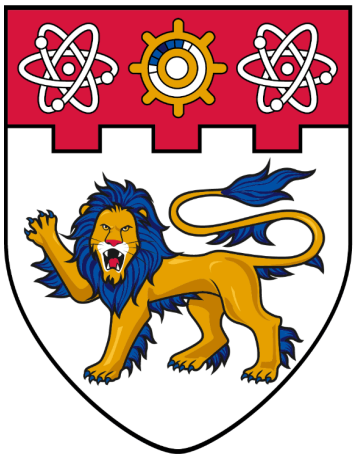

NANYANG TECHNOLOGICAL UNIVERSITY

SINGAPORE

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

CZ4071 / SC4022 - NETWORK SCIENCE

Group Project: Network Science-Based Analysis of Collaboration Network of Data Scientists

Project Report

Contributors:

Name	Matriculation Number
CHOLAKOV KRISTIYAN KAMENOV	U2123543B
DHANYAMRAJU HARSH RAO	U2023045C
CHUA WEE SIANG FRASER	U2122535E

Table of Contents:

1. Introduction

1.1 Background: Collaborative Networks

A collaborative network is a network of various entities that represent the nodes in the graph and are connected via their collaboration in different jobs. Typically, the network is well distributed in terms of the characteristics of the individuals in it. Analyzing collaborative networks explains their structure, behaviour, and evolving dynamics of networks. Focusing on the collaborative network of data scientists, the developing technological advancement has resulted in the increase of scientific papers being published every year and the diversity of the collaboration between their authors. Thus, evaluating such networks can provide significant insights into the effectiveness of collaborative practices, the geographical and institutional distribution, and the patterns in the collaboration.

1.2 Project Abstract

In this project, we constructed a network based on the crawled data for a selection of around 1000 data scientists. By studying how data scientists collaborate on different projects and papers, we identified the structure of the network and its characteristics. We performed this analysis by visual and statistical inspections of the collaborations between the various scientists. For a better understanding of the network, also, we compared it with a randomly generated one with the same features, and monitor its evolution over time.

2. Data Crawling

2.1 Given Data Analysis

The starting point for the data collection is the given table with more than 1000 records with the information about various data scientists. The information consists of:

- “name” - the name of the scientist
- “country” - the country of the institution of the data scientist
- “institution” - the institution the data scientist is part of
- “dblp” - the link to the DBLP page of the scientist
- “expertise” - empty column that is randomly assigned later

2.2 Data Collection

Despite the different columns in the given collection, none of them represents information about the collaboration between the individuals. Thus, we used the DBLP link to extract all recorded data about each scientist and use it to construct the collaborative network.

In the first place, the dataset was inspected for duplicates and it was found that there were several records with duplicated DBLP links but different country and institution values. This was due to the fact that some scientists are part of multiple institutions in various countries. These records were combined and the institutions and countries data was stored in the same columns but in the form of a set.

After further inspecting the URLs and the pages to which they point, we noticed that every author in the DBLP collection is identified by a unique “pid” identifier. Also, DBLP provides the functionality to extract all data about each person in an XML file. By accessing each link in the “dblp” column from the provided dataset, we can extract the “pid” of the scientist from the part of the final URL after redirecting (Figure 1).

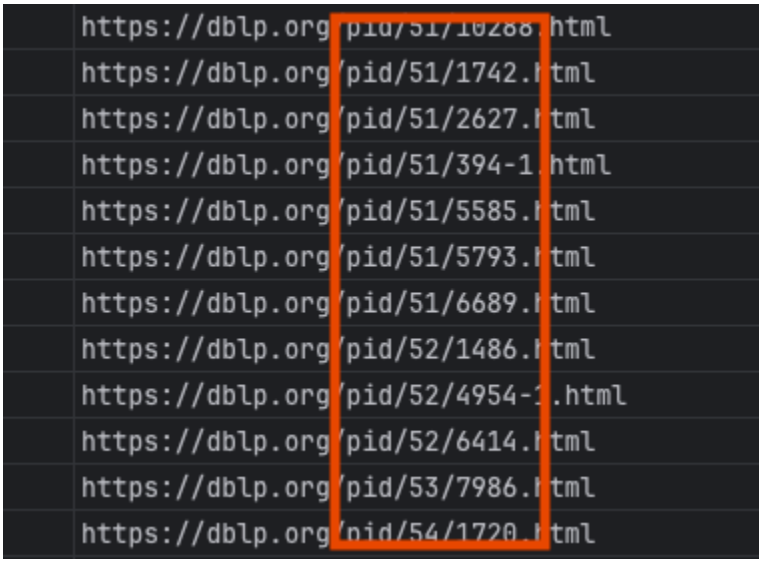

Figure 1. “pid” in the DBLP URL

Then, the link to the XML file is derived by simply replacing the “.html” extension with a “.xml” one. Retrieving the XML can be achieved by sending a request to the new XML link and saving the response as an XML file. During this process of crawling the data from the DBLP collection, we found that some of the links were broken, and after further investigation, it appears that the pages of the scientists were deleted. To maintain a

clean dataset, we removed these entities. The XML files for the data scientists were stored in a separate directory.

2.2 Data Preprocessing

After the data collection, theoretically, all data needed for constructing the network is available. However, the XML files are too broad and present the collaborative information in a very unfriendly format for network construction.

In order to utilize the data from the XML, we converted it to a table with all papers (articles, proceedings, inproceedings, etc.). We preprocessed the data by iterating through all XML files and searched for the specific tags that stored the information we needed. Each XML file consists of multiple “<r>” tags that record the data for each paper, each “<r>” tag has a subtag with the type of the paper, and the specific data for each paper is located inside these tags. The title, year and author of the paper were extracted by looking for the following tags in the XML structure:

- “<title>” - stores the title of the paper
- “<year>” - stores the year of the paper
- “<author>” or “<editor>” - stores the authors of the paper

Every paper from each scientist’s XML file was converted to a single row in a CSV table. The table had the following columns:

- “Title” - stores the extracted title of the paper
- “Year” - stores the extracted year of the paper
- “Key” - stores the unique identifier of the paper
- “Authors” - stores the authors of the paper that are part of the given scientists set
- “External Authors” - stores the authors of the paper not in the given scientists set
- “Publication Type” - stores the type of the paper (article, book, etc.)
- “File” - stores the name of the XML file the data was extracted from

Now, this table holds all the data needed for constructing the collaborative network. Each paper represents a link between every pair combination of its authors.

2.3 Data Cleaning

Although the data is converted in the wanted format, it still needs to be cleaned from duplicates or unmeaningful records. Also, the data was collected from multiple files, which resulted in recording the same paper several times for each scientist’s XML file it occurs in. Duplicated records and ones with errors corrupt the data and may result in an unrealistic network with invalid connections. To overcome this problem several cleaning rules were applied.

First, we dropped all duplicated records, the result of collecting the papers from multiple scientists' XML files. However, this was not enough as some of the duplicated publishing remained. After a closer inspection, it was observed that these records had the same title and authors but they had different keys. This was because some papers were published in several conferences or journals. Thus, the duplicated records with the same title and authors were removed. Although, there were still records with identical titles, most of them had strange titles with less than 2 words. Also, many of these duplicated papers had only one author. They were not significant for the network because they could not form any links with only one author. Finally, the only duplicates left were papers with the same titles but their authors' arrays were subsets of each other. As these records were not invalid, the ones with the same title were combined into one record where the authors array consisted of all unique authors from the duplicates.

The only thing left was to randomly assign the expertise column of the scientists with values from 1 to 10. As expected, this formed a uniform distribution that can be seen in the pie chart in Figure 2.

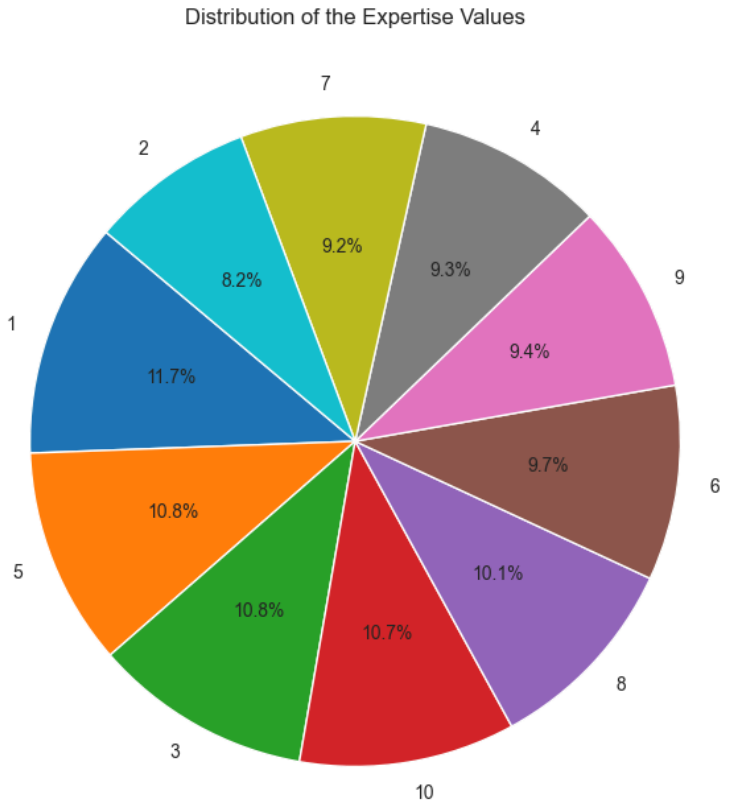

Figure 2. The Distribution of the Expertise Values

2.4 Analysis on the Collected Data

After the cleaning, there are 90 837 publishings in the papers collection and 1 052 scientists identified by their unique “pid” from the cleaned data scientists collection. These statistics are calculated from the collections that include the external authors too.

The distribution of the countries is presented in Figure 3. As we can see, most of the scientists are associated with the United States, with the second and third places taken by China and India. As expected, the countries with the most data scientists in the given collection are developed countries, known for investing in education and technology.

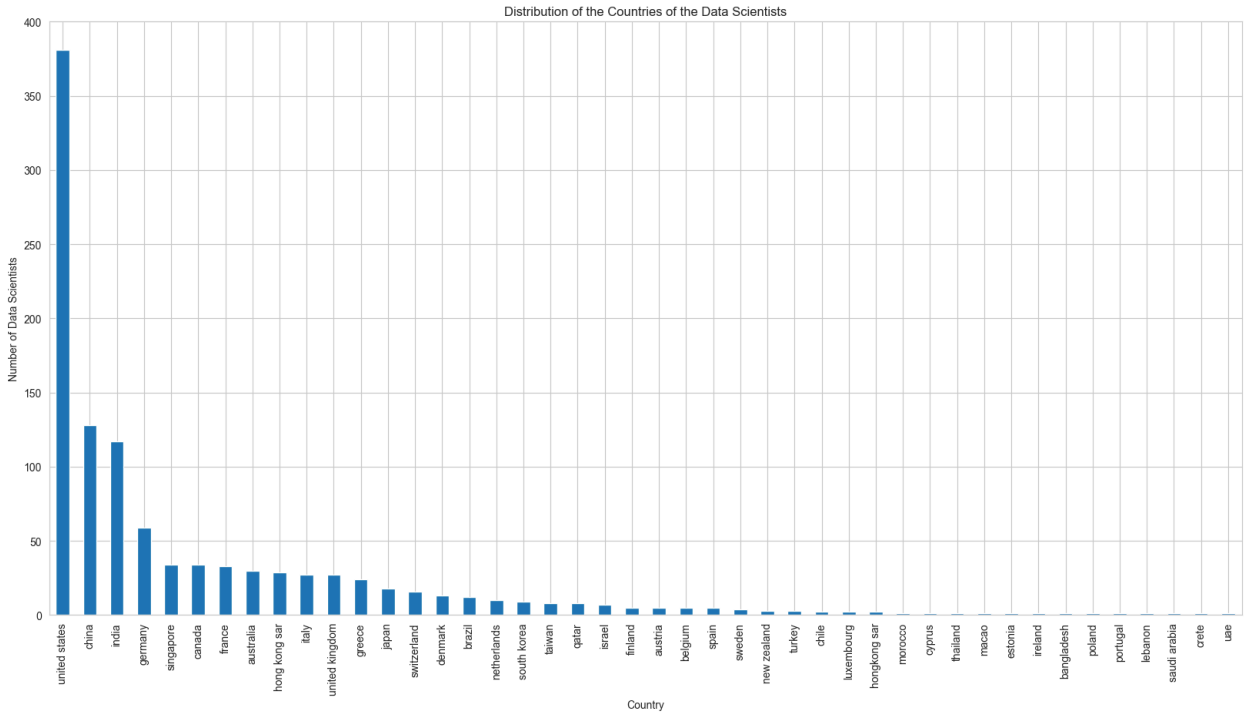

Figure 3. The Distribution of the Scientists’ Countries

As for the distribution of the institutions, there were too many unique values in the institution column, thus, only the top 20 of them were displayed in the bar chart in Figure 4. Again, the majority of the scientists are from industry-leading companies, associated with prestigious research and significant technological innovations. The first two places are taken by Microsoft and Microsoft Research, followed by IBM, Google, and universities like NUS, NTU, MIT and Tsinghua University, known for their focus in Data Science.

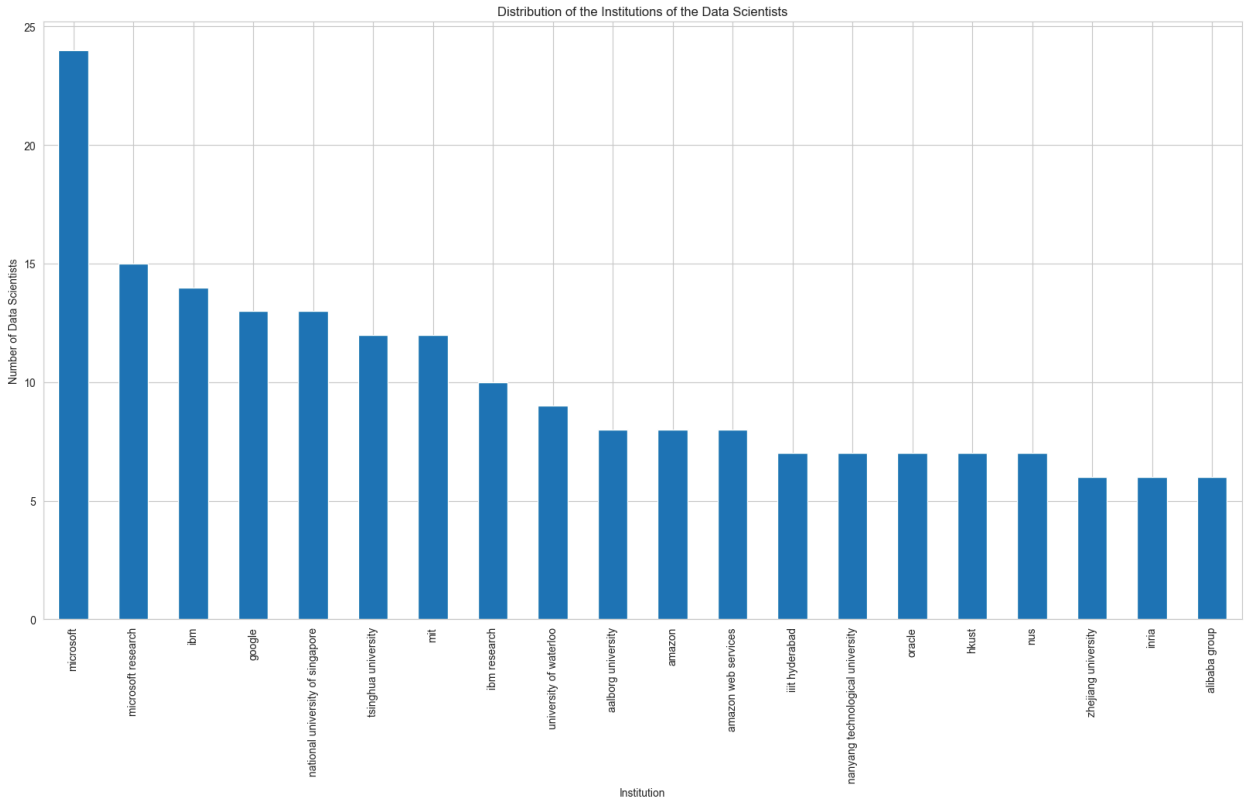

Figure 4. The Distribution of the Top 20 Scientists' Institutions

3. Network Analysis

In this section of the report, we will analyze the structure and properties of the network, using only the data scientists from the initial dataset as nodes. External authors will not be considered in this analysis. For the network analysis component of our project, we have selected the Python library NetworkX. NetworkX is one of the most well-known packages for working with network structures. Our decision to choose it was motivated by its ease of use, diversity of functions, and ability of dynamical network manipulation. These features make it an ideal tool for detailed network analysis of networks such as the collaborative network.

3.1 Network Construction

The network construction phase is a significant beginning for the network analysis because it builds the NetworkX network graph that is later used for computing the properties of the collaborative network of data scientists.

First, the output from the data preprocessing part (the papers collection) has to be converted to a more suitable format, so it can be used for constructing the graph. We

have decided to do this using a custom dictionary data structure similar to an adjacency list. The keys of the dictionary will be represented by the scientists' "pids" and their corresponding values will be all other scientists they have ever collaborated with (and some information about their collaboration - year, number of collaborations, etc.). The collaborated data is extracted by iterating through the papers collection.

After converting the papers collection to the custom dictionary format, the keys will be added as nodes using the NetworkX functionality to the graph, and for all keys and their corresponding scientists (they have collaborated with) an edge will be added to the graph. Having converted the data into the adjacency list format and adding the nodes and links from it, the constructed graph is displayed in Figure 5.

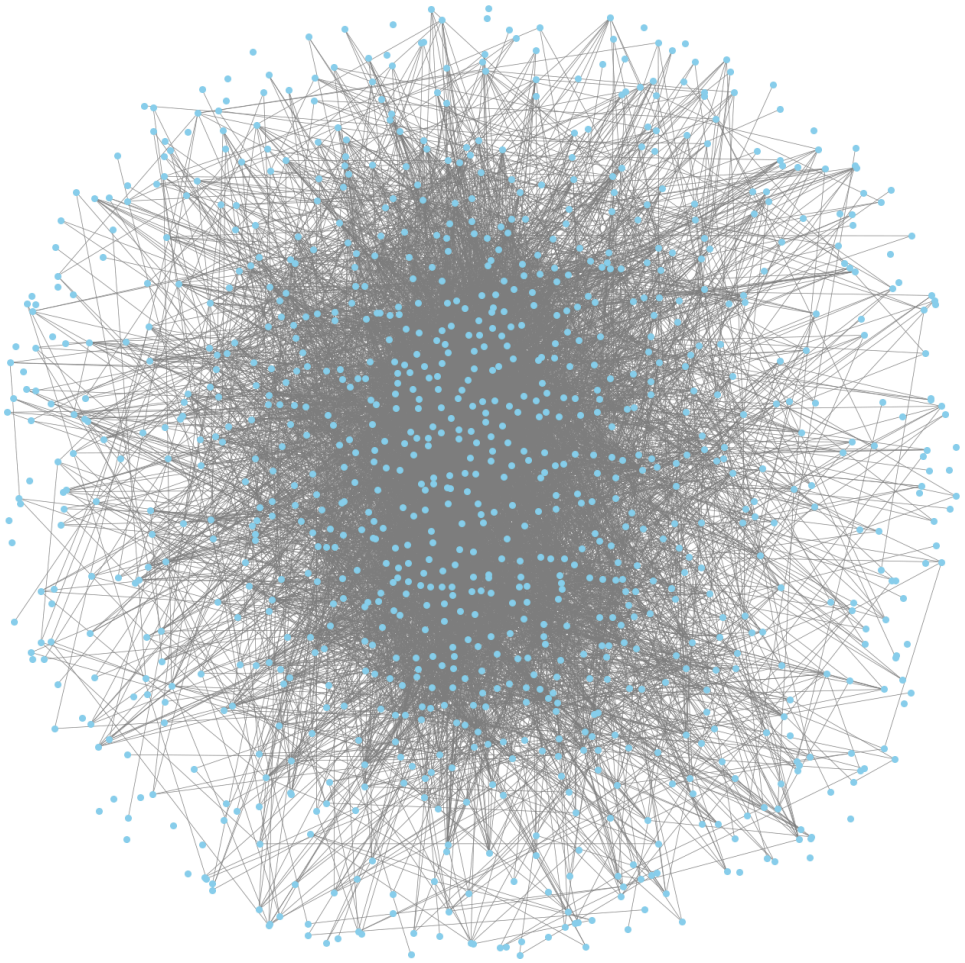

Figure 5. The Collaboration Network of Data Scientists

3.2 Network's Basic Properties

The constructed network has 1 052 nodes and 9 356 edges. As seen in Figure 5, the network is dense with many nodes with several connections and some nodes (the ones located closer to the center) with many connections. There are also nodes with 0 links. The average degree of the network is $< k >= 17.79$.

3.3 Network's Degree Distribution

The degree distribution of a network is the distribution of the number of connections (degrees) of the nodes in the network. It is an important measure of the network because it shows the connectivity patterns in it and gives information about the structure type. The degree distribution of the network can be seen in Figure 6 and Figure 7.

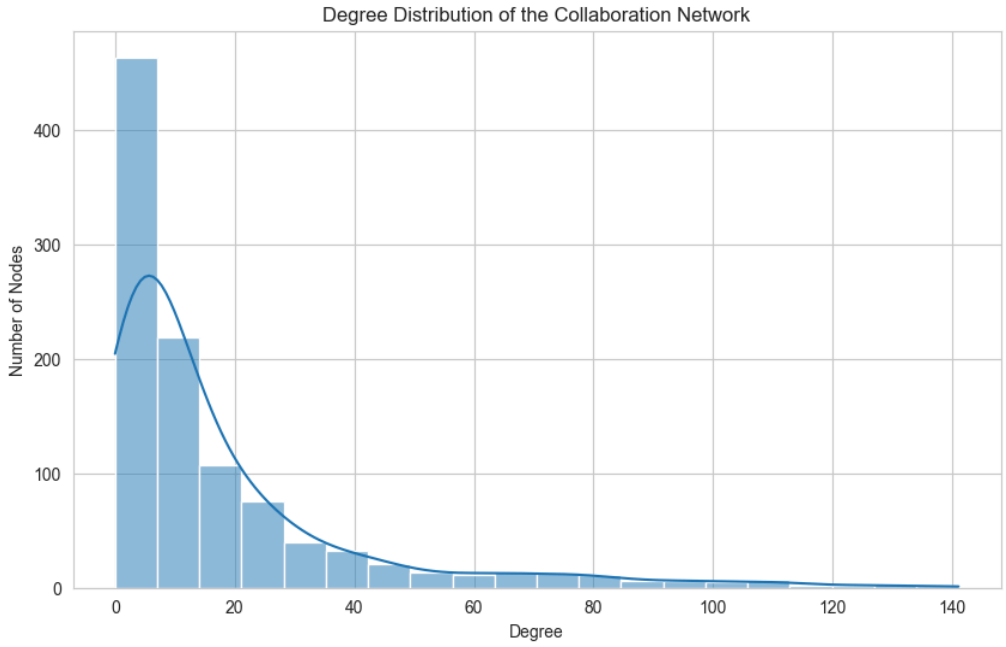

Figure 6. Network's Degree Distribution - Histogram

Figure 6 and Figure 7 show that most of the nodes in the network have degrees less than 50. However, the distribution ends with several nodes with degrees bigger than 100. Having this in mind and considering the shapes of the distributions in the two figures, we can make the conclusion that the degree distribution of the collaboration network of the data scientists follows a power law. This indicates that the network is a scale-free network and the nodes with the highest degrees are the hubs in the scale-free network.

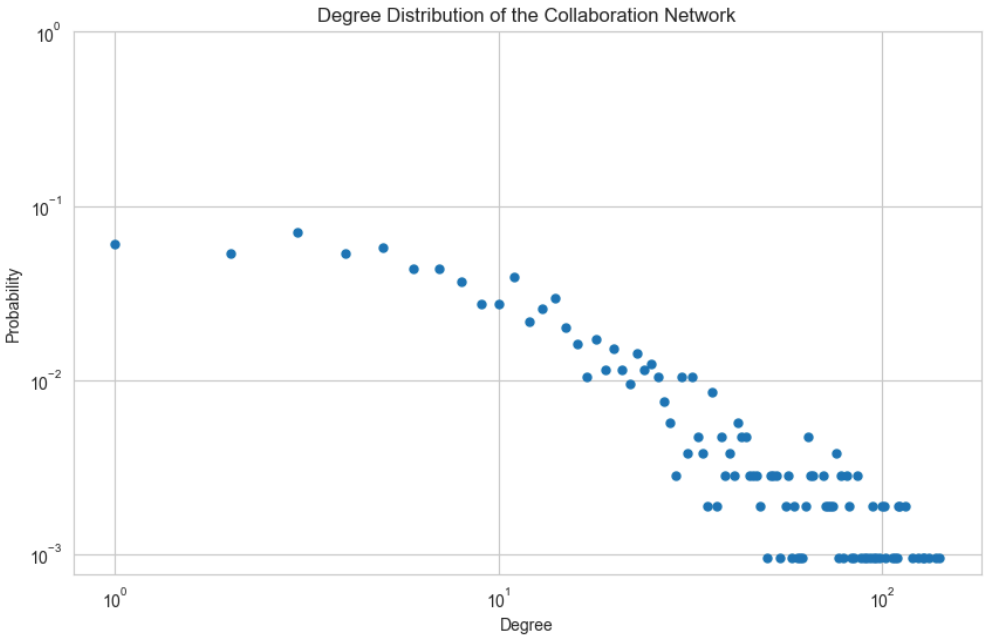

Figure 7. Network's Degree Distribution - Probability Scatter Plot

3.4 Network's Distance and Giant Component

After further analysis of the network, it appears to have one giant component (GC), several smaller ones and many isolated nodes with degree 0. This can be spotted easily if the disconnected nodes (70 in total) from the GC are colored in red like in Figure 8.

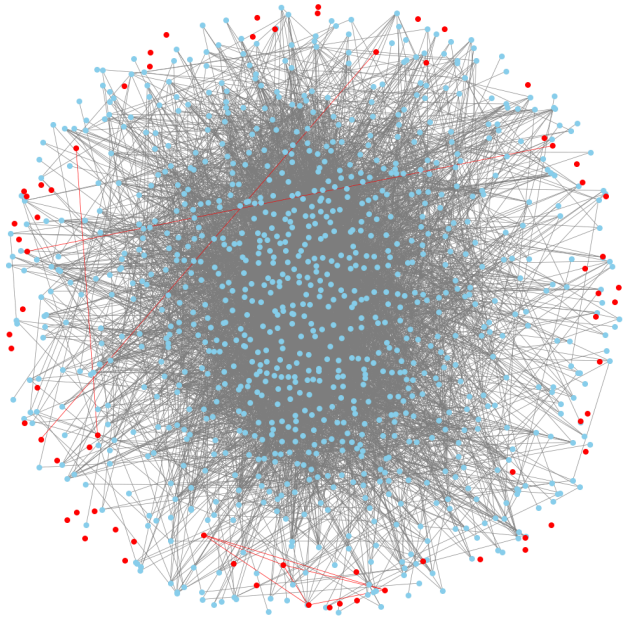

Figure 8. Network's Isolated Nodes and Components

As the network is not fully connected, the computations of the diameter and the average shortest path (distance) are based on the giant component. The giant component has a diameter $d_{max} = 9$ and average distance $< d >= 3.12$. These results show that the small-world property (average distance shorter than 6) is applicable for the network because the average distance is short despite the number of nodes.

3.5 Power-Law Exponent Estimation

Power-law exponent estimation is important as it can show us in which regime the network is located in. The estimation of the power-law exponent for the network was performed using the "powerlaw" Python package for analysis of heavy-tailed distributions. More specifically, the "powerlaw.Fit" functionality was used, it relies on statistical methods developed in [1]. The estimation, seen in Figure 9, gave us an exponent $\gamma = 2.4$. This result means that the network is located in the ultra-small-world regime $2 < \gamma < 3$.

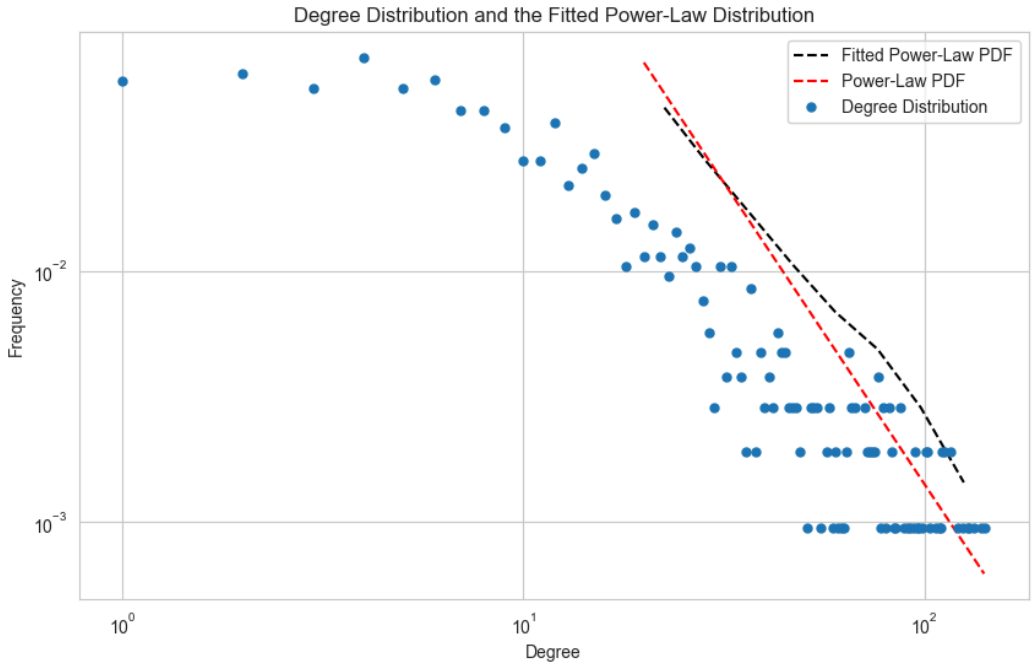

Figure 9. Power-Law Exponent Estimation

3.6 Network's Clustering Coefficient

The clustering coefficient measures the degree to which nodes in a graph tend to cluster together. Analyzing it gives a better understanding of the local connectivity in the network. The collaborative network of data scientists has an average clustering coefficient $C = 0.3178$, meaning that almost one third of all neighbors' links are present.

As for the individual clustering coefficients, their distributions can be seen in Figure 10 and Figure 11. As can be seen in Figure 11, the nodes with higher degrees have lower clustering coefficients. This indicates that the data scientists with more collaborations have lower local connectivity. This is expected as the data scientists with more collaborations are more likely to have collaborations with data scientists from different groups. Also, it is harder for all of their neighbors to be connected to each other.

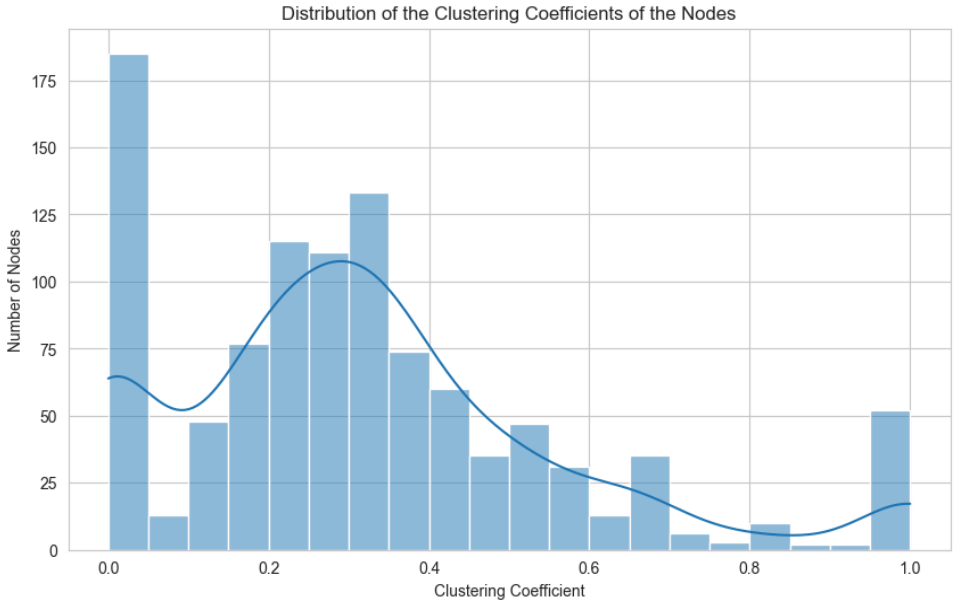

Figure 10. Network's Nodes' Clustering Coefficients Distribution

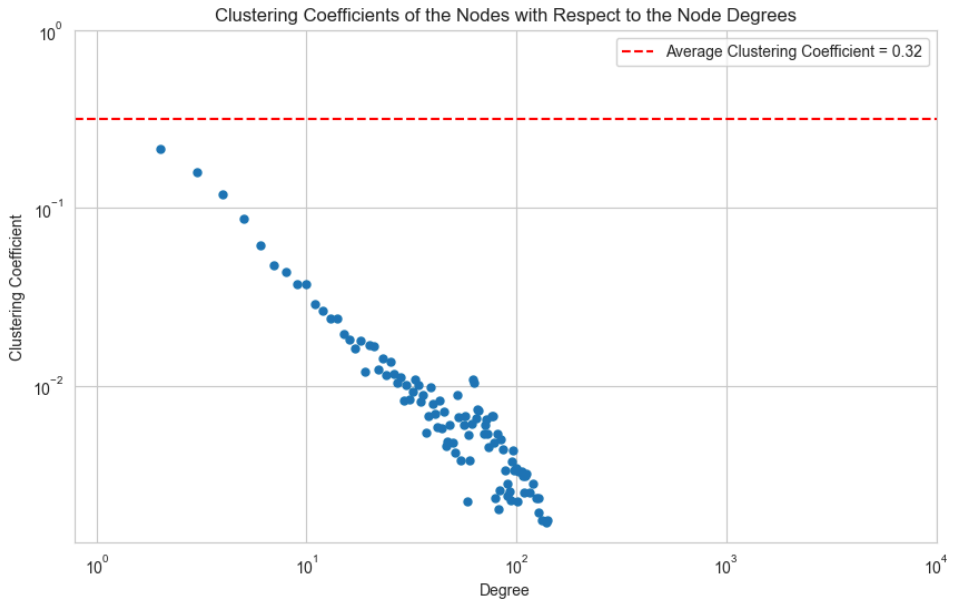

Figure 11. Network's Degrees' Clustering Coefficients Distribution

3.7 Network's Centrality Measures

The centrality of a node measures the centralization of the node's position in the network,thus, it can be used as a measure of node importance. We computed the centrality for every node using 4 different methods:

- Degree centrality
- Closeness centrality
- Betweenness centrality
- Eigenvector centrality

If we plot the top 20 nodes with the highest centrality measures like in Figure 12, we can indeed see that the nodes are located in the center of the network. They are important as they have many connections with other scientists. These nodes are some of the hubs in the network because they have high degree and are connected with other hubs.

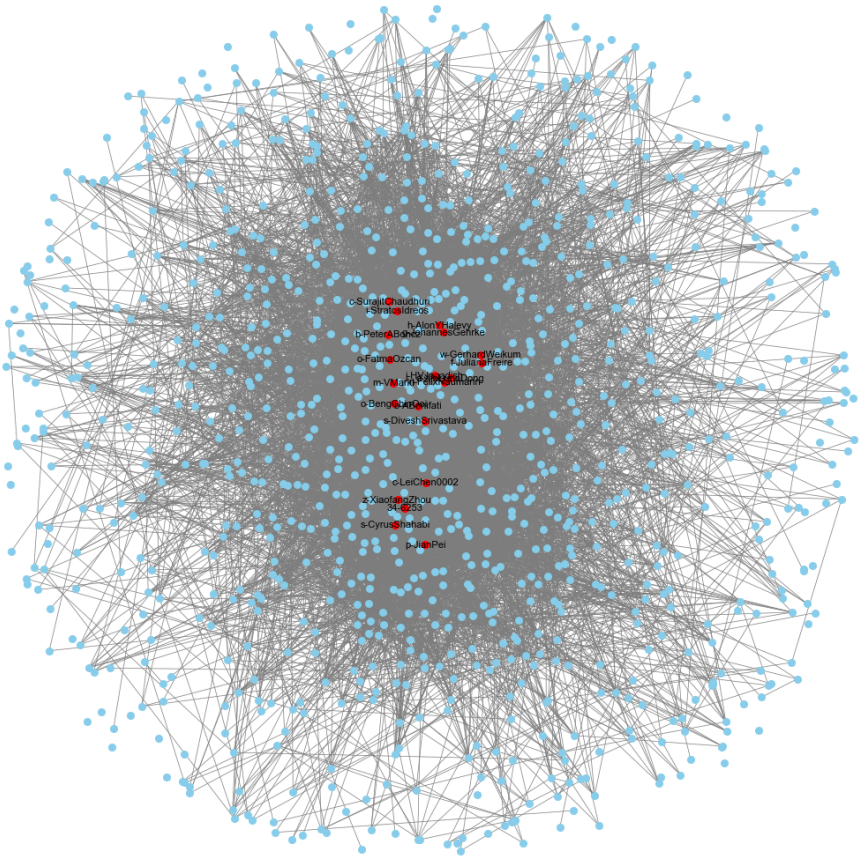

Figure 12. Top 20 Nodes with Highest Closeness Centrality

As we can see in Figure 13, the 3 plots representing the degree, betweenness, and eigenvector centralities are left-skewed. This indicates that the majority of the nodes have low values for these centrality measures. In the closeness plot, we can see that

the distribution is more right-skewed. Also, we can see a peak at the beginning of the distribution. This peak is caused by the isolated notes in the network that are not connected to the main component. Overall, there are not any anomalous nodes with uniquely high centrality, this means that there are no nodes that are extremely important for the collaborations in the network. There are no nodes whose removal will cause the giant component to split.

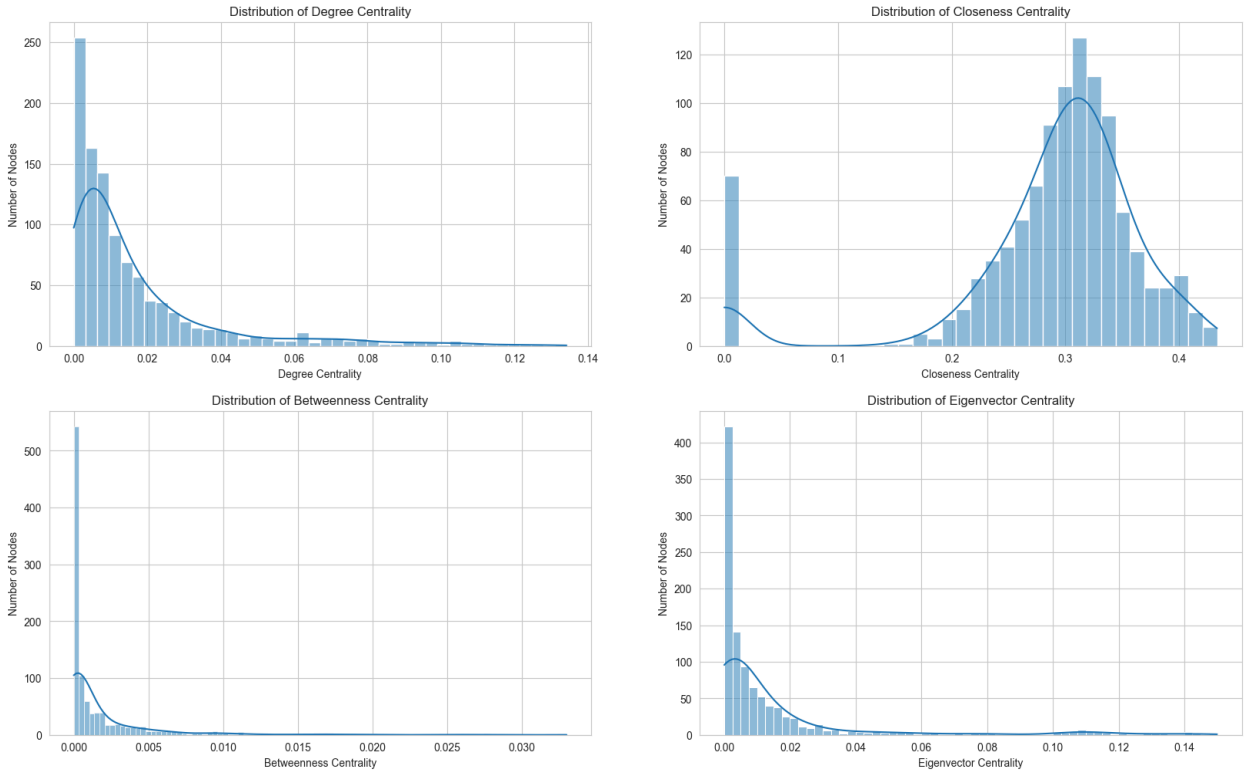

Figure 13. Different Centrality Measures' Distributions

3.8 Network's Cutoffs

The network's cutoffs (natural and structural) are important measures of the network as they help in understanding the limitations and the scalability of the network's structure. The natural cutoff by definition is k_{max} and in the case of the collaborative network of data scientist $k_{max} = 141$. If we refer to the estimated power-law exponent in section 3.6, we can compute the theoretical natural cutoff defined by the following function $k_{max} \sim N^{\frac{1}{\gamma-1}}$, which in our case gives the result $k_{max} \sim 145.08$. We can also compute the structural cutoff, defined by the formula $k_s \sim (< k > N)^{\frac{1}{2}}$, $k_s \sim 136.79$.

The closeness of these cutoff values suggests that the network is quite cohesive and efficient in terms of connectivity. The network's structure is such that it approaches its maximum potential connectivity, given the constraints of network topology and the limits imposed by the network's scale-free nature.

A high natural cutoff indicates the presence of very highly connected hubs, which can be both a strength and a weakness. It implies robustness against random failures (as removing a few non-hub nodes doesn't significantly disrupt the network) but vulnerability to targeted attacks (as removing the hubs could dismantle the network's connectivity).

The proximity of the structural cutoff to the natural cutoff suggests that the network's growth or evolution is closely aligned with its inherent structural constraints.

3.9 Countries Network Analysis

In this section, we will analyze the network formed from the countries of the data scientists. In order to get the data in the wanted adjacency list format, we will just replace each scientist's "pid" with the country they are associated with, then we will combine the duplicated countries. After converting the data in the wanted format, the network is constructed and can be seen in Figure 14.

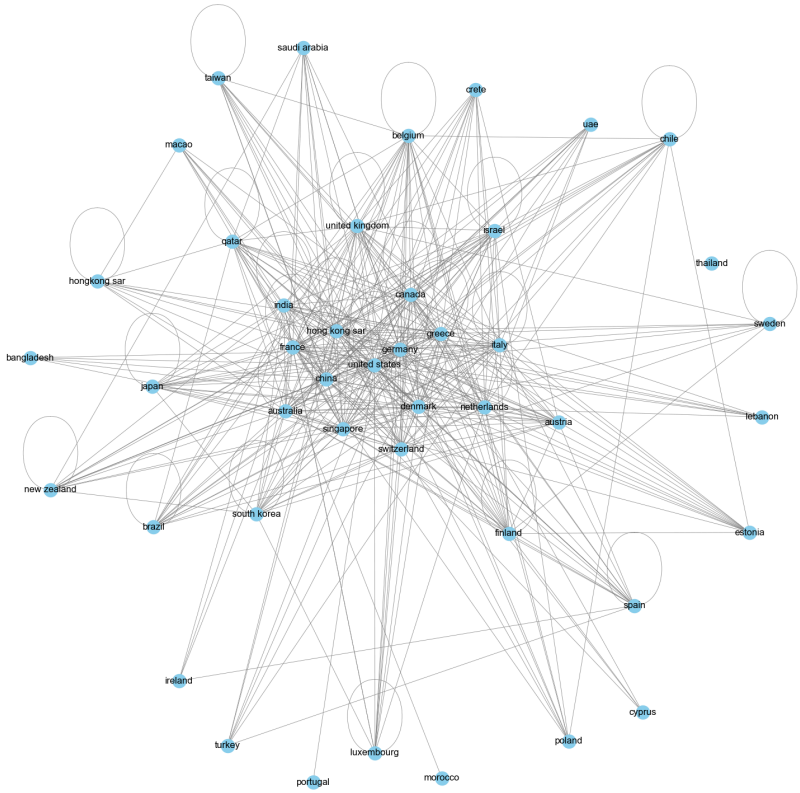

Figure 14. Scientists' Countries Network

As we can see in Figure 14, the collaboration network of the data scientists based on their countries is again quite dense in the center with many connections. This indicates that this network follows the original network's scale-free structure, despite its smaller size. Also, if we focus on the nodes in the center we can see that these are the biggest and most developed (powerful) countries in the world. It makes sense for them to be the hubs in the network, as their financing in the technological and educational systems is higher than the one of the other countries.

We can also use the Plotly Python package to display the network on the actual world map. The result can be seen in Figure 15.

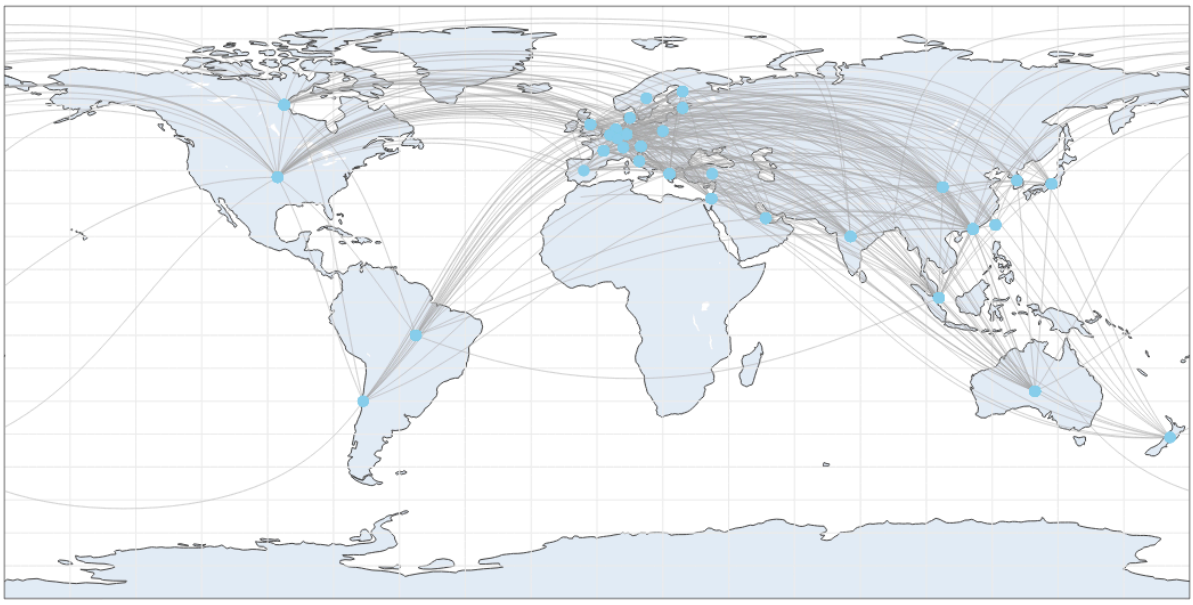

Figure 15. Scientists' Countries Network on the World Map

Finally, the degree distribution seen in Figure 16 does not clearly show that it follows a power law. Although, we can see that it is more skewed to the left than to the right. The initial set of scientists is not of a great size, this limits the diversity of countries in the network. Also, Data Science is a very specific field that is mostly popular in well-developed countries, so this may be the reason we do not have many nodes with low degrees in the countries network.

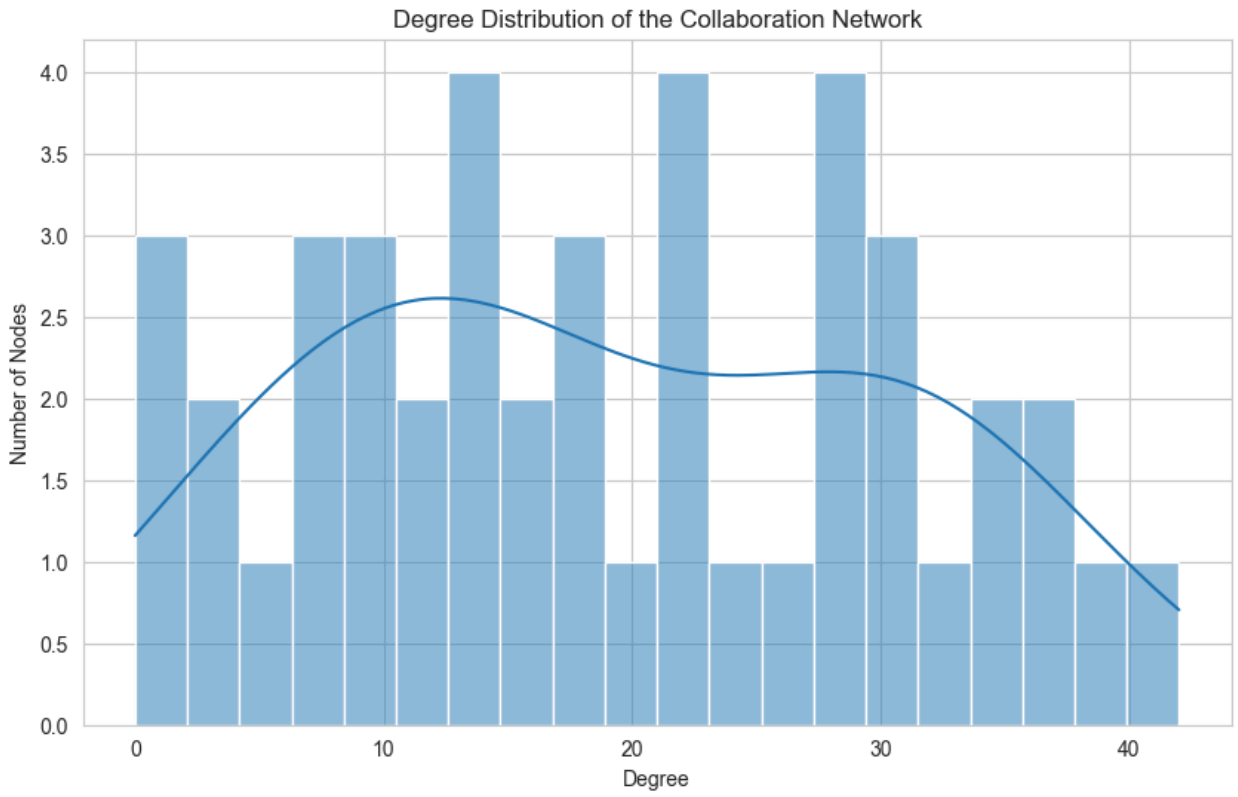

Figure 16. Scientists' Countries Network Degree Distribution

4. Network's Evolution over Time

In this section we will analyze the evolution of the network over time, we will use the year of the publications to monitor the collaboration over the years. Inspecting the evolution of the various network's properties is important as it shows the dynamics and the speed of expanding the network.

4.1 Network Graph Evolution

The evolution of the network over time can be seen in Figure 17 (Figure 17 is a GIF image, if it is not playable, use this link to access it: Figure 17). As seen in Figure 17, in the beginning, the network was constructed of only isolated components and nodes as there were not enough papers in the Data Science field. It was the year 1999 when the giant component first began to emerge in the network. Since then, it has grown in size and started including more hubs within it. This significant evolution (after 1999) of the network can be explained with the technological advancements in the early 2000s. Furthermore, the early 2000s were the time when the term "big data" emerged as industries and academia recognized the value of data-driven approaches.

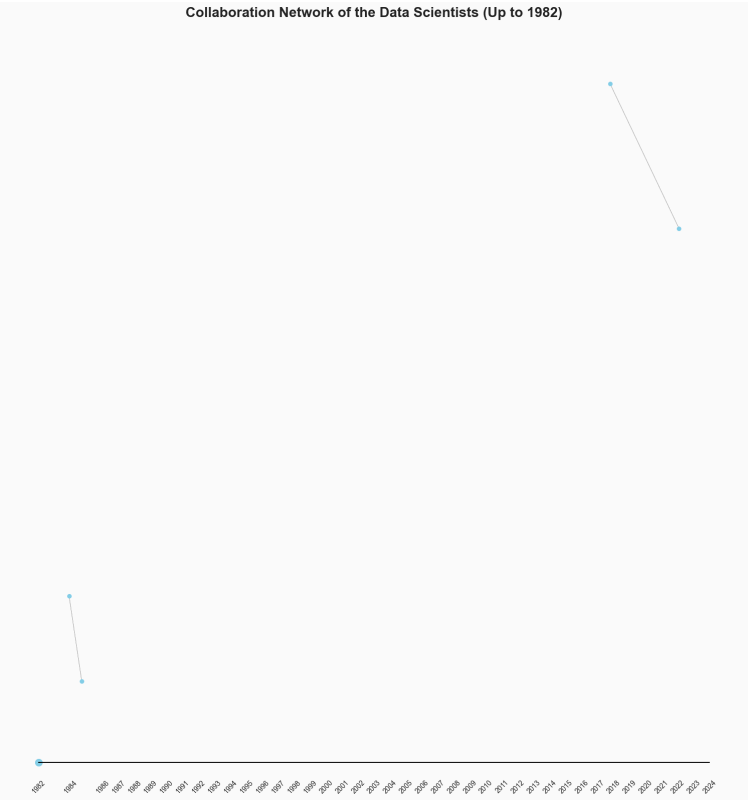

Figure 17. Network's Evolution over Time

The evolution can also be monitored in the evolution in the number of edges (links) and nodes presented in Figure 18.

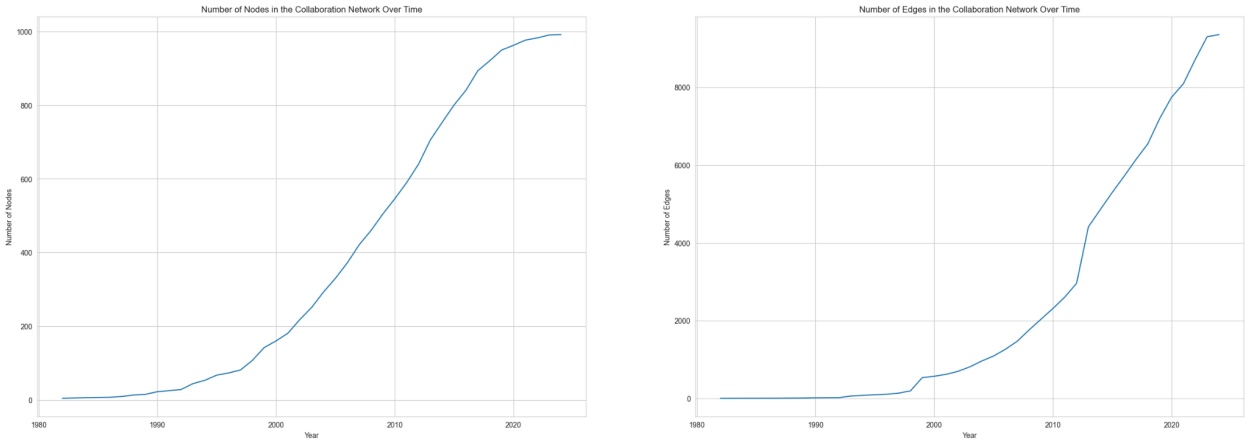

Figure 18. Number of Nodes and Links Evolution

4.2 Average Degree and Maximum Degree Evolution

Analyzing the average degree and maximum degree evolutions in comparison will help understand the importance of the nodes and the formation of the hubs in the network. As seen in Figure 19, the maximum degree k_{max} starts increasing significantly around the year 2000 and continues with similar pace until nowadays. In comparison, the average degree of the network remains almost the same over the years. This shows that hubs started to form around the early 2000s, this proves our observation from the previous section. The k_{max} deviation from $< k >$ represents the formation of hubs because the newly added nodes tend to connect to already well-recognized scientists with many collaborations. As a result, the degree of several nodes (the hubs) has significantly increased, while the degree of the others remains the same.

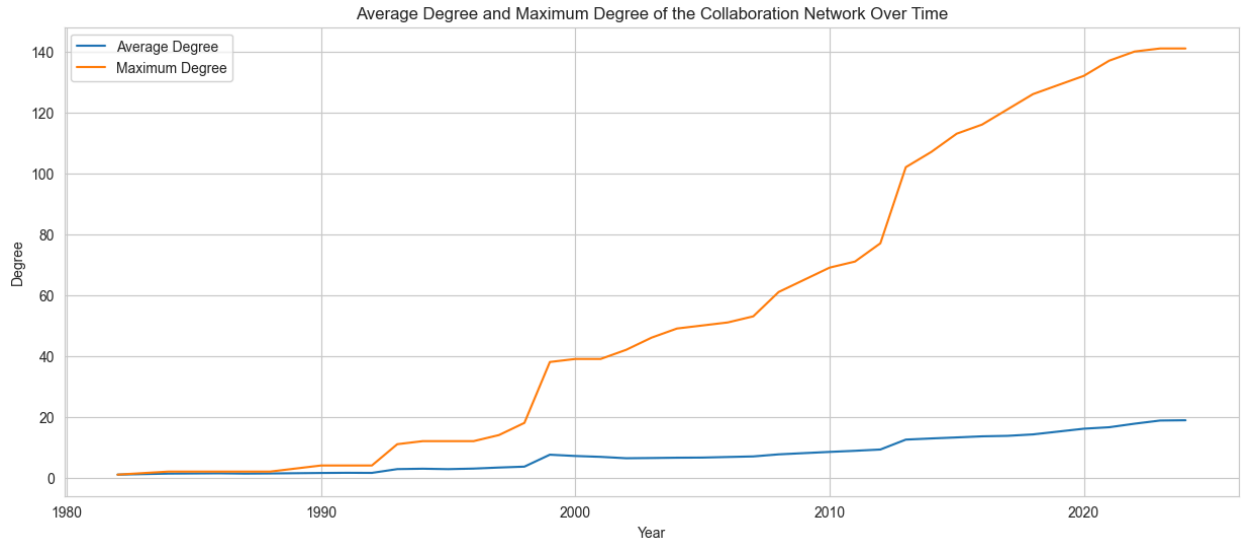

Figure 19. Average Degree and Maximum Degree Evolution

4.3 Diameter and Average Distance Evolution

Figure 20 shows the evolution of the diameter and the average distance of the network. There is a significant period where the distance and the diameter experienced a noticeable drop. Again, this period (early 2000s) matches with the one from the previous sections. Drops in the distance and the diameter are often caused by the emergence of hubs in the network. However, the diameter's plot experiences a significant increase around the year 2010, this can be explained with the introduction of many new edges (publications), also seen in Figure 18.

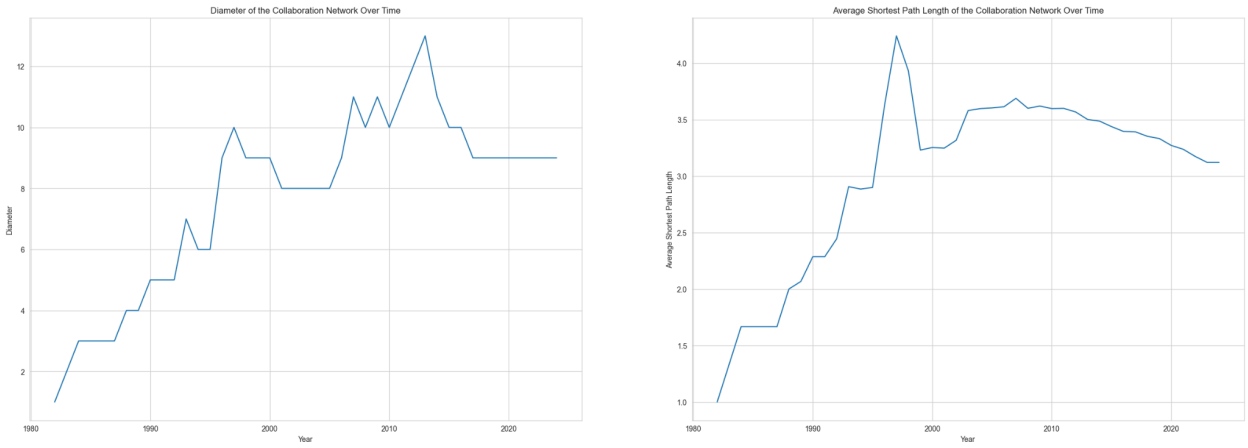

Figure 20. Diameter and Average Distance Evolution

4.4 Clustering Coefficient Evolution

Inspecting Figure 21, we can observe that during the initial years, the clustering coefficient was low and was increasing over time up to the early 2000s when it started to stabilize. This shows that the data scientists were not well-connected in local groups in the beginning, but as time passed, they started to form groups and collaborate more. Then, the clustering coefficient stabilized, indicating that the data scientist started to connect to the most important scientists in the network (hubs) rather than to their local groups.

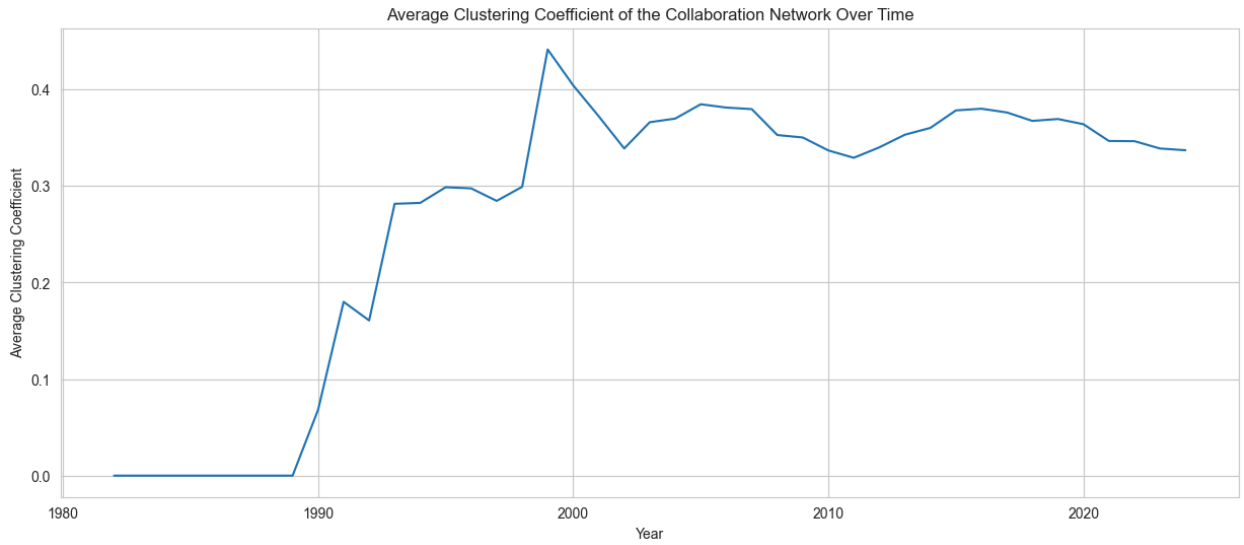

Figure 21. Average Clustering Coefficient Evolution

4.5 Connected Components Evolution

Analyzing the number of connected components of the network will help us understand how the smaller groups of data scientists have merged over time. As presented in Figure 22, the number of connected components in the network increased at first because many new nodes were entering the network, but they were not connecting to each other. Then the number of connected components started to decrease as the data scientists started to collaborate more and more. The little components started to merge into the main component of the network and only some isolated very small components remained.

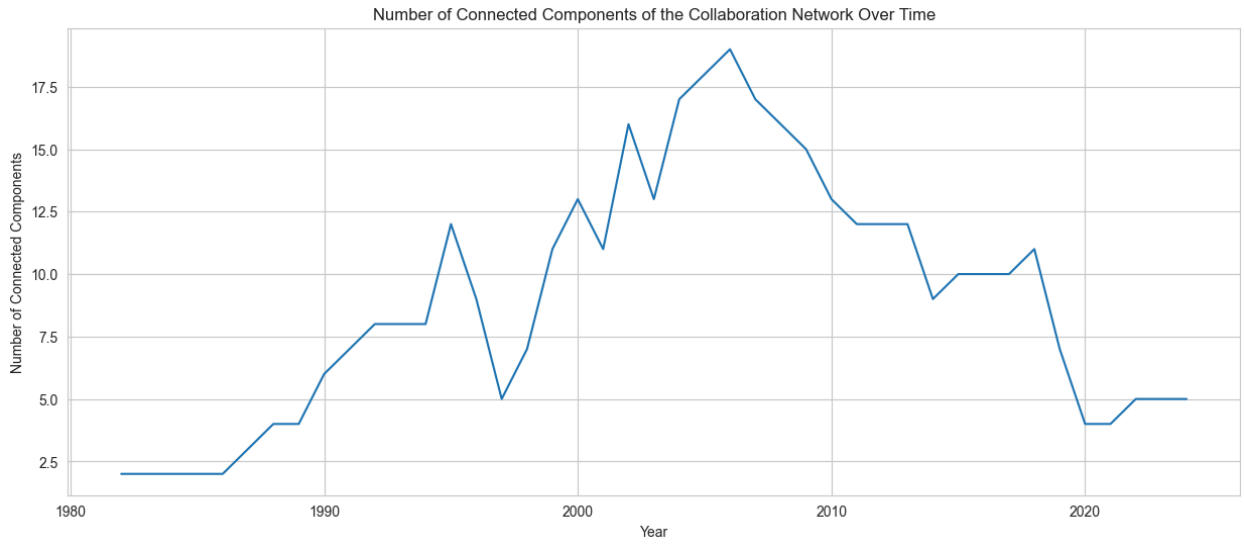

Figure 22. Number of Connected Components Evolution

4.5 Connected Components Evolution

Looking at the plot at Figure 23, we can observe that both the average and the maximum closeness centrality of the network follow a similar trend. At first, the average and maximum closeness centrality of the network were high and dropped drastically. Then after reaching the minimum, they started to increase again and continued to increase over time. This indicates that the data scientists were not well-connected in the beginning (because there weren't many scientists), but as time passed (and new nodes were added to the network and hubs emerged) the closeness centrality measure started to increase, the scientist needed some time to start collaborating and forming hubs which connect to each other.

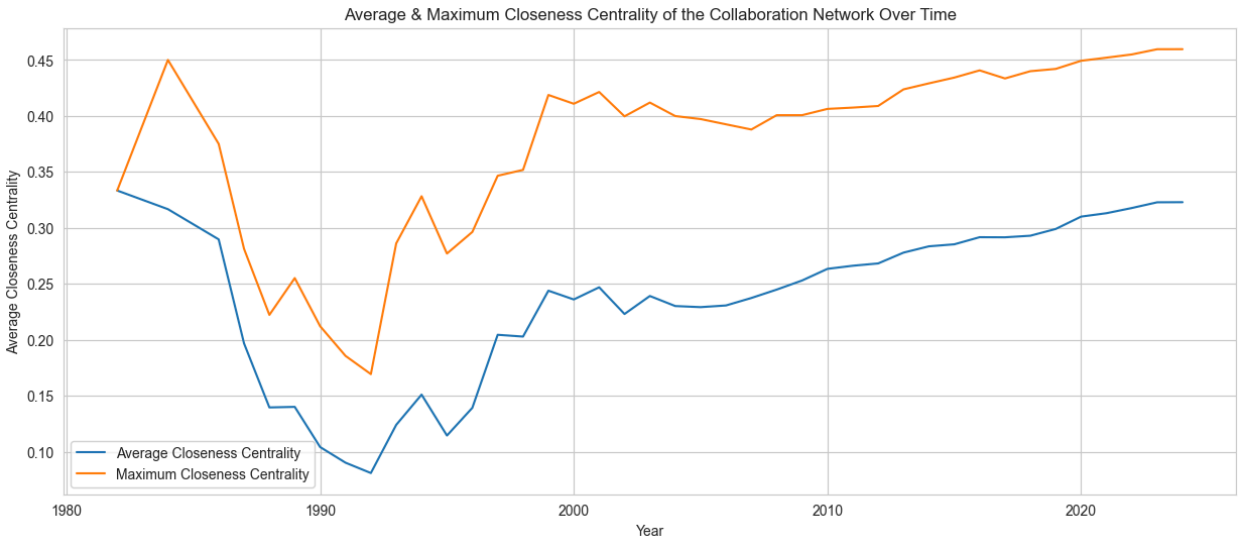

Figure 23. Average & Maximum Closeness Centrality Evolution

5. Comparison with Random Network

5.1 Calculating Edge Probability

When generating a random network using the Erdős-Rényi model the edge probability is needed, edge probability p is calculated based on the following formula: $p = \frac{2L}{N(N-1)}$, where L is the number of edges and N is the number of nodes in the network. The formula is derived from the fact that the total number of possible edges in an undirected graph with N nodes is $\frac{N(N-1)}{2}$, and the edge probability p is the ratio of the number of edges L to the total number of possible edges.

5.2 Erdős-Rényi model

Random network is generated with the NetworkX random graph generator. The number of nodes N and the number of edges L being used is the same as the collaboration network in section 3.1, in order to have a fair comparison. The argument N and p is being passed through as arguments to generate the model.

The difference can be seen in Figure 24, when the Erdős-Rényi random networks are distributed more uniformly compared to the collaboration network of the data scientists due to the independent creation of edges with the probability p. While the collaboration network is a scale free network in which the degrees follow a power law where the central degrees are higher. Hence, the main difference would be a random network

lacking a well defined center but spreads evenly across all nodes while collaboration networks have a well defined center and a scale free structure.

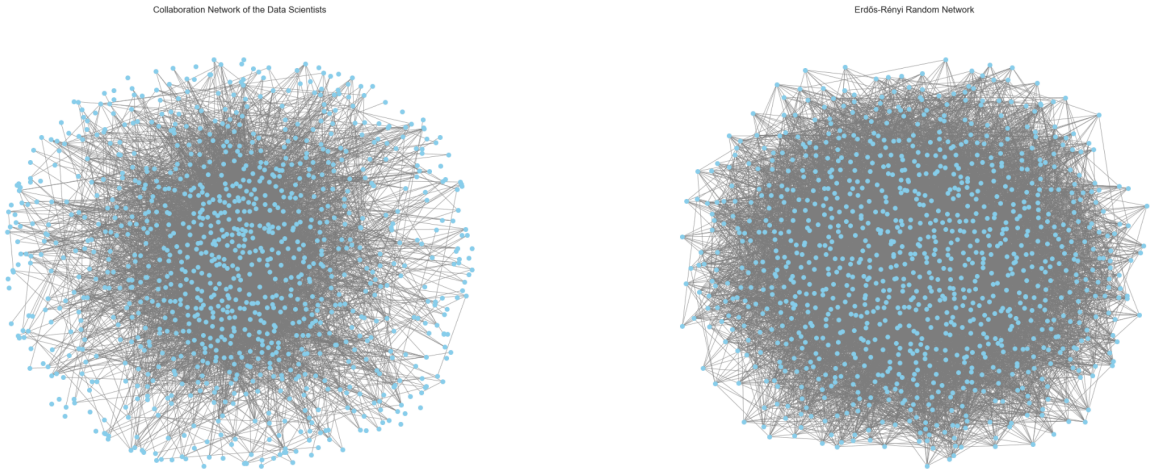

Figure 24. Original vs Random Network Comparison

5.3 Degree distribution comparison

The difference can be seen in Figure 25 and Figure 26 that the degree distribution of random networks follows a poisson distribution, while the one of scale free network follows a power law distribution.

When the degree of nodes follow a power law distribution in a typical real world network there are fewer nodes with a high degree and more nodes with low degree. While in a random network the degree of nodes are uniformly distributed due to each edge being created independently with the same probability.

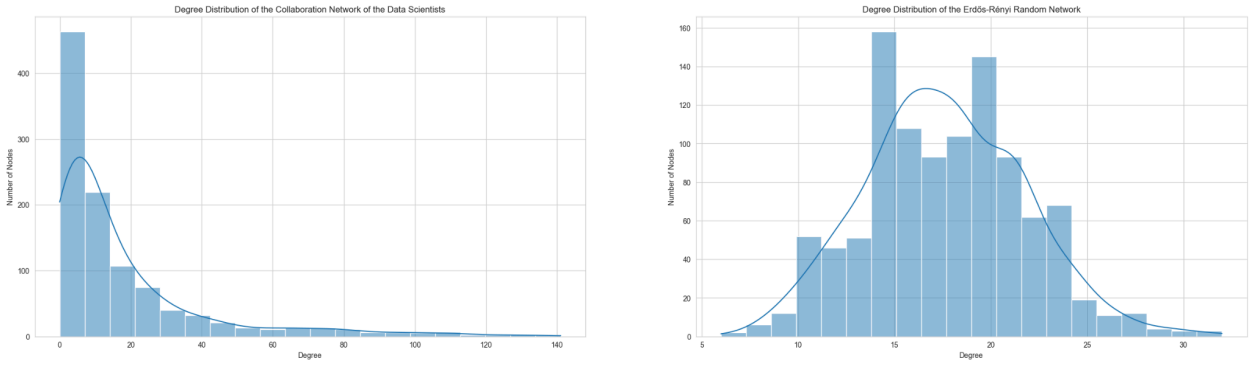

Figure 25. Original vs Random Degree Distribution Comparison - Histogram

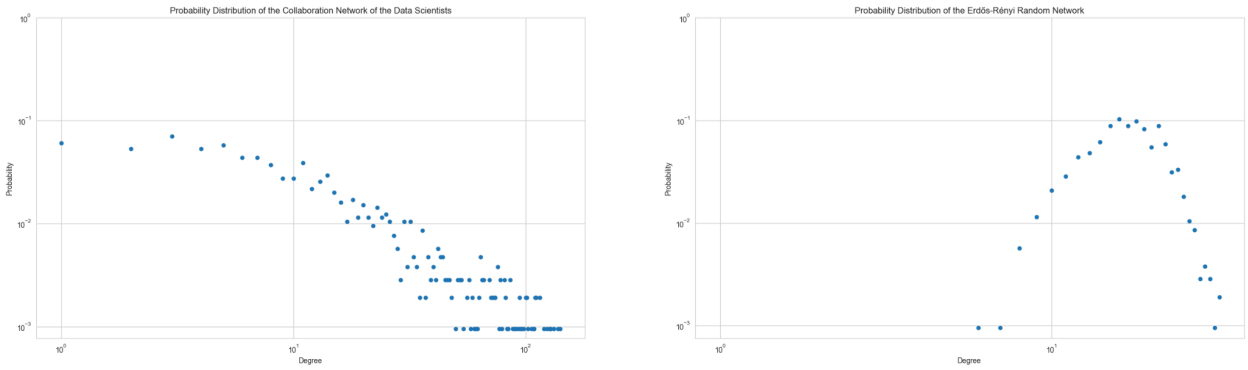

Figure 26. Original vs Random Degree Distribution Comparison - Probability

5.4 Distance & Diameter Comparison

As it can be seen in Figure 27, both the average shortest path length and the diameter of the collaboration network of the Erdős-Rényi random network are lower than the collaboration network of the data scientists. This is because in the random network, the nodes are connected more uniformly, while in the real collaboration network we have hubs but also many nodes with low degrees. Because of the fair connection in the random network, we can see the difference in the number of isolated nodes from the giant component. The Erdős-Rényi random network has 0 isolated nodes, meaning that the whole network is fully connected, while in the real scientists collaborative network there are 70 nodes, not connected to the giant component.

```
The average shortest path length of the collaboration network of the data scientists is <d>=3.1201
The diameter of the collaboration network of the data scientists is d_max=9
The number of nodes isolated from the main component of the collaboration network of the data scientists is 70

The average shortest path length of the Erdős-Rényi random network is <d>=2.7285
The diameter of the Erdős-Rényi random network is d_max=4
The number of nodes isolated from the main component of the Erdős-Rényi random network is 0
```

Figure 27. Original vs Random Distance & Diameter Comparison

5.5 Clustering Coefficient Comparison

The clustering coefficient will help us understand the local connectivity of the networks, the average clustering coefficient for the Erdős-Rényi random network is lower than the scale-free network. As seen in Figure 28 the scale-free network has many anomalies and is not uniformly distributed due to the hubs in the networks and isolated nodes. When compared to the random network it shows a difference as there is a more uniform distribution of clustering coefficient.

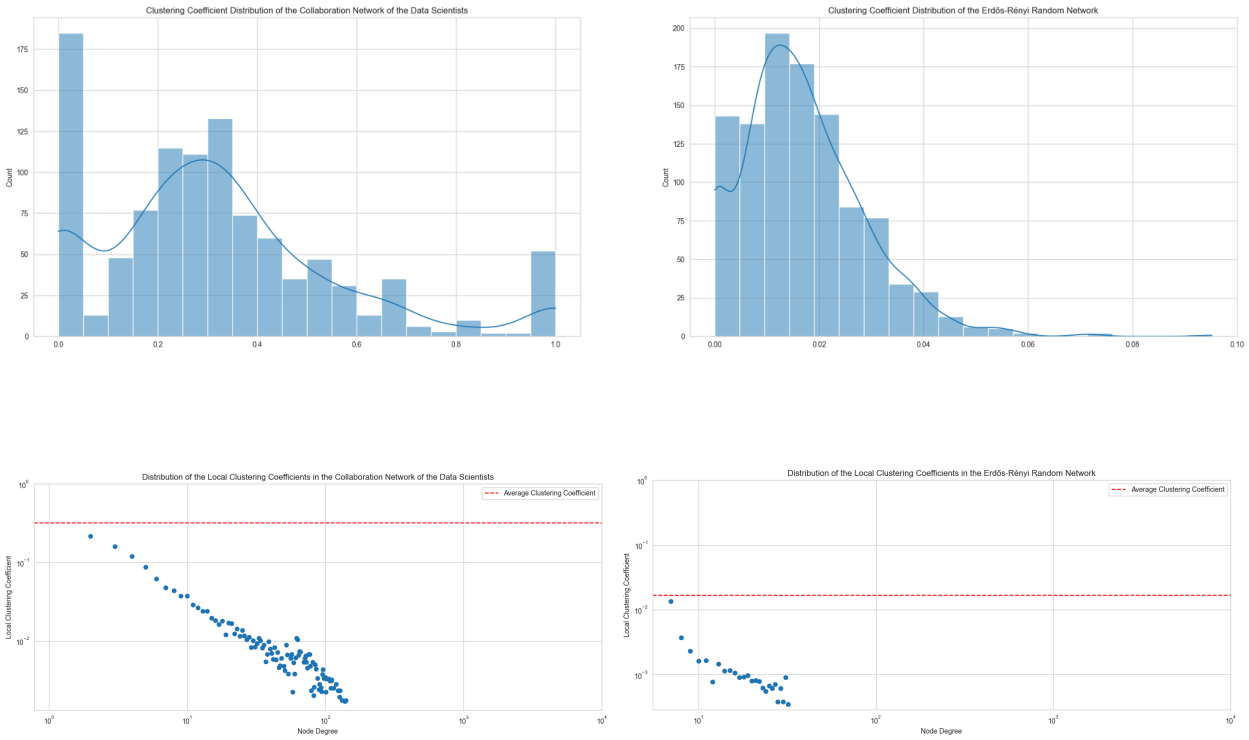

Figure 28. Original vs Random Clustering Coefficient Comparison

5.6 Closeness Centrality Comparison

Figure 29 shows the distribution of closeness centrality for the real collaboration network and the random network. As expected from the average distance comparison $< d >$, the mean closeness centrality of the random network is higher than that of the real collaboration network due to its homogeneous distribution. The real collaboration network has a larger range of values in closeness centrality with some being very high (probably hubs) and some being zero (probably isolated nodes), in the context of data scientist the hubs are scientist who are influential in the field or have played significant roles while the isolated nodes are probably scientist that are new to the field or have limited interactions with others. In Figure 30, all 4 centrality measures for the random network are displayed.

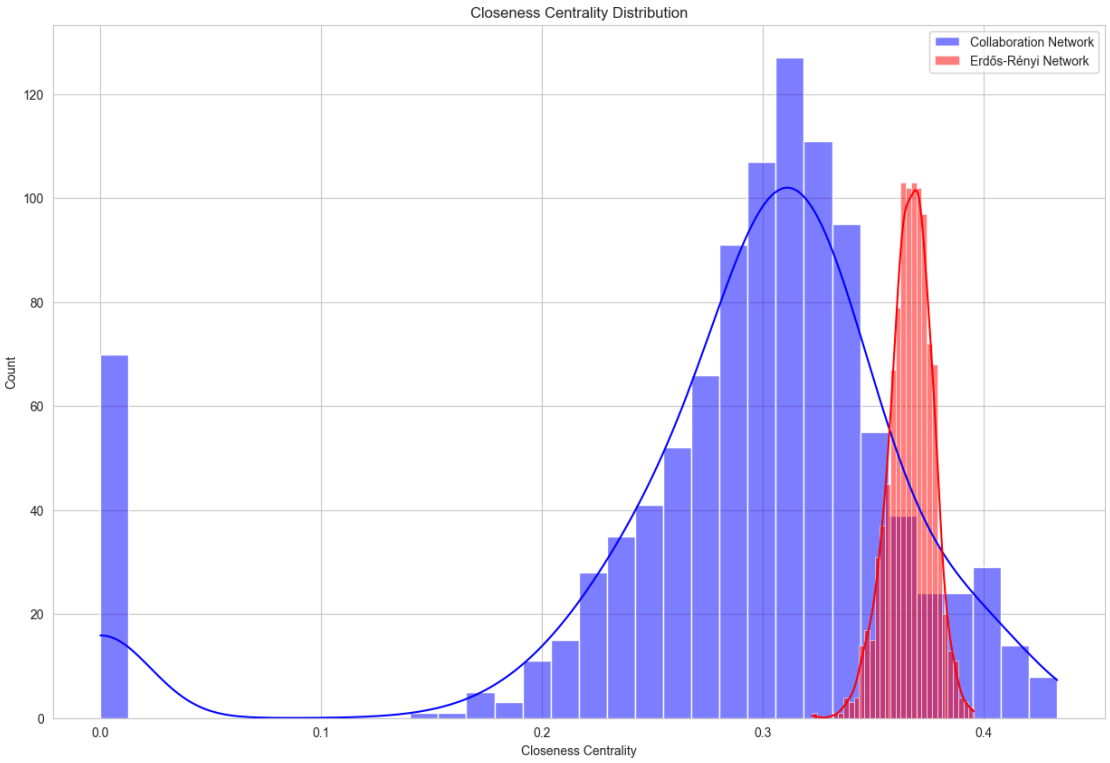

Figure 29. Original vs Random Closeness Centrality Comparison

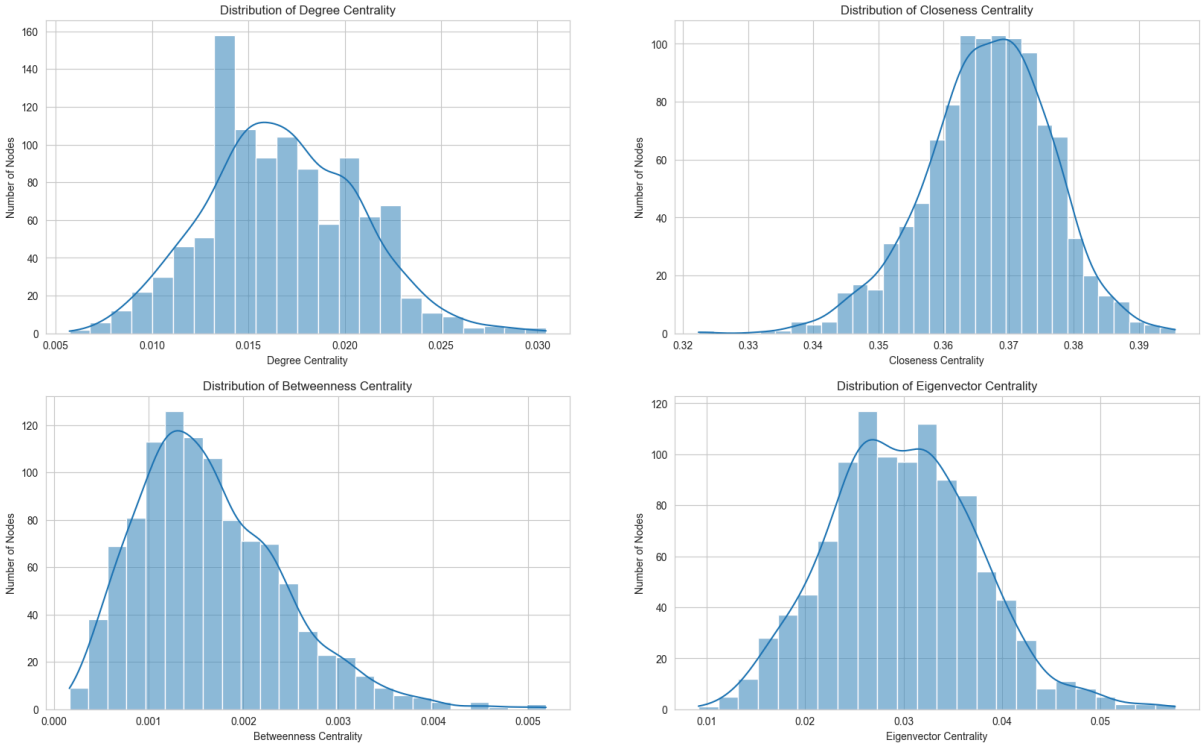

Figure 30. Original vs Random Centrality Measures Comparison

5.7 Country Comparison

The country network in Figure 31 displays the same differences between the real network and random network model, despite having fewer nodes. The real network has a central core of nodes which are highly connected, while there are fringe nodes which are only connected to a few of their neighbors and to the central core. It resembles a hub and spoke topology. The random network, on the other hand, has all nodes connecting to many other nodes as it has no systematic pattern of assortativity on the nodes attribute.

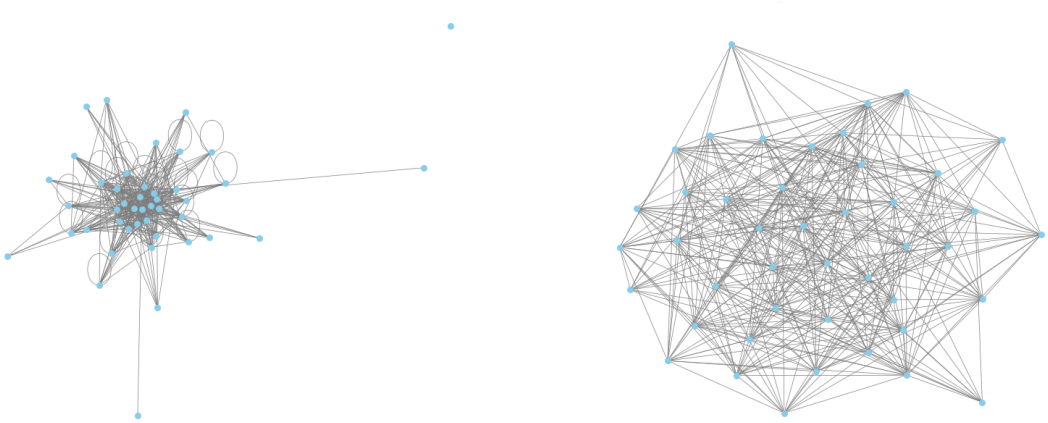

Figure 31. Original vs Random Country Network Comparison

Figure 32 shows the degree distribution of the real country network and the random network. The degree distribution of the real country network is wider and more uniform than that of the random network, indicating more variation in the degree of nodes which is similar to that seen in a hub and spoke topology. While the random network is narrower and not uniform due to the heterogeneous distribution of connection as expected from the characteristics of a random network having a less structured and more evenly distributed network by having most nodes having a similar number of connections.

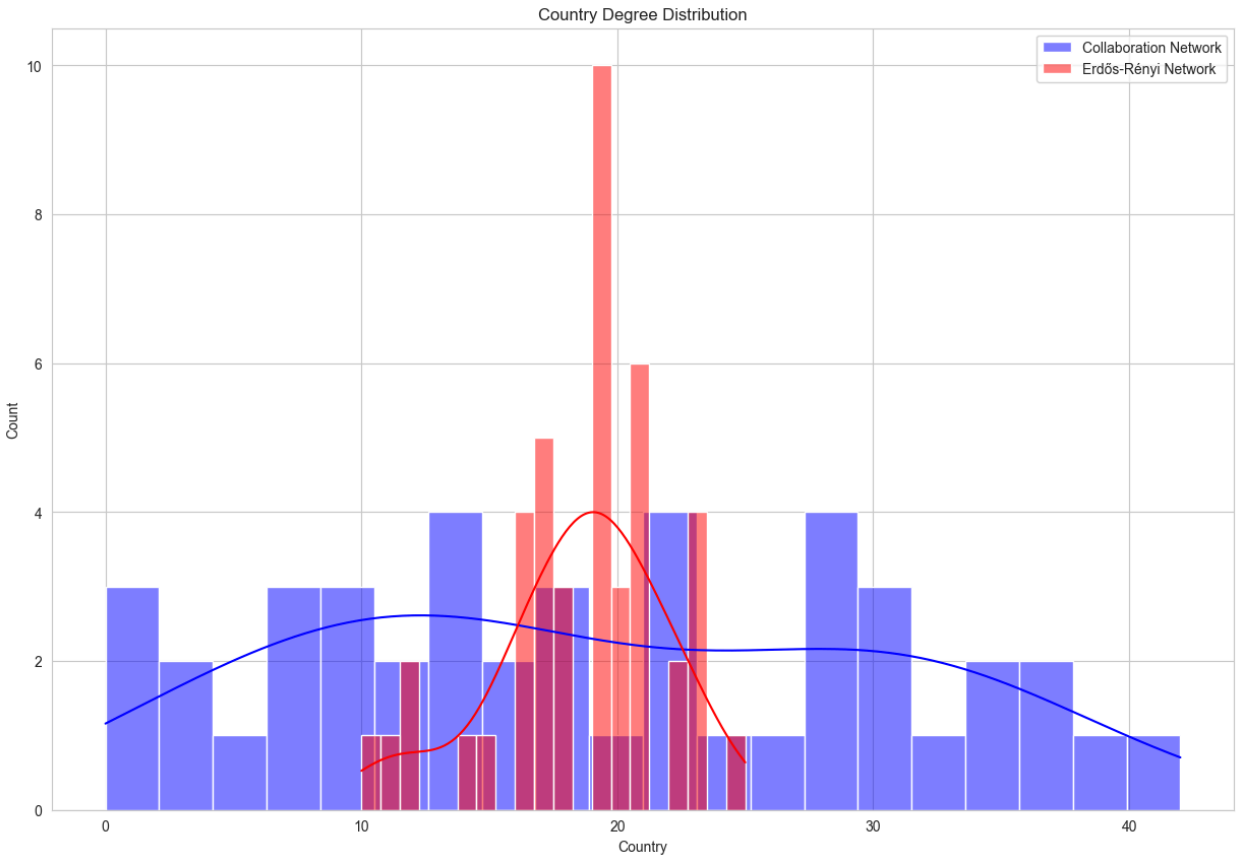

Figure 32. Original vs Random Country Degree Distribution Comparison

6. Reducing Network Size

In this section, we will create a function that will transform the existing collaboration network of the data scientists into a new network. The new network will be transformed based on the following goals:

- The maximum degree of any node will not exceed beyond a user-specified k_max, referred to as collaboration cutoff, which will be smaller than the degrees of hubs.
- The transformed network will have a smaller giant component and larger number of isolates than the original collaboration network.
- The transformed network will be built to identify links between Famous Data Scientists (High Degree) & their less famous co-authors (Low Degree nodes).

To maintain the diversity of nodes in the network, we decided against removing any nodes from the network. Therefore, the transformed network will have the same number

of nodes as the untransformed network. We instead focus on removing unwanted edges from the network. In general every network transformation technique had 2 parts:

1. Define a metric of edge importance
2. Define a policy to prune edges of less importance

6.1. Best Algorithm - Hard Cutoff for Normalised Degree Difference

In this section we discuss the algorithm which was the most successful at reducing the size of the giant component - Hard Cutoff for Normalised Degree Difference. This algorithm has two parts:

1. **Edge Importance Metric** - To measure the importance of each edge we defined a metric Normalised Degree Difference as:

$$Normalised\ Degree\ Difference\ \left(E_{n1,n2}\right) = \frac{|C_D(n1) - C_D(n2)|}{\max\left(C_D(n1), C_D(n2)\right)}$$

   where n1 and n2 are nodes in the graph, $E_{n1,\ n2}$ is the edge connecting n1 & n2, and $C_D(x)$ is the degree of node x. The expected values of Normalised Degree Difference for different edges are:

   a. Hub - Hub Edges => Small Numerator and Large Denominator => Very Small - Small Value
   b. Hub - Normal Edges => Large Numerator and Large Denominator => Medium - Large Value
   c. Normal - Normal Edges => Small Numerator and Small Denominator => Small - Medium Value

   Hence, edges with a smaller Normalised Degree Difference are more likely to be edges between hubs or between normal nodes, rather than those between hubs and normal nodes.

2. **Edge Pruning Policy** - To prune the edges in the network, we define a hard cutoff, calculated based on the user specified k_{max}. All edges with an edge importance below this threshold will be removed from the network.

Figure 33 shows the distribution of Normalised Degree Difference in the network. The uniform nature of the distribution makes this metric ideal for being used with a hard threshold and multiple values of k_{max}. Metrics with exponential distributions do not work

as well with a hard threshold as they either remove too many edges or too few depending on the value of k_{max} specified by the user.

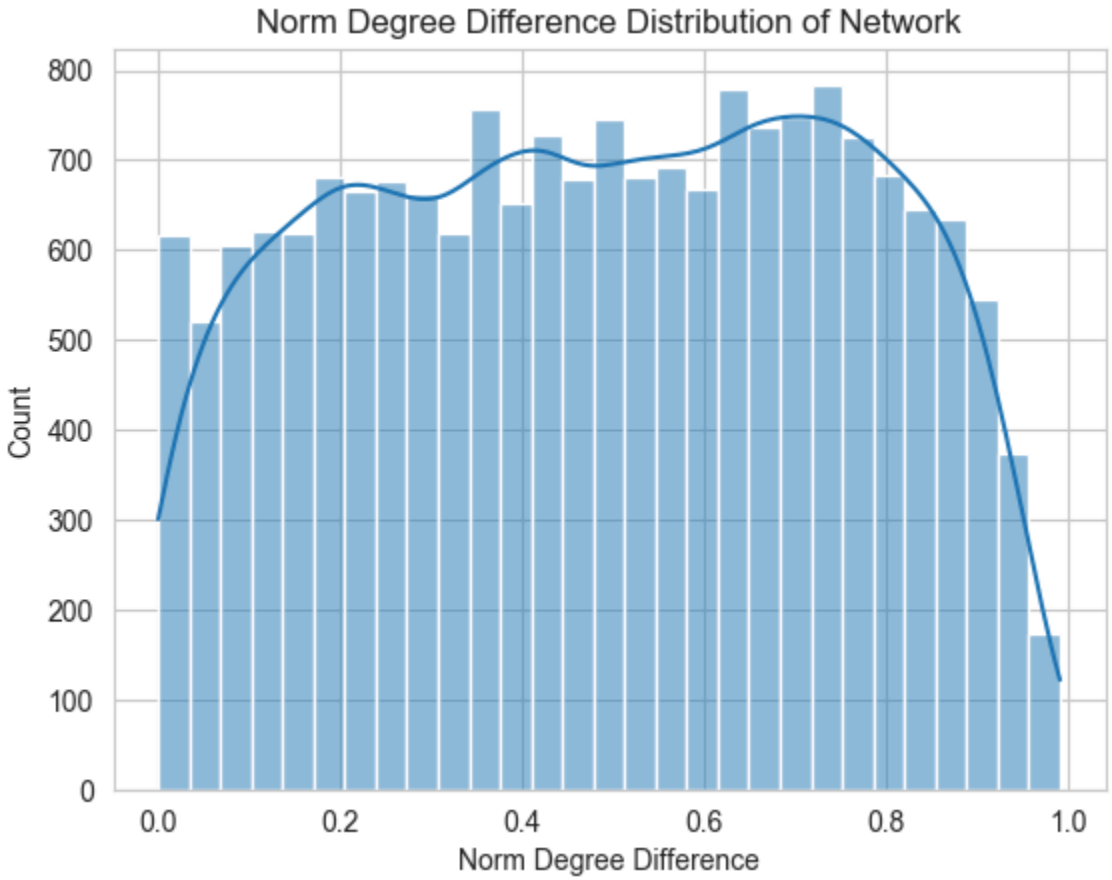

Figure 33. Normalised Degree Difference Distribution

Figure 34 depicts the relation between the size of the giant component of the transformed network and the user-specified value of k_{max}, while using the Hard Cutoff for Normalised Degree Difference algorithm. We can observe that the algorithm provides a range of values for the size of the giant component depending on the input k_{max}.

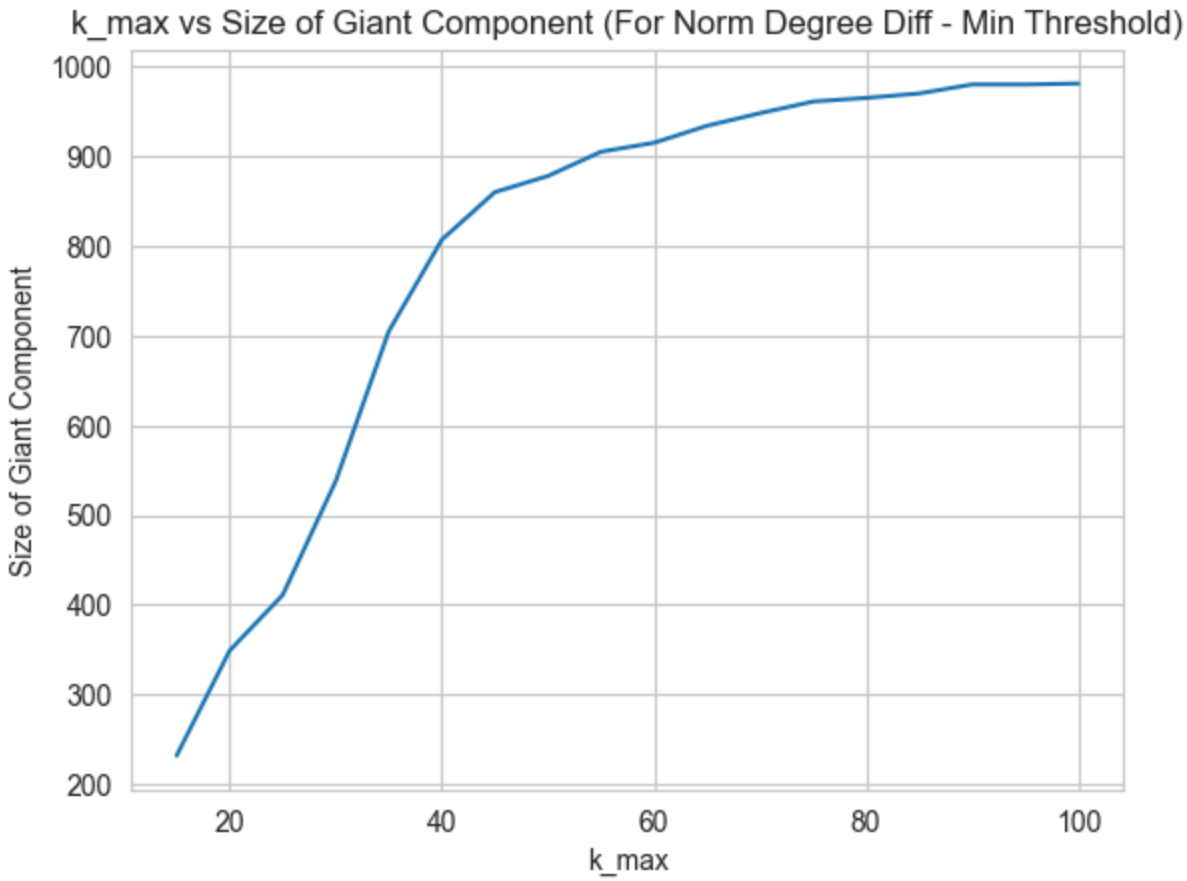

Figure 34. k_{max} vs Size of Giant Component

In Figure 35 we can see a transformed network with k_max = 35. It has a giant component with 705 nodes, compared to the 993 in the original network.

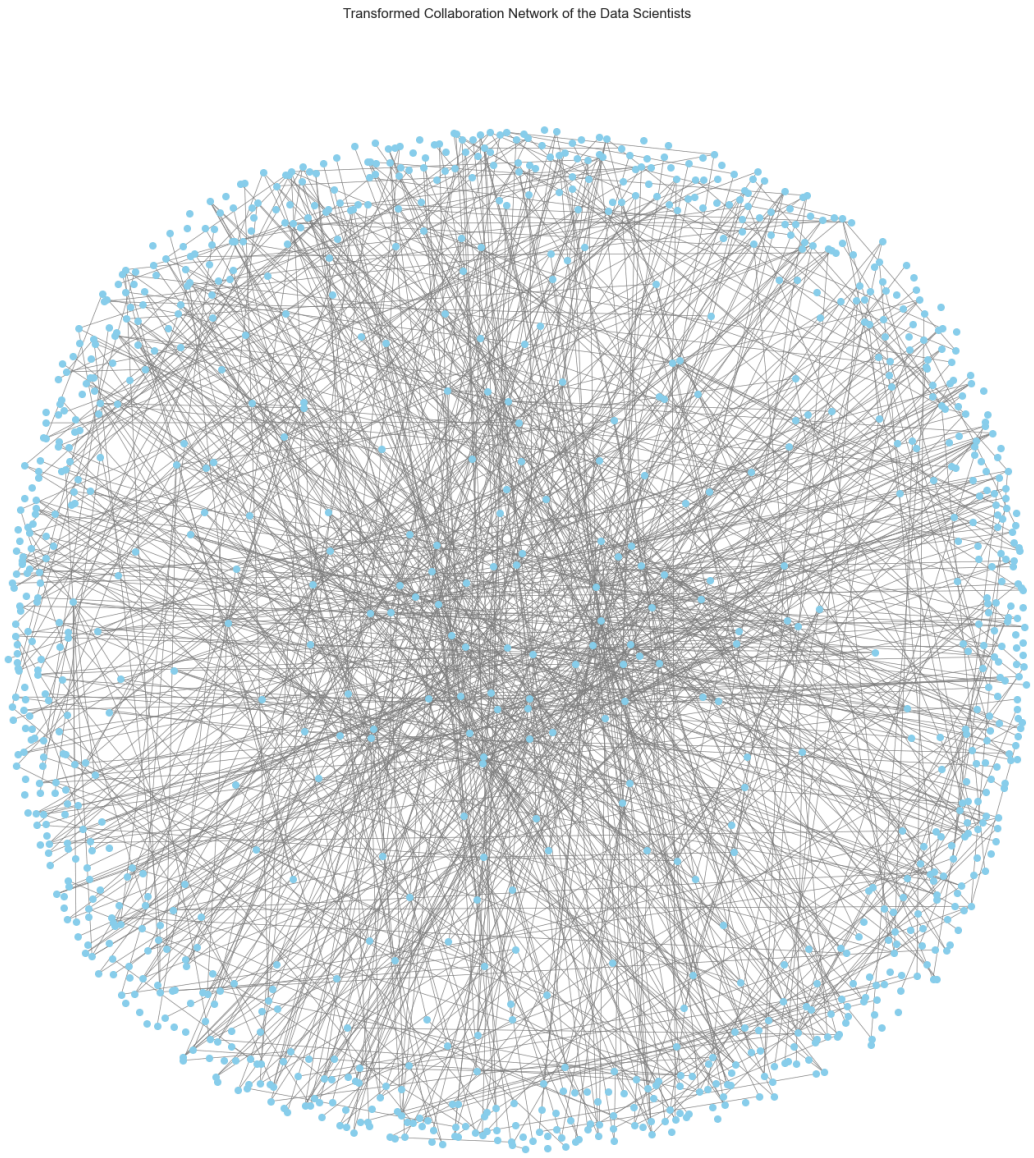

Figure 35. Transformed Network

Lastly, we will compare the node degree sum for the country, expertise and institution of the scientists to ensure that the newly transformed network follows the same diversity as the original one.

In Figure 36, we can see that the country's node degree sum distribution remained almost the same with the change of reducing the sum 7 times. This shows that we successfully reduced the node degrees in the network but maintained the original distribution in terms of the countries of the scientists.

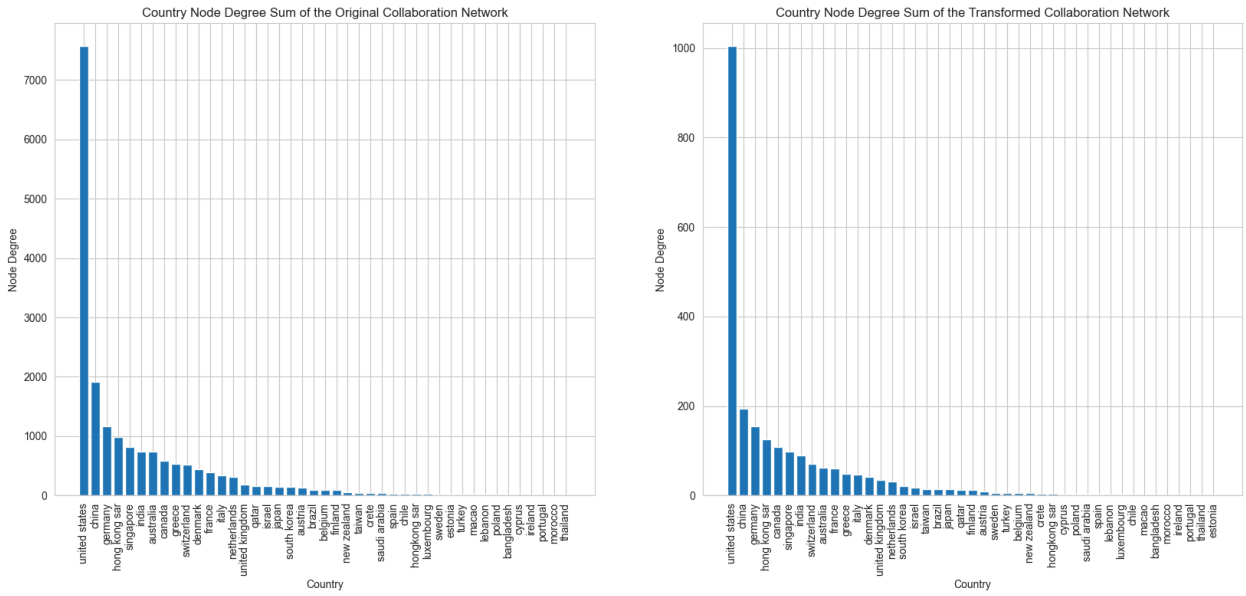

Figure 36. Countries Node Degree Sum - Transformed vs Original Network Comparison

The same thing can be seen in Figure 37 too, again, the distribution of the scientists' expertise remained the same with just the sum being reduced significantly.

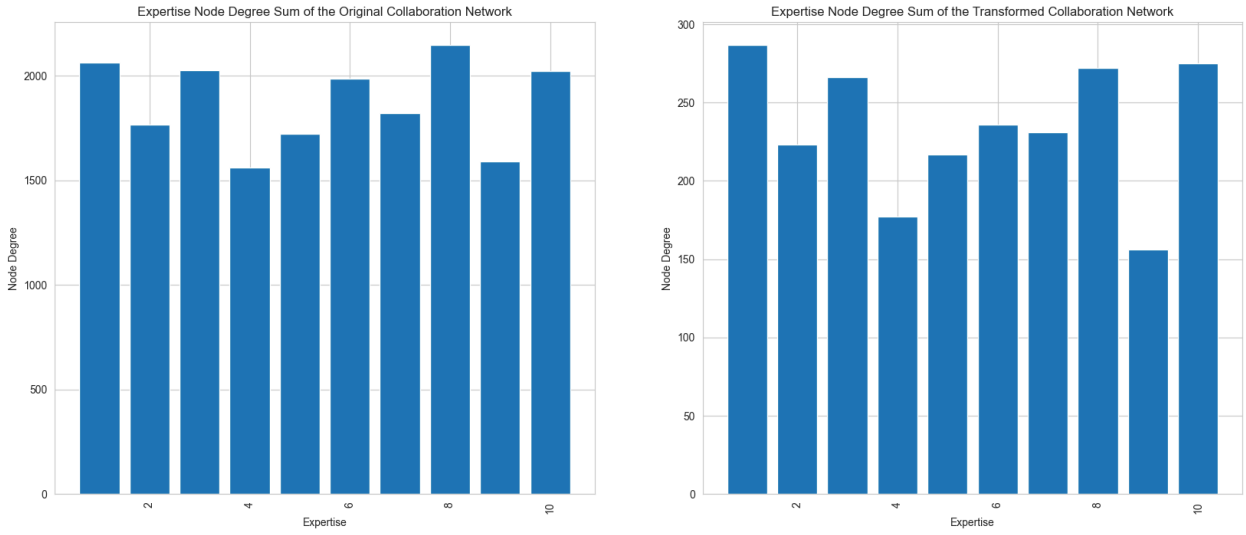

Figure 37. Expertise Node Degree Sum - Transformed vs Original Network Comparison

Finally, we also managed to maintain a similar distribution of the institutions of the scientists like in the original network. As we can see in Figure 38, despite some positions changed, most of the institutions with highest degree sums remained the same. This distribution is the hardest one to be maintained as it has much more discrete

values compared to the other two distributions. The values of the sums were reduced too, following the requirements of the network transforming model.

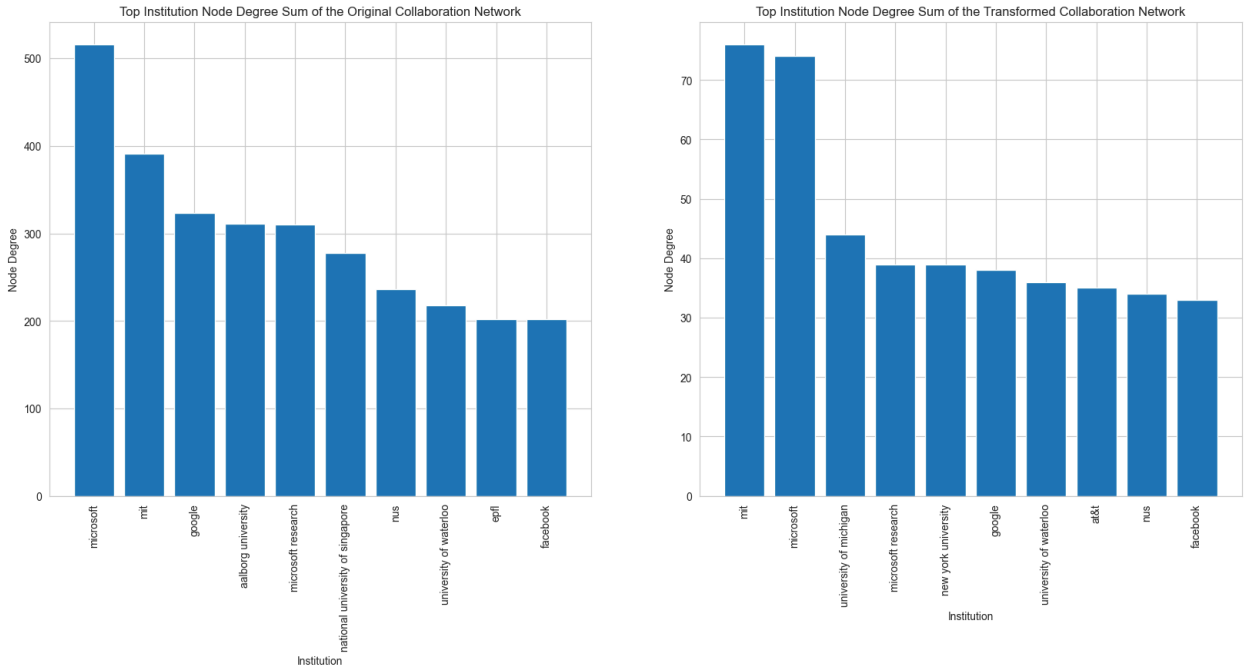

Figure 38. Top Institutions Node Degree Sum - Transformed vs Original Network Comparison

In conclusion, judging by the 3 figures above, we can guarantee that our algorithm successfully manages to reduce the k_{max} and the node degrees k_i in the network while maintaining all distributions and diversities from the original network.

6.2. Alternate Algorithm - Scale Free Reduction using Closeness Centrality Difference

This section goes through a quick introduction to an alternate reduction algorithm. While this model is not able to reduce the size of the giant component greatly it is better at maintaining the degree distribution of the original network. This technique has two parts:

1. **Edge Importance Metric** - To measure the importance of each edge we defined a metric Closeness Centrality Difference as:

$$Closeness\ Centrality\ Difference\ (E_{n1,n2}) = \left|\frac{1}{C_C(n1)} - \frac{1}{C_C(n2)}\right|$$

where n1 and n2 are nodes in the graph, $E_{n1,\,n2}$ is the edge connecting n1 & n2, and $C_C(x)$ is the closeness centrality of node x.

2. **Edge Pruning Policy** - To prune the edges in the network, we use the user defined k_{max} to calculate a new γ for the network. We then reduce the degree of each node to the degree with the same probability in the new scale-free network by removing edges with the greatest difference in each node.

$$\gamma_{new} = 1 + \frac{\ln(N)}{\ln(k_{max})}$$

$$k_{new} = k_{old}^{\frac{\gamma_{old}}{\gamma_{new}}}$$

Figure 39 shows the distribution of Closeness Centrality Difference in the network. The linear decreasing nature of the distribution is not ideal for being used with a hard threshold and multiple values of k_{max}. as it may either remove too many edges or too few depending on the value of k_{max} specified by the user.

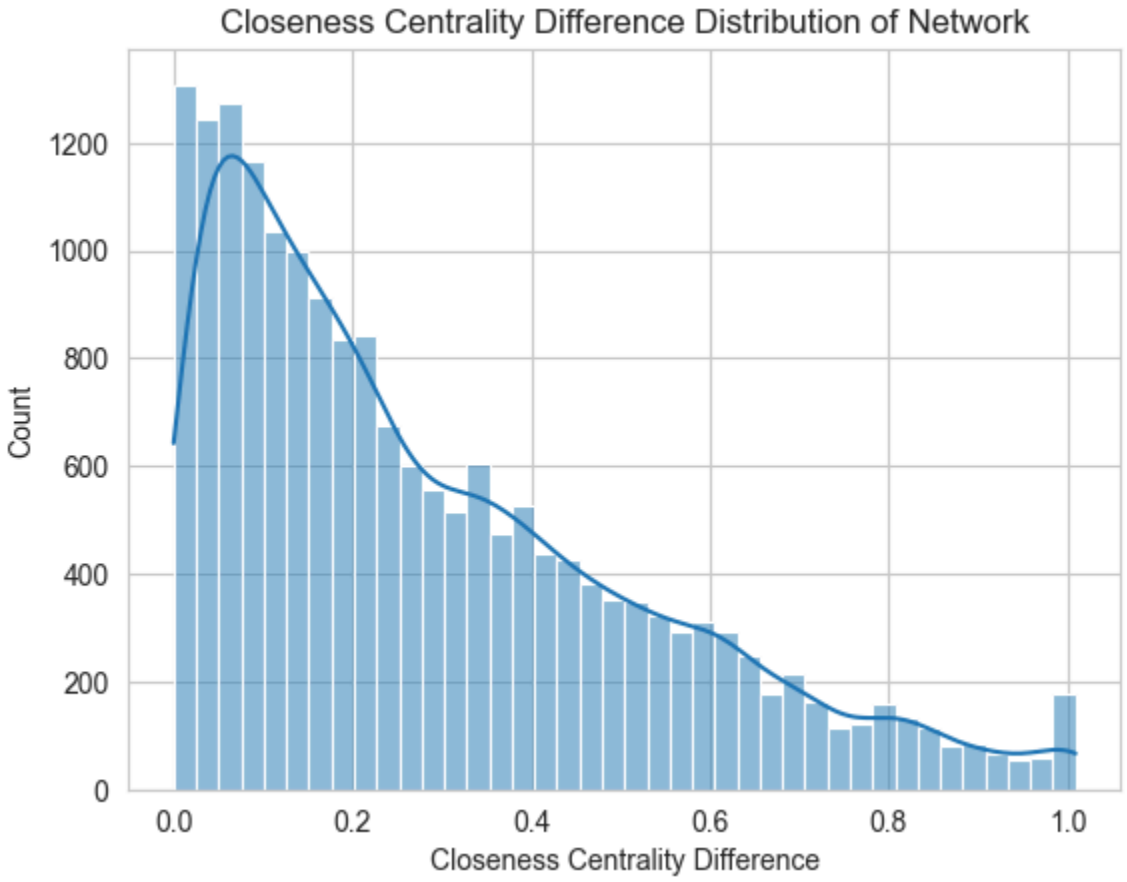

Figure 39. Closeness Centrality Difference Distribution

The relation between the size of the giant component of the transformed network and the user-specified value of k_{max}, while using the Scale Free Reduction with Closeness Centrality Difference algorithm is displayed in Figure 40. We can observe that while the algorithm provides a range of values for the size of the giant component depending on the input k_{max}, the range is limited to a small range between 820 and 950.

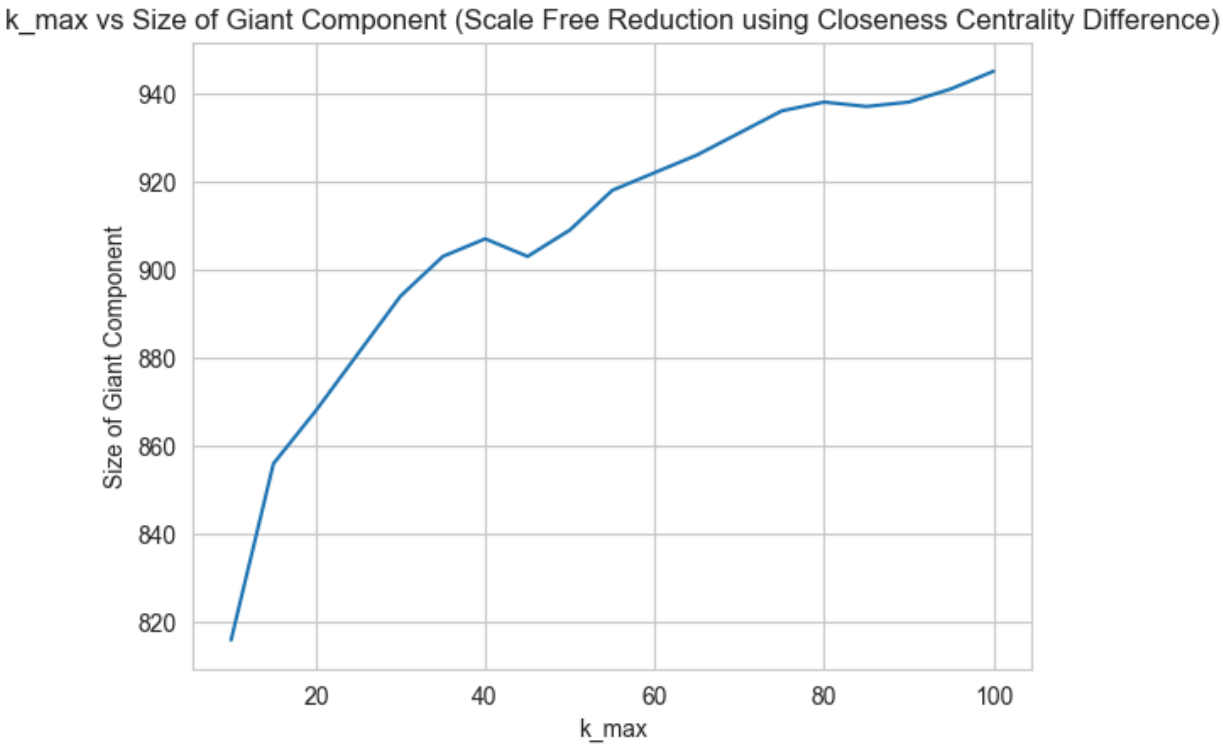

Figure 40. k_{max} vs size of giant component

This algorithm performs worse than the previous in reducing the size of the giant component. It does a better job at preserving the degree distribution of the original network. Networks transformed by this algorithm also have significantly fewer nodes with a degree of zero than the previous algorithm. Figure 41 shows the degree distribution of a network transformed using this algorithm with k_max = 75. It has a giant component containing 935 nodes compared to 993 in the original network.

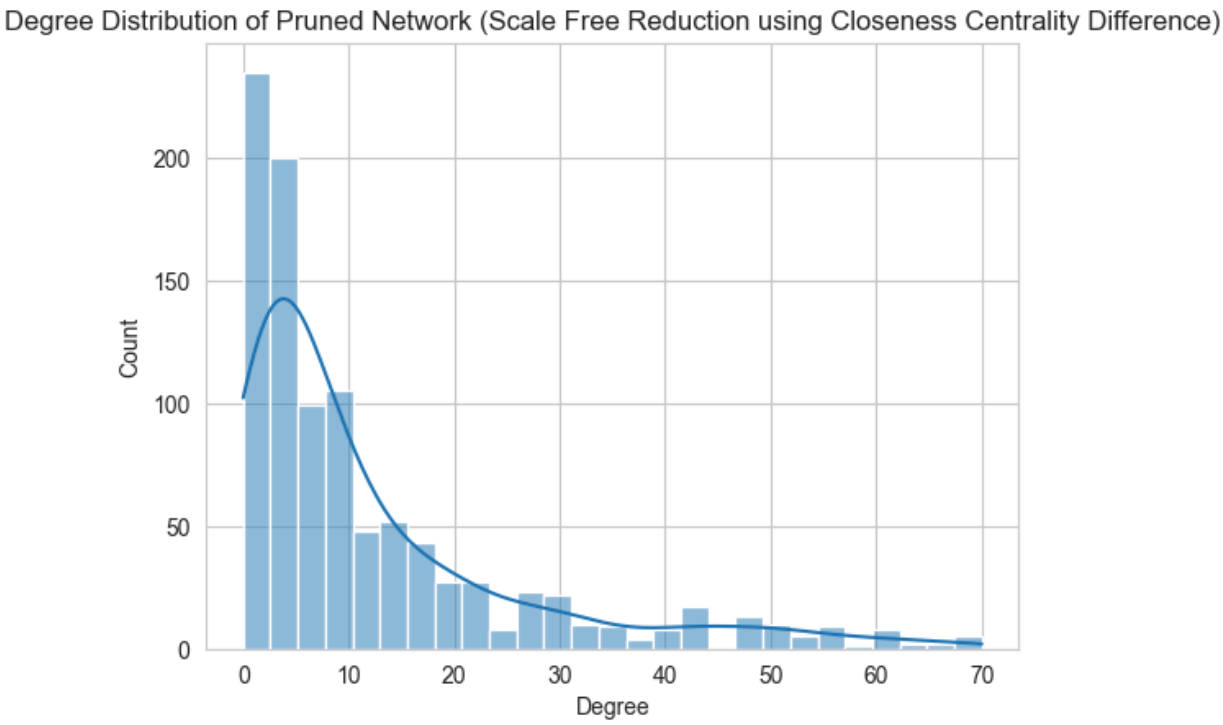

Figure 41. Degree Distribution of Pruned Network

6.3. Other Algorithms

The table below (Table 1) contains all the edge importance metrics we tried before arriving at the two methods above. We tried each importance metric with both pruning techniques (Hard Threshold and Scale Free Reduction).

Table 1. Alternate Edge Importance Metrics Tried

Metric Name	Formula
Net Normalised Weight (Defined)	$Net\ Norm\ Weight(E_{n1,n2}) = \frac{1}{2} \cdot \left(\frac{Weight(E_{n1,n2})}{\sum_{i \neq n1} Weight(E_{n1,i})} + \frac{Weight(E_{n1,n2})}{\sum_{i \neq n2} Weight(E_{i,n2})} \right)$
Net Normalised Edge Importance (Defined)	$Net\ Norm\ Edge\ Imp(E_{n1,n2}) = \frac{1}{2} \cdot \left(\frac{1}{C_D(n1)} + \frac{1}{C_D(n2)} \right)$
Max Normalised Edge Importance (Defined)	$Max\ Norm\ Edge\ Imp(E_{n1,n2}) = max\left(\frac{1}{C_D(n1)}, \frac{1}{C_D(n2)} \right)$

Edge Betweenness Centrality	$Edge\ Betweeness\ Centrality\ (e) = \sum_{n1,n2 \in V} \frac{\sigma(n1, n2\ \|e)}{\sigma(n1, n2)}$

7. Network with External Authors (Additional Analysis)

In this section, we will analyze the collaboration network of the data scientist again, but this time we will also include the connections between the given ones with their external collaborators.

7.1 Network Visualization

Including the external authors to the network increases its size substantially. Because of the increased size, we visualized the network using the Cosmograph online tool. The Cosmograph tool is used for visualizing complex networks, and it provides a user-friendly interface for exploring the network. We have shown the visualization of the network in the notebook using the screenshot of the Cosmograph tool (Figure 42). The Cosmograph tool can be accessed at https://cosmograph.app. The external authors are colored in red, while the original ones remain in blue. The red clusters represent the external authors that collaborated to a single scientist from the original dataset.

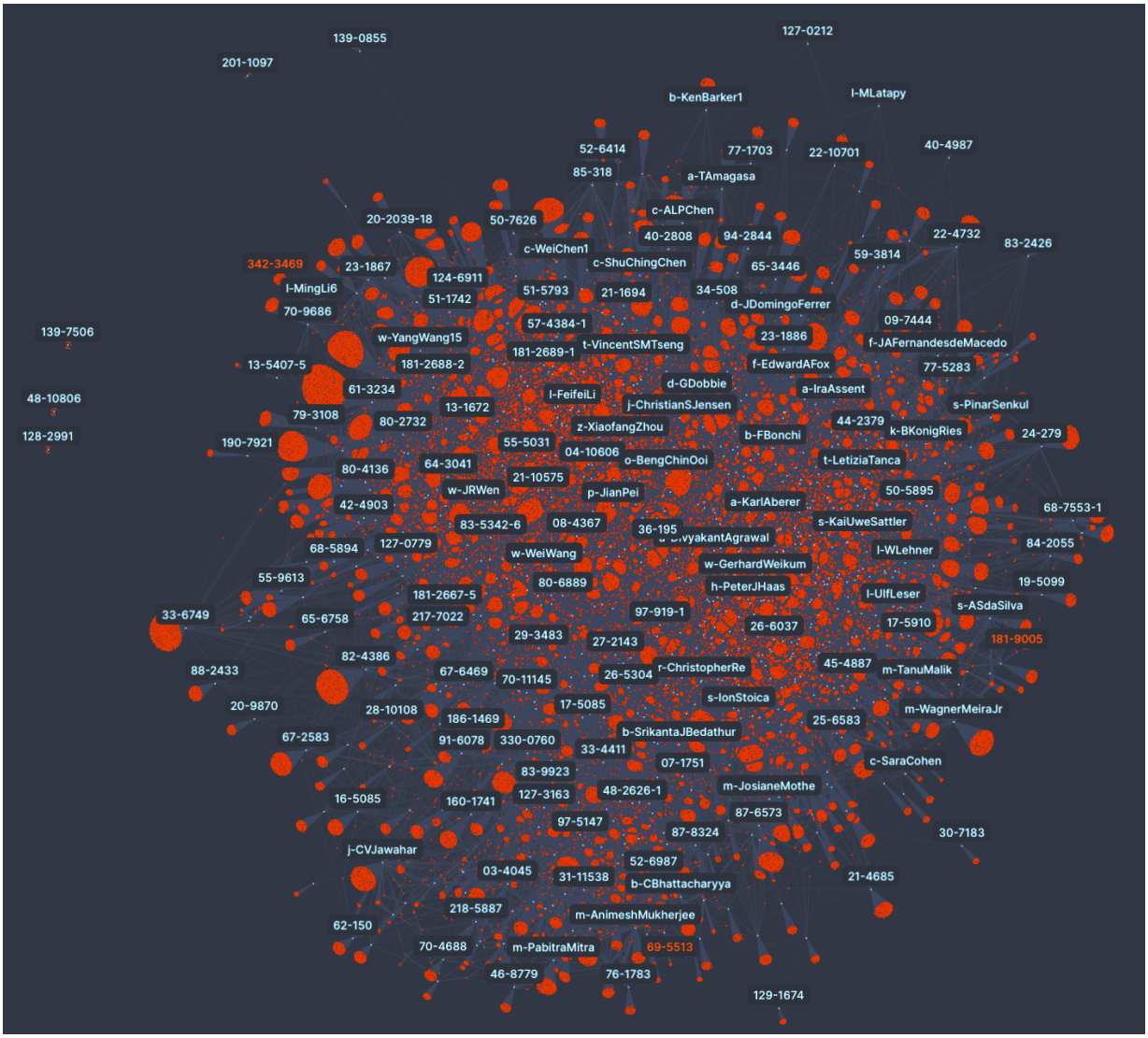

Figure 42. The Network with External Authors

7.2 Network's Basic Properties

The new network consists of 81 975 nodes and 148 335 edges. After computing the average node degree, its value is $< k >= 3.62$. Despite the fact that the network expanded with a lot of external authors, the average degree of the nodes in the network became lower. This is because we have recorded the connections between the original scientists and their external collaborators, we are not considering the connections between the external collaborators. As a result, most of the external authors will be connected to only a few original scientists, and none will be connected to each other. The clustering coefficient of the network has the value of $C \sim 0$. As expected, the average

clustering coefficient of the network is extremely low because we added tens of thousands of nodes without connections between them.

7.3 Degree Distribution Analysis

When we compare the degree distribution of the network with the external authors (in Figure 43) to the one of the original network, we can see that both follow the power law. Thus, the new network with the external authors is a scale-free one too.

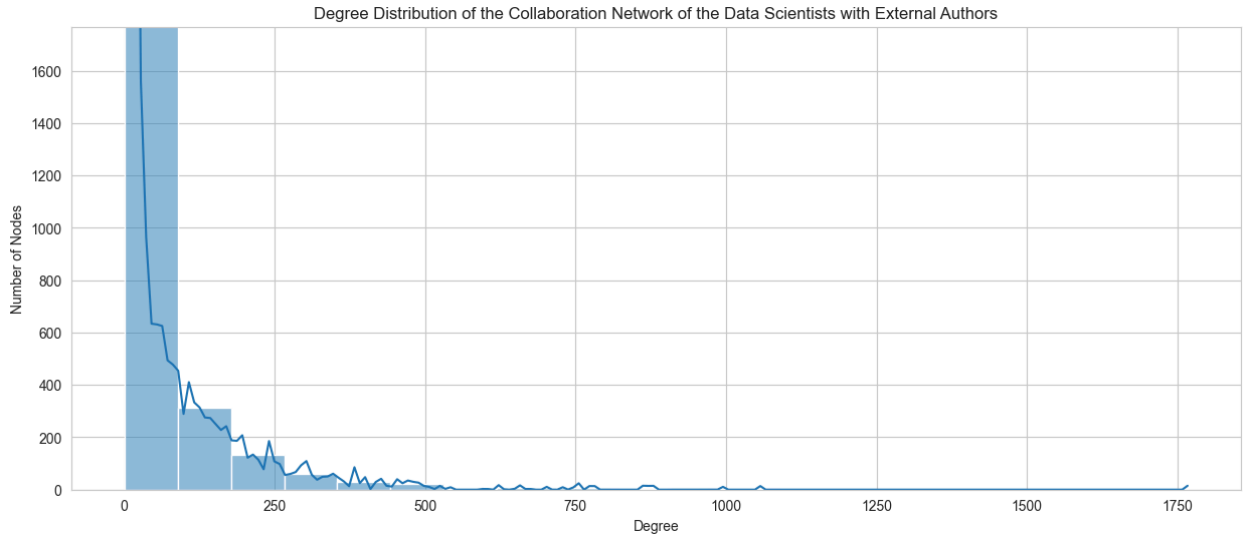

Figure 43. The Degree Distribution of the Network with External Authors

7.4 Degree Centrality

When we compare the degree centrality of the network (Figure 44) with the node degree distribution (Figure 403, we can see that they are very similar. This is because the degrees of the nodes with the high degree are the ones from the original network because only they can have connections between them, the new ones have low degree as they cannot connect to each other. Thus, the degree centrality of the nodes follows the same power-law distribution as the one for the node degrees. The newly added nodes can be considered as the leaves of the network, so their degree centrality is very low (near 0).

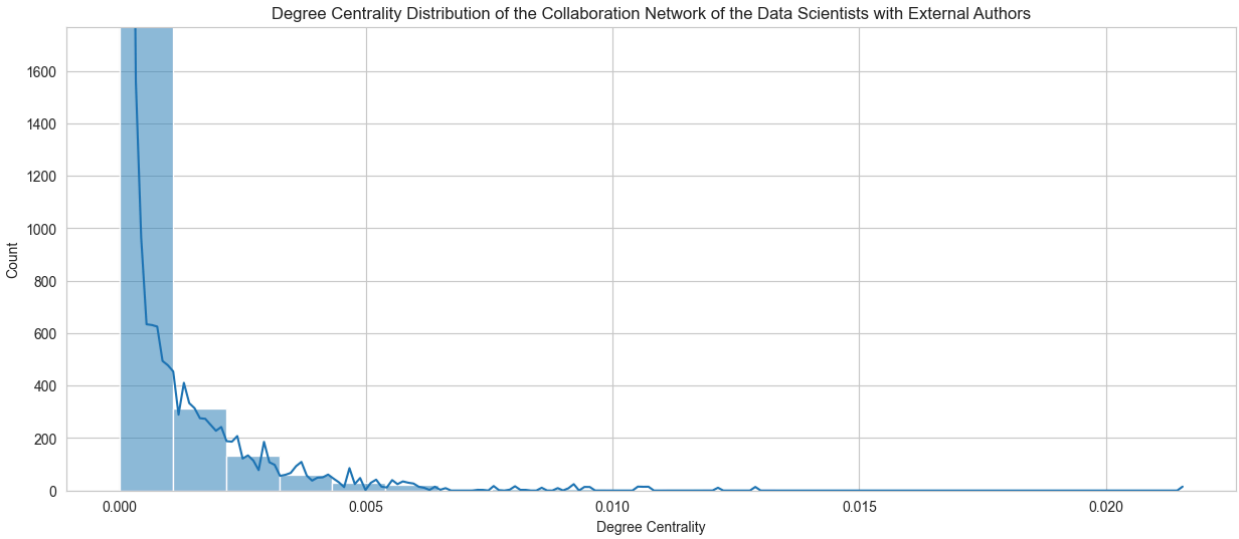

Figure 44. The Degree Centrality of the Network with External Authors

References

[1] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-law distributions in empirical data,” SIAM Rev., vol. 51, no. 4, pp. 661–703, Nov. 2009, doi: 10.1137/070710111.