NETWORK SCIENCE-BASED ANALYSIS OF COLLABORATION NETWORK OF DATA SCIENTISTS

SC4022 Project Report

Authors: Jiang Yunjun (U2120350B), Jameerul Kader Faizan (U2023863D)

Question 1:

What are the network properties of the collaboration network?

The DBLP collaboration network in DataScientists.xls has n=976 authors as nodes. The publications of these authors were crawled to retrieve all co-authors that have collaborated with the author. The detailed code for crawling and data manipulation can be found in *crawling.py*. The detailed code for analysis can be found in *analysis.py.*

Summary statistics of the collaboration network:


```
{
        'average_node_degree': 15.687693898655636,
        'max_node_degree': ['o/BengChinOoi', 97],
        'degree_assortativity': 0.2345366735165963,
        'average_clustering_coeff': 0.294941661575456,
        'giant_clique_size': 25,
        'giant_component_size': 895,
        'giant_component_clustering_coeff': 0.31531685669661014,
        'graph_density': 0.016239848756372292,
        'node_highest_degree_centrality': [{'o/BengChinOoi'}, 0.10144927536231885],
        'node_highest_betweeness_centrality': [{'s/DiveshSrivastava'}, 0.03074248846573167],
        'node_highest_closeness_centrality': [{'o/BengChinOoi'}, 0.40596980354628515]
}
```


The collaboration network is sparsely connected with a very low graph density, but has a relatively large clustering coefficient, indicating some tendency to form clusters. This is expected of a research network as researchers who have collaborated on previous papers are more likely to collaborate with the same group on subsequent papers.

The positive degree assortativity of the graph indicates the tendency for higher degree nodes to connect to other higher degree nodes. This may be expected due to the increased perceived credibility of researchers as they publish more papers. Researchers may tend to work with other more credible researchers.

Figure 1 shows the degree distribution of the collaboration network. The degree distribution reflects a power-law distribution with a few large clusters and many small isolated nodes.

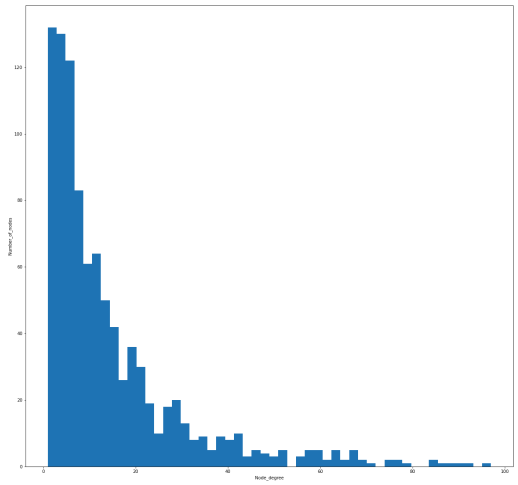

Figure 1: Degree distribution of the collaboration network.

A prominent giant component exists with size n=895, connecting 91.7% (895 out of 976 nodes) of nodes in the network. The average clustering coefficient within the giant component is not significantly larger than the average clustering coefficient, indicating it is also a sparse component.

Figure 2 shows a prominent clique, fully connected with n=25 nodes existing within the network. This is the most prominent publishing group of researchers.

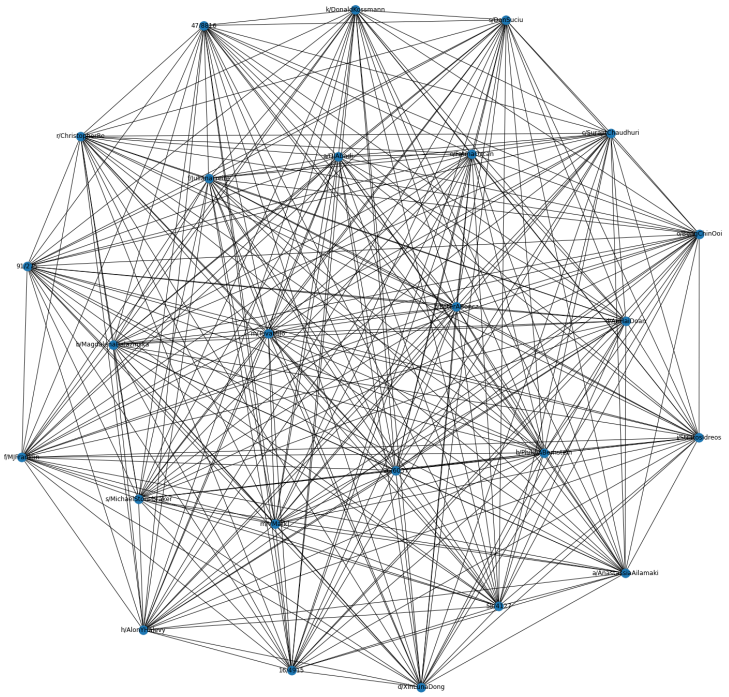

Figure 2: Largest clique in the collaboration network.

Within this clique, we can see the most prominent researcher nodes, 'o/BengChinOoi' and 's/DiveshSrivastava' which have the highest degree centrality and betweenness centrality, respectively. 'o/BengChinOoi' also had the highest closeness centrality, indicating that they were central to many different collaboration groups and appeared as the shortest connecting node. A quick search on Google Scholar revealed these researchers to be very active and credible who have been cited on an enormous number of papers.

Figure 3 shows the scatterplot correlating degree centrality to degree distribution. It is perfectly correlated. A perfect correlation between degree centrality and degree distribution intuitively implies that the collaboration network is such that researcher nodes with higher degrees naturally emerge as being more central.

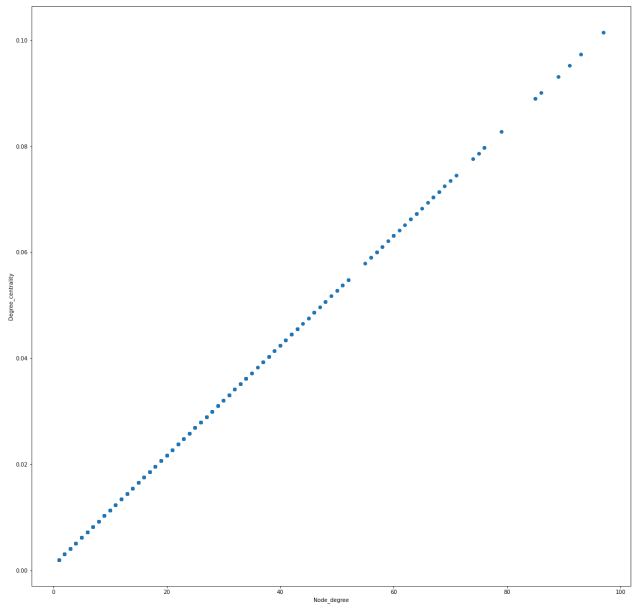

Figure 3: Scatterplot correlating degree centrality to degree distribution.

Figure 4 shows the scatterplot correlating degree centrality to betweenness centrality. The positive correlation indicates that nodes with higher degree also lie on the shortest paths connecting other nodes. In other words, more collaborative researchers have greater influence in acting as bridges for other researchers in collaborations.

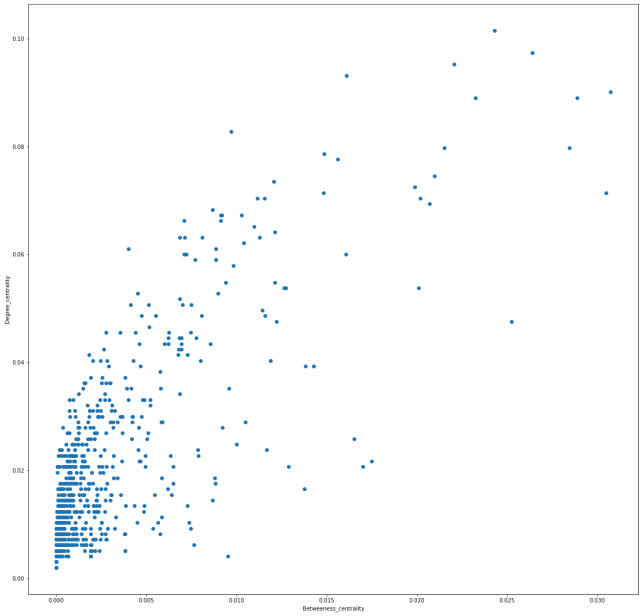

Figure 4: Scatterplot correlating degree centrality to betweenness centrality.

Question 2:

How has the collaboration network and its properties evolved over time?

Summary statistics for 2010:

{
'average_node_degree': 0.07377049180327869,
'max_node_degree': ['77/5283', 3],
'degree_assortativity': 1.0000000000000013,
'average_clustering_coeff': 0.0030737704918032786,
'giant_component_size': 3,
'giant_component_clustering_coeff': 1.0,
'graph_density': 7.566204287515763e-05,
'node_highest_degree_centrality': [{'77/5283', 'g/PaoloGarza', '60/7439'}, 0.0041025641025641026],
'node_highest_betweeness_centrality': [{'25/6583', '94/10440', 'a/LyublenaAntova', '78/5231', '67/1328-10',
'r/ChinyaVRavishankar', ……..'}, 0.0],
'node_highest_closeness_centrality': [{'77/5283', 'g/PaoloGarza', '60/7439'}, 0.0020512820512820513]
}

Summary statistics for 2024:

{
'average_node_degree': 4.813524590163935,
'max_node_degree': ['s/DanSuciu', 68],
'degree_assortativity': 0.5301161777877768,
'average_clustering_coeff': 0.2288127192351919,
'giant_component_size': 472,
'giant_component_clustering_coeff': 0.4390986171755946,
'graph_density': 0.004936948297604035,
'node_highest_degree_centrality': [{'s/DanSuciu'}, 0.07076923076923076],
'node_highest_betweeness_centrality': [{'07/1181'}, 0.03282580843960593],
'node_highest_closeness_centrality': [{'s/DanSuciu'}, 0.16000649139889647]
}

Visualisation of the collaboration network reflects a clear difference between 2010 and 2024 (Figure 5).

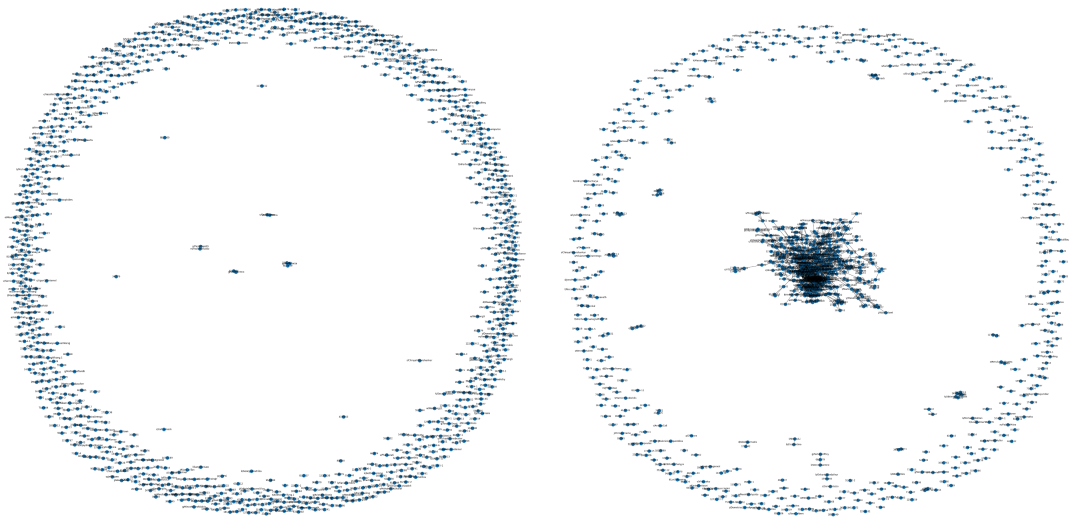

Figure 5: Graph visualisation of collaboration network in 2010 (left) and 2024 (right)

As depicted in the summary statistics, the drastic increase in average node degree, giant component size, graph density and all centrality measures indicate an increase in overall publishing activity.

The size of the giant component in 2010 was 3. As the clustering coefficient within the giant component was 1.0, indicating that the giant component is also fully-connected. This implies a triangular graph.

Over time, the number of large clusters increased and the number of small isolated nodes decreased. For example, see the change from 2010 to 2024 (Figure 6). The number of 0 degree nodes decreased significantly from ~700 in 2010 to ~440 in 2024. This may be expected as more researchers publish papers after 2010. Additionally, the maximum node degree increased from 2.9 to 43 from 2010 to 2024. Therefore, the collaboration network is still a sparsely connected graph, but has increased in number of connections and clusters.

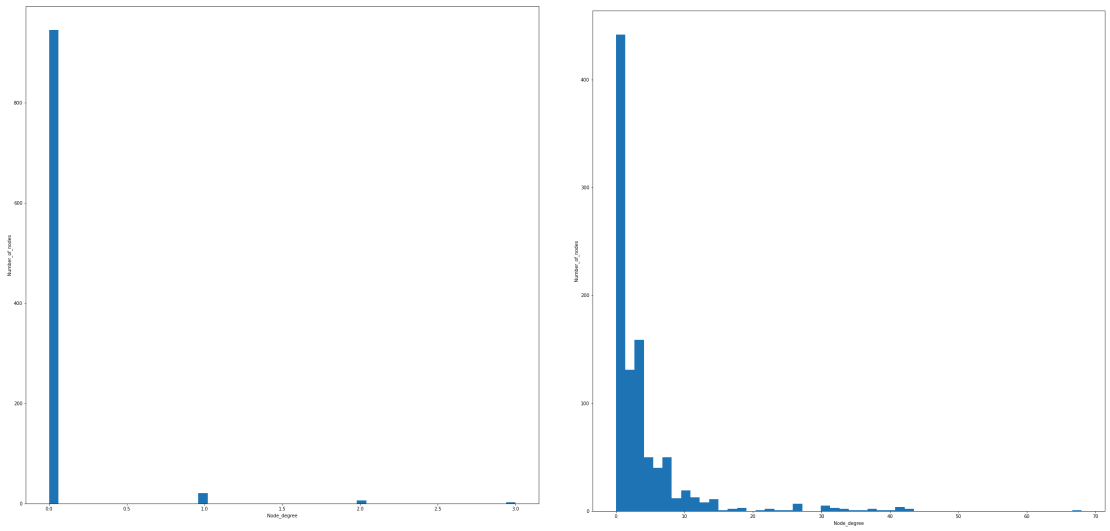

Figure 6: Degree distribution of the collaboration network in 2010 (left) and 2024 (right).

The overall increase in degree centrality naturally followed, as shown in Figure 7.

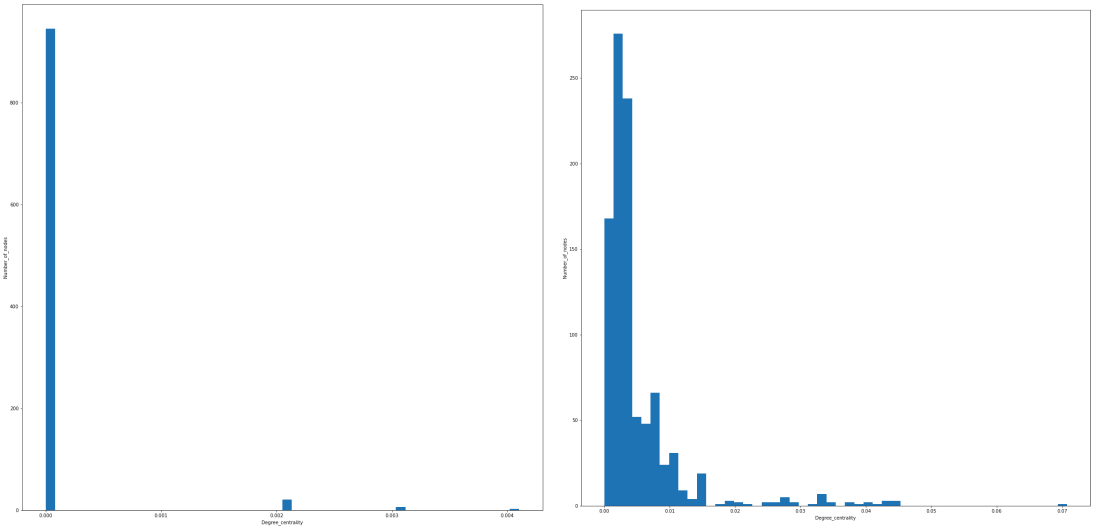

Figure 7: Degree centrality of the collaboration network in 2010 (left) and 2024 (right).

Figure 8 shows the most prominent cliques in 2010 (left) and 2024 (right). Although both are fully connected, there was an increase from n=3 to n=25 nodes within the network, indicating that the most prominent publishing group of researchers has grown in number over time. In 2010, the largest clique was only a triangle graph.

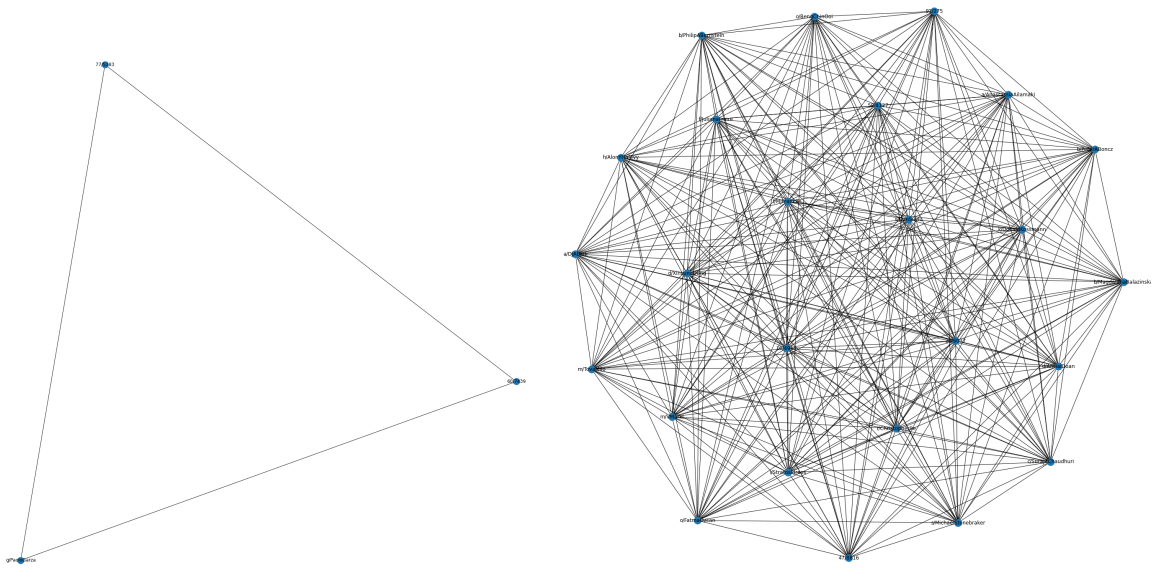

Figure 8: Most prominent cliques in 2010 (left) and 2024 (right).

The non-linear positive correlation in 2010 becomes an almost perfectly linear, positive correlation (Figure 9). Therefore, the researcher nodes with higher degrees increasingly emerge as being more central as time passes on.

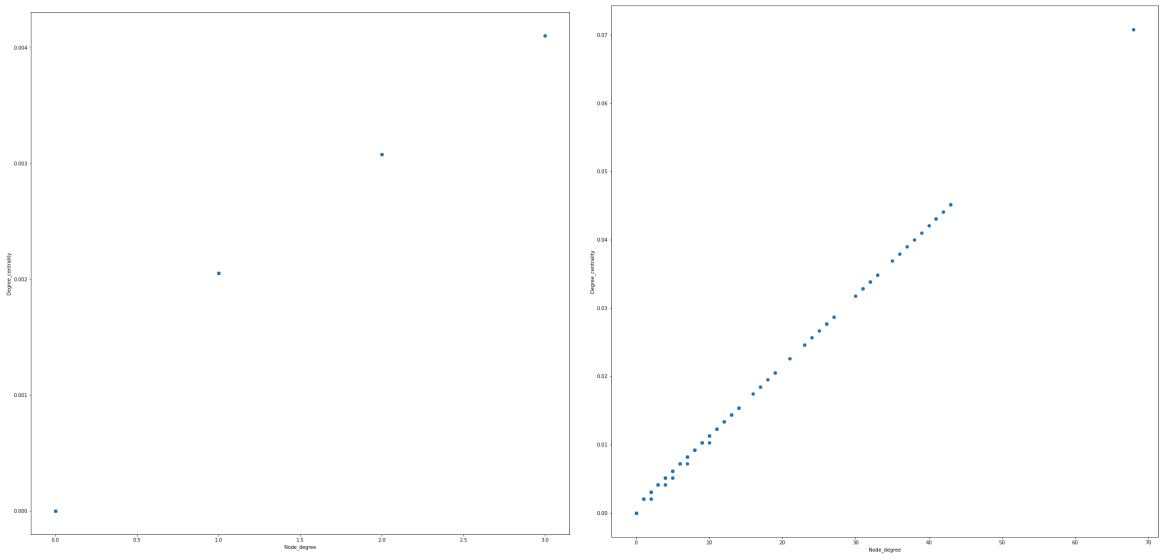

Figure 9: Scatterplot correlating degree centrality to degree distribution from 2010 (left) to 2024 (right).

Figure 10 shows the changes in the scatterplot correlating degree centrality to betweenness centrality. There is an increase in the max betweenness centrality and max degree centrality. This means that there are increasingly more collaborative researchers over time that have more influence as bridges for other researchers.

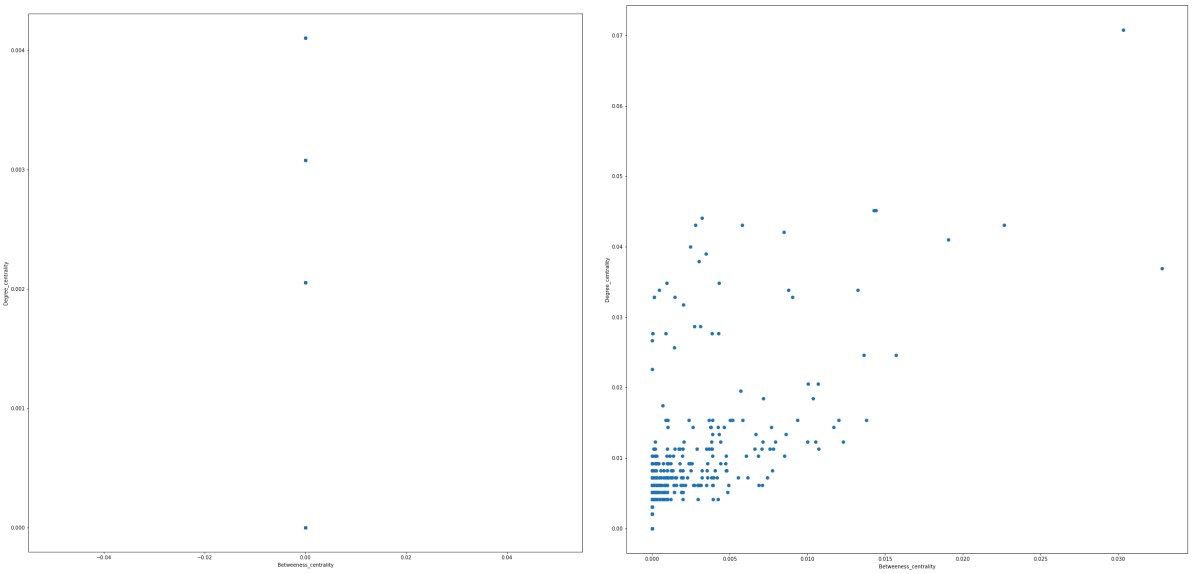

Figure 10: Scatterplots of correlating degree centrality to betweenness centrality.

Question 3:

Assume that we create a random network from the set of individuals in the input file. How does the properties of this network differ from the real collaboration network in (1)

1. Degree Distribution

Real Network	Random Network
- a few nodes (hubs) have a very high number of connections (co-authors), while most nodes have a relatively low number of connections	- The number of connections for most nodes follows a bell curve around the average degree. There would be fewer hubs and a more even distribution of connections.

2. Clustering Coefficient:

Real Network	Random Network
- Collaboration networks often have a higher clustering coefficient.	- Random networks generally have a lower clustering coefficient

3. Average Path Length:

Real Network	Random Network
- Collaboration networks tend to have a small-world property	- Random networks have a slightly longer average path length compared to a real collaboration network due to the lack of clustering

4. Community Structure:

Real Network	Random Network
- Collaboration networks often exhibit communities or groups of researchers who frequently collaborate with each other within a specific field or topic	- A random network would not have any inherent community structure

Question 4:

Algorithm to transform the Collaboration Network:

1. Identify high-degree nodes (hubs) in the original collaboration network.
2. Remove nodes connected to high-degree nodes, as well as some bridges, while prioritizing diversity.
3. Ensure that the transformed network has a smaller giant component and a larger number of isolates.
4. Enforce a maximum degree constraint (collaboration cutoff) for any node in the transformed network.
5. Implement strategies to maintain diversity in terms of country, expertise, and institutions.

Here's a high-level algorithm to achieve this:

1. **Identify High-Degree Nodes**:
   - Determine the degree of each node in the original collaboration network.
   - Identify nodes with degrees exceeding a certain threshold (e.g., `kmax`).

2. **Remove Nodes Connected to High-Degree Nodes**:
   - Remove nodes directly connected to high-degree nodes.
   - Optionally, remove some bridges to reduce the size of the giant component.

3. **Ensure Network Properties**:
   - Ensure that the transformed network has a smaller giant component and a larger number of isolates.

4. **Enforce Collaboration Cutoff (kmax)**:
   - For each remaining node, if its degree exceeds `kmax`, selectively remove some of its connections until the degree constraint is satisfied.

5. **Finalize Transformed Network**:
   - After applying the above steps, the transformed network should exhibit the desired characteristics.