Social Network analysis and community detection for the physics co-authorship network

1. **Network Generation:**
   The network used in this analysis is the physics co-authorship which features 14289 papers written in physics. The network generation process featured 2 steps: data cleaning and forming three network of the co-authorship data.

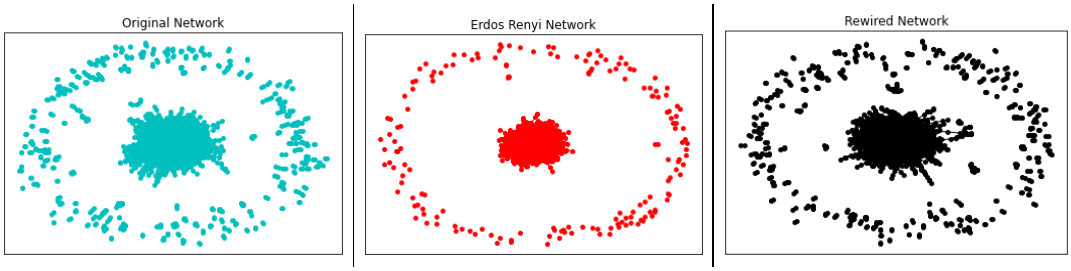
   Data Cleaning:
   The data was initially provided in a format:
   <year> , <volume> , <journal>, <author(s)>, <title>
   So the data cleaning steps were:

   1- Splitting the authors column using the & character
   2- Forming a pairs dataset that represents the edges betwen auhtors
   3- Removing single author paper by removing pairs with "Nan" or "None"
   4- Converting names into lower case to unify the author names
   5- Removing special characters from names to unify the names
   6- Removing self-edges (Existent if multiple authors share the same last name)
   7- Removing the direction of the edges
   8- agreggating multiple edges (collaborations) to calculate the weight of the edge.
   9- Removing names with empty strings

   The final original network included 5402 unique authors with 5402 weighted edges.

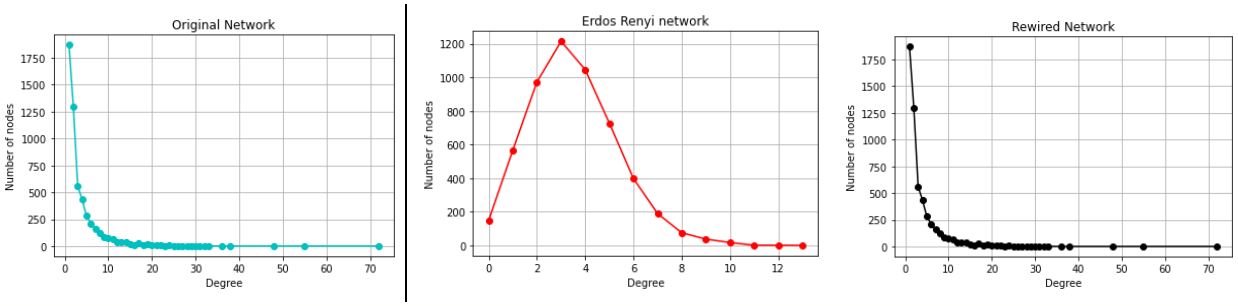
   Three networks:

   After the generation of the edges list three networks were formed: the original network, the Erdos Renyi network , and a rewired network based on the specifications the following is a plot of the three networks.



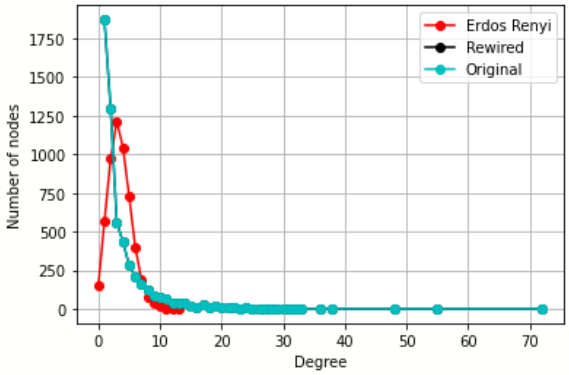
]

2. Plotting the degree distribution:

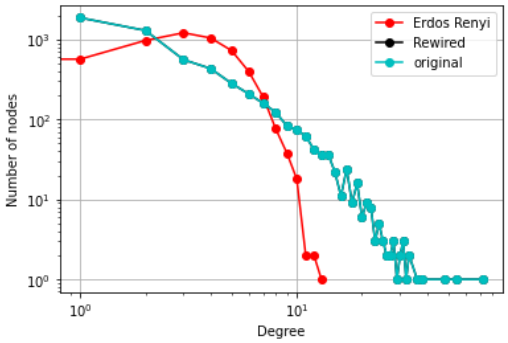


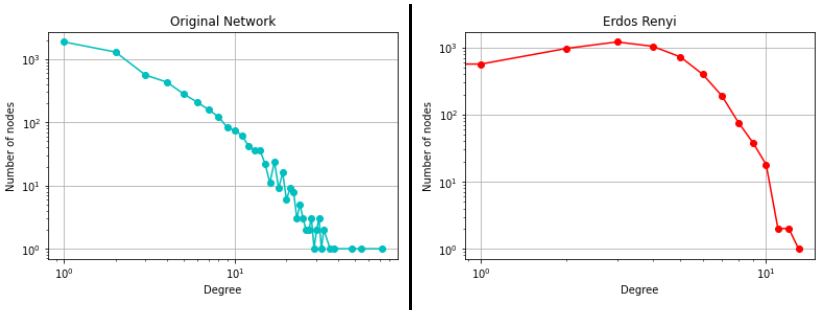
The original network degree distribution follows a power law as expected with smaller degree nodes forming the majority of the graph nodes and the larger degree nodes are fewer in nuimber, the same thing applies to the rewired graph because the rewireing process was designed to preserve the degree distribution. The Erdos Renyi model on the other hand has a binomial degree distribution as expected which is slightly skewed to the left. The value of p used to build the Erdos renyi network is 0.00065 (to preserve the number of edges in the original network. The skewness in the degree distribution shows that larger number of nodes have lower degrees and smaller number of nodes have higher degrees.

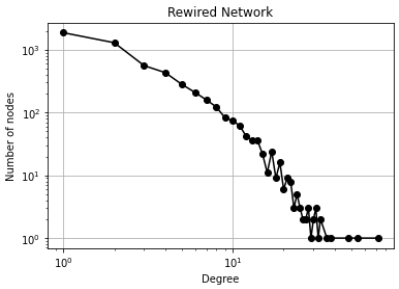
Below is the superimposed plot of degree distributions



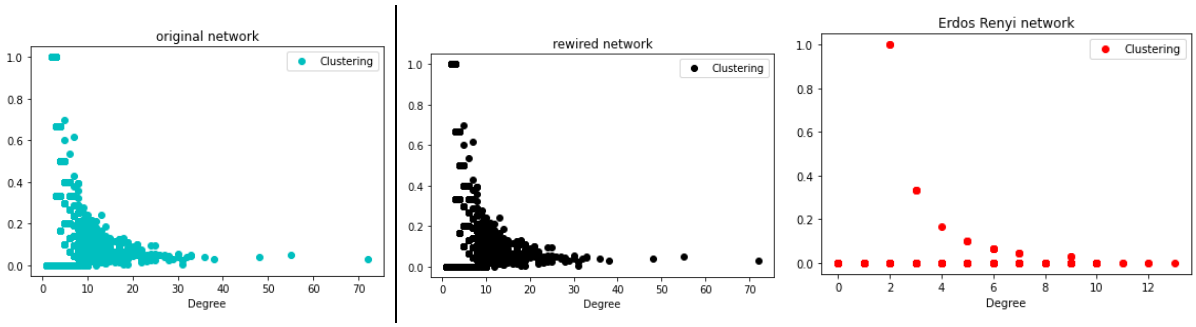
This plot shows that unlike the original network and the rewired network, the Erdos Renyi model doesn't capture the existence of very high degree nodes which will be shown in more details on the log-log scale plots below which shows that the Erdos renyi network doesn't preserve the heavy tail phenomena.







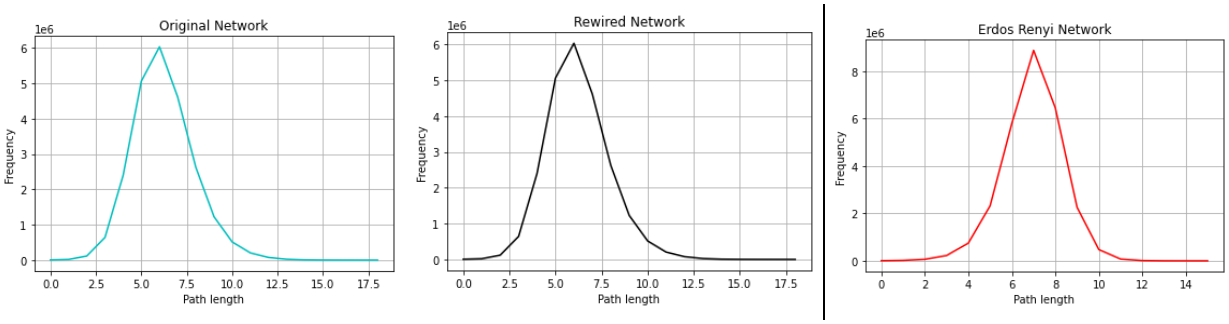
3. Plotting the Clustering coefficient



The clustering coefficient of the original and rewired network is similar showing that smaller degrees have high clustering coefficient while the Erdos Renyi model tends to have low clustering coefficient overall due to the randomness. However, the relation is between the node degree and the clustering coefficient doesn't look linear.

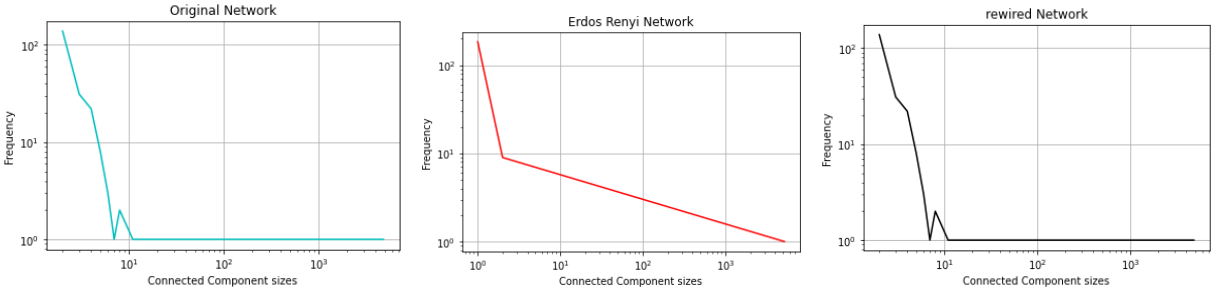
The average clustering coefficient of the original network, the rewired network, and the Erdos Renyi network respectively is: 0.122, 0.1125463, and 0.000815. The original network has the highest clustering coefficient which is justified by virtue of the fact that it represents the real-life, while the rewired network shows lower clustering coefficient because of the randomness aspect and finally the Erdos Renyi has the lowest clustering cooficient which is a result of it being random and ignoring the degree distribution.

4. Path Length plot



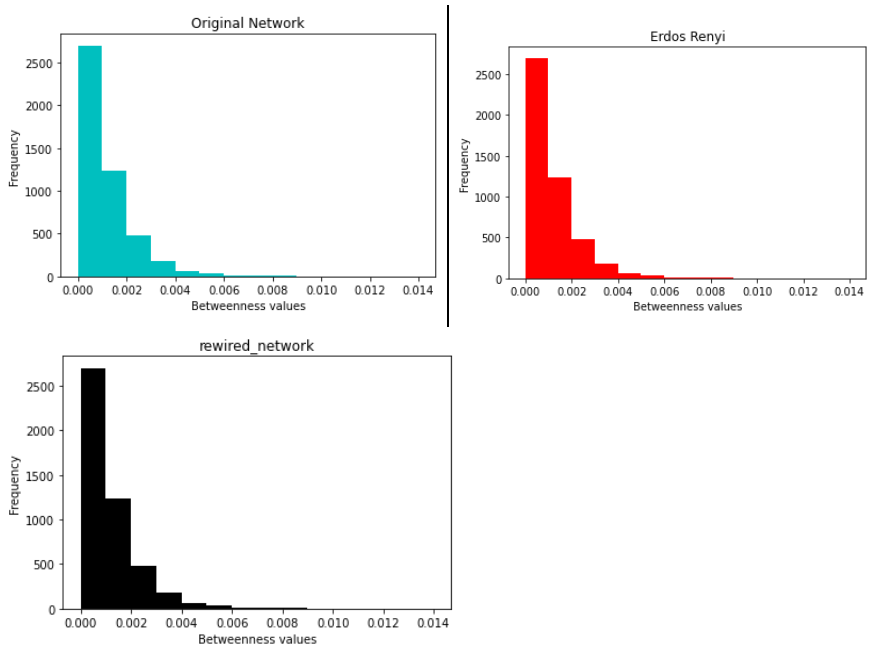
The three networks show that the average path length is similar and very short in all the networks however the Erdos Renyi model has shorter lengths of paths while the rewired network and the original network have higher path lengths.

5. Connected Components analysis



The Previous graph shows the log-log plot of the connected components sizes distribution all the networks have one big connected component of size approximately 5000 and smaller connected components of sizes 1, 2 , 3 ..8. The log log plot is used for more readability.

6. Betweenness Measures



The three networks show very similar betweeness values distribution. There are very few nodes with large betweenness values which means that the connect different communities within the network and very large number of nodes with small betweenness values which means they belong inside a community. The three networks show similar betweenness distribution which can be justified be the structure of the networks plotted in 1. Even though the sampling and wiring is different in the three networks, all networks possess a core periphery structure and hence the betweenness distribution.

7. Community detection

Girven Newman

	Original Network	Erdos Renyi	Rewired Network
Modularity	**0.718**	**0.5679**	**0.7203**
Number of communities	**243**	**167**	**245**
Average community size	**22.2**	**29**	**21.9**
Largest community	**456**	**237**	**510**

Greedy Modularity Maximization

	Original Network	Erdos Renyi	Rewired Network
Modularity	**0.67**	**0.56**	**0.59**
Number of communities	**272**	**169**	**262**
Average community size	**19.1**	**30**	**20.6**
Largest community	**954**	**341**	**742**

Both The modularity maximization algorithm and the Girvan Newman have great modularity measures but has a problem with the core peripheral structure of the network because more than 50% of the communities in the three networks in both models are of size 3 or less which means they are part of the peripheral nodes.
The number of the communities in the Erdos Renyi network in comparison to the two other networks is small which is a natural result of the randomness of the Erdos Renyi model resulting in less clustering and less independent communities, whereas in real network communities are formed easily.
The Girvan Newman algorithm performs better than the Modularity maximization algorithm but with a very high computational cost.