南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY | 统计与数据科学系 Department of Statistics and Data Science

STA404 Final Project
Collaboration Network Analysis

Group Number: 10

Members:

12110633 Vincent William Hadiasali
12111925 Luo Lizhuo

Completion Date: June 16, 2024

Abstract

This paper analyzes the network property that a collaboration paper network has. The reason why the dataset was chosen is because there are plenty amount of complete subgraphs in the network which is special. Fundamental analysis includes degree distribution, clustering coefficient, comparison with degree nodes and clique. As the introduction in the website says that the dataset is undirected while the inside of the txt. file said that it is directed, we decided to choose undirected version. The paper also includes some meaningful deep explorations which are community detection by some algorithms along with their comparison, hierarchial structure, null model, simulation and prediction by using some proportional of the original dataset. It was found out that the network follows a scale-free network and rather has sophisticated properties. Community detection using algorithms like Louvain, Girvan-Newman, and Kernighan-Lin is performed, revealing diverse community structures. Hierarchical structure analysis and null model comparisons further highlight the network's characteristics. Simulations predict future changes in collaboration strength and paper publication rates. The findings indicate that the network follows a scale-free model with mixed assortative and disassortative properties and is densely populated with cliques, demonstrating its dynamic and robust nature.

1 Introduction

In this project, we want to actively explore the High Energy Physics - Phenomenology collaboration network and try to analyze its characteristics and rules. We conduct a simple analysis of the network from Basic Statistics and Structure, and conduct a more in-depth exploration of community detection, hierarchical structure, Null models and simulation to understand the information of this network.

Network Background: Arxiv HEP-PH (High Energy Physics - Phenomenology) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to High Energy Physics - Phenomenology category. If an author i co-authored a paper with author j, the graph contains a undirected edge from i to j. If the paper is co-authored by k authors this generates a completely connected (sub)graph on k nodes. This logic reasoning led to the decision of using the undirected version of the data.

2 Basic Statistics

Some fundamental informations about the network are given below:

- Number of nodes: 12,008
- Number of links: 118,521
- Hubs: node 482, 486 and 491 are selected as hubs
- Average degree: 19.740339773484344

Also, the informations of the giant component are given below:

- Average Distance: 4.67
- Diameter: 13

The network visualization of Python and Gephi are given below

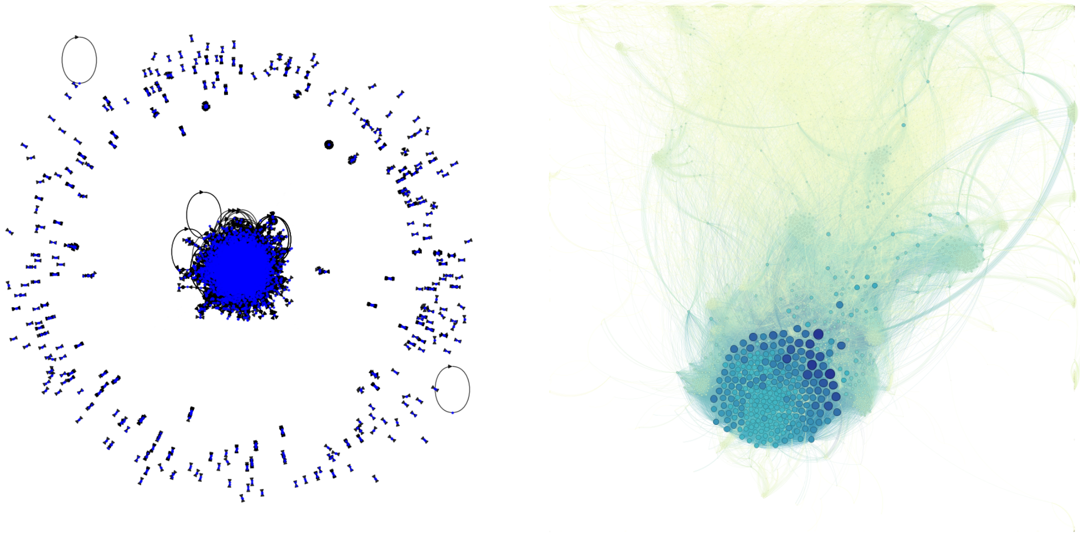

Figure 1: Network Visualizations using Python and Gephi

According to Python plotting, the network has a topology of a mesh of network beingz surrounded by a ring of some isolated nodes. In Gephi, nodes with relatively large degrees shown by a bigger size and color getting closer to blue tends to clump together which means that those are nodes in the giant component.

3 Structure

3.1 Connectedness

By definition, a graph is connected if a graph only consist of one giant component. Otherwise, we call it a disconnected graph. Since the graph consists of some splitted node groups, we have a disconnected graph. It is reported that there are 276 connected components and there is 11,204 nodes in the giant component.

3.2 Degree Distribution

The degree of all nodes in the network ranges from 2 to 491 and the average degree with the average of 19.7403397734843444. The degree distribution graphs are given below:

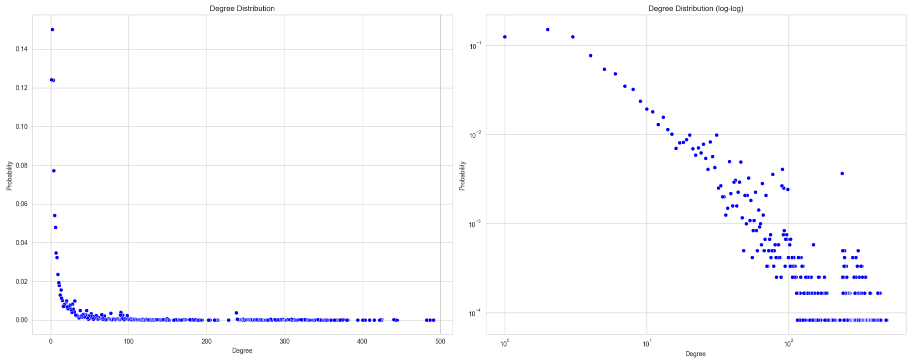

Figure 2: Degree Distribution on regular and log-log scale.

The left plot depicts that the network follows degree distribution higher degree of a node means a lower probability which is proportional to the number of such node in the network. It is also essential to also analyze the degree distribution in log-log scale for detecting a scale-free network property. As the plot on the right graph shows there is a decreasing linear pattern between the two variables, it also follows scale free network.

3.3 Hubs

Our motivation of conducting such comparison analysis is to observe its assortativity property. The average assortativity coefficient is 0.632 indicating assortative.

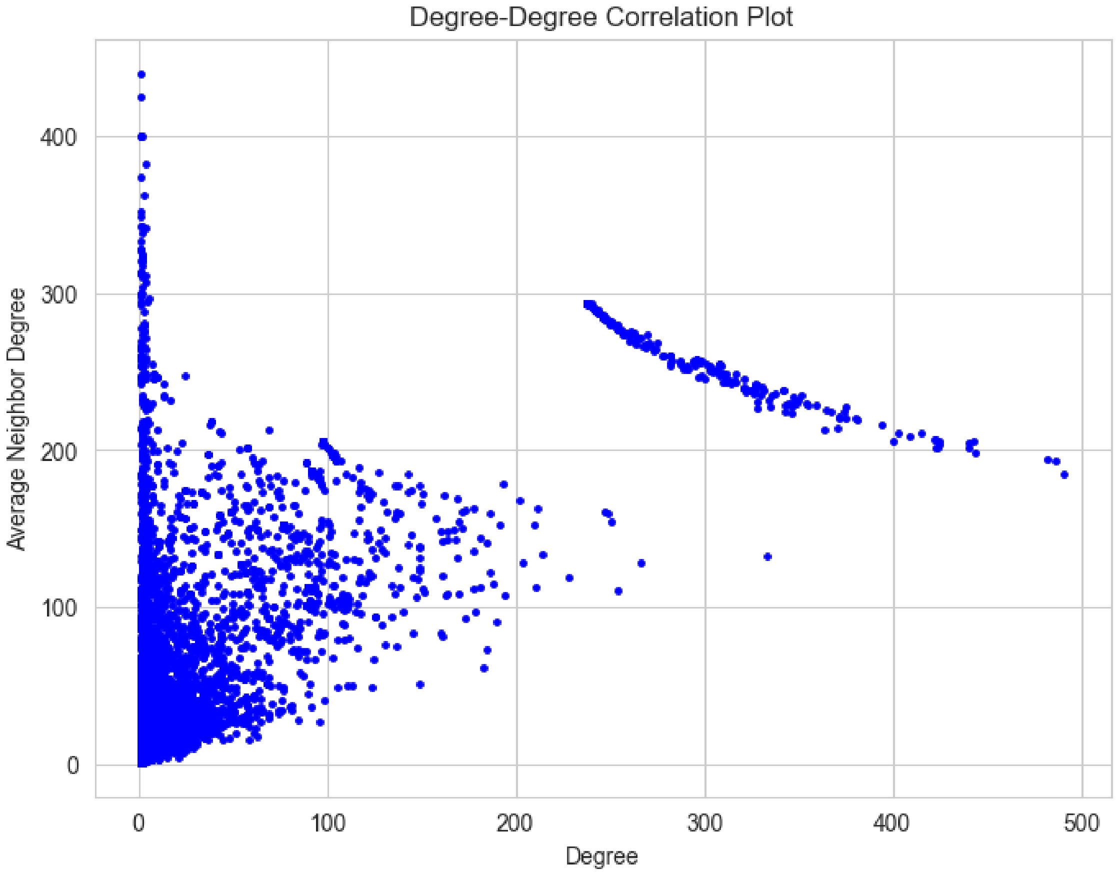

Figure 3: Degree and average neighbor degree Comparison Plot.

However, the plot shows the average neighbor degree goes smaller as the degree goes bigger. Specifically, there is a shifting trend starting from degree 250 where it is clear that almost all nodes will have a decreasing pattern in average neighbour degree as the degree node is higher. Hence, it is ‘Mixed’.

3.4 Clustering Coefficient

The clustering coefficient measures the local link density of a network: The more densely interconnected the neighborhood of node i, the higher is its local clustering coefficient. The value of clustering coefficient ranges from 0 to 1.

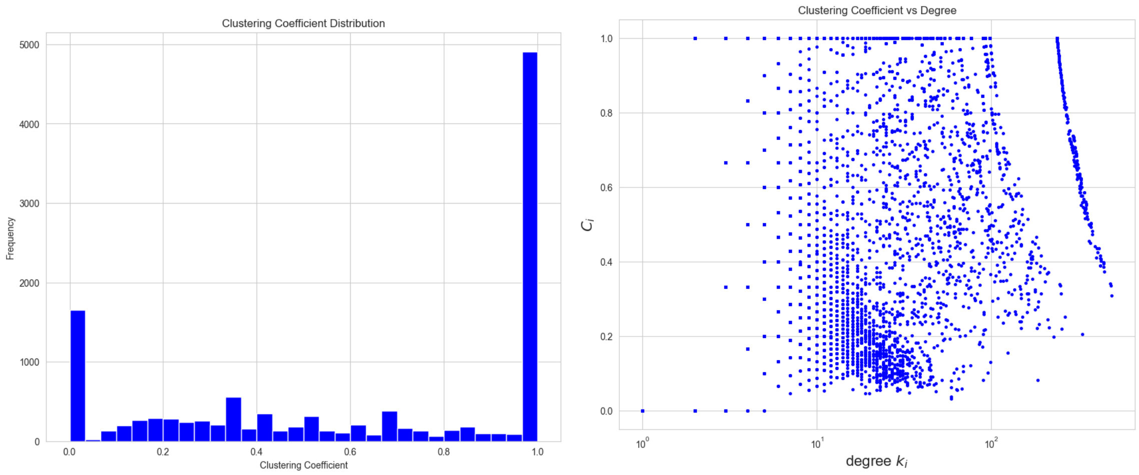

Figure 4: Clustering Coefficient distribution and its comparison with degree.

Considering the fact that k-authors in a publications implies a k-clique, the majority of nodes reaches the maximum value of clustering coefficient which is 1.0. While the second most is nodes with 0.0 clustering coefficient. We proceed on analyzing the relationship between clustering coefficient and degrees. Nodes with low degrees (on the left side of the right plot) exhibit a wide range of clustering coefficients, from 0 to 1. Nevertheless, there is an interesting pattern going on from a degree of 0 to around 100: most points at one relatively low value clustering coefficient group are clumping together which impose a potential decreasing clustering coefficient. On the other hand, nodes with high degrees ,which is partitioned by the wide curve gap in the graph from the left side, tend to have lower clustering coefficients, often clustering around lower values. Hence, high-degree nodes share the following properties:

- Hierarchical Structure: the decreasing trend in clustering coefficient with increasing degree suggests a hierarchical structure in the network. High-degree nodes serve as hubs connecting different clusters, while low-degree nodes are more likely to be part of these clusters.

- Small-World Nature: The high clustering coefficient for low-degree nodes combined with the low clustering coefficient for high-degree nodes is a common feature of small-world networks. This means that the network exhibits high local clustering while maintaining short average path lengths between nodes.

- Disassortative: The high-degree nodes connect with low-degree nodes. This is indicated by the low clustering coefficients of high-degree nodes, which are likely connecting different clusters of low-degree nodes.

3.5 Clique

As the network consists of numerous cliques, we decided to do a structure analysis on the clique. The existence of 14,939 cliques in a network means there are 14,939 publications in the network. The size of cliques ranges from 2 to around 230. Furthermore, more than a half of the publications proportions has more than 3 co-authors (8,093).

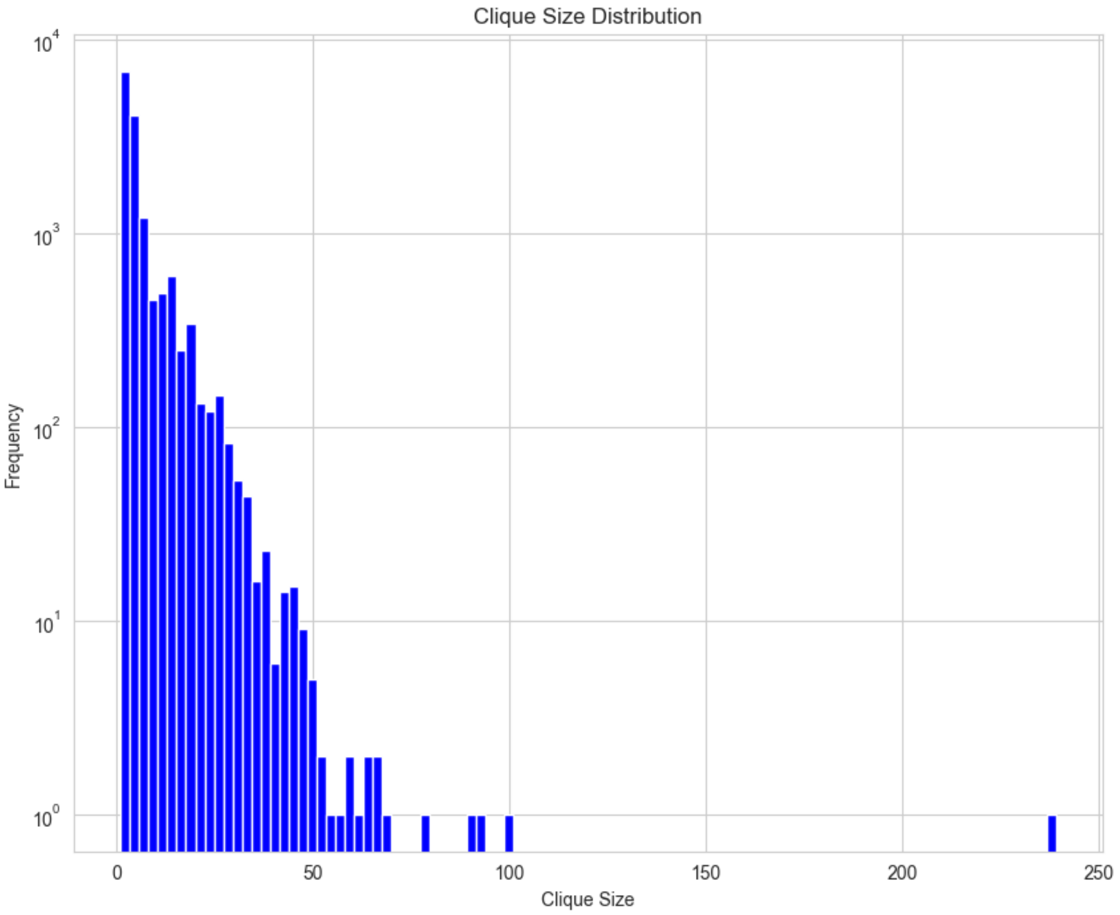

Figure 5: Clique Size distribution.

As we can see, this graph shows that even the degree distribution of clique almost follows power law distribution as the frequency generally decreases when the clique size increases. With a discovery that the average clique Size is 6.168819867461008, this means that each publication in the network will have 6 co-authors in average which is usually a big project.

4 Deep Exploration Topic

Because the original dataset is too large to conduct a more exploratory analysis, we chose a substitute - a network with similar properties but fewer nodes for the following study. It can be regarded as a compressed version of the original network. Because they have common properties, we will not introduce them in detail here.

4.1 Community Detection

4.1.1 Louvain Algorithm

The Louvain algorithm is a popular and efficient method for community detection in large networks. It works by iteratively maximizing modularity, which measures the strength of division of a network into communities. The algorithm has two main phases: first, it assigns each node to its own community and then merges communities that result in the highest gain in modularity. This process is repeated hierarchically, producing a multi-level decomposition of the network that reveals its community structure. The Louvain algorithm is known for its speed and ability to handle large networks effectively.

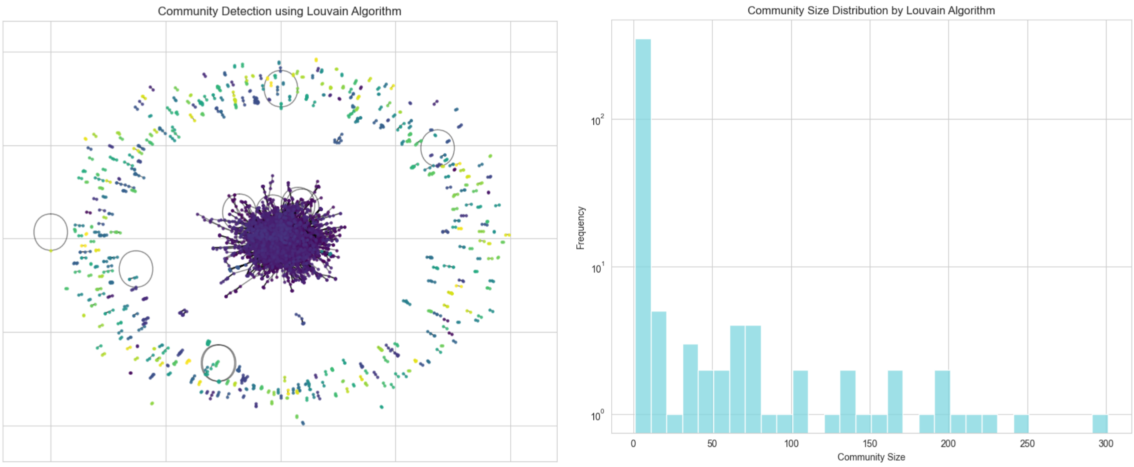

Figure 6: Network visualizations with Community Detection by Louvian Algorithm.png

From the figure, we can see that the area with densely distributed nodes in the middle is clearly divided into a community, and the surrounding nodes are scattered into some communities. From the distribution of communities, we can observe that the size of communities is widely distributed, but small communities are still the majority.

4.1.2 Girvan-Newman Algorithm

The Girvan-Newman Algorithm is a method used in network analysis to detect community structure within a network. It works by iteratively removing the edges with the highest betweenness centrality, which measures the number of shortest paths passing through an edge. This process eventually separates the network into distinct communities, revealing its underlying structure.

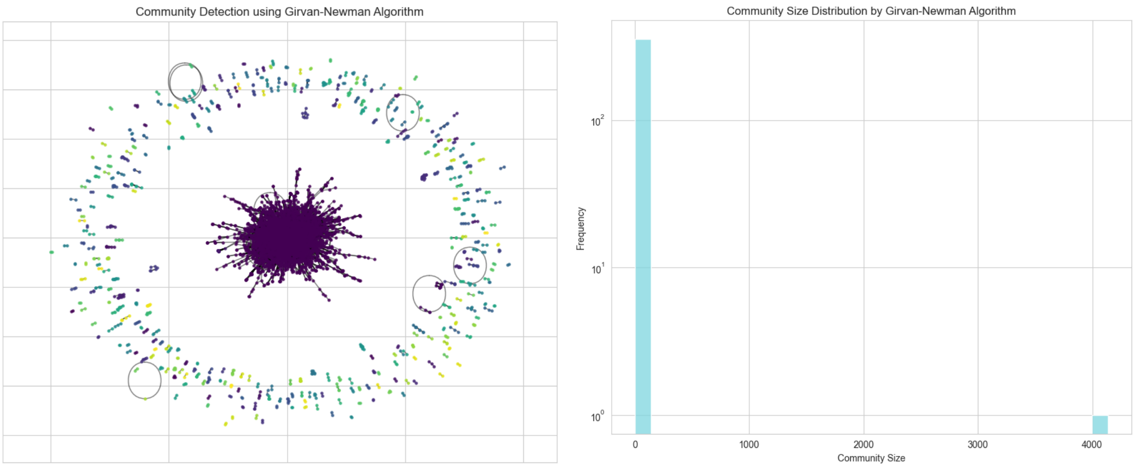

Figure 7: Network visualizations with Community Detection by Girvan-Newman Algorithm.png

If we simply look at the network visualization, we will find that the results presented by the Girvan-Newman algorithm are roughly the same as those of the Louvain algorithm. However, the results are quite extreme when looking at the distribution of community size. In the left figure, the center part is divided into a whole, while the rest of the surrounding nodes are divided into scattered communities.

4.1.3 Kerninghan-Lin Algorithm

The Kernighan-Lin Algorithm is a heuristic method used for graph partitioning. It aims to divide a graph into two equally sized subsets while minimizing the edge cut between them. The algorithm works by iteratively swapping pairs of vertices between the subsets to reduce the cut size, using a local search strategy to find an optimal partition. It is particularly useful in applications such as VLSwe design and parallel computing.

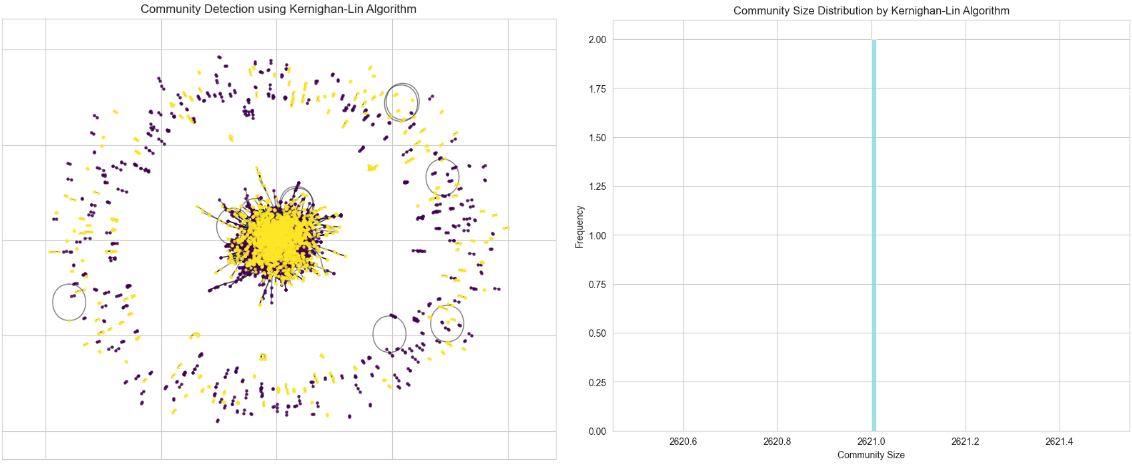

Figure 8: Network visualizations with Community Detection by Kerninghan-Lin Algorithm.

This is a community detection algorithm that tends to be binary. From the distribution of the network, we can clearly see that the network is divided into two communities. Overall, this distribution seems to be unreasonable. And from the distribution diagram on the right, we can also be surprised to find that this algorithm tends to divide the community into two even parts.

4.1.4 Modularity Comparison

The comparison on modularity among Louvain, Girvan-Newman, Kernighan Lin and Informap algorithm we haven't mentioned before is given below

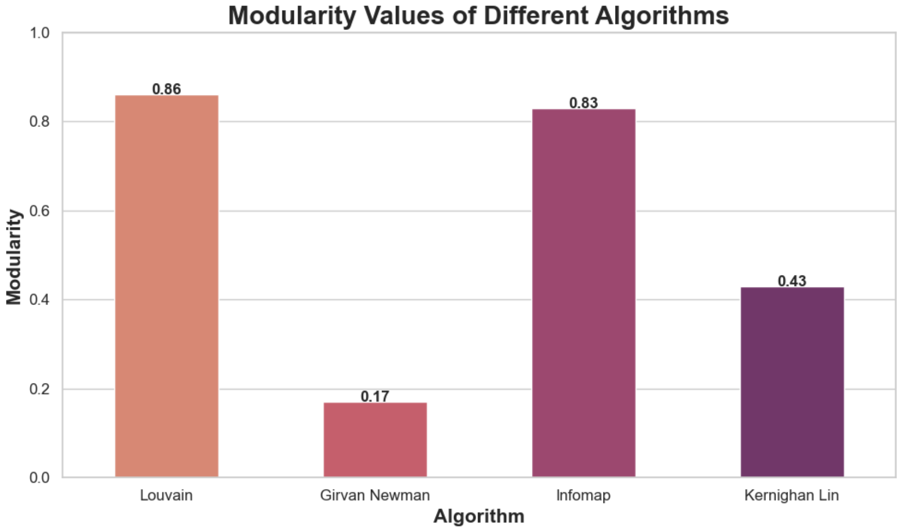

Figure 9: Network with Community Detection.

From the figure, we can clearly see that the modularity of the Louvain algorithm is the largest, and the modularity of the Girvan Newman algorithm is the smallest. In general, the Louvain algorithm is still a superior method when performing community detection.

4.2 Hierarchial Structure

Hierarchical structure in network analysis organizes a network into multiple nested levels of communities, revealing both local and global patterns. From the figure we can clearly see the hierarchical relationship in the entire network. We can find that the green and purple volumes are larger, while the orange and red volumes are relatively small.

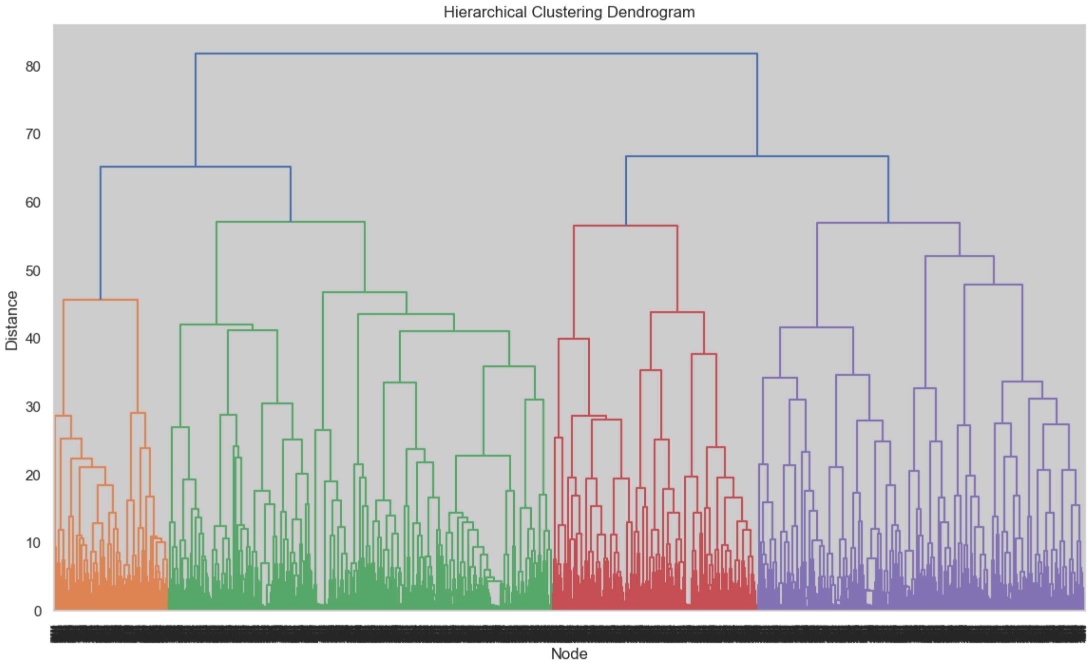

Figure 10: Hierarchial Clustering Dendrogram.

4.3 Null Model

Because there are many complete subgraphs in our original graph, which is a special property, we decided to examine the relevant properties of the null model with the same degree distribution to better infer the special properties in our network.

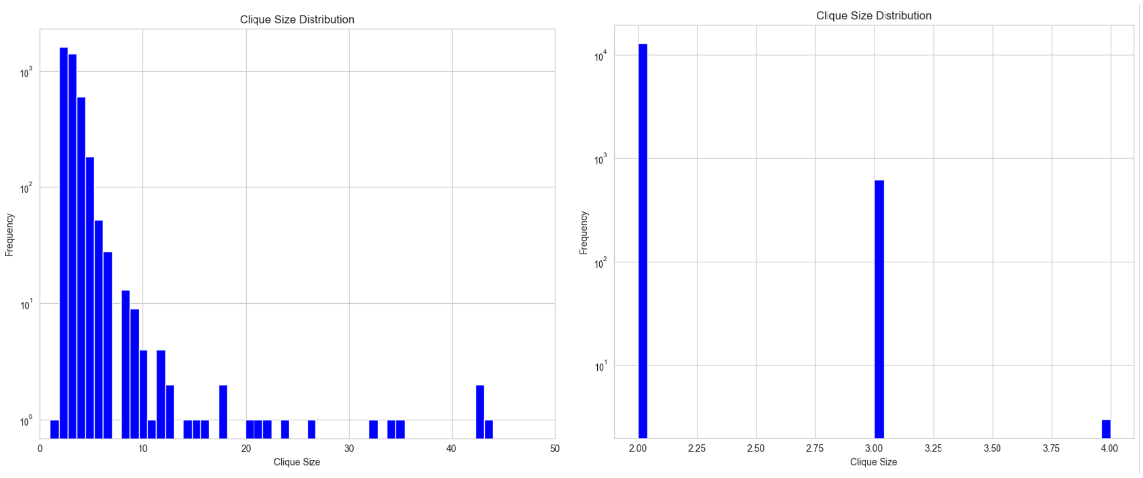

Figure 11: Clique Size Comparison

From the above figure, we can clearly see that the zero model is still very different from our original network. It is difficult to have a relatively large complete subgraph, and subgraphs with more than three nodes are very rare. This is enough to show that multiple complete subgraphs are a special property of the network we are studying.

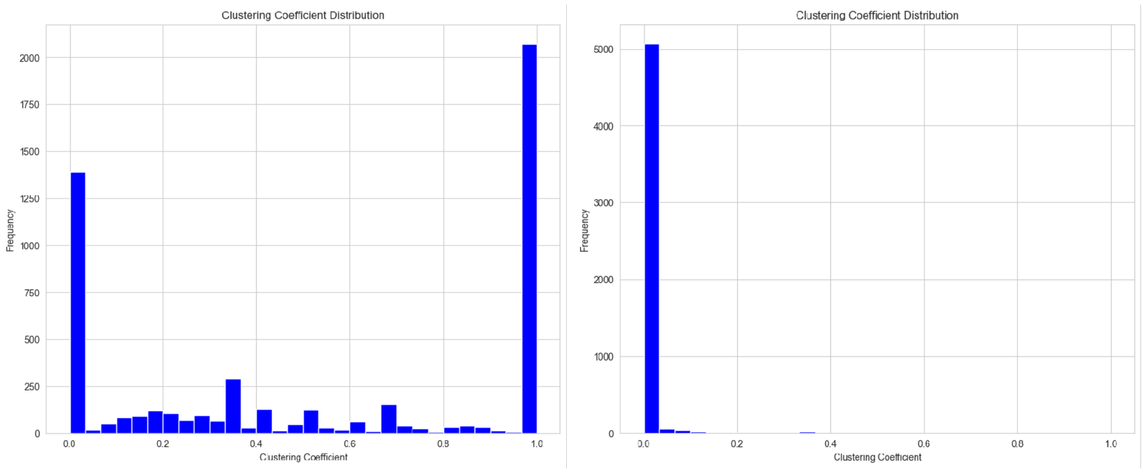

Figure 12: Clustering coefficients comparison.

Next, we compare the clustering coefficients and find that there are also huge differences. The clustering coefficients of the null model are mostly close to 0, and there are very few nodes with high clustering coefficients. This also reflects that the property of our network's multi-complete subgraphs is not due to degree distribution. The existence of special patterns in dense networks is still due to the inherent mechanisms and laws of the network.

4.4 Simulation and Prediction

Because our network is static, in order to better study the changes of our network over time, we conducted the following simulation. Because the entire network is large, we only selected some communities for simulation. The simulation process was done in the following ways:

1. Make sure there are no duplicate edges.

2. Define the author activity by the number of papers he published estimated by degree and average degree and define the connection strength by the number of collaboration between to authors initialed by 1.

3. Choose authors randomly and update their author activity based on their current activity and Chose edges randomly and update their connection strength based proportion to strength

4. Remove some nodes and edges temporarily in random, indicating any pair of coauthors will not contact for some time. And add some nodes with new edges indicating the joint of new researcher.

5. After some iterations, fit a linear model to predict the future.

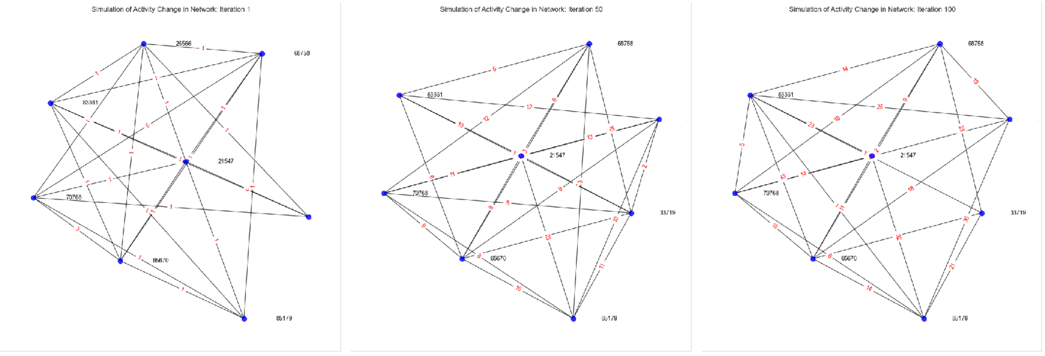
Simulation of Activity Change in Network: Iteration 1
Simulation of Activity Change in Network: Iteration 50
Simulation of Activity Change in Network: Iteration 100
26566
68758
83361
21547
70768
65670
85179
33719

Figure 13: Simulation Sample Desplay

To fit and predict node activity and edge weights in a network using linear regression, historical activity and weight records are first converted into feature matrices (X) and target matrices (y). Separate linear regression models are then trained for node activities and edge weights. These models are subsequently used to predict future node activities and edge weights based on the most recent historical data.

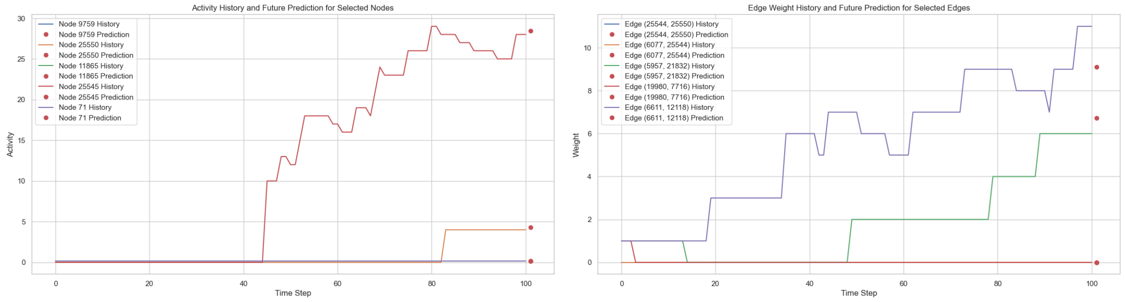

Figure 14: Activity History and Edge weight history along with their prediction.

In the two figures above, the left figure predicts the activity of the author, and the right figure predicts the strength of the connection between collaborators. From the left figure, we can see that the activity of all authors is still on the rise. However, we will find that the points with high activity will become higher and higher. We analyze that they may be scholars who are committed to research. However, there are also those whose activity does not change after increasing. This may be because they stopped after a short period of academic research. The right figure also shows an upward trend, which also well reflects that relationships in the academic community are often difficult to break. Similarly, we will also find that the weights of some edges will rise rapidly, which also shows that the academic connection between these two people is getting closer.

5 Visualization

All the visualization have been displayed in former parts.

6 Conclusion

To sum everything up, the network is nearly a scale free network. The degree distribution almost follows power law. With a pattern of nearly both assortative and disassortative, It is ‘mixed’. Due to the properties of collaboration, there are lots of cliques in the graph which leads to a dense network. There are lots of typical community in the network, and the Louvain algorithm have the best partition. The simulation part show that the collaboration strength goes larger as time goes and authors published more papers with different speed.