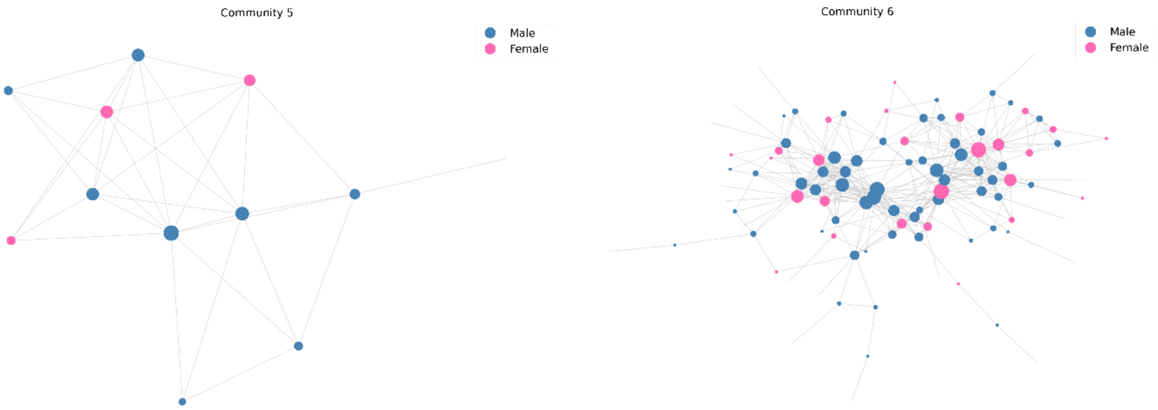

Fig. 5.4. Louvain Communities 5 & 6

If the previous communities were intersecting and communicating between them, communities 5 and 6 are at opposite sides. Community 5 has only **12 nodes**, from which 9 males (75.00%) and 3 females (25.00%), with **33 connections** between them, and a **relatively high density** of 0,5, giving the low number of nodes. On the other hand, Community 6 has **100 nodes**, from which 73 males (73%) and 37 females (27%) and **396 edges**, with a **low density** of 0.08.

Out of all communities, the fifth one has the **minimum degree** of 5,5, being the **smallest community**, and the **minimum shortest path length** of 1,59, due to the minimum number of nodes. In contrast with the sixth community with the **minimum clustering coefficient** of 0,472, suggesting that the nodes are less interconnected and the network is more loosely connected; and the **maximum shortest path length** of 3,164, being observable as well in the sparse aspect of the community.

Comparing the three types of centralities measures: **closeness**, **eigenvector** and **betweenness**, Fig. 5.5 represents each highest centrality measure, based on the 6 communities.

For Communities 1, 2 and 5 there is only **one central node**, with all 3 maximum centralities, all 3 of them being males. One thing to note is that the most central node in Community 2 is not the same as the network coordinator.

For community 4, **2 central nodes** are observable, both male, having equal closeness centrality, one of them having the maximum eigenvector centrality and the other the maximum betweenness centrality.

Community 6 also has **2 central nodes**, 1 male and 1 female, the male having the maximum closeness and eigenvector centrality, and the female having the maximum betweenness centrality.

Lastly, Community 3 has **3 different central nodes** for each centrality measure, given the sparse aspect captured in Fig 5.3, 2 male nodes, one having the maximum closeness and the other maximum eigenvector centrality, and 1 female node has the highest betweenness centrality.

Out of all 10 central nodes corresponding to each community, just 2 of them are female nodes, highlighting again the **significant underrepresentation of women** in the network.

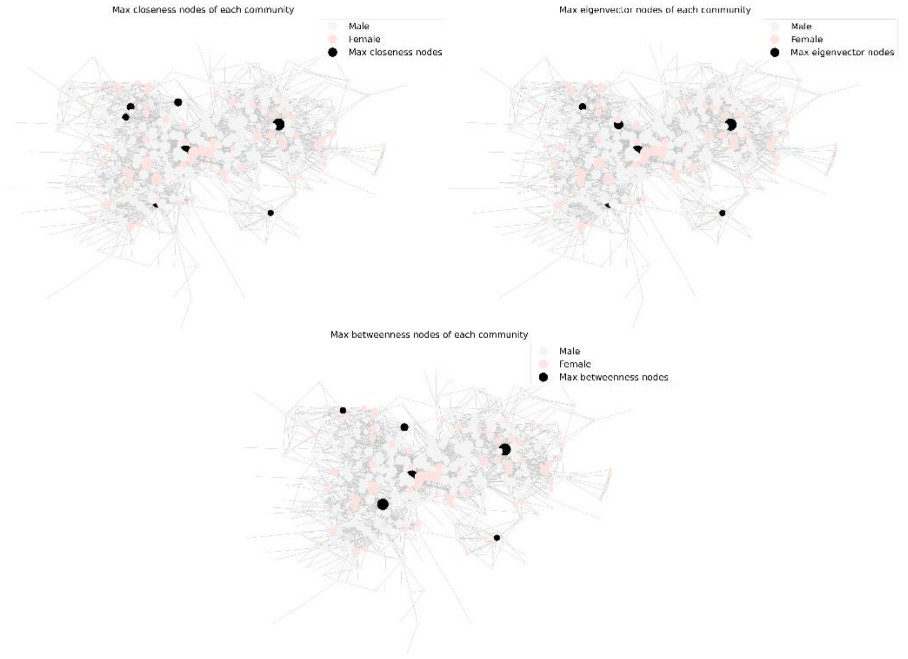

Fig. 5.5. Highest centrality measures on communities

Overall, the analysis of the resulting communities can provide insights into the structure and organization of the network, including identifying nodes that are more central or influential within their community and highlights the importance of considering community-level characteristics in network analysis.

6 Conclusions

In conclusion, this paper provides a thorough analysis of the effects of policy intervention on a network cluster of individual agents, with a particular focus on gender differences. The findings suggest that **women are significantly underrepresented in the network**, **coordinator cluster** and **each Louvain community**, with an approximate **7:2 male-to-female ratio**, in **both periods** before and after treatment. However, there is a **slight improvemen**t in their representation over the time, the major increase seen in their **connectivity** (degree) and **betweenness centrality**, in the second period, compared to the first one. These results highlight the need for continued efforts to address the gender imbalance in the scientific community. It is also important to note the **limitations** of this report, as only one cluster cannot have a final word for the whole network, and also limited access to information about the data.

References

1. Trudeau, R. J..: Introduction to Graph Theory. 4th edn. Dover Publications, Inc., New York (2013)
2. NetworkX documentation, https://networkx.org/documentation/stable/reference/generated/networkx.classes.function.density.html, last accessed 11/05/2023
3. NetworkX documentation, https://networkx.org/documentation/stable/reference/classes/generated/networkx.Graph.degree.html, last accessed 14/05/2023
4. NetworkX documentation, https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.average_clustering.html, last accessed 14/05/2023
5. NetworkX documentation, https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.transitivity.html, last accessed 14/05/2023
6. NetworkX documentation, https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.closeness_centrality.html, last accessed 14/05/2023
7. NetworkX documentation, https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.eigenvector_centrality.html, last accessed 14/05/2023
8. NetworkX documentation, https://networkx.org/documentation/networkx-1.10/reference/generated/networkx.algorithms.centrality.betweenness_centrality.html, last accessed 14/05/2023
9. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment (2008)
10. Newman, M. E. J.: Networks: An Introduction. 1st edn. Oxford University Press, Oxford (2010).
11. NetworkX documentation, https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.quality.modularity.html, last accessed 14/05/2023