NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

NTU Research

Network Science-Based Analysis of Collaboration Networks of Data Scientists

Done By:

Alvin Tang

Edan Kang

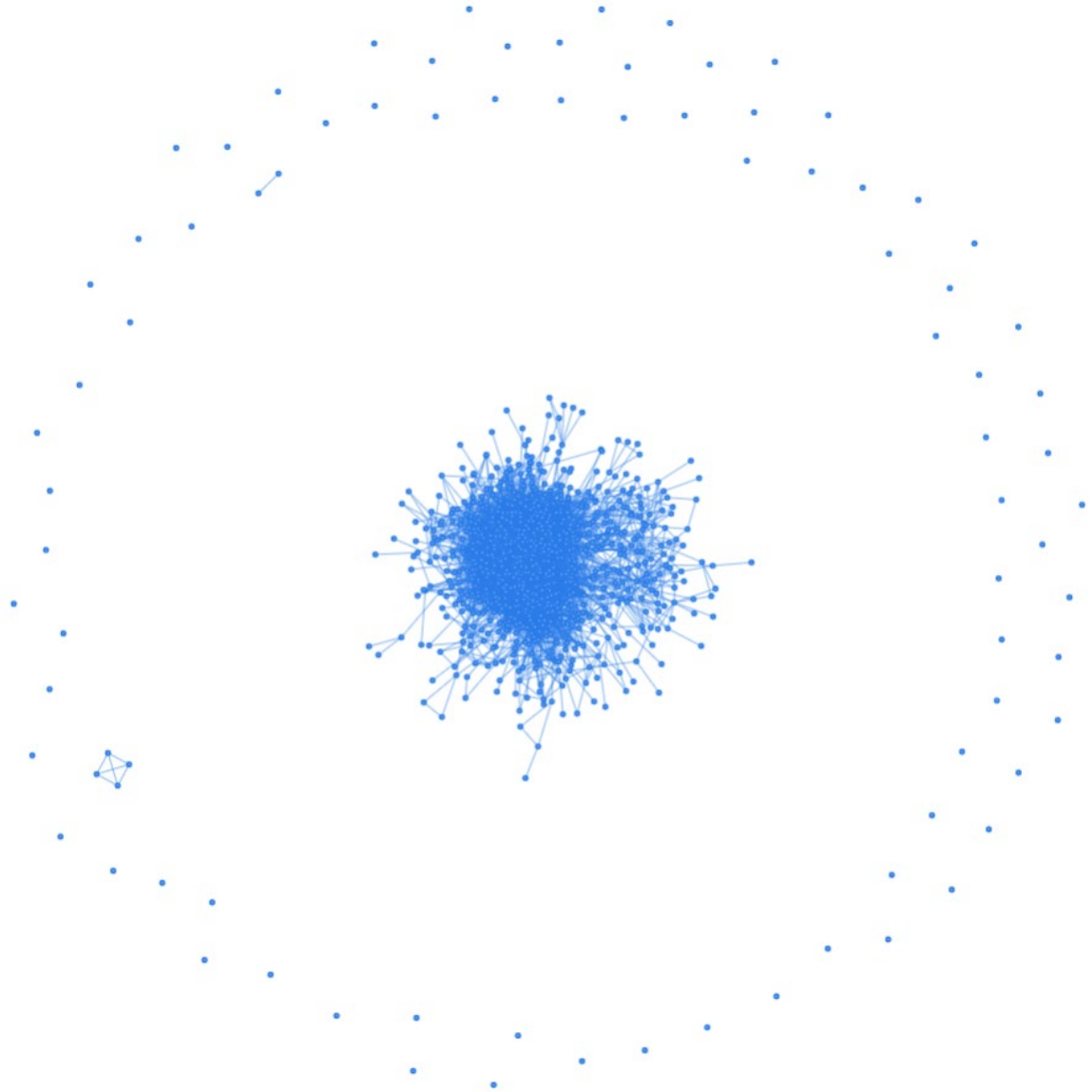
Goh Nicholas

Ivan Chay

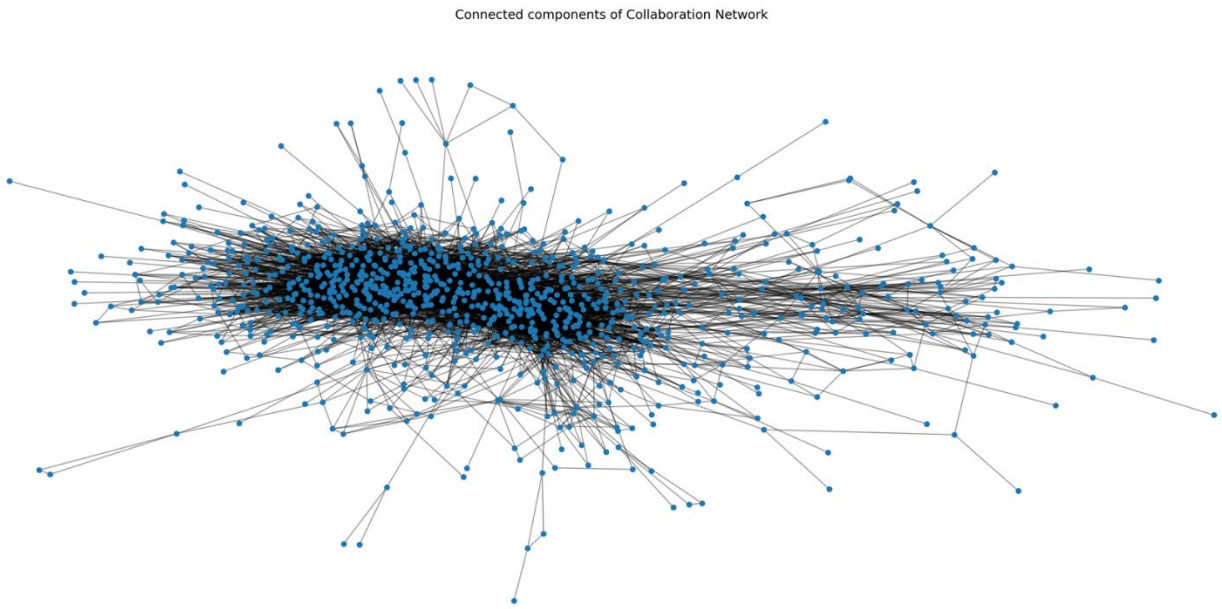
Network Properties of Collaboration Network

General Properties of Collaboration Network	
Information	Graph with 1068 nodes and 6454 edges
Density (3.s.f)	0.0113
Total number of nodes	1068
Total number of edges	6454
Average Degree	12.086
Average clustering coefficient (3.s.f)	0.302

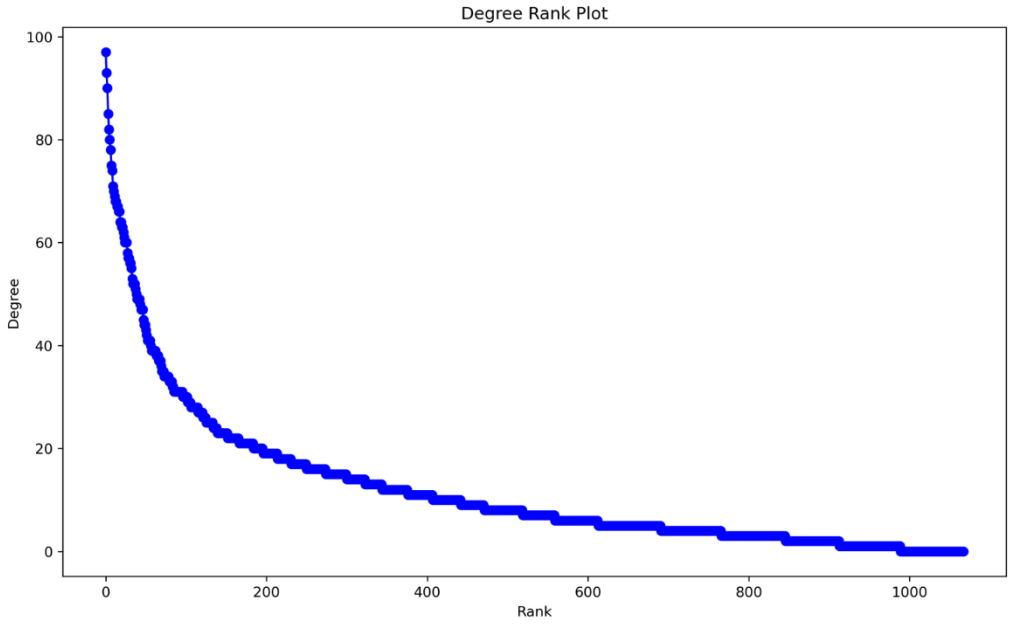
The subgraph of connected components of the collaboration network



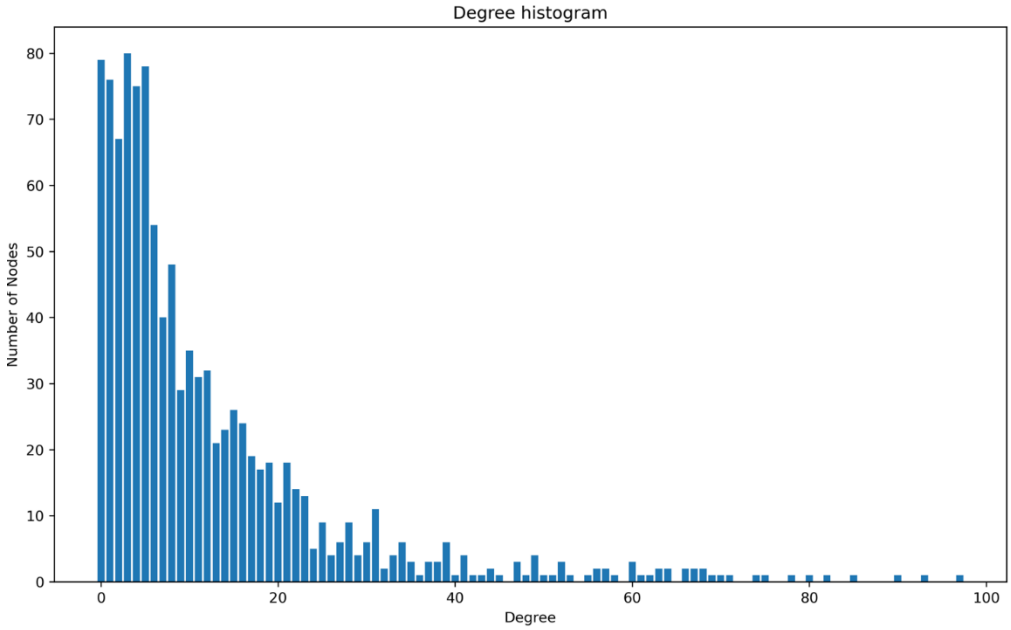
Giant Component of collaboration network



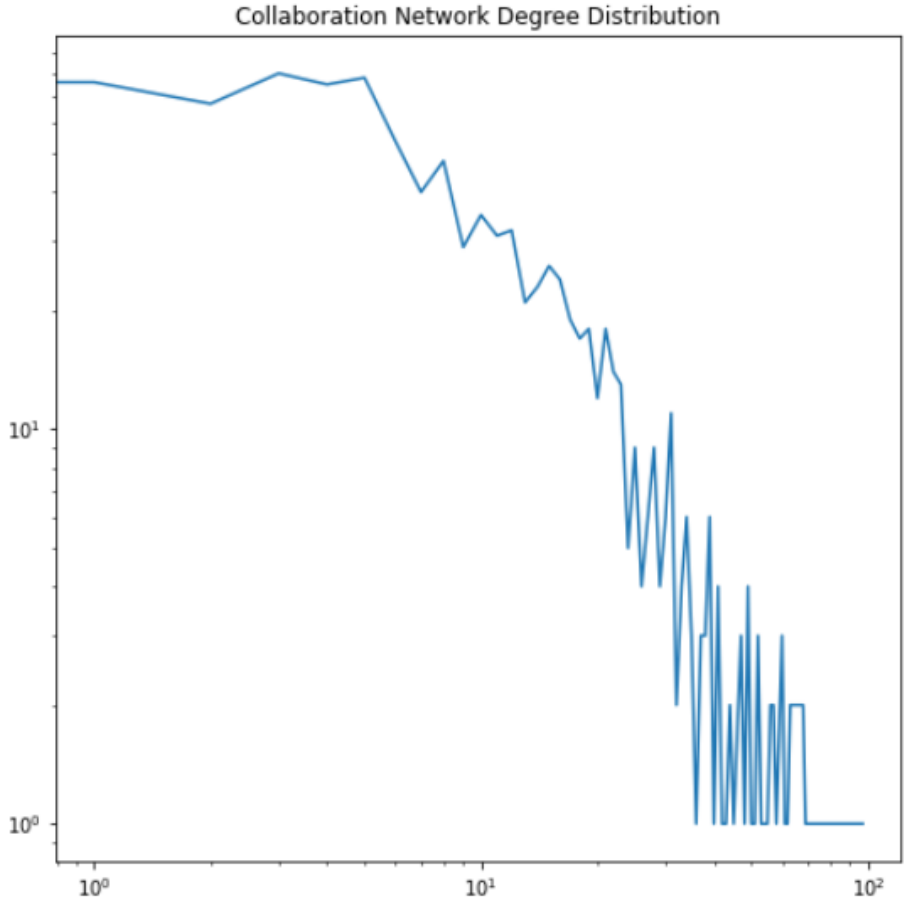
The degree-rank plot of the collaboration network



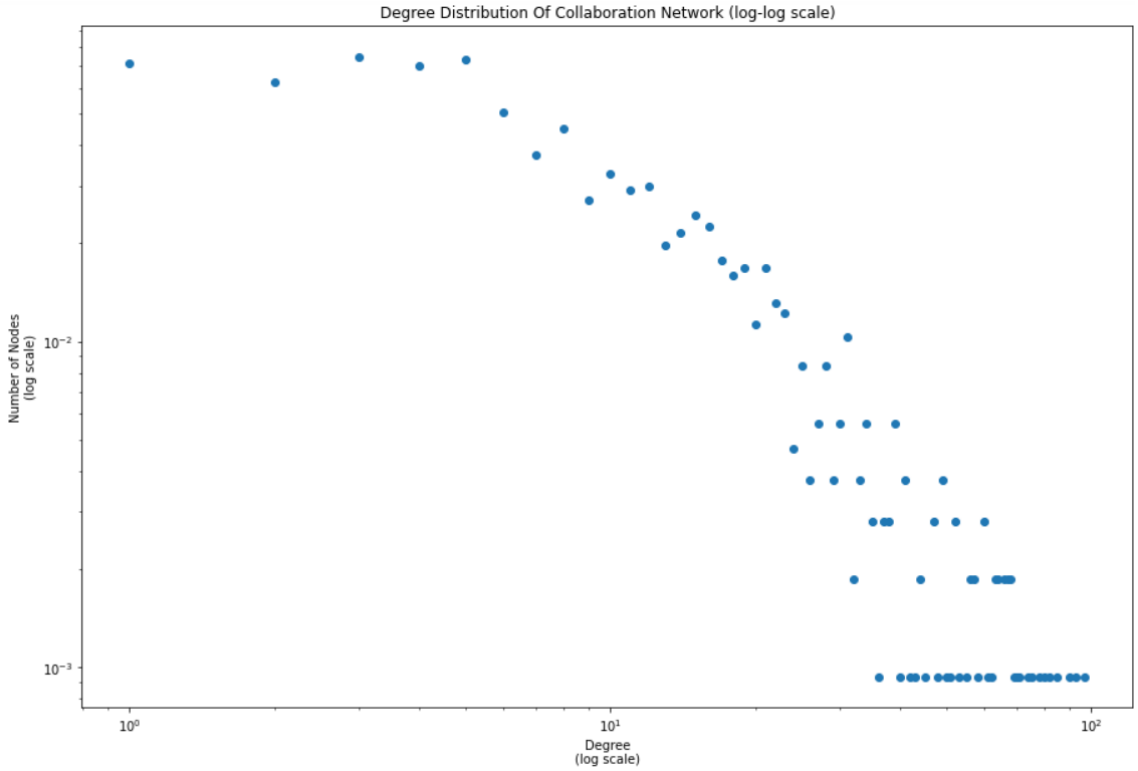
The degree histogram of the collaboration network



The degree distribution of the collaboration network

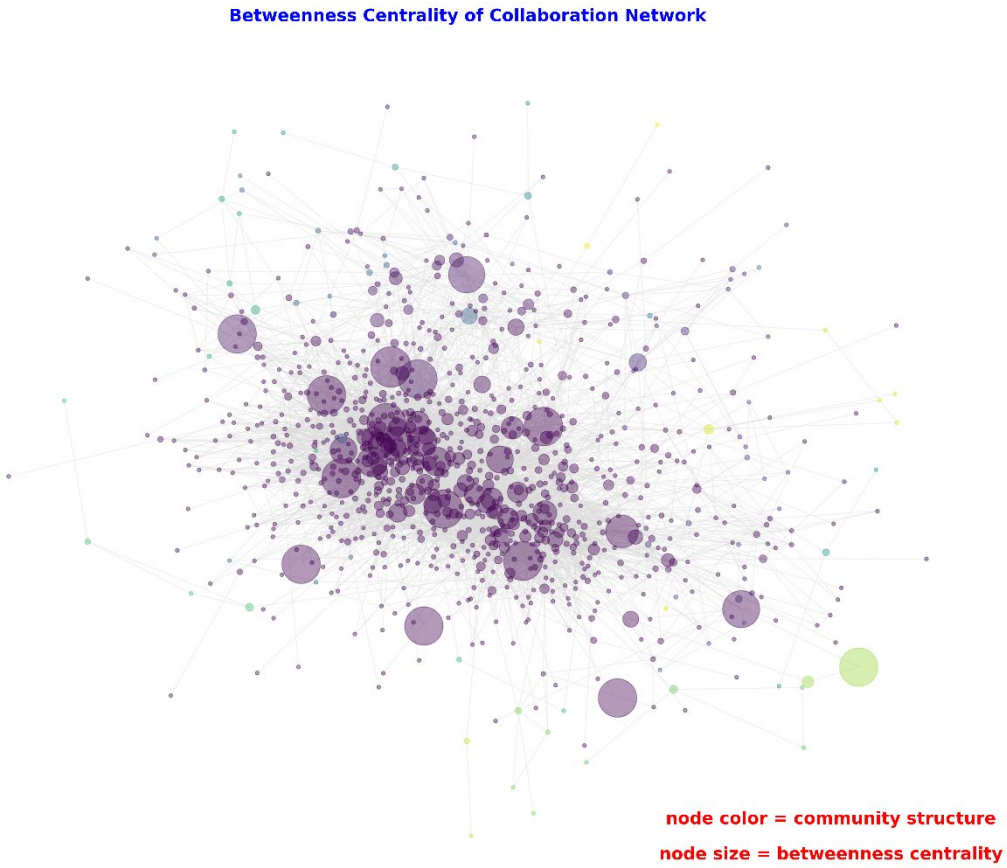


The degree distribution of the collaboration network (log-log scale)

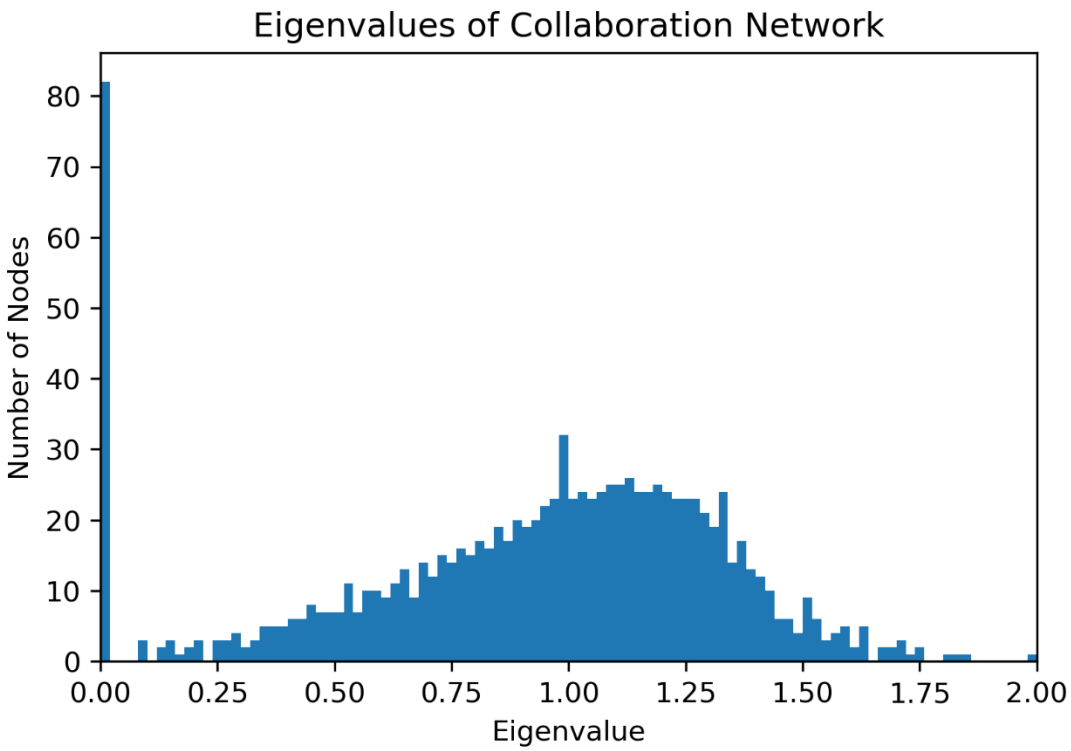


Below will be the network centrality properties

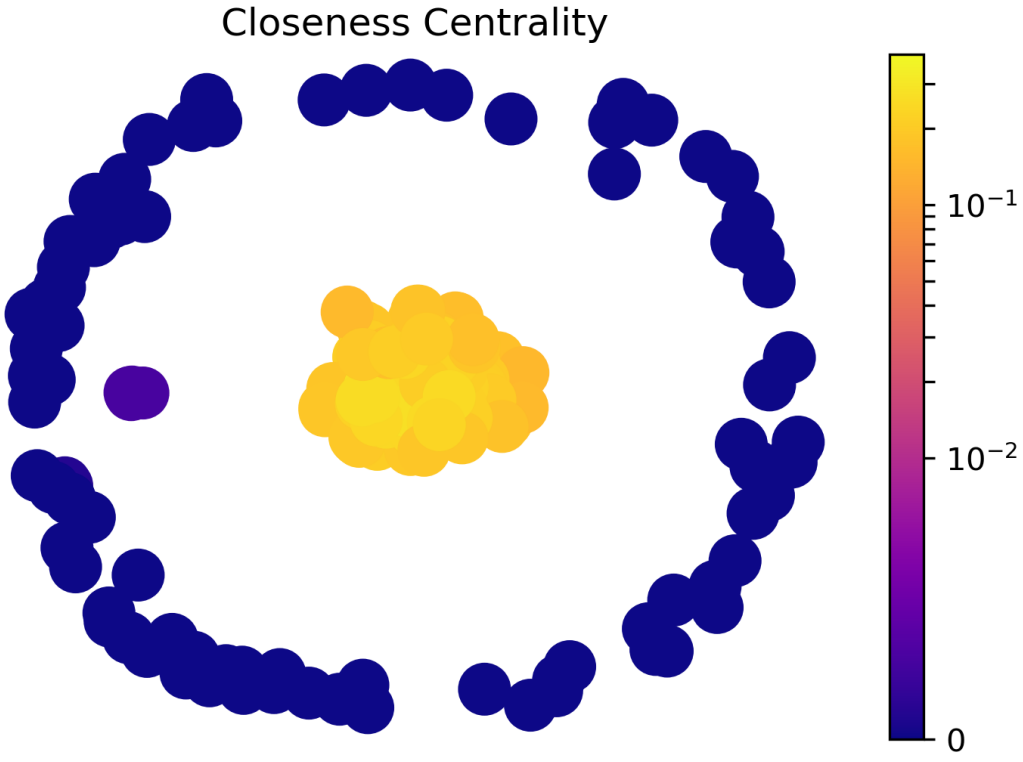
- Betweenness Centrality



➢ Eigenvector Centrality



➢ Closeness Centrality



The table below contains further Analysis on the Collaboration Network's properties

Average Shortest Path Length	3.41	1	0
Diameter	9	1	0
Total number of subgraphs with corresponding values	1	2	79

The table below contains nodes with special properties

Network Properties	PID	Score
Highest Degree Centrality	o/BengChinOoi	0.09091
Highest Betweenness Centrality	w/GerhardWeikum	0.03982
Highest Eigenvector Centrality	o/BengChinOoi	0.18969
Highest Closeness Centrality	o/BengChinOoi	0.39432

The analysis for part 1 looks at the entire collaboration network. The temporal dimension of the network will be considered in part 2.

From the above illustrations, the clearest distinction about the collaboration network is with regards to its degree distribution. The degree distribution appears to follow a power law distribution from the charts above as observed from the tail of the distribution. The long tail end with the many nodes having only a few links as well as a few hubs with large number of nodes suggests that the Collaboration Network follows a power law distribution and characterises the scale-free property of the Collaboration Network.

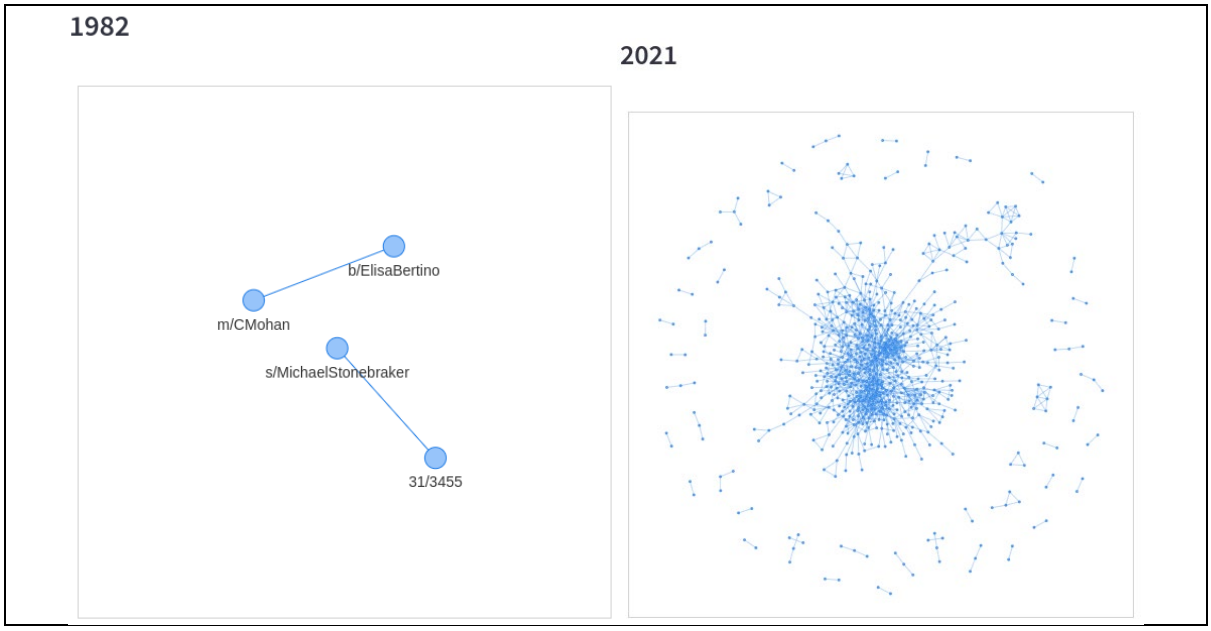
The average degree of the network was found to be around 12. The existence of nodes with degrees close to 100, which is much larger than <k> of 12, denotes the existence of hubs. <k> is also much smaller than N – 1 (1068 – 1 = 1067), which signifies that the network is sparse.

The actual clustering coefficient of 0.302 is relatively high. This shows the cliquishness of nodes within the network. Combined with the low average distance within the network, the small world property can be observed from the Collaboration Network.

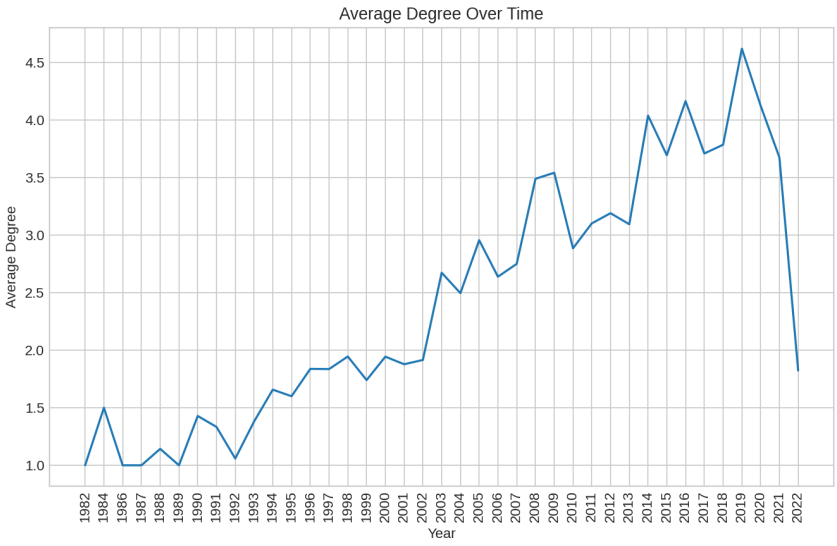
o/BengChinOoi appears to be the most central node with highest centrality values across three out of the four measures. Despite this, his betweenness centrality value falls on the lower end. This could suggest that there are multiple paths in the network, and that the ego is near many people, but so are many others. This can be observed from many other nodes having similarly high closeness centrality values as seen below:

```
PID with highest closeness centrality:
 o/BengChinOoi  with value:  0.39432

PID with second closeness centrality:
 j/HVJagadish  with value:  0.38038

PID with third closeness centrality:
 s/DiveshSrivastava  with value:  0.38006

PID with forth closeness centrality:
 w/GerhardWeikum  with value:  0.37563

PID with fifth closeness centrality:
 h/AlonYHalevy  with value:  0.37485
```

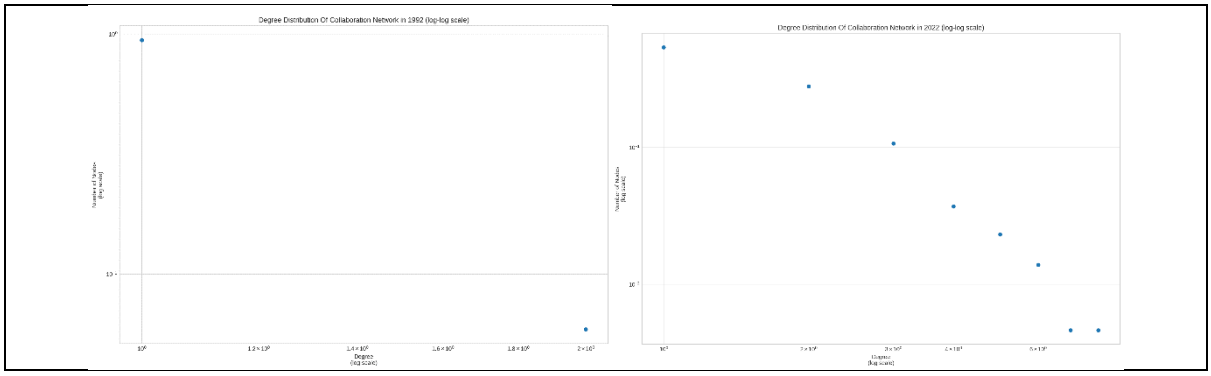
1. Evolution of Collaboration Network and its Properties Over Time



Progression of network



The network starts out small and gradually becomes more connected as the years increases. This can be seen in the average degree over time, which increases over time as more papers are co-published per year. With the advent of technology to allow people to collaborate more with ease regardless of where they live, it explains why the average degree is higher as compared to the years before. Between 2021 and 2022, we can observe a huge dip in average degree. This is due to 2022 being only 3 months in, skewing the average degree data. Our team predicts that by the end of 2022, the average degree for this current set of data scientists will be around or higher than 4.5

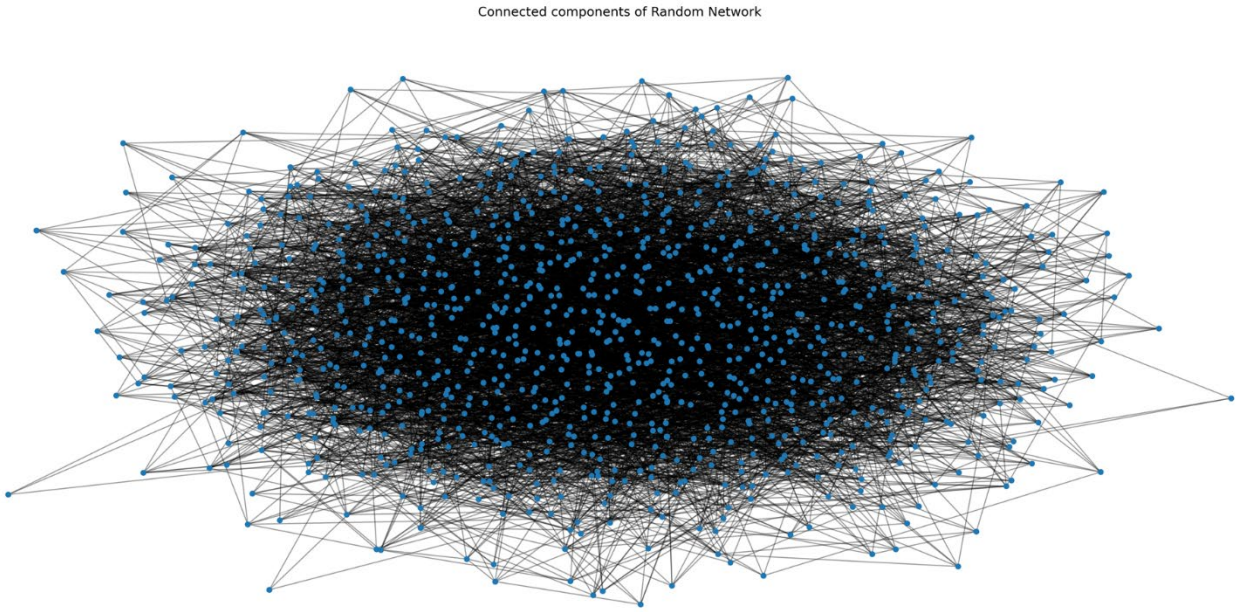


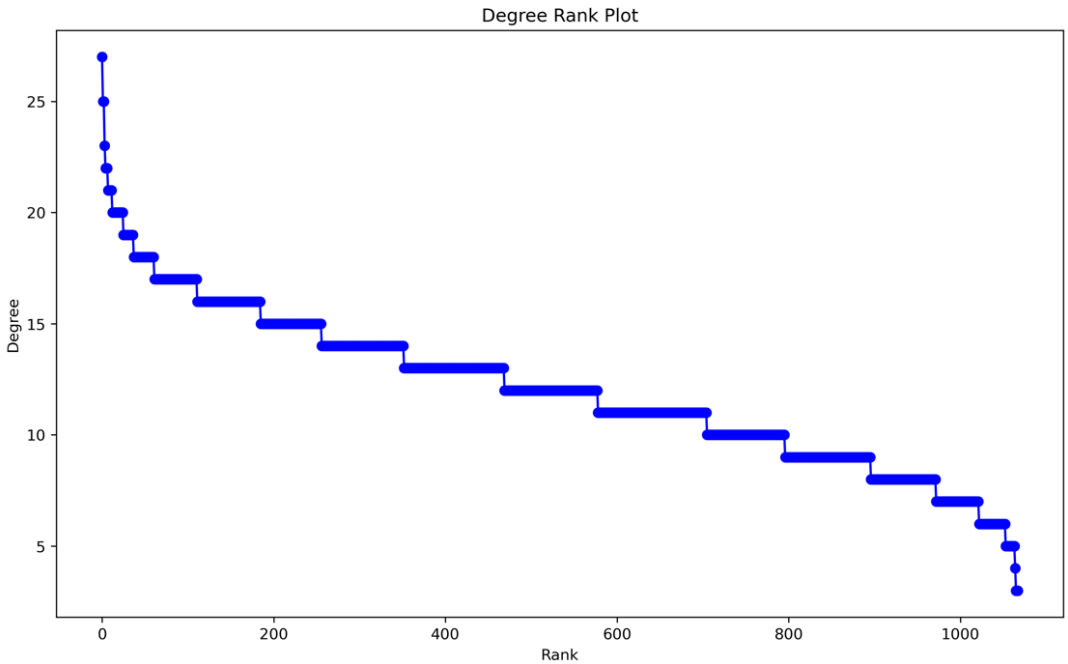
Degree distribution over time

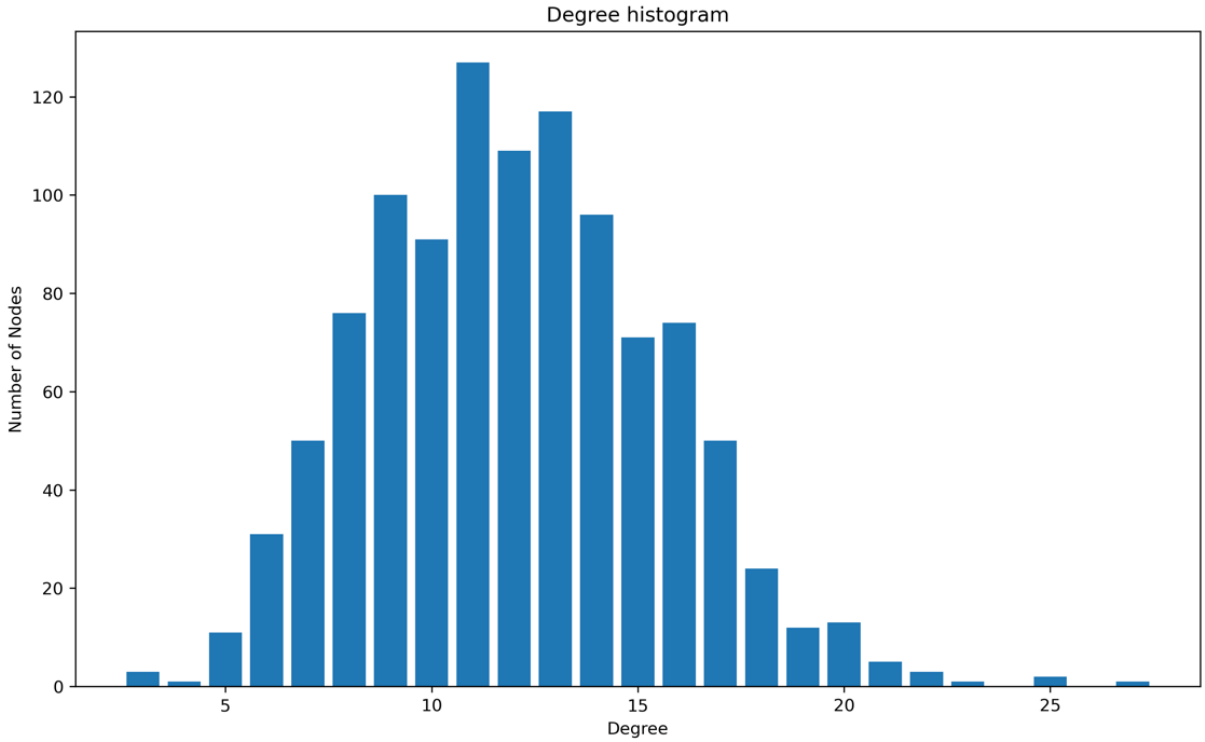
The collaboration network per year also appears to follow a power distribution as the seen in the straight line in the log log plots above. This follows from the power distribution of the whole network discussed in part 1.

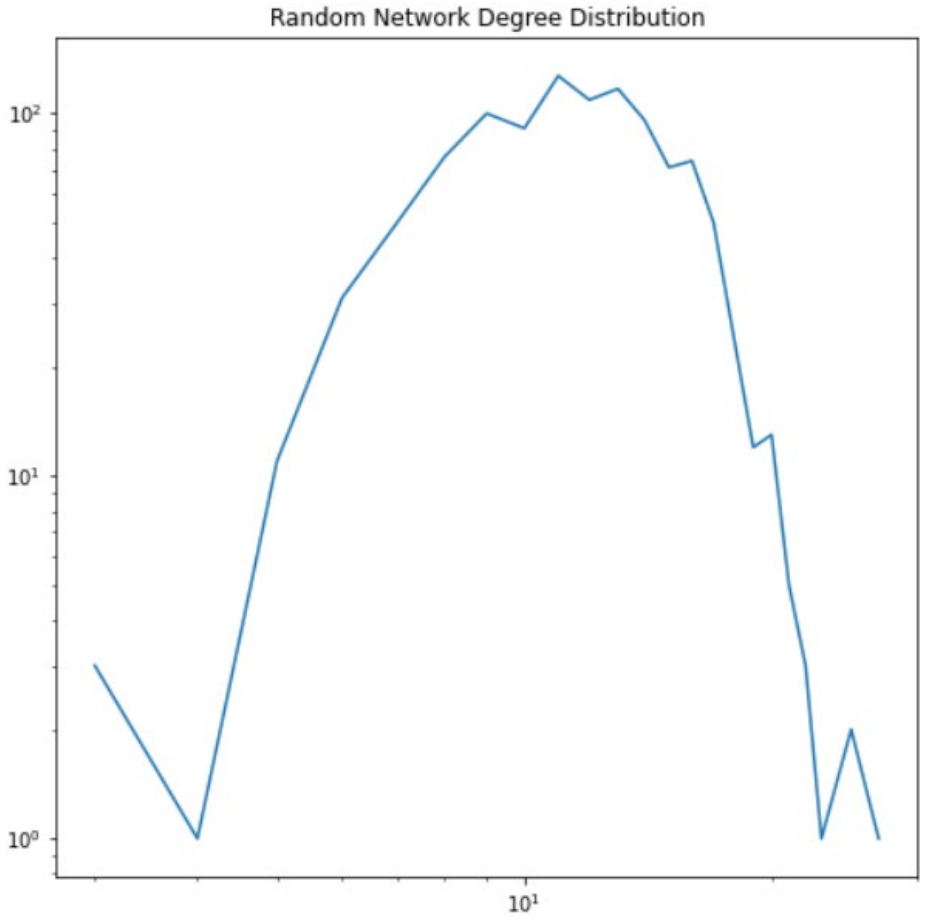
2. Random Network and its Comparison with the Collaboration Network

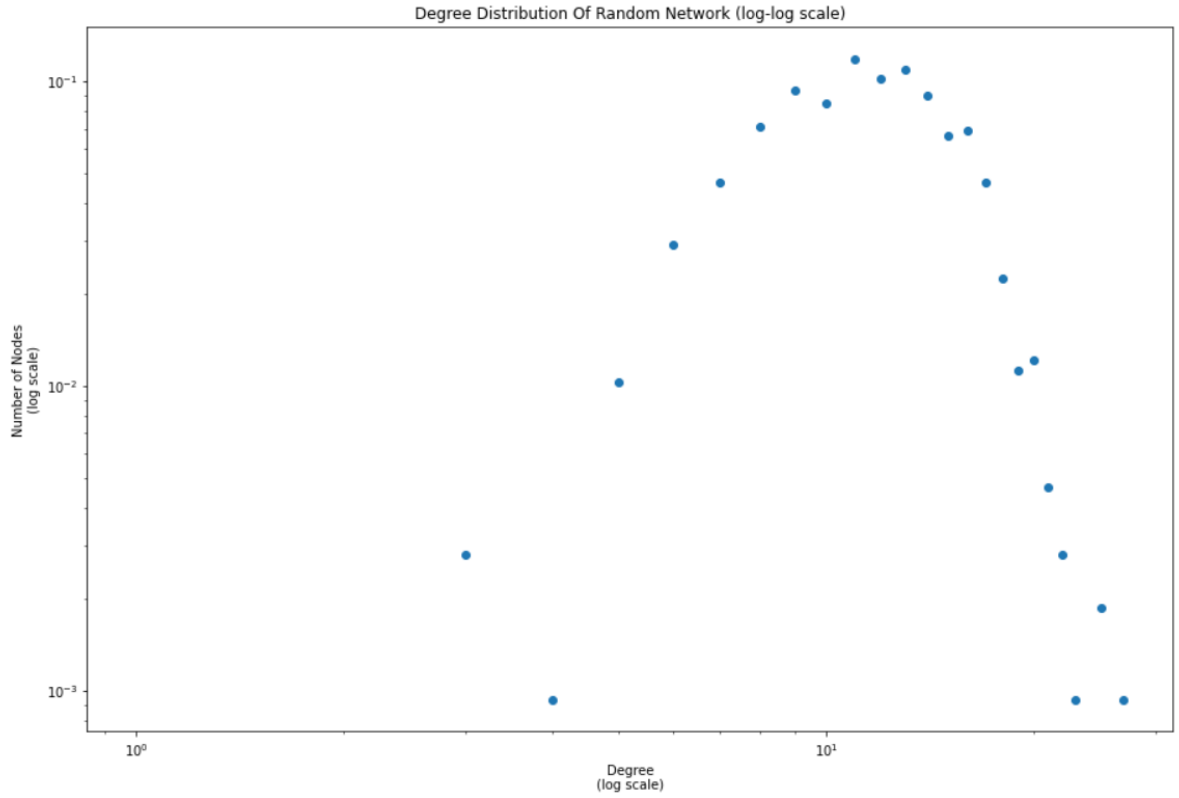
General Properties of Collaboration Network	
Information	Graph with 1068 nodes and 6441 edges
Density (3.s.f)	0.0113
Total number of nodes	1068
Total number of edges	6441
Average Degree	12.068
Average clustering coefficient (3.s.f)	0.011



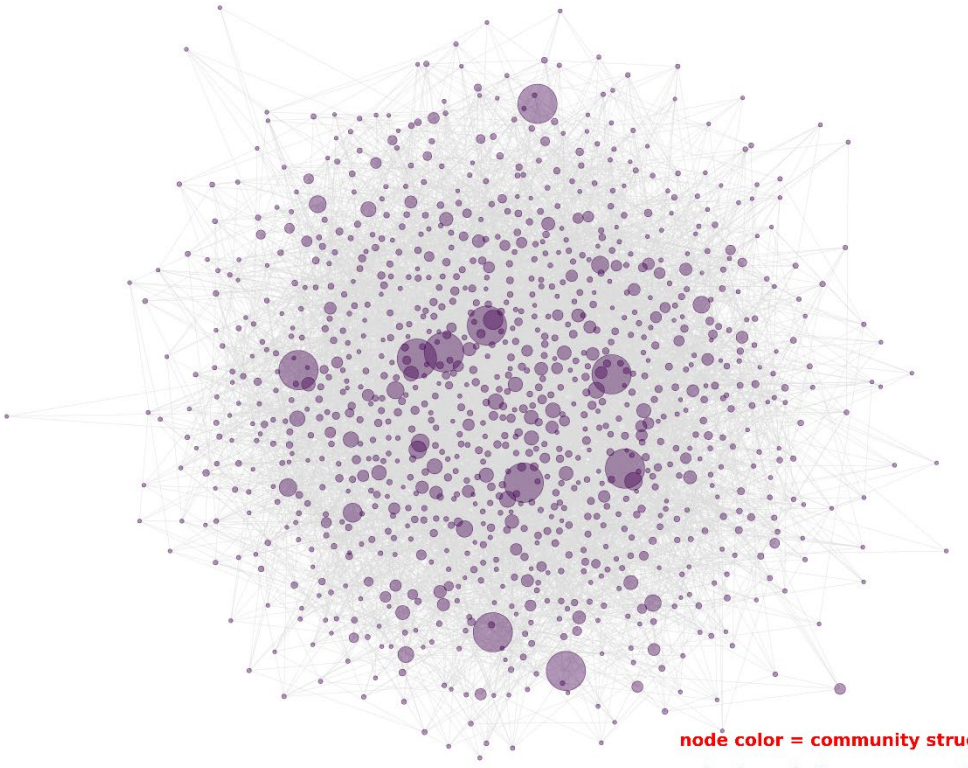


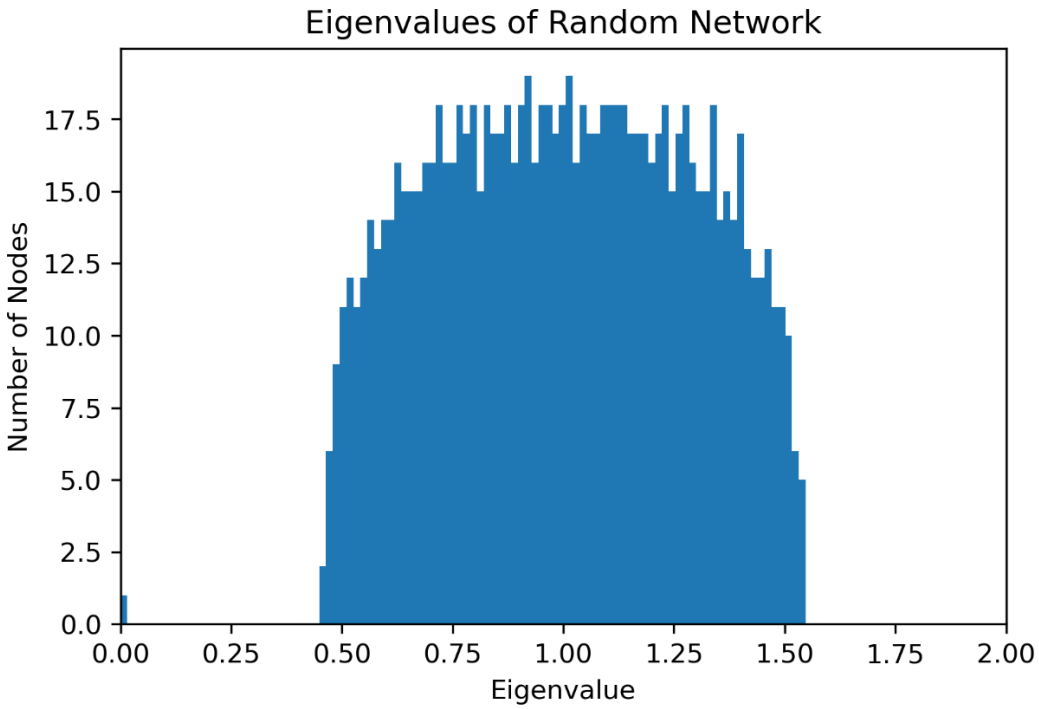

Degree histogram
Number of Nodes
Degree
0
20
40
60
80
100
120
5
10
15
20
25

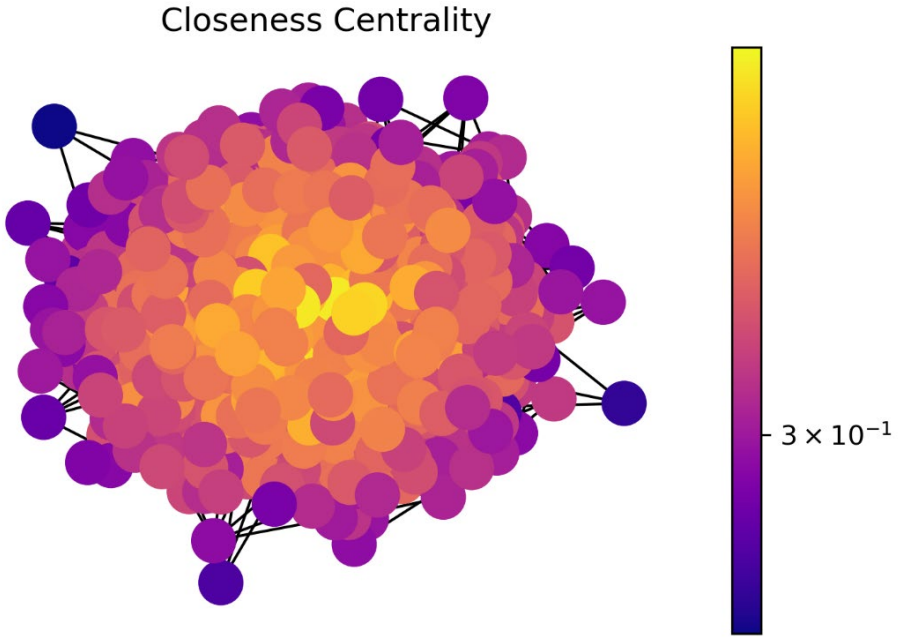

Random Network Degree Distribution
10^2
10^1
10^0
10^1



Betweenness Centrality of Random Network







The table below contains further Analysis on a particular Random Network's properties

Average Shortest Path Length	3.06
Diameter	5
Total number of subgraphs with corresponding values	1

The table below contains nodes with special properties

Network Properties	PID	Score
Highest Degree Centrality	472	0.02343
Highest Betweenness Centrality	472	0.00734
Highest Eigenvector Centrality	634	0.06578
Highest Closeness Centrality	634	0.36367

To facilitate a fairer comparison between the collaboration network and the random network, we designed the random network in such a way that the number of nodes, number of links, and average degree are similar to that of the collaboration network.

To do so, the probability p of a connection being formed between two arbitrary nodes must be set specifically as:

$$\langle k \rangle = p(N - 1)$$

$$p = \langle k \rangle / (N - 1)$$

Substituting <k> and N from the actual collaboration network, we get:

$$p = 12.086 / 1067$$

The fast_gnp_random_graph function from the NetworkX Python library was used to generate the Erdős-Rényi random graph used in this comparison, using number of nodes = 1068 and probability of edge creation = 12.086/1067 as parameters.

Similarities

Both the random graph and the actual collaboration network share similarly small average distances of 3.06 and 3.41 respectively.

Differences

A key difference between the actual collaboration network and the random network is the degree distribution. As can be seen in the charts above, for the random graph, there is an umbrella shaped distribution about the mean value <k>, suggesting a Binomial/Poisson distribution. This contrasts with that of the collaboration network which seems to follow a long-tailed power law distribution. The implication of this is that large hubs connected to many other nodes exist in the actual collaboration network, but such hubs do not exist in the corresponding random network.

Also, the clustering coefficient of the random network (0.011) is much smaller than that of the actual collaboration network (0.302). This suggests that the random network is unable to capture the cliquishness of nodes present in the collaboration network. This is likely due to the fixed probability of edge formation between pairs of nodes for the random network.

Conclusion

Ultimately, from the unreconcilable differences between the actual collaboration network and the random graph, it is quite conclusive that the collaboration network is not random.

3. Algorithm for Transformed Network

The requirement of this portion is to transform the network in such a way that it reduces the size of the giant component (1), reduces the degree of every node below a user-specified maximum value (2), as well as preserves country, expertise, and institution diversity (3).

To fulfil tasks 1 and 2, nodes connected to highly connected hubs can be removed. Doing so, the degree of such hubs will decrease to a value below the maximum specified value. Removing such nodes can also help break up the giant component into smaller components.

To satisfy task 3, the selection of the nodes to be removed must preserve overall diversity. Our algorithm attaches weights to individual nodes based on the frequency of appearance of their country, institution, and expertise level among the individuals in the input file. A node belonging to a common country, institution, and/or expertise, will have a higher weight. This means that nodes with lower weights will be preserved to ensure diversity.

Outlined below is the pseudocode of the algorithm.

Pseudocode

1. Count occurrences of each country, institution and expertise, and attach a weight to each node that scales with the number of occurrences. Nodes with high weight implies that it belongs to a common country, institution and/or expertise level.
2. From a hub with degree above user-specified kmax, select a node connected to said hub with the highest weight to be removed.
3. Update the weights of nodes with similar properties (I.e., same country, institution, expertise) as the removed node to reflect the reduced occurrence of such properties.
4. Repeat steps 2 and 3 until all nodes have degree equal to or below kmax.

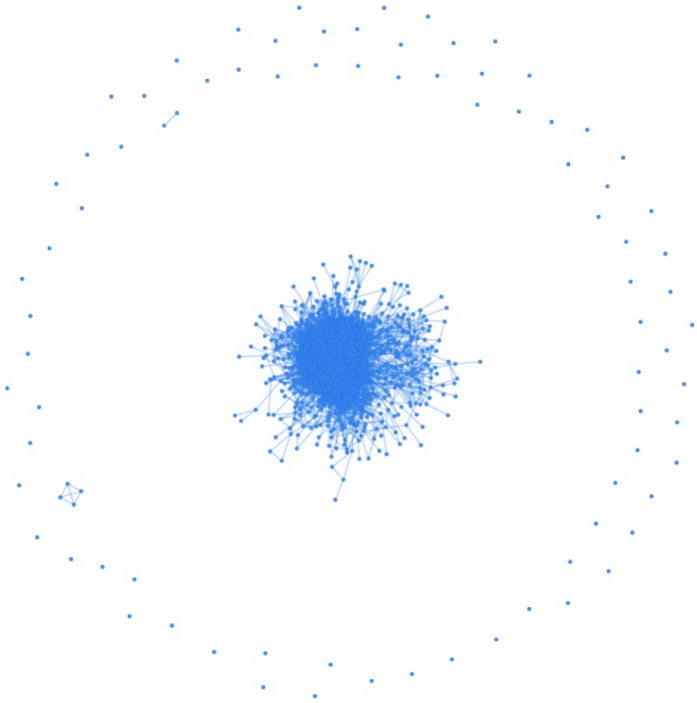
Results

After performing the algorithm for transformed network, the user will see a printed message:

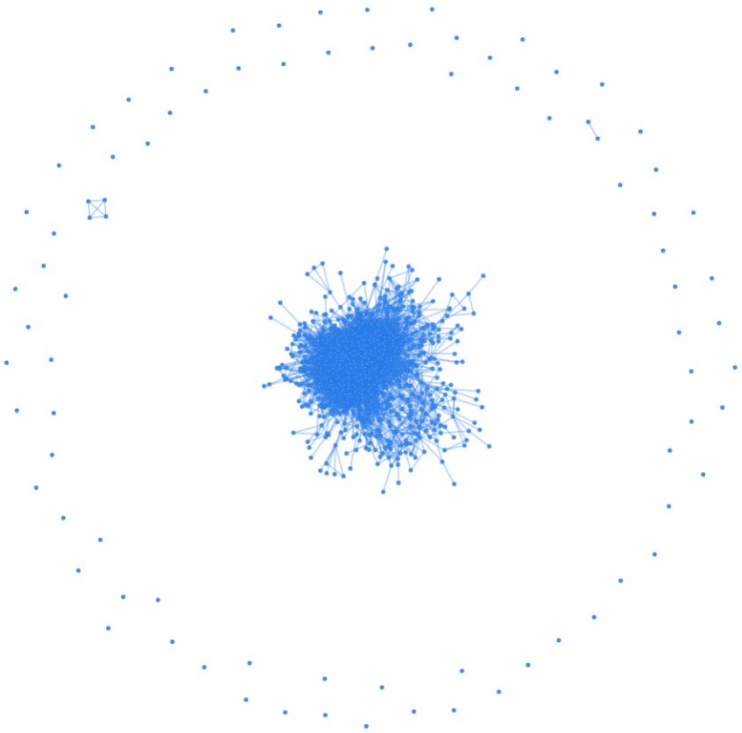
```
Network has been transformed with a maximum degree of: 50
```

The below contain the visualized comparison between Collaboration Network and Transformed Network with maximum degree of 50

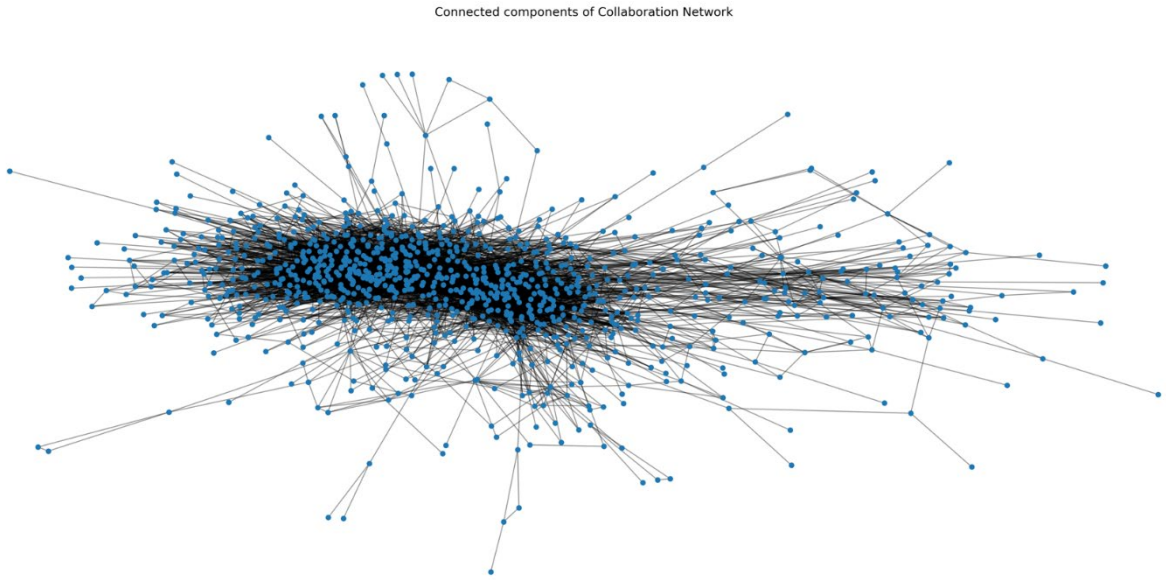
Collaboration Network:

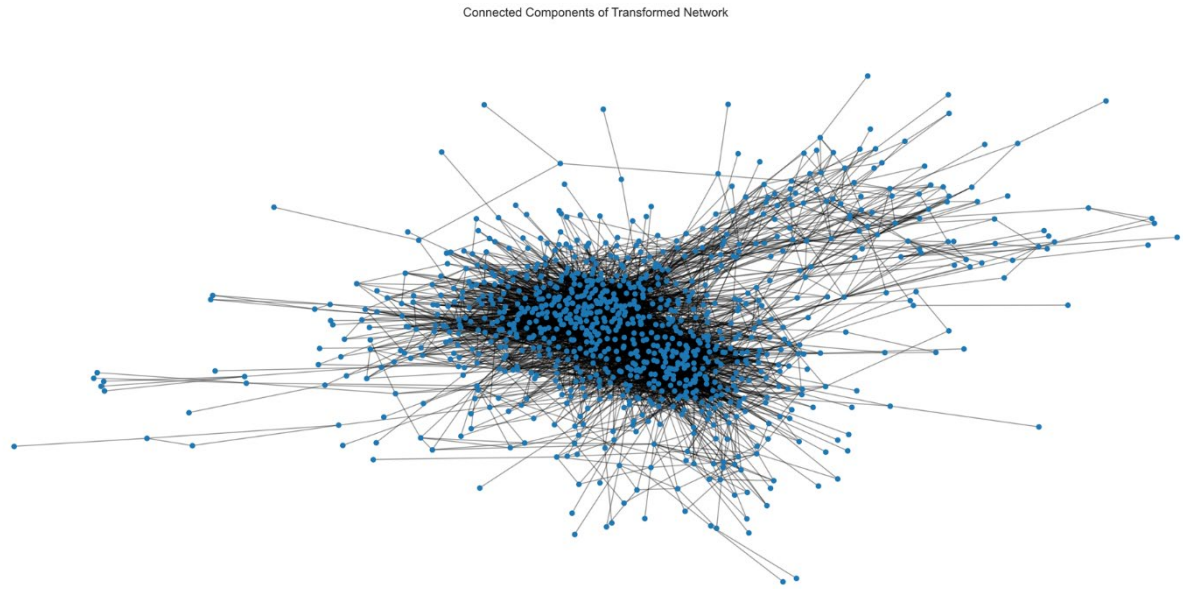


Transformed Network:



The below contain the visualized comparison between size of giant component in collaboration network and transformed network





For better clarity on the reduction in size of the Giant Connected Component (GCC), information on GCC's nodes and edges are printed as well:

Collaboration Network's GCC Information

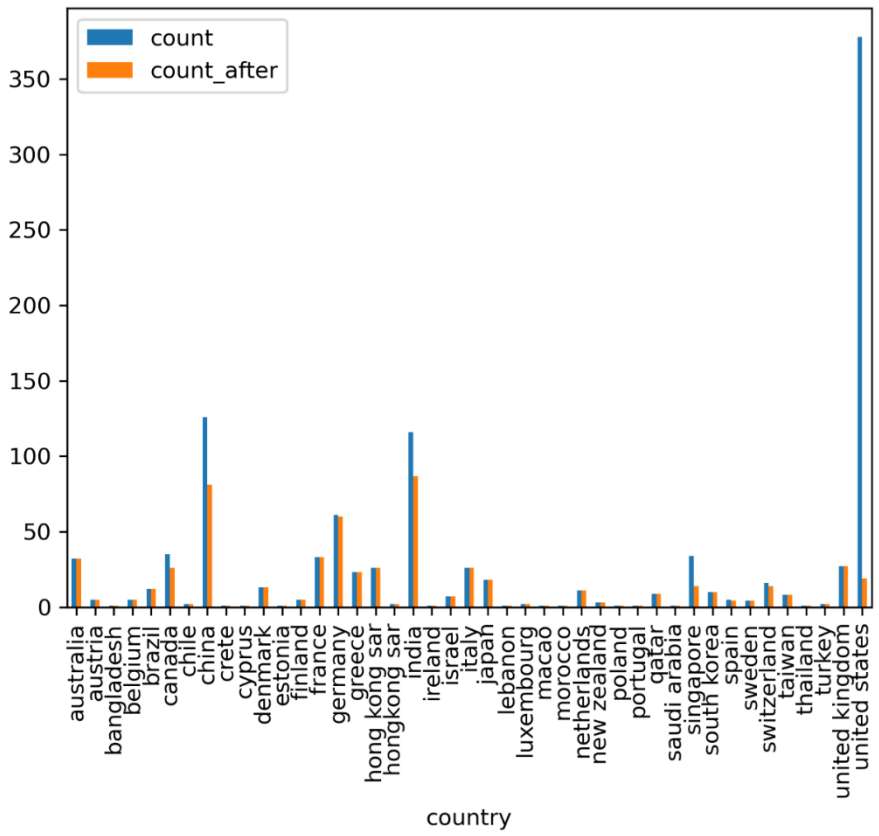
```
Info of Gcc
Graph with 983 nodes and 6448 edges
```

Transformed Network's GCC Information

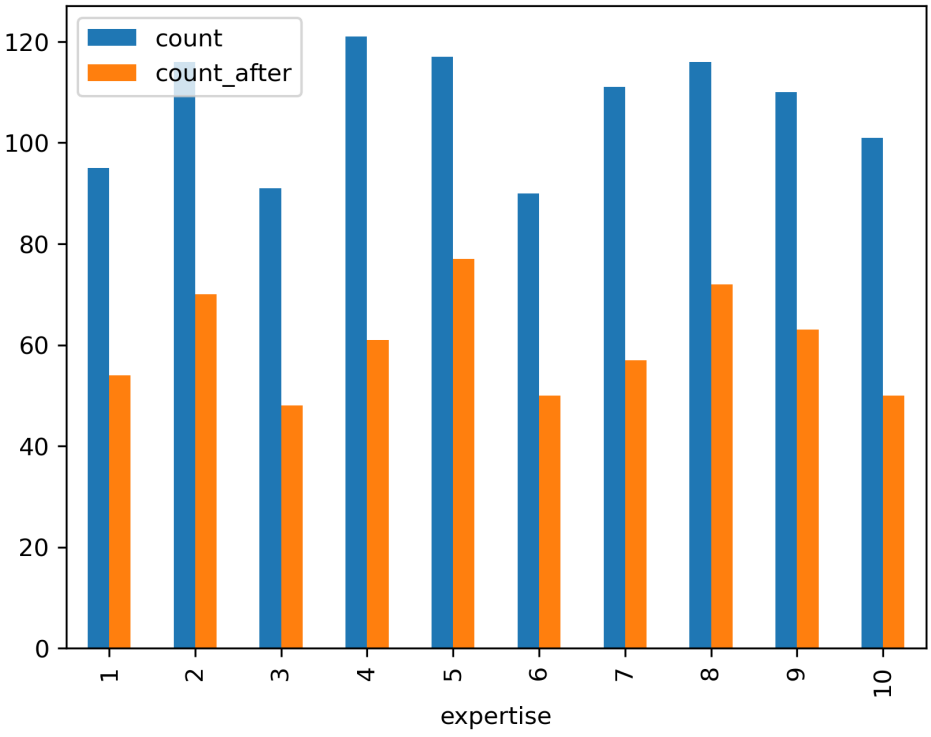
```
Info of Gcc
Graph with 499 nodes and 1988 edges
```

To display the preservation of diversity, a comparison will be done on the before and after comparison between the Collaboration Network and Transformed Network's countries, expertise, and institutions.

Countries before and after comparison:

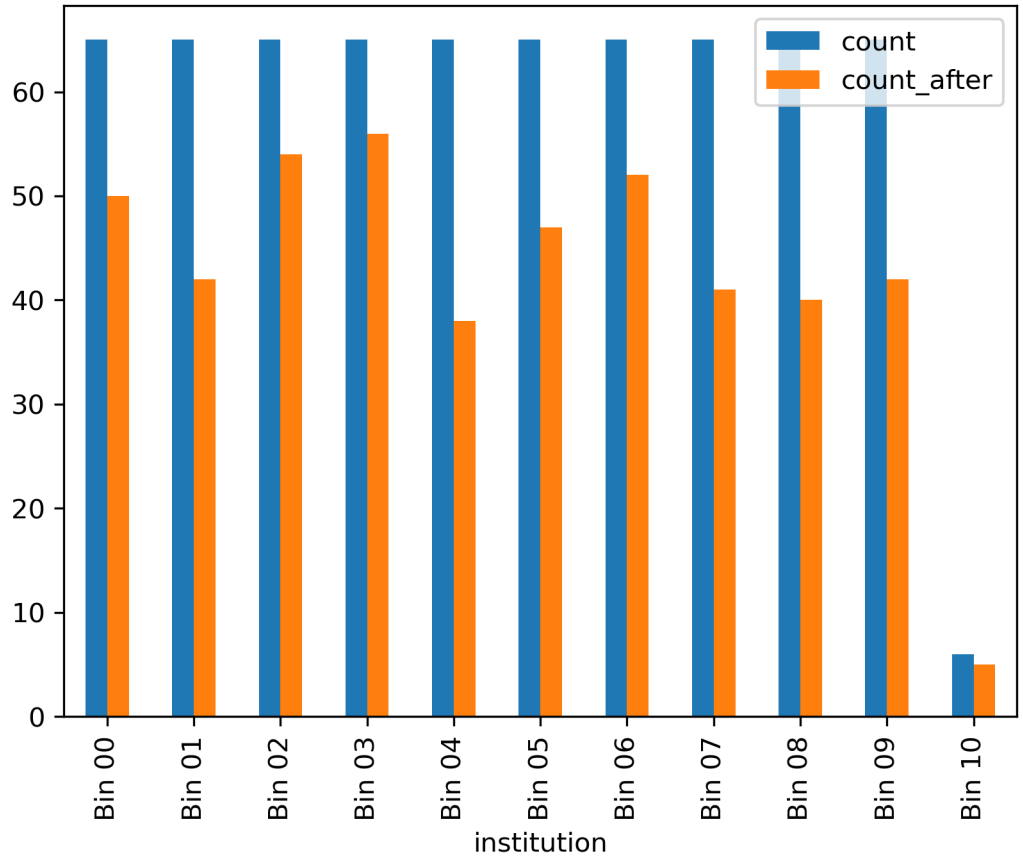


Expertise before and after comparison:



As there are too many different institutions amongst the nodes, our team has sorted the institutions alphabetically and grouped them 10 of them together to compare the differences between each bins to see the difference before and after the transformation

Institutions before and after comparison:



4. Limitations

There are three people with no institution information in the input file.

They are:

1. Chen wang
2. Niketan pansare
3. Sambudda roy

There is no straightforward way to assign them institutions, so we left them as NaN.

In the input file, the naming of countries is inconsistent. For example, there are individuals with country value “hongkong sar” as well as “hong kong sar”. We could not identify a safe way to rectify such inconsistencies, and so we left them as is. This could have an impact on our transformation algorithm since both values are seen as separate entities by our algorithm and thus, our algorithm will try to preserve both as separate countries.

Our transformation algorithm outlined in part 4 focuses on ensuring that all countries, institutions, and expertise levels are preserved in the transformed network, rather than keeping the percentages of each diversity feature (expertise, country, institution) the same as the original non-transformed network. For example, referring to the diversity chart in the **Results** section of part 4, it is apparent that the proportion of nodes belonging to the country ‘united states’ has been reduced in the transformed network.

For our transformation algorithm, we removed nodes with the highest weight, where our weight is defined as:

Weight = 50 * Institution Count + 1 * Country Count + 1 * Expertise Count

For example, taking an individual belonging to an institution that appears 4 times in the input file, a country that appears 100 times in the input file, and an expertise that appears 300 times in the input file, the weight assigned to this individual will be:

Weight = 50 * 4 + 1 * 100 + 1 * 300 = 600

As can be seen from the weight definition above, an arbitrary 50x multiplier is applied to the institution frequency whereas original frequencies of country and expertise are used in the calculation of the node’s weight. A 50x multiplier was chosen for the institution frequency because of the high number of institutions represented in the input file, resulting in institution frequencies being low across the board for all institutions. To force a more equal institution representation in the calculation of the weight, a multiplier was used.

However, the multiplier approach also becomes a limitation of our algorithm for two reasons. Firstly, the multiplier used is arbitrary and non-scientific. It could very well be that a different multiplier would give better results for the transformation. Secondly, the multiplier is static. Given a different input file with higher institution frequencies, such a large multiplier could warp the weight calculation overly in favour of the institution frequencies.

Furthermore, since the weight calculation is an aggregation of the country, institution, and expertise frequencies, it is possible for anomalous values in any of the three components to influence the overall weight. For instance, an individual belonging to an institution unique to all other individuals in the input file could belong to a common country and expertise level. The country and expertise features could skew the weight calculation upwards and make the individual a target for removal, despite belonging to a unique institution. This will result in the unique institution no longer being represented in the final transformed network, leading to a loss of diversity.