Introduction :

We have a network consisting of individual agents, specifically scientists. The cluster has been observed during two subsequent periods of time - before and after a policy intervention. In the first period, no individuals were treated, while in the second period, all individuals were treated. The data for this analysis consists of two datasets that can be linked through individual identifiers. The first dataset, provides time-invariant information on individual agents, including their identifier, coordinator status, core member status, and gender. The second dataset, provides time-varying information on dyads, including the source and target node identifiers, and the nature of their scientific collaboration before and after treatment. All bilateral scientific collaborations are undirected, with each dyad appearing twice, once in each direction. Our objective in this analysis is to explore the impact of the policy intervention on the scientific collaboration of the network cluster, using a social network analysis approach. In this analysis we will try to answer the following questions:

- *How many nodes and edges are in the network before and after the policy?* This could give you a sense of whether the policy had an effect on the number of individuals or interactions in the network.

- *What is the proportion of edges to possible edges in the network before and after the policy?* This could give you a sense of how connected the network is overall and whether the policy affected the strength of connections between nodes.

- *What is the clustering coefficient of the network before and after the policy?* This could give you a sense of the extent to which nodes in the network tend to form clusters or groups, and whether the policy affected this tendency.

- *What is the distribution of the number of connections (degree) among nodes in the network before and after the policy?* This could give you a sense of whether the policy affected the way individuals are connected within the network.

- *What is the centrality (betweenness, eigenvector, closeness) of nodes in the network before and after the policy?* This could give you a sense of which nodes are most important within the network and whether the policy affected their centrality.

Properties of the data :

First, we will try to understand the properties of the data we are working with before the analysis. The properties of what we found are the following:

- We have 247 nodes and we have 1512 edge.
- Before the policy there was 844 connections between the scientists and after the policy that number increased to 1100. We can say that the policy had an effect but we don't know the details of the effect on each network, so we have to do some analysis on the networks.

- The females represent approximately 30% of the scientists, with the network being composed of 54 females and 190. We will see if the policy will have an effect on the role of the females in the network.
- We have 50 core members and 1 coordinator. Their role in the network will also be analysed.
- Only 10 females are core members, which is 20 % of the total number of core members.

The connectivity of the graph before and after the policy :

We created two graphs one before the policy which we called G_pre and the other after the policy G_post, both of these graphs are undirected.

For the pre graph:

- Number of nodes: 244
- Number of edges: 422

For the post graph:

- Number of nodes: 244
- Number of edges: 550

The pre graph had a ***density*** of 0.014 while the post graph had a density of 0.0185. In a graph with 244 nodes indicates that both graphs are relatively sparse. This suggests that the nodes in the graph are not very strongly connected to one another, and there may be many disconnected subgraphs or isolated nodes, but the post graph is more connected the pre graph because it has more edges.

For ***the average degree***, we note that a graph with n = 244 nodes, the maximum degree of any node is 243. The pre graph has an average degree of 3.459 and 4.508 for the post graph, which is relatively small compared to the maximum degree of 243, which suggests that the nodes in the graph are not very strongly connected to one another. So, the average collaborations between scientists are approximately 3 in the pre graph and 4 in the in the post graph.

The next measure is the ***average local clustering***, where we found an average of 0.412 for the pre graph and 0.480 for the post graph. These averages suggests that the nodes in the graph tend to be relatively well-connected to one another, forming relatively tightly knit clusters. a high average local clustering coefficient can indicate that the graph has a number of tightly-knit subgroups or communities, with relatively few connections between them, which we theorised before, that there are subgraphs of scientists that work very closely with each other. We can conclude again that the post graph is more tightly connected than the pre graph.

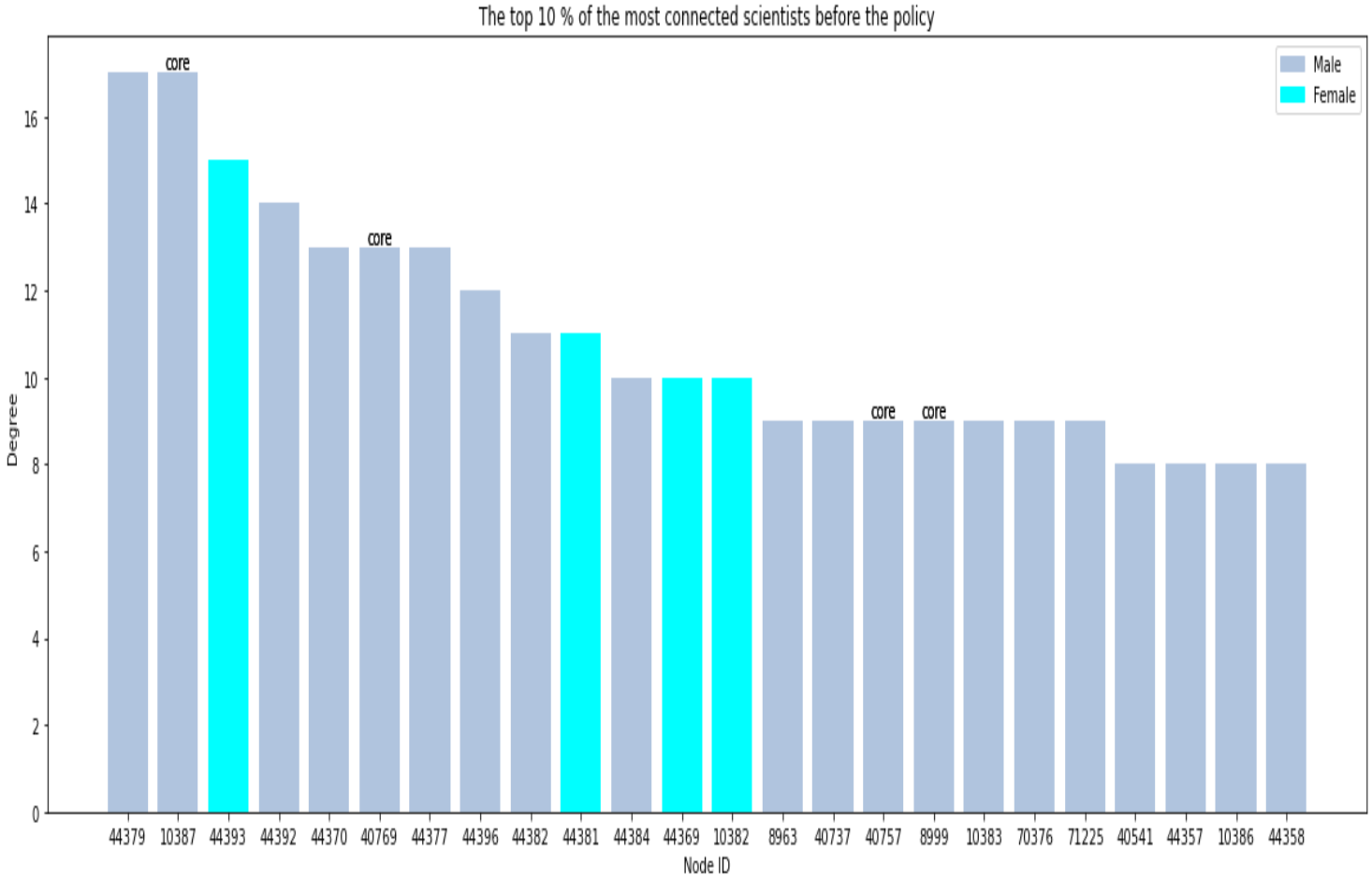
For the transitivity, we got 0.4409 for the pre graph and 0.449 for the post graph, a slight increase. A transitivity value of 0.4409 indicates that the proportion of connected triples in the graph that actually form triangles is around 44.09%. This means that the graph has a moderate degree of clustering, where nodes tend to be

connected to one another through common neighbors or shared edges. Meaning that scientists tend to work more with their neighbors.

The most connected scientists :

In both graphs we got the top 10 % of scientists with the highest degrees meaning, the most collaborations.

In the pre graph we got:



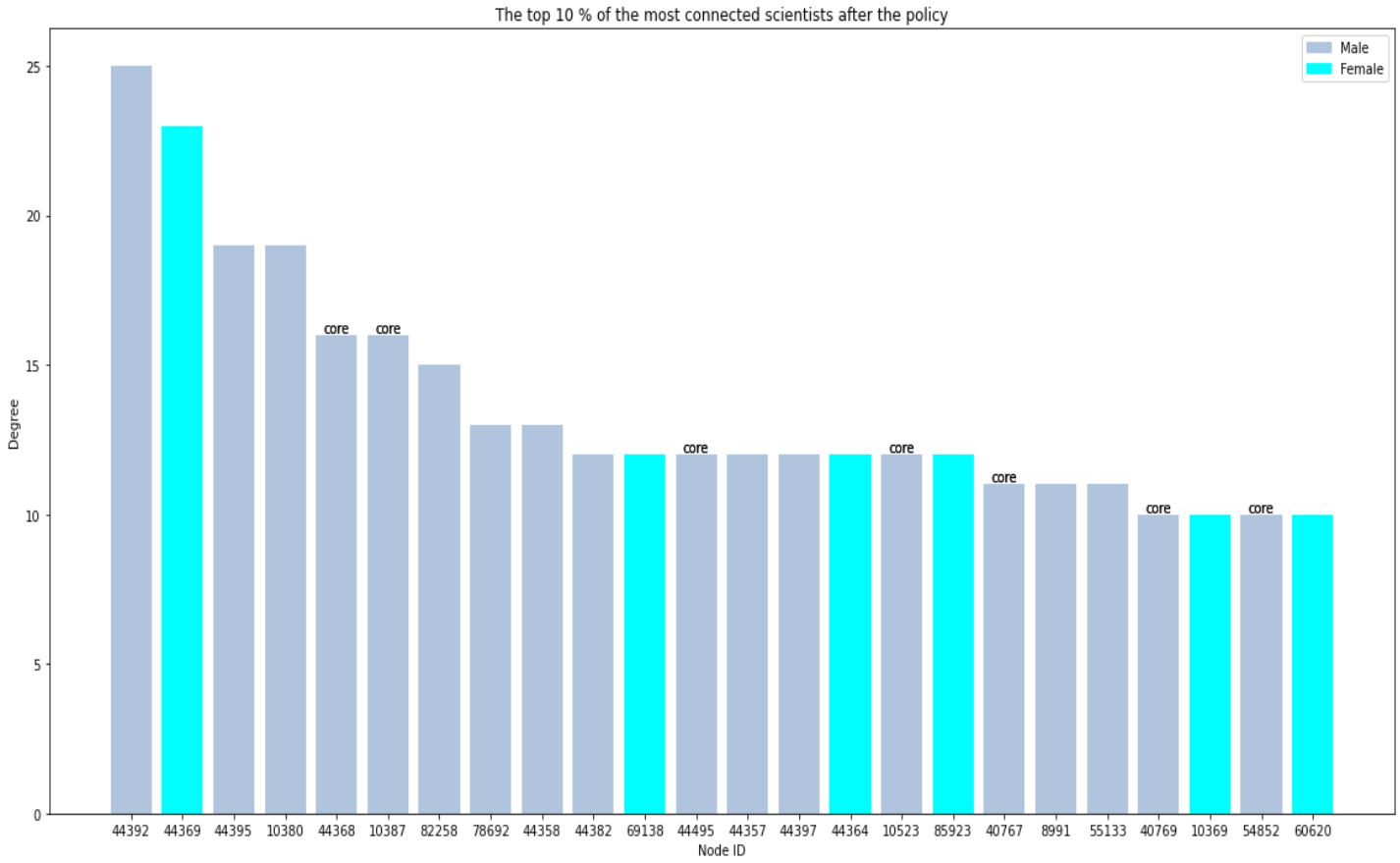
The first thing we can see is the low number of the females in the top 10 % of the nodes that are most connected and we also have 4 core members of this list, none of them are female.

The percentages are the following:

- 7.4 % of the females are in the top 10%.
- 10.52 % of the males are in the top 10%.
- 8 % of the core members are in the top 10%.

We can see that the males and core members are more represented than females in the top 10 % of the nodes that are most connected. We will see if those percentages will change after the policy.

Now we do the same for the post graph:



The number of the females has increased by 2 members, on the other hand, the number of core members in this list is 7, which is an increase from the pre graph that had 4 core members, but still, we don't see a core member who is a female.

The new percentages are the following:

- 12.96 % of the females are in the top 10%.
- 9.47% of the males are in the top 10%, a decrease from before.
- 14 % of the core members are in the top 10%, an increase of 6 %.

We can see that the policy had some effect on the properties and the measures of the network. For example, the nodes in the list changed, so some scientist became more connected than others, while some became less connected.

Centrality measures :

- Closeness centrality :

Closeness centrality is a measure of how "central" a node is in a network, meaning, how easily and quickly it can access other nodes in the network. A node with a high closeness centrality is one that is able to reach many other nodes in the network with a small number of steps, while a node with a low closeness centrality is one that is relatively isolated from the rest of the network.

We calculated this measure for both to compare the results and we got the following:

Pre_link Top 10 withCloseness centrality	
Node	**Closeness centrality**
44377	0.134489099
40767	0.132146083
40771	0.132146083
44369	0.128562461
44393	0.123940934
44370	0.123537218
40769	0.121169092
44382	0.119263918
44381	0.117417727
10380	0.117055327

Post_link Top 10 withCloseness centrality	
Node	**closeness centrality**
44369	0.196447112
10380	0.18813189
40767	0.187141723
44395	0.185837599
44392	0.185514403
44495	0.184551526
40771	0.181412894
44382	0.179580441
44357	0.178678027
69138	0.176607255

We have the top 10 nodes with the highest closeness centrality in both graphs, all of which have very similar centrality scores. This suggests that these nodes are very central in the network and can access a large number of other nodes with only a small number of steps. Between those nodes there are 3 females and equally 3 core members in the pre graph and we only have two females and the same number of core members in the post graph. So, there is a decrease in the number of females in the category of the central nodes after the policy, an increase in the values of the measure and also a change in the scientists who take that role. For example, the node number 44377 was the most central node in the graph but now it doesn't appear in the post list. On the other hand, the node number 10380 was the last in the pre list but become the second highest in the post list.

We can also see if there is an overlap between the nodes with the highest closeness centrality and the nodes that are the most connected. In the pre list, there is an overlap of 7 nodes and in the post list there are 9 nodes, an increase of 2 nodes.

- Eigenvector centrality :

Eigenvector centrality is a measure of the influence of a node in a network. It takes into account not only the number of connections a node has, but also the importance of the nodes that it is connected to. Therefore, a high eigenvector centrality indicates that a node is connected to other important nodes in the network.

We got the following lists :

Pre_link Top 10 with Eigenvector Centrality	
Node	**eigenvector centrality**
10387	**0.294075368**
44393	**0.285955343**
44379	**0.261840526**
44396	**0.258585874**
44370	**0.252468831**
44382	**0.251328655**
44384	**0.236918105**
44392	**0.216544274**
10383	**0.211299261**
71225	**0.202491428**

Post_link Top 10 with Eigenvector Centrality	
Node	**eigenvector centrality**
44392	**0.332455217**
44395	**0.268819897**
82258	**0.254303315**
44368	**0.25393039**
10387	**0.249585963**
10380	**0.239044898**
44358	**0.212812513**
55133	**0.210013585**
44364	**0.201941897**
44397	**0.201226982**

Here, Node 10387 has the highest eigenvector centrality in the pre graph, indicating that it is connected to other important nodes in the network. From the 10 nodes we only have one female and one core member, so these two categories of scientists are less connected to important nodes in the network. We still have only one female in the post graph, but the number of core members has increased by one. As for the closeness centrality, the positions of nodes in the list have changed after the policy.

For the eigenvector centrality there are 10 nodes that exist in the top 10 % nodes that are most connected, the same in the post graph.

- Betweenness centrality :

Betweenness centrality is a measure of the importance of a node in a network, based on how many shortest paths between pairs of nodes pass through that node. A node with high betweenness centrality means that it lies on many of the shortest paths between other nodes in the network, and as a result, it has more control over the flow of information or resources within the network.

Pre_link Top 10 with betweenness centrality	
Node	**betweenness centrality**
40769	**0.071321197**
44377	**0.055238823**
40767	**0.038266099**
40771	**0.038266099**
40757	**0.036016733**
44369	**0.025945562**
10387	**0.019313537**
44370	**0.018356002**
44393	**0.014556849**
44379	**0.0136724**

Post_link Top 10 with betweenness centrality	
Node	**betweenness centrality**
40769	**0.071321197**
44377	**0.055238823**
40767	**0.038266099**
40771	**0.038266099**
40757	**0.036016733**
44369	**0.025945562**
10387	**0.019313537**
44370	**0.018356002**
44393	**0.014556849**
44379	**0.0136724**

We ca see that node 40769 has the highest betweenness centrality in both graphs, indicating that it lies on many of the shortest paths between other nodes in the network and has a high level of control over the flow of information or resources.

Nodes 40767 and 40771 have the same betweenness centrality, suggesting that they are equally important in terms of their control over the flow of information or resources within the network.

This time the nodes in both lists are the same, meaning that the policy didn't affect this aspect of the network.

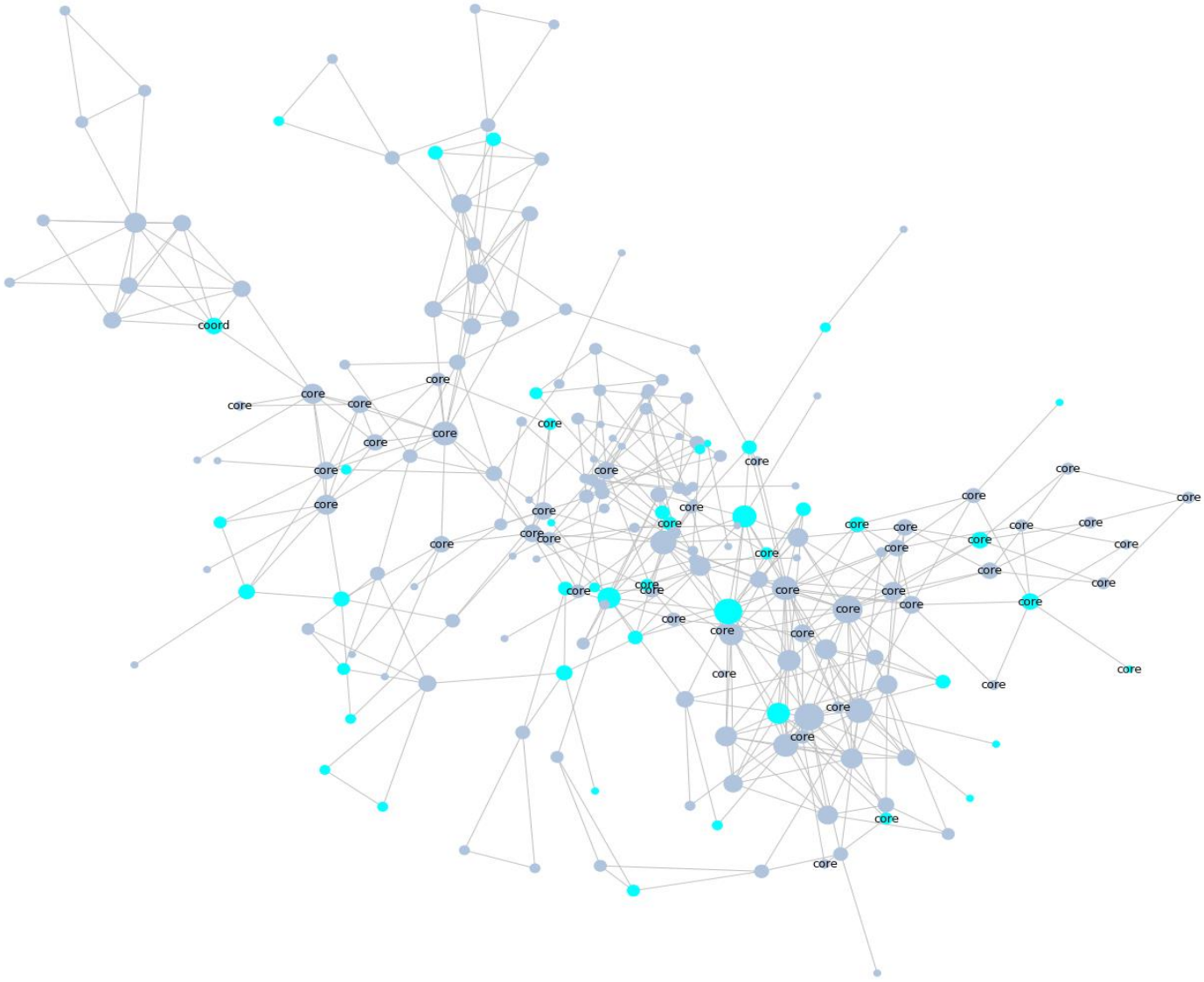
From the 10 nodes in the top, we have 2 females and 5 core members in both periods, which means that core members are the most important in the flow of information or resources within the network and none of those core members are female even though they represent 20 % of them.

For the betweenness centrality the overlap decreased from 8 to 4 nodes that exist in the top 10 % nodes that are most connected.

Drawing the pre and the post graph :

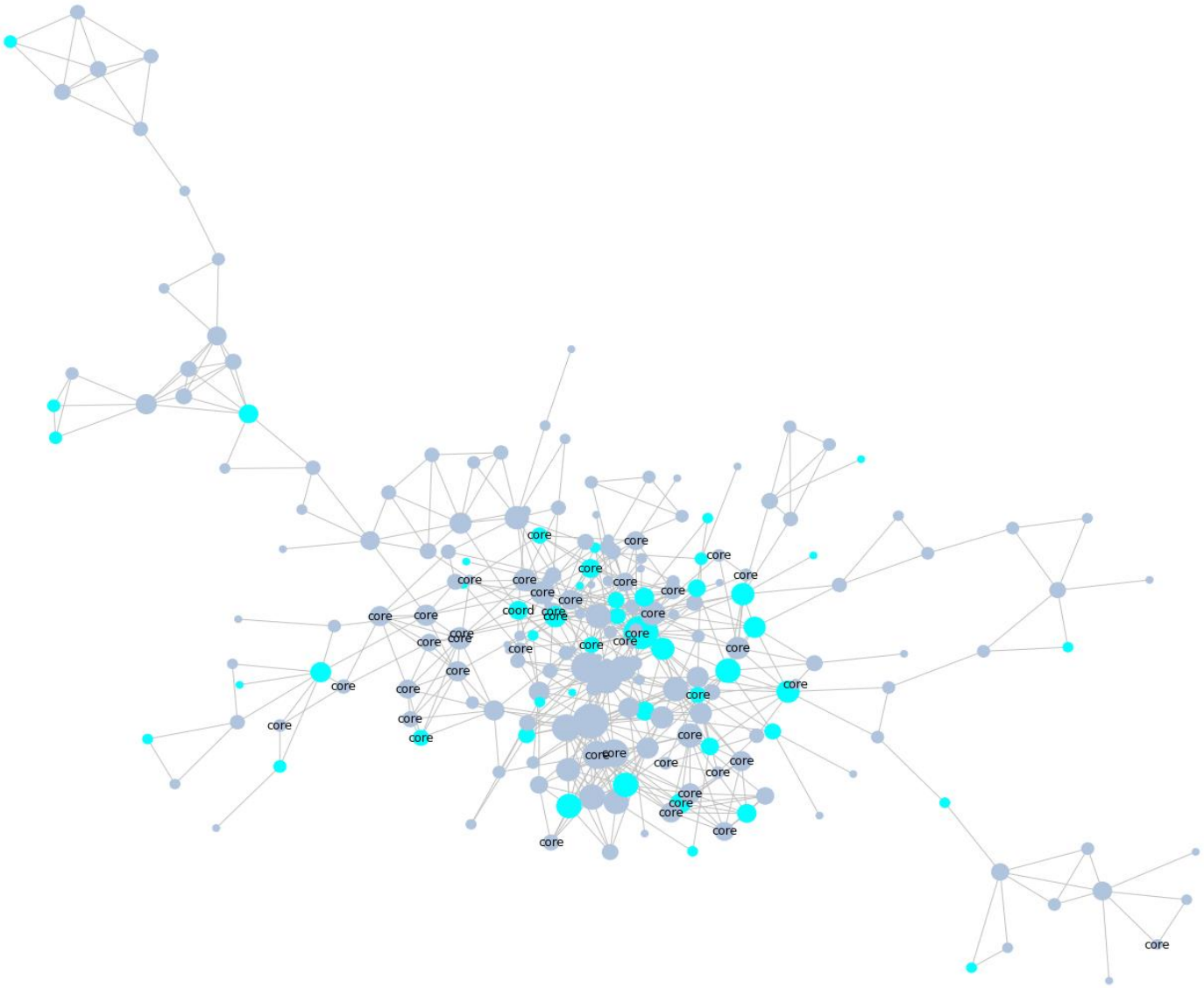
Now that we analysed the properties of each graph, we can use the visualization of the networks to see the changes. We will plot the size of each node based on the number of degrees, the colour of the node will depend on the gender of the scientist and we will put a label on a node if it's a core member or a coordinator.

The pre graph :



We can see that the graph is spares, with many subgraphs. The core members are more concentrated in the middle. There are also more collaborations around the central nodes, but overall, the graph is less dense.

The post graph :



After the policy, the graph seems more connected and the size of many nodes has gotten bigger, meaning that there are more collaborations between scientists and there are also less subgraphs.

Conclusion :

Based on these findings, we can conclude that the policy intervention had an overall positive effect on the connectivity of the network, as indicated by the increase in the number of collaborations and the increase in the clustering coefficient. However, more detailed analysis is needed to determine how the policy intervention affected specific nodes or subgroups within the network, as well as the role of gender and core members in the network.