NETWORK SCIENCE-BASED ANALYSIS OF COLLABORATION NETWORK OF DATA SCIENTISTS

SC4022 Project Report

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

Contributions

Questions	Contributor
Q1	Jiang Yunjun (U2120350B)
Q2	Jiang Yunjun (U2120350B)
Q3	Jameerul Kader Faizan (U2023863D)
Q4	Jiang Yunjun (U2120350B)

The brief description of the project files can be found in *readme.md* or Appendix.

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.

Summary statistics of the collaboration network:

{
'number of nodes': 967,
'number of edges': 7585,
'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'}, 0.030742488465731663],
'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.

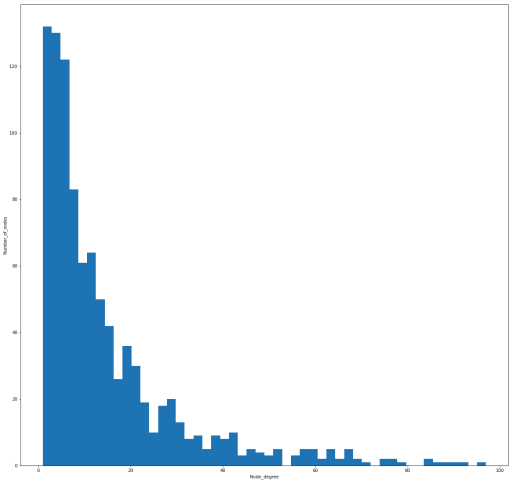


Figure 1: Degree distribution of the collaboration network.

A prominent giant component (Figure 2) 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.

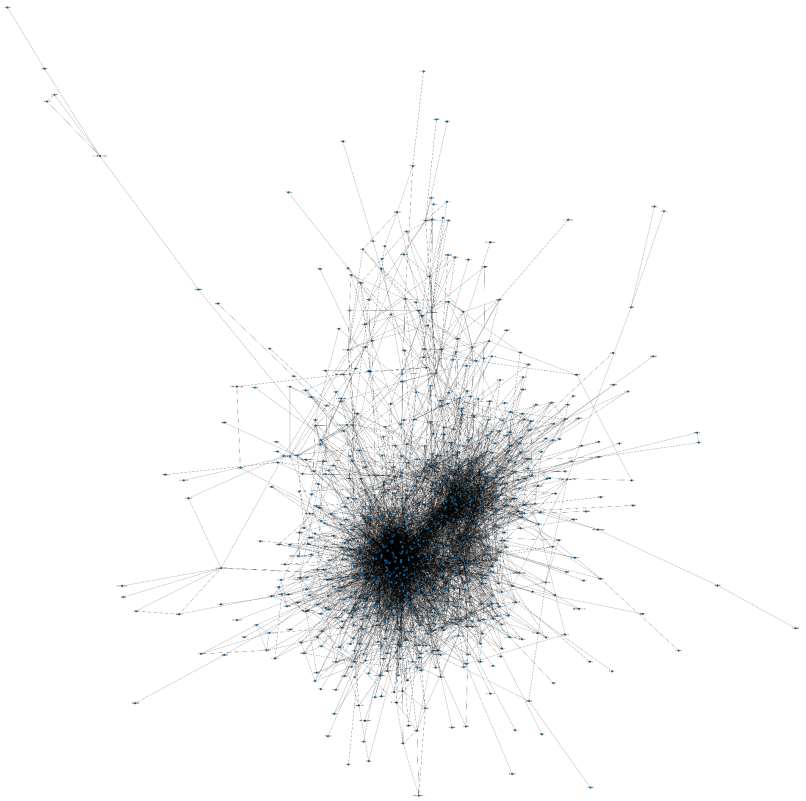


Figure 2: The giant component in the collaboration network.

Figure 3 shows a prominent clique, fully connected with n=25 nodes existing within the network. This is the most prominent publishing group of researchers.

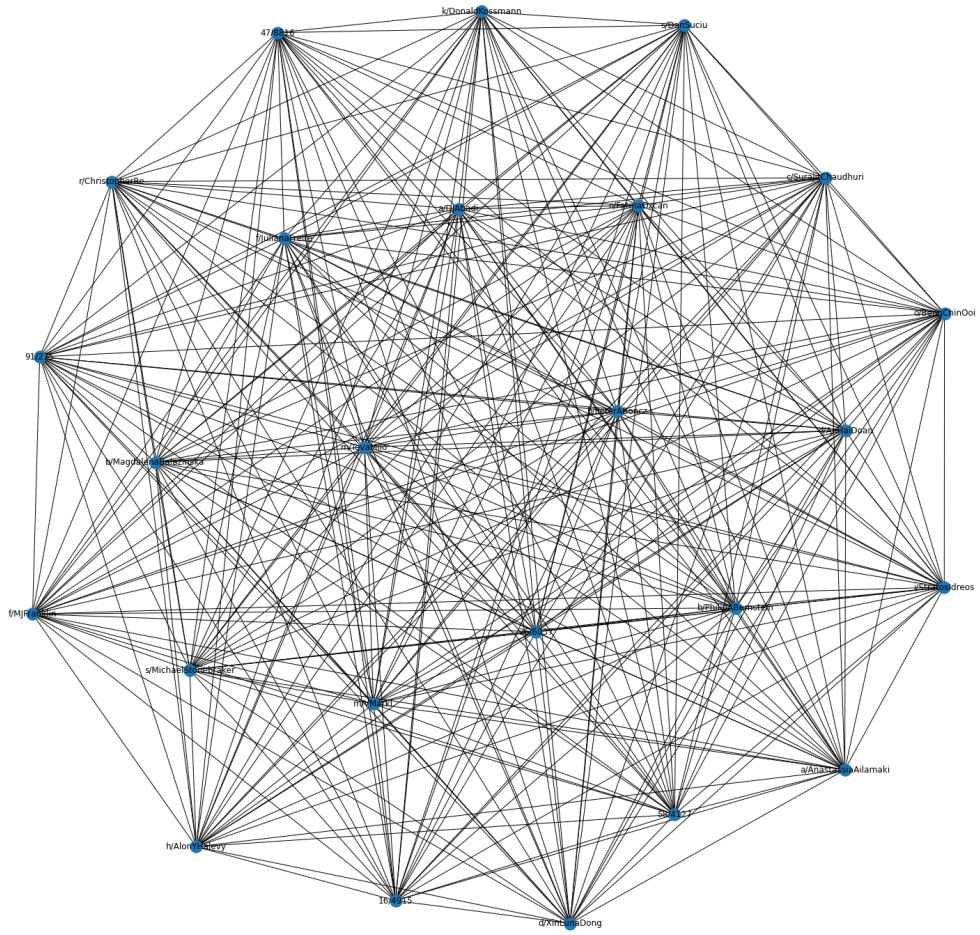


Figure 3: Largest clique in the collaboration network.

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 4 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.

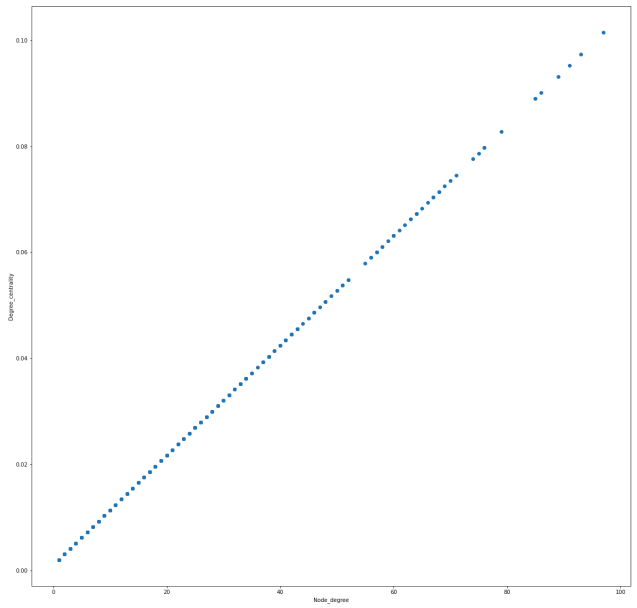


Figure 4: Scatterplot correlating degree centrality to degree distribution.

Figure 5 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.

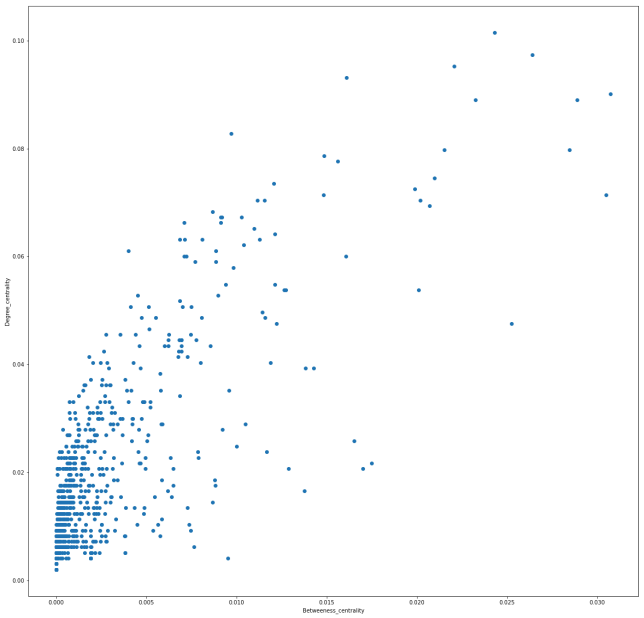


Figure 5: Scatterplot correlating degree centrality to betweenness centrality.

Question 2

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

Summary statistics for 2010:

{
'number of nodes': 50,
'number of edges': 36,
'average_node_degree': 1.44,
'max_node_degree': ['77/5283', 3],
'degree_assortativity': 1.0000000000000013,
'average_clustering_coeff': 0.06,
'giant_component_size': 3,
'giant_component_clustering_coeff': 1.0,
'graph_density': 0.029387755102040815,
'node_highest_degree_centrality': [{'g/PaoloGarza', '77/5283', '60/7439'}, 0.08163265306122448],
'node_highest_betweeness_centrality': [{'a/WGAref', 'k/MasaruKitsuregawa', 's/KaiUweSattler', '08/4367', '40/2808', 't/AnthonyKHTung.1', '96/3834', '77/5283', 'o/FatmaOzcan', '11/1466', 't/JensTeubner', 'b/IlariaBartolini', 'g/SergioGreco', '66/3005', '65/3446', 'p/PaoloPapotti', 'n/FelixNaumann' …}],
'node_highest_closeness_centrality': [{'g/PaoloGarza', '77/5283', '60/7439'}, 0.04081632653061224]
}

Summary statistics for 2024:

{
'number of nodes': 809,
'number of edges': 2349,
'average_node_degree': 5.80716934487021,
'max_node_degree': ['s/DanSuciu', 68],
'degree_assortativity': 0.5301161777877768,
'average_clustering_coeff': 0.27604599996730217,
'giant_component_size': 472,
'giant_component_clustering_coeff': 0.4390986171755948,
'graph_density': 0.00718709077335422,
'node_highest_degree_centrality': [{'s/DanSuciu'}, 0.0853960396039604],
'node_highest_betweeness_centrality': [{'07/1181'}, 0.047807288000833933],
'node_highest_closeness_centrality': [{'s/DanSuciu'}, 0.19307713999248025]
}

Visualisation of the collaboration networks reflects a clear difference between 2010 and 2024 (Figure 6).

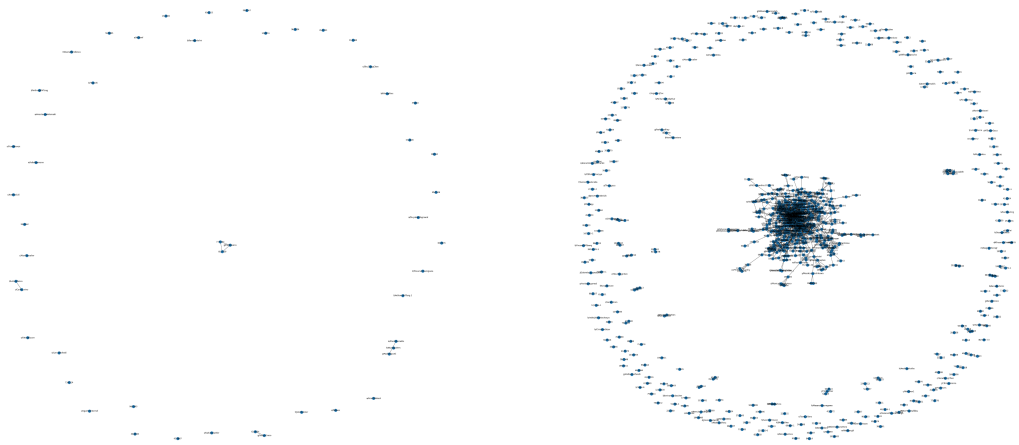


Figure 6: Graph visualisation of collaboration network in 2010 (left) and 2024 (right)

As depicted in the trend graphs, the drastic increase in number of nodes and edges (Figure 7), giant component size (Figure 8) and average node degree (Figure 9) all indicate an increase in overall publishing activity. Specifically, a sharp increase in activity is observed after 2015.

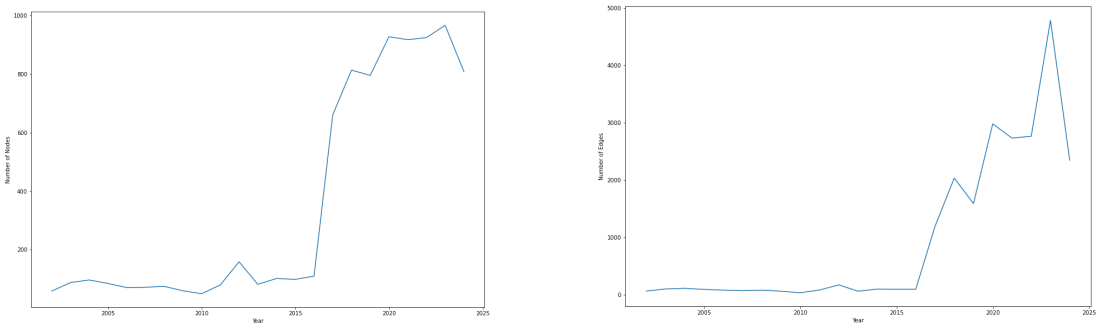


Figure 7: Number of Publishing researchers (left) and number of collaborations (right) over time.

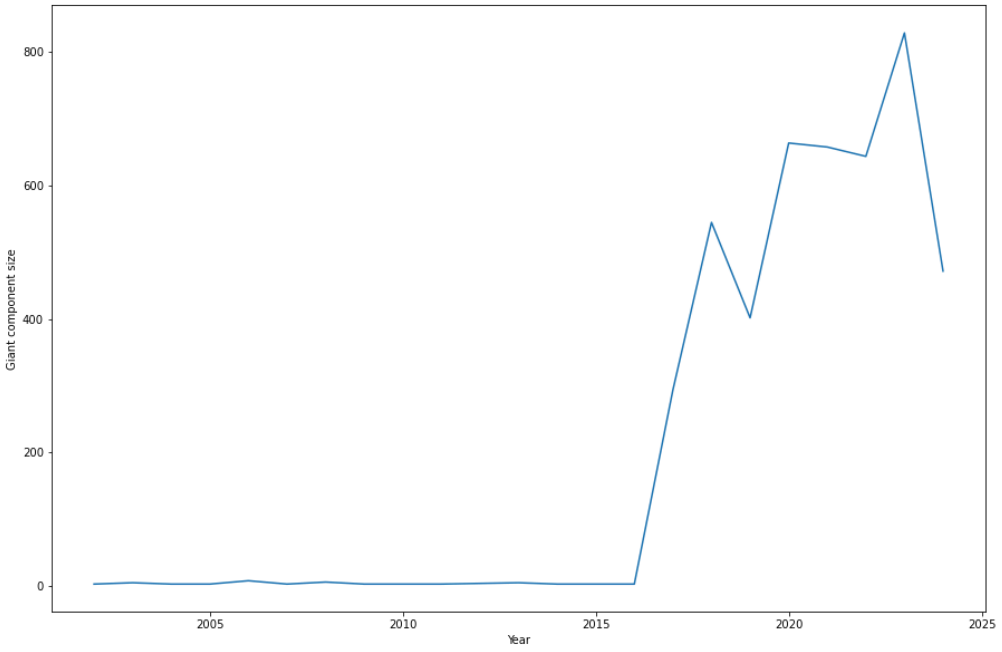


Figure 8: Size of the giant component over time.

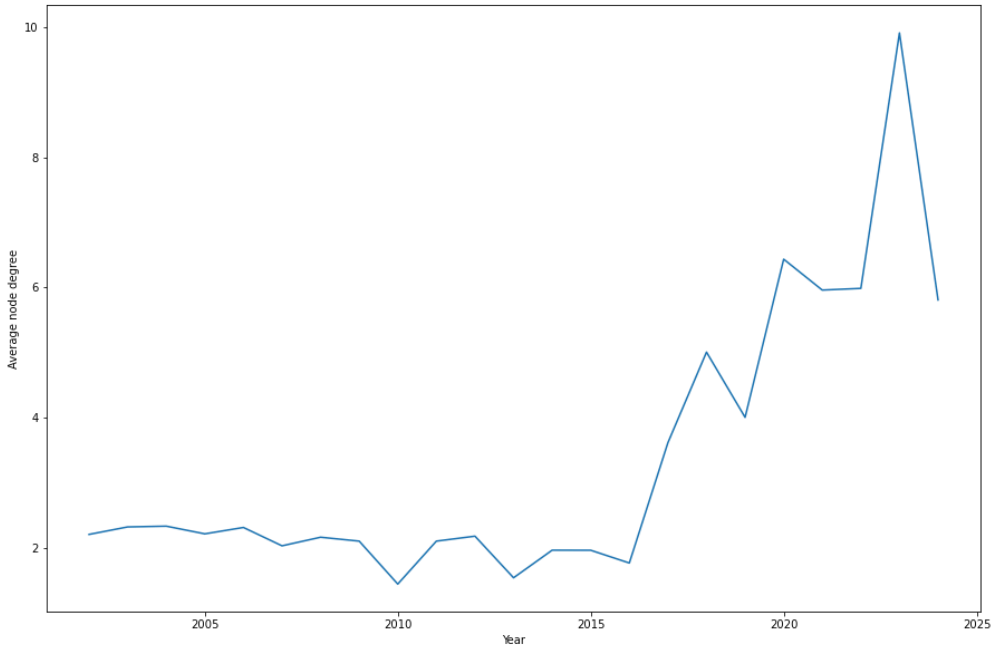


Figure 9: Average node degree over time.

Over time, the collaboration network increasingly resembles a power law distribution. For example, see the comparison of degree distribution from 2015 to 2024 (Figure 10).

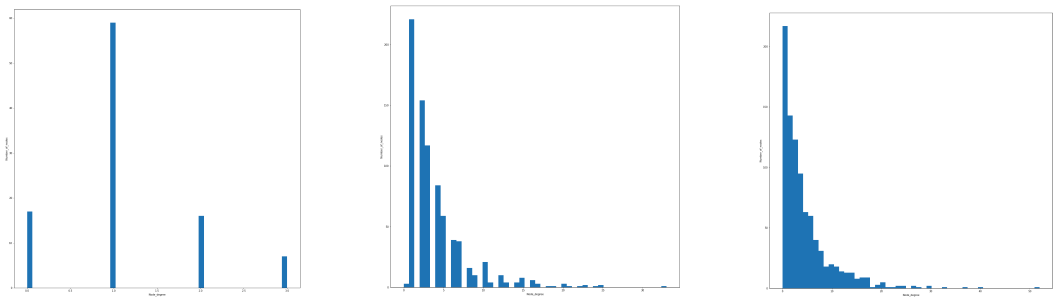


Figure 10: Degree distribution of the collaboration network in 2015 (left), 2018 (middle), 2024 (right).

Additionally, an increase in clustering coefficient is observed (Figure 11), indicating an increase in tendency for researchers to cluster. While this can be partially explained by the increase in size of the giant component, it is also possible that individual publications are being co-authored by a larger number of researchers.

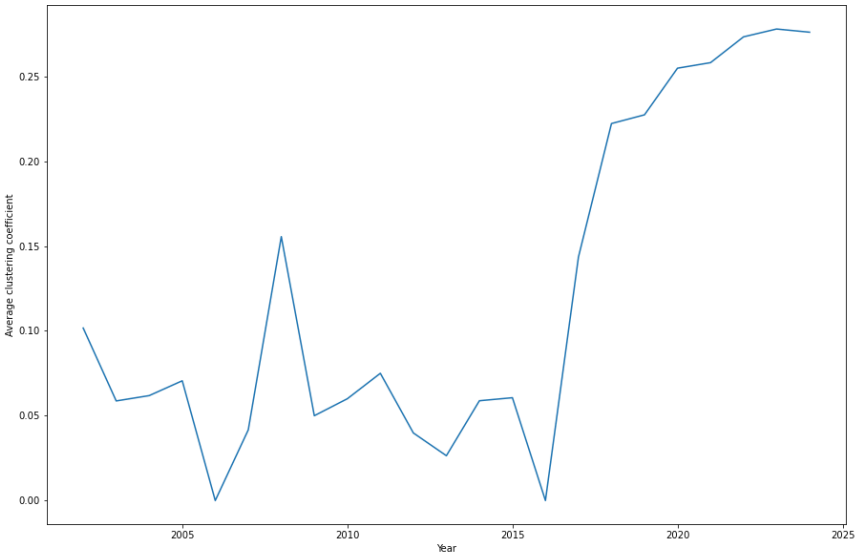


Figure 11: Average clustering coefficient of the collaboration over time.

However, the collaboration network has decreased in both density (Figure 12) and degree assortativity (Figure 13). This indicates that although more publications are being published, they tend to be written by new and small clusters of researchers.

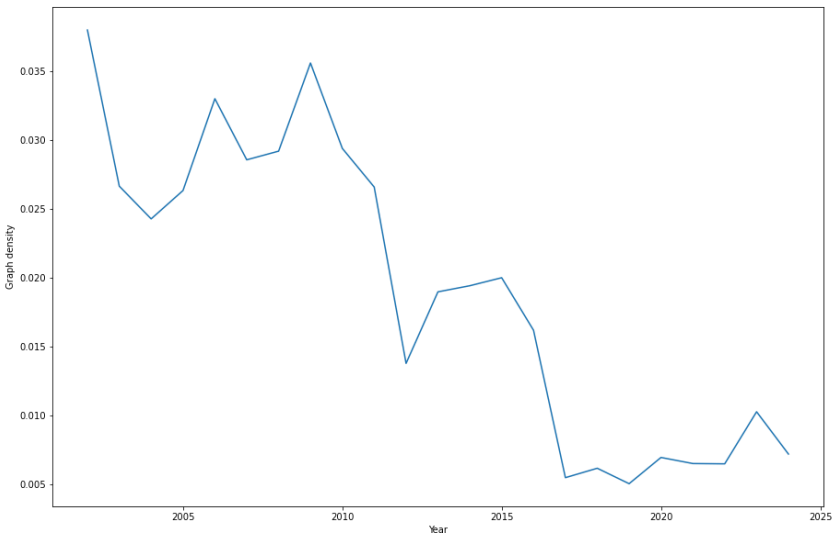


Figure 12: Density of the collaboration network over time.

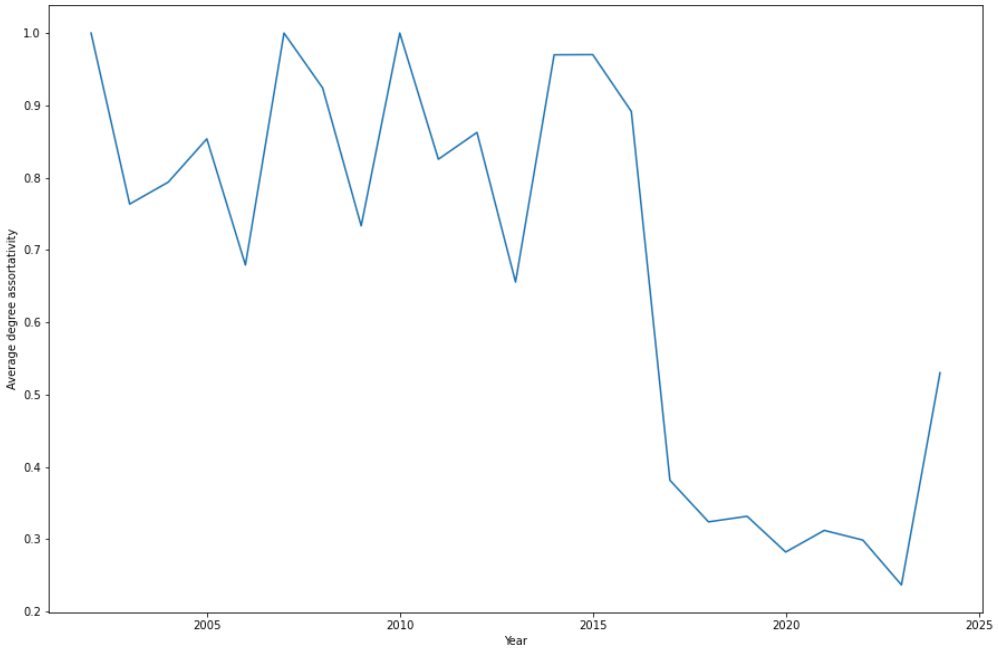


Figure 13: Degree assortativity of the collaboration network over time.

Figure 14 shows the most prominent cliques in 2010 (left) and 2024 (right). There was an increase from n=3 to n=25 nodes within the clique, 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.

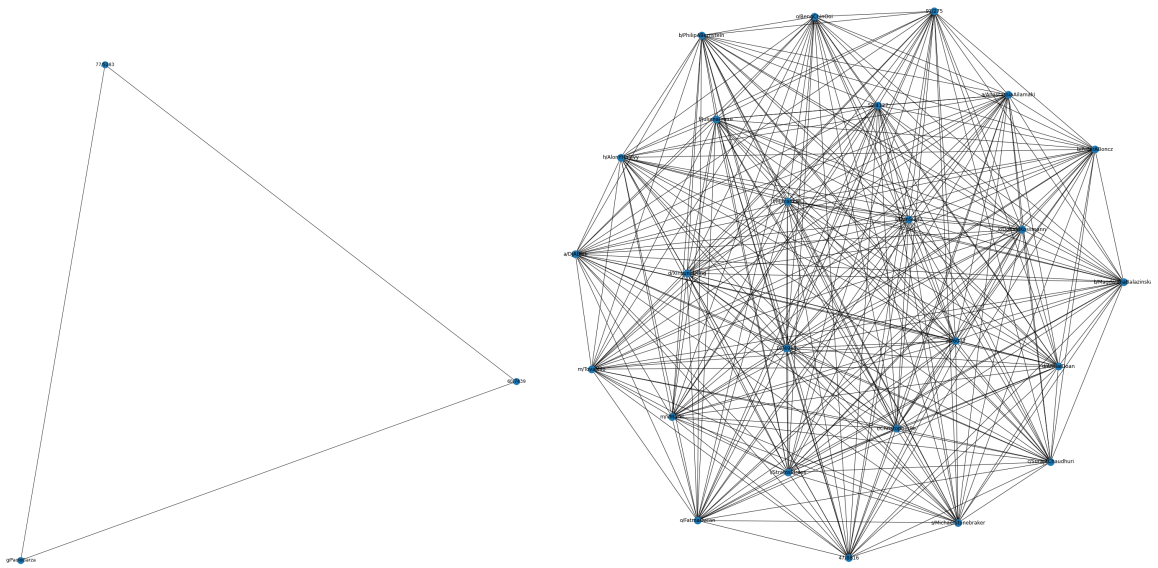


Figure 14: Most prominent cliques in 2010 (left) and 2024 (right).

The non-linear positive correlation in 2010 becomes an almost perfectly linear, positive correlation (Figure 15). Therefore, the researcher nodes with higher degrees increasingly emerge as being more central as time passes on.

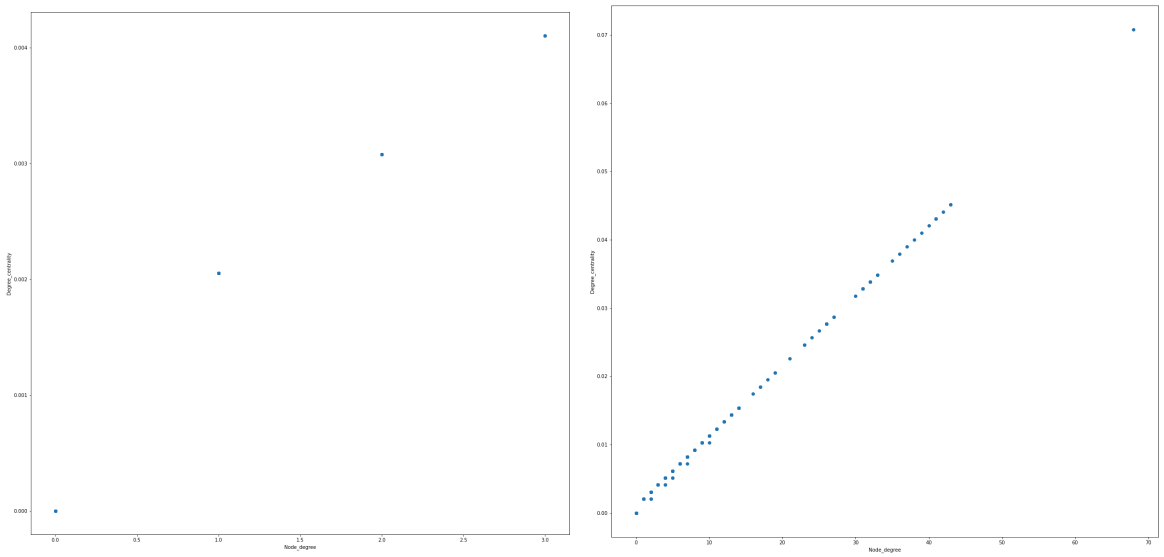


Figure 15: Scatterplot correlating degree centrality to degree distribution from 2010 (left) to 2024 (right).

Figure 16 shows the changes in the scatterplot from 2015 to 2020 correlating degree centrality and betweenness centrality. There is an emergence in positive correlation. There is also 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.

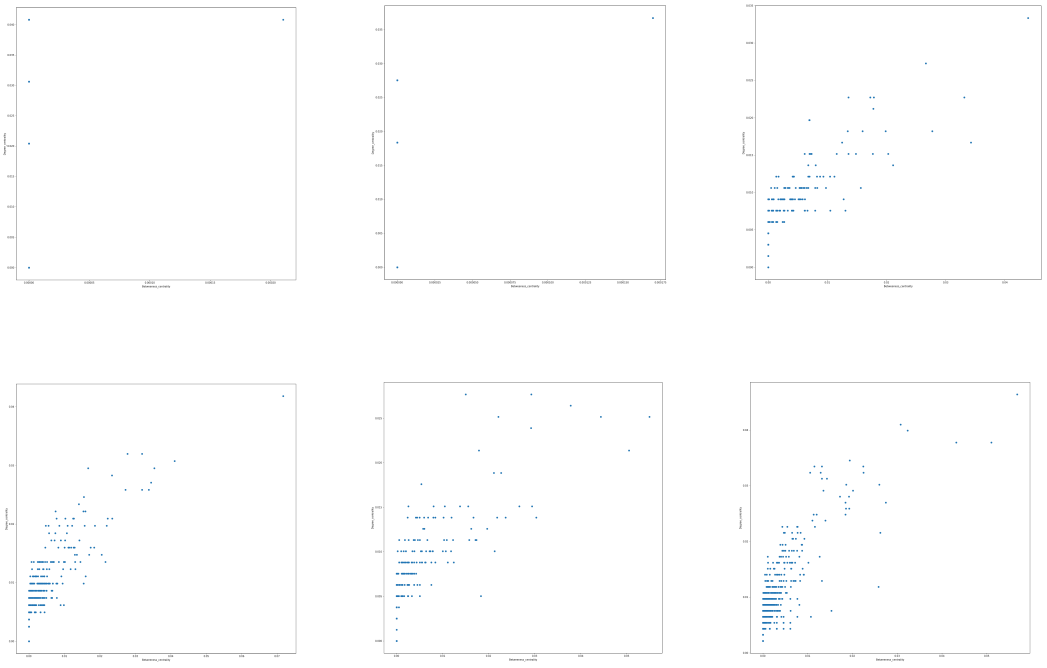


Figure 16: Scatterplots from 2015 (top left) to 2020 (bottom right) correlating degree centrality to betweenness centrality.

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)?

If we are to create a Random Network from the set of individuals in the input file, the following properties will be different from the real collaboration network in (1);

1. Number of Nodes
   - This indicates the total number of nodes (or vertices) in the network. In the collaboration network, there are 967 authors, while in the random network, there are 1072 nodes, which could represent any arbitrary entity.

2. Number of Edges
   - This represents the total number of edges (or connections) between nodes in the network. The collaboration network has 7585 edges, reflecting actual collaborations between authors. In contrast, the random network has 448234 edges, which are generated randomly based on a specified probability distribution.

3. Graph Density
   - Graph density measures the proportion of actual connections to potential connections in the network. A value closer to 1 indicates a denser network. The collaboration network has a low graph density (0.0162), reflecting sparsity in collaborations. The random network has a much higher graph density (0.7808), indicating a densely connected structure.

4. Degree Density
   - Degree density represents the average degree (number of connections) of nodes in the network. In the collaboration network, each author collaborates with an average of 15.69 other authors. In contrast, in the random network, each node has an average degree of 836.26, indicating a highly connected network.

5. Clustering Coefficient
   - The clustering coefficient measures the extent to which nodes in the network tend to cluster together. In the collaboration network, the clustering coefficient is 0.2949, indicating some tendency for authors to form clusters of collaborations. In the random network, the clustering coefficient is much higher (0.7851), indicating a highly clustered structure.

6. Size of Giant Component
   - The giant component is the largest connected subgraph in the network. In the collaboration network, the giant component comprises 895 authors, representing a significant portion of the network. In contrast, in the random network, the entire network forms a giant component, with all 1072 nodes connected to each other.

Question 4

Algorithm to transform the Collaboration Network:

To realise the algorithm for question 4, we implement a check on the coauthorship background, node degree as well as the degree of neighbour nodes.

The steps to graph generation is as follows:

1. Load all author nodes into the graph.

```
614
615     G = nx.Graph()
616     G.add_nodes_from(collab_network.index)
617
```

2. For each author node, we perform the following:
    2.1. Check for each coauthor (neighbour node) attributes:
        2.1.1. Country;
        2.1.2. Institution;
        2.1.3. Expertise (randomly assigned value in [1, 10])

```
for coauthor in coauthors:
    if coauthor in G.nodes():  #keep track of coauthorship stats

        country = collab_network.loc[coauthor, "country"]
        countries.append(country)

        institute = collab_network.loc[coauthor, "institute"]
        institutions.append(institute)

        expert = collab_network.loc[coauthor, "expertise"]
        expertise.append(expert)
```

    2.2. Based on the list of countries, institutions and expertise of all coauthors, we compute the weights of each class within the attributes (e.g. the weight of “Germany” in Countries; the weight of “CMU” in Institutions); and

    2.3. Normalise the weights to obtain an index for each attribute.

```
unique_countries = list(np.unique(countries))
class_weights = compute_class_weight(class_weight = "balanced", classes=unique_countries, y = countries)
countries_prob_distribution = dict(zip(unique_countries, class_weights / np.sum(class_weights)))

unique_institutions = list(np.unique(institutions))
#use a list to avoid key clashing
class_weights = compute_class_weight(class_weight = "balanced", classes=unique_institutions, y = institutions)
institutions_prob_distribution = np.array([unique_institutions, class_weights / np.sum(class_weights)])

unique_expertise = list(np.unique(expertise))
class_weights = compute_class_weight(class_weight = "balanced", classes=unique_expertise, y = expertise)
expertise_prob_distribution = dict(zip(unique_expertise, class_weights / np.sum(class_weights)))
```

2.4. The coauthor_index of a coauthor is the product of the three indices.

```
coauthor_index_list=[]

for coauthor in coauthors:
    if coauthor in G.nodes():
        coauthor_country_index = countries_prob_distribution[collab_network.loc[coauthor, "country"]]

        coauthor_institute = collab_network.loc[coauthor, "institute"]
        coauthor_institute_index = get_index(institutions_prob_distribution, coauthor_institute)

        coauthor_expertise_index = expertise_prob_distribution[collab_network.loc[coauthor, "expertise"]]

        #print(type(coauthor_country_index), type(coauthor_institute_index), type(coauthor_expertise_index))
        coauthor_index = coauthor_country_index * coauthor_institute_index * coauthor_expertise_index
        coauthor_index_list.append([coauthor, coauthor_index])
```

2.5. We sort the list of authors in ascending order by coauthor_index; therefore prioritising diversity in all three attributes.

```
coauthor_index_list = sort_list(coauthor_index_list)
#print(coauthor_index_list)
```

2.6. We then enforce Kmax by slicing the coauthor list if it is longer than Kmax.

```
if len(coauthors) > degree_threshold: #limit node degree
    coauthors = coauthors[:degree_threshold]
```

3. Finally, we load the transformed coauthor list as neighbours of the author. On loading, we check the node degree of the coauthor; and skip the coauthor if it is already greater than Kmax.

```
for coauthor in coauthors:
    if coauthor in G.nodes() and G.degree[coauthor] < degree_threshold:
        G.add_edge(author, coauthor)
```

Brief analysis of the transformed network


```
{
        'number of nodes': 967,
         'number of edges': 664,
        'average_node_degree': 1.3733195449844882,
        'max_node_degree': ['117/3757', 10],
        'degree_assortativity': -0.18875322947517165,
        'average_clustering_coeff': 0.02588352128165986,
        'giant_component_clustering_coeff': 0.058676720896172256,
        'giant_component_size': 401,
        'graph_density': 0.0014216558436692423,
        'node_highest_degree_centrality': [{'117/3757'}, 0.010351966873706004],
        'node_highest_betweeness_centrality': [{'86/9726'}, 0.02512880380573654],
        'node_highest_closeness_centrality': [{'86/9726'}, 0.08390651974635059]
}
```


Based on the summary statistics, we can observe the following:

1. The average node degree of the transformed network is much smaller than the original network.
2. The giant component size is much smaller than the original network.
3. There is a higher number of isolated nodes.

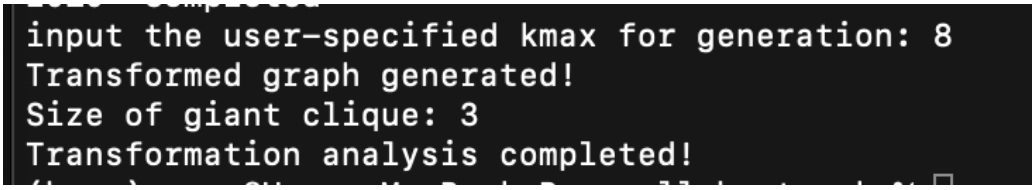
Appendix: Brief description of code

The program code is located in *project.py*.

The program requires the file *DataScientists.xls* to be placed in the *Inputs* folder before running.

To run the program, simply run python project.py.

For question 4, the program requests for an input to Kmax. Simply type the value of Kmax node degree.



The program generates two folders, *output* and *Results*.

The *output* folder contains three items:

1. *collab_network_csv.csv* contains the input data required to perform analysis for Question 1. It is the original data crawled from DBLP.
2. *problem_list_csv.csv* contains the list of authors that were unable to be crawled from DBLP and thus disregarded for the analysis.
3. *year_granularity_df.csv* contains the input data required to perform analysis for Question 2. It is the original data in yearly granularity crawled from DBLP.

The *Results* folder contains 4 groups of items:

1. q1_ prefix indicates the file is relevant to answering Question 1.
2. q2_ prefix indicates the file is relevant to answering Question 2.
3. Folders labelled by years (e.g. *2024*, *2004*) indicate it is the analysis of the DBLP collaboration network of that year.
4. *Transformed* contains the analysis of the transformed collaboration network as requested in Question 4.