MULTIVARIATE ANALYSIS AND NETWORK ANALYSIS FINAL PROJECT

Chun-Yi Chiang, Nagraj Deshmukh, Sanjay Katyal, Sahana Reddy

1. Background and Problem Statement

The datasets our team chose are derived from arXiv data from Standford University, in which a node represents a person who authored at least one arXiv paper within a certain academic field, and an edge connects people who are collaborating on the paper. This data is undirected data written in a directed format, which needed to be modified. From this data, our team plans to obtain insights about the popularity of certain individuals within the academic field.

There are affiliation networks available for the five largest categories in the arXiv (ASTRO–PH, HEP–TH, HEP–PH, COND–MAT and GR–QC), and we plan to do our analysis on two sets - the GR-QC (Collaboration network of Arxiv General Relativity) data and CA-HepTh (Collaboration network of Arxiv High Energy Physics Theory).

It should be noted that the original data (Graph Evolution: Densification and Shrinking Diameters by Leskovec et al.) contained time-series data which is not present in this dataset. But the data available at https://snap.stanford.edu/data/ is just the edge list. Thus, we won't be able to do the temporal analysis, and we will focus on network descriptive and ERGM analysis.

The main aim of our study is to understand the relationships between this real-world network of researchers who have published work in General Relativity and try to model this network using ERGM to gain more insights from it and compare our findings with the findings from the network of researchers who have published work in High Energy Physics.

We hypothesize the following:

1. If author i and author j work together, and author j and author k work together, the odds of author i working with author k are high.
2. There will form many small clusters of researchers for example in specific universities
3. There will be a few key opinion leaders who are well connected (have written papers with many other researchers) showing "superstar" phenomena.
4. We also hypothesized small world phenomena should apply in the network. As the academia is known for its tight community and highly connected, we also assume the node needed to create small world phenomena can be lower than the rule of 5.
5. Between different fields of academia, such as High Energy Physics and General Relativity, there will be similarities in network structure patterns.

The project is to identify the correctness of the hypotheses mentioned above.

2. Significance and Impact of Analysis

This research could be useful to observe patterns of social stratification and clique behavior in academic settings and could be used in further research about its impact in academia. Comparisons can also be drawn between different fields of academia to observe similarities or differences in patterns of behavior. With further research and information, more data and features could be obtained on the nodes and conclusions could be drawn about the reasons for popularity. One theory to test would be to check if the nodes with the highest number of edges within clusters are professors and their connected nodes are their students.

3. Operational/Analytic Goals and Plan for Analysis

For this project, our goal is to perform network descriptive analysis network simulations to

To analyze this dataset our team plans to take the following steps:

- **Network Descriptive Analysis**

  The data was presented as a directed network, showing duplicate edges, so the team modified and cleaned the data to be used as an undirected network. The network descriptive analysis by analyzing the centrality, betweenness, transitivity, geodesic, and community detection was based on the cleaned data. Specific analysis will be conducted here to test the hypotheses outlined.

- **Network Simulation**

  The team will conduct network simulation such as ERGM and small-world network analysis. Specific analysis will be conducted here to test the hypotheses outlined, specifically the small-world phenomena hypothesis.

4. Network Descriptive Analysis

As mentioned above, the data was cleaned and transferred to an undirected network prior to conducting descriptive analysis. Before diving into the descriptive analysis, we visualized the network as benchmark for the accuracy of the descriptive analysis.

For both networks, we can see there seems to be one densely connected component and few other small clusters (components), which vaguely support the hypothesis we made on “superstar” phenomena and the small clusters.

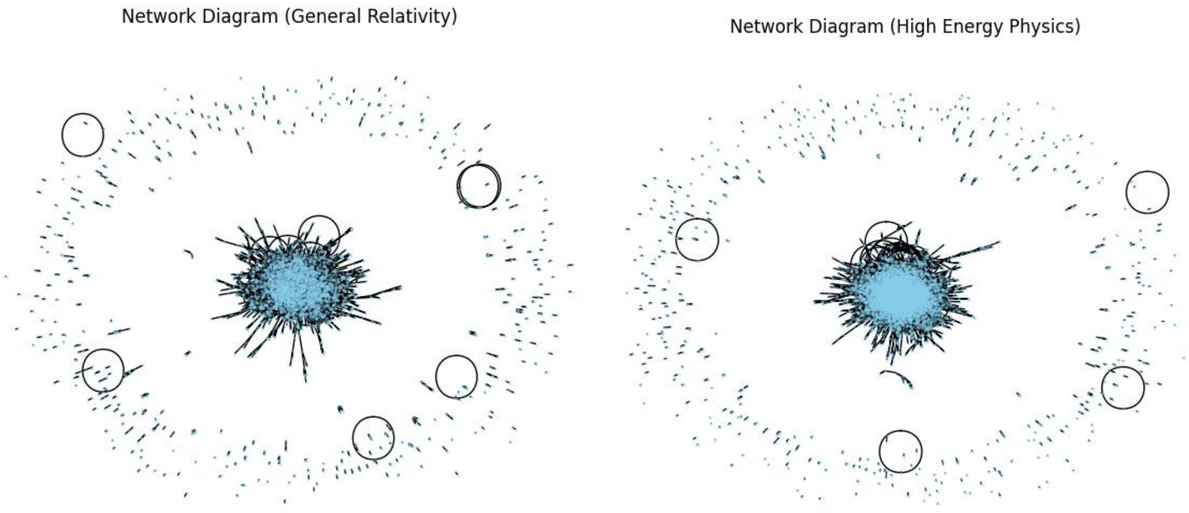
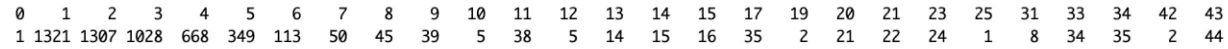

Network Diagram (General Relativity)
Network Diagram (High Energy Physics)

Figure 1: This figure shows the network diagrams for General Relativity and High Energy Physics.

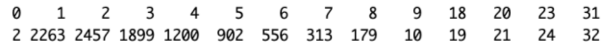
All the small clusters raised a question of whether there is someone that is left out of the collaboration with anyone. In another words, is the academic collaboration network, as mentioned in the hypothesis, tightly connected? By conducting the k-cores, we find the number of isolated nodes is very low, suggesting the collaboration on academic paper is very common phenomenon, even may refer as preferred to collaborate than to work alone.

General relativity



```
 0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   17   19   20   21   23   25   31   33   34   42   43
 1 1321 1307 1028  668  349  113   50   45   39    5   38    5   14   15   16   35    2   21   22   24    1    8   34   35    2   44
```

High Energy Physics



```
 0    1    2    3    4    5    6    7    8    9   18   20   23   31
 2 2263 2457 1899 1200  902  556  313  179   10   19   21   24   32
```

Then we further conduct analysis of degree of node, transitivity, density and diameter. With network descriptive analysis, it is expected to have the initial understanding of hypotheses 2- 5. Below table shows the result of analysis on both datasets.

	General Relativity	High Energy Physics
Nodes	5,242	9,877
Edges	14,496	25,973
Degree of Node	(max/ mean) 81 / 5.526135	(max/ mean) 65 / 5.259289
Betweenness	(max/ mean) 1016871 / 16649.92	(max/ mean) 2311650 / 37355.74
Transitivity	0.6298425	0.2839997
Density	0.001054405	0.0005325323
Diameter	17	18

Although both academic collaboration networks show a pattern of one major component with few small clusters, the density between fields differs. In this specific case, the larger network has lower density. This is also evident by the lower rate of transitivity and k-core decomposition, the ratio of two-degree open ends components is higher.

With both networks close to fully connected, we can also analyze whether the small world phenomena exist in academia. The rule of thumb for the small world phenomenon is that by connecting to 5 nodes, we should be able to connect to the whole world, and with the mean degree of nodes set around 5 for both networks also match the phenomenon. However, it is noticeable how the degree distribution shows the mode of degree of nodes sets around 2, suggesting that most of the author collaborate with only one author. Our speculation towards this circumstance is that superstars, which in this case is the professor, collaborate with their students, and hence, the network is formed around the superstar.

To provide more evidence of our speculation on “the larger networks is, the lower popularity single node gain,” the betweenness distribution is conducted. In the context of this network, high measures of degree centrality show a researcher with high degree centrality has collaborated with many others. This indicates they are highly collaborative and well-connected in their field. A high closeness centrality shows researchers can reach all other researchers more quickly than others can, on average. These nodes could likely be central figures like professors, who can access and distribute information quickly, but more data would be needed to draw that conclusion. A node with high eigenvector centrality is not just well-connected but also connected to other influential researchers. This could be a leading researcher or professor in the field who is connected to other leaders. Their collaborations could carry more weight. A researcher with high betweenness centrality lies on many shortest paths between other researchers. This means that they may be a key collaborator between different researchers and research groups. They would be able to spread information and ideas to otherwise disconnected research groups.

Aside from the mean for degree of nodes, we can also identify that the larger network doesn't necessarily lead to higher centrality for degree of nodes when referring to the maximum degree of nodes. In contrary, it might be with larger network where more researchers are involved, less experienced professionals (students or young professional) have more choices for their mentor for paper, and thus, the popularity for single node decreases. From the distribution plots conducted on both networks, we can see a positive right skew indicating that many researchers are less collaborative and well-connected in the field. This could indicate a superstar phenomenon where few researchers are the most well-connected.

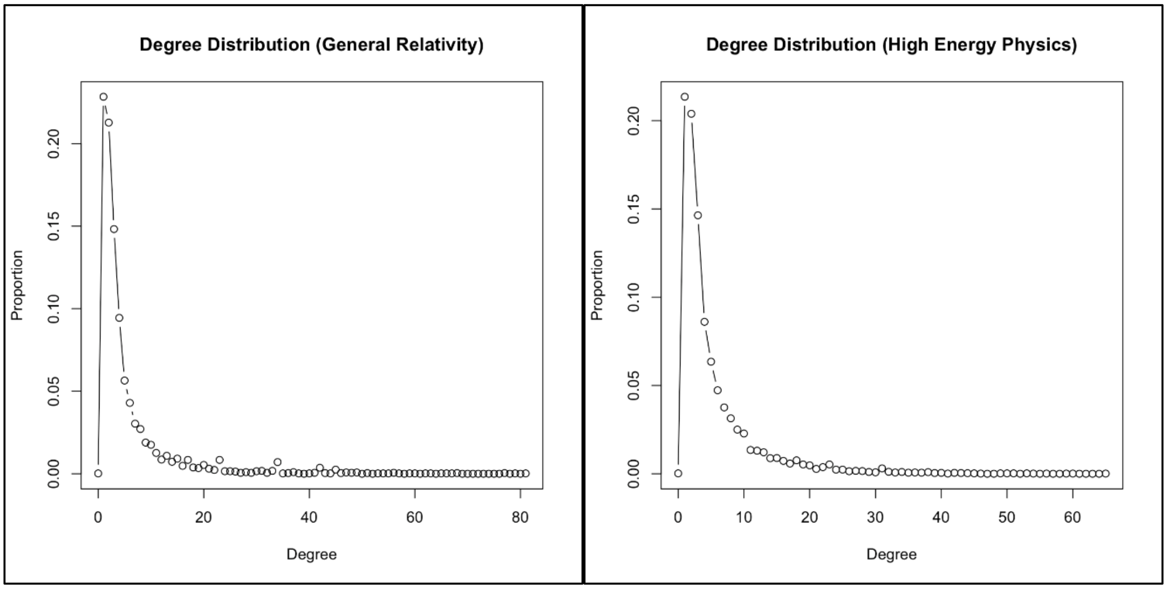


Figure 2: This figure shows the degree distribution plots for General Relativity and High Energy Physics.

From the plots conducted on betweenness centrality on the networks for both General Relativity and High Energy Physics, there is a similar observable pattern with key differences. We can see that generally the plots show many nodes with low betweenness with fewer nodes showing high betweenness with a heavy skew to the right. This signifies a centralized network and may indicate a superstar phenomenon which was hypothesized. This may or may not necessarily indicate several small clusters of researchers, however.

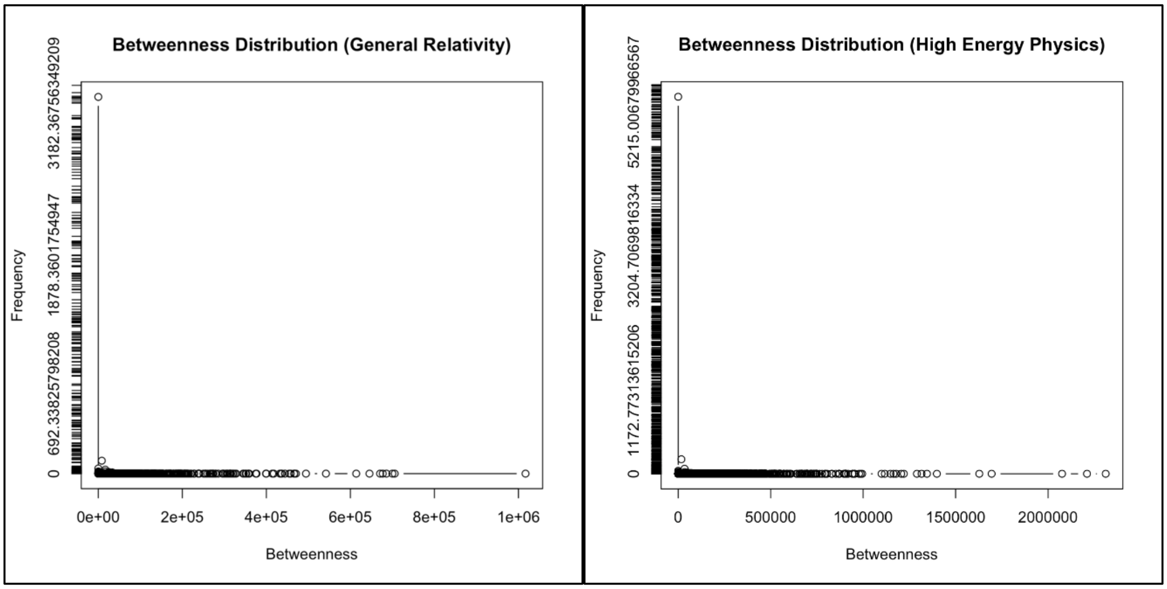


Figure 3: This figure shows the betweenness distribution for the General Relativity and High Energy Physics networks.

From the plots on closeness distribution, we can see that a high closeness centrality indicates that a node can quickly interact with all other nodes. Please note that the code to perform this closeness centrality was normalized for computational ease. From these plots, we can infer that there are indeed many nodes which show high connectivity and could be important connectors in communication with the other researchers.

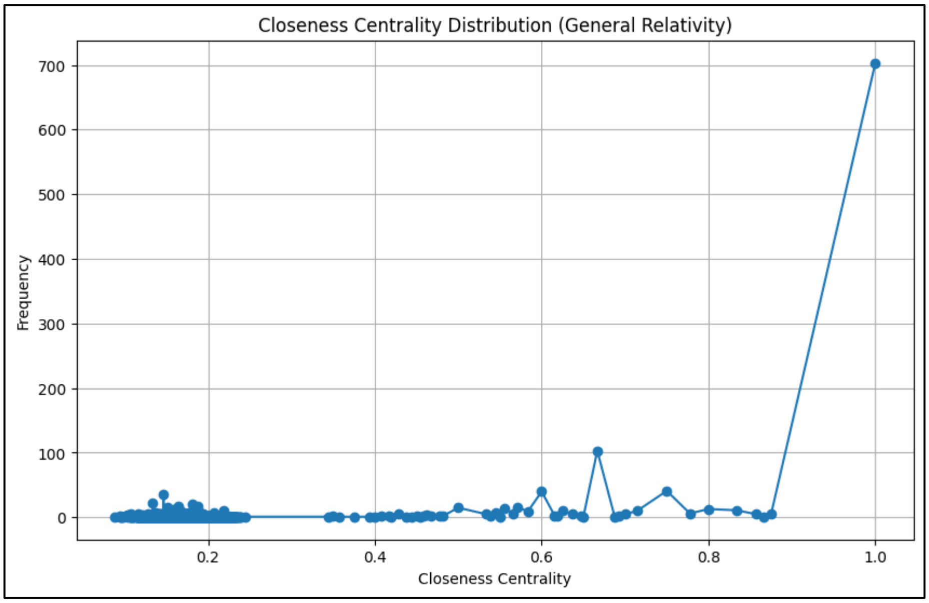


Figure 4: This figure shows the closeness centrality distribution for the General Relativity network.

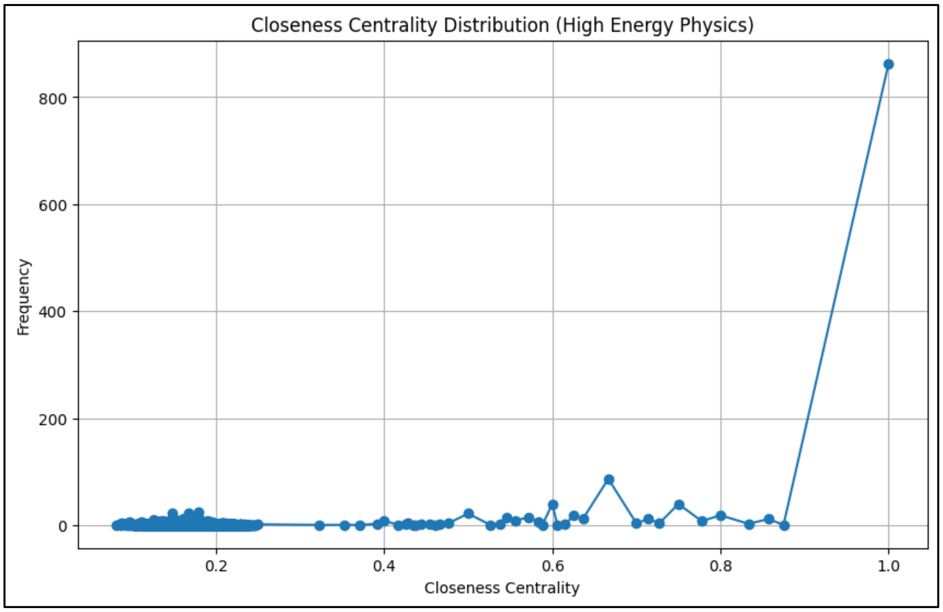
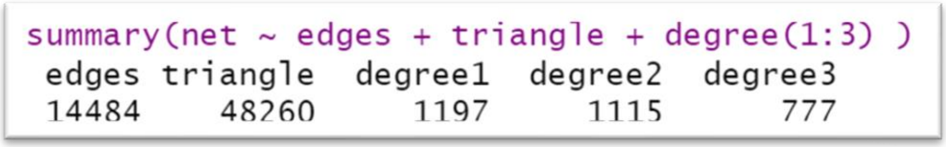


Figure 5: This figure shows the closeness centrality distribution for the High Energy Physics network.

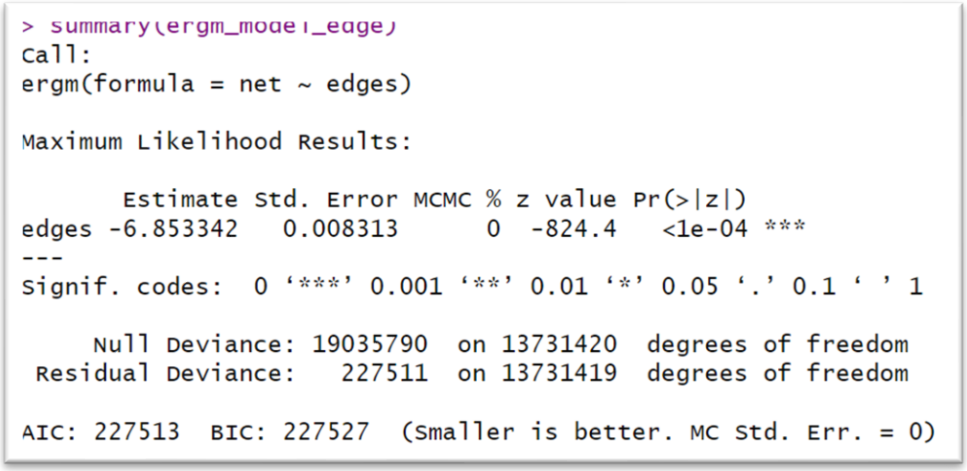
5. Exponential Random Graph Model (ERGM)

First, we try building ergm with the "General Relativity" data.



```
summary(net ~ edges + triangle + degree(1:3) )
 edges triangle  degree1  degree2  degree3
 14484    48260     1197     1115      777
```

Note: When using any parameter out of the ones shown above other than the number of edges, we were facing an issue during the model estimation. The MPLE (Maximum Pseudolikelihood Estimation) which provides initial estimates of the parameters ran without convergence issues, but MCMLE (Monte Carlo Maximum Likelihood Estimation) which refines these estimates was running into errors. The error message indicated a problem known as "model degeneracy" which occurs when the estimated model tends to produce extreme network configurations that do not reflect the complexity or the structure of the observed network data. Hence, we stuck to using just the number of edges as our only parameter.



```
> summary(ergm_model_edge)
Call:
ergm(formula = net ~ edges)

Maximum Likelihood Results:

      Estimate Std. Error MCMC % z value Pr(>|z|)
edges -6.853342   0.008313      0  -824.4   <1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 19035790  on 13731420  degrees of freedom
 Residual Deviance:   227511  on 13731419  degrees of freedom

AIC: 227513  BIC: 227527  (Smaller is better. MC Std. Err. = 0)
```

1. The coefficient for `edges` is -6.853342, which is quite negative. This suggests that the presence of additional edges in the network is less likely than in a random network, assuming no other parameters are affecting the edges. The more negative the coefficient, the less likely an additional edge is, holding other factors constant.

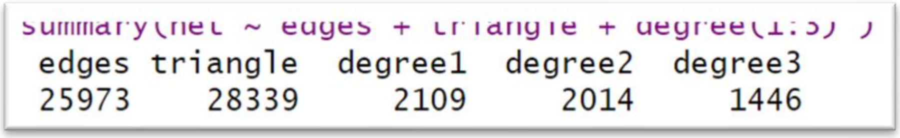
2. The standard error of the estimate is very small (0.008313), indicating a high level of precision in the estimate of the coefficient.

3. The p-value is `<1e-04` (less than 0.0001), indicating that the results are highly statistically significant, meaning the likelihood of observing such a coefficient by chance (if there truly were no effect) is extremely low.

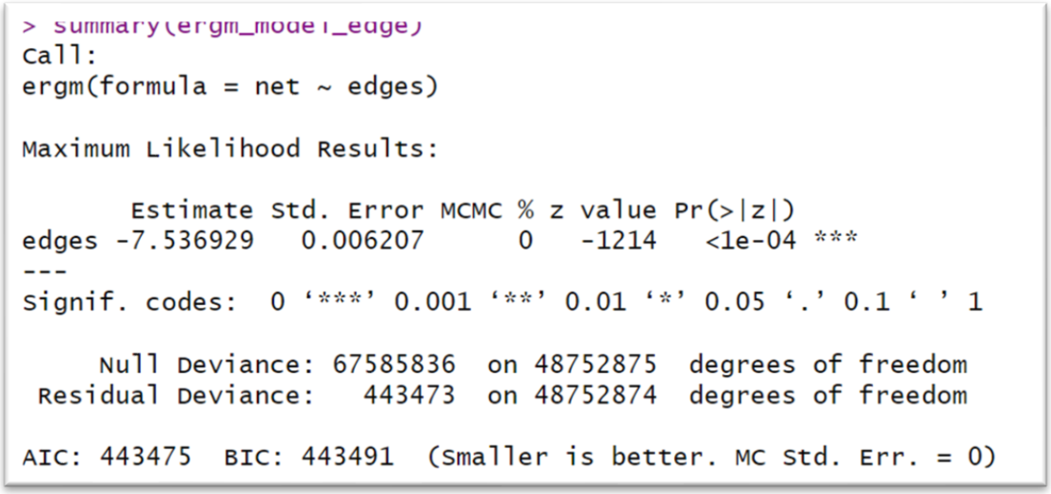
4. Deviance Information:

- Null Deviance: This represents the deviance of a model with no predictors, and here it's very high, showing that a model without predictors fits very poorly.
- Residual Deviance: This is much lower than the null deviance, indicating that your model (including the edge parameter) fits the data significantly better than an empty model.

In summary, our ERGM model strongly indicates that the formation of edges in our network is significantly less likely than in a random network, assuming no other factors are influencing the network formation. This tells us that the observed network has a significantly smaller number of connections, potentially leading us to the similar conclusion that these connections might be localized to some extent. Also, we can see similar results for the “High Energy Physics” data.



```
summary(net ~ edges + triangle + degree(1:3) )
 edges triangle  degree1  degree2  degree3
 25973    28339     2109     2014     1446
```



```
> summary(ergm_model_edge)
Call:
ergm(formula = net ~ edges)

Maximum Likelihood Results:

       Estimate Std. Error MCMC % z value Pr(>|z|)
edges -7.536929   0.006207      0   -1214   <1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 67585836  on 48752875  degrees of freedom
 Residual Deviance:   443473  on 48752874  degrees of freedom

AIC: 443475  BIC: 443491  (Smaller is better. MC Std. Err. = 0)
```

6. Community Detection

In our problem's context, community detection offers many useful insights as to how the network of researchers is structured. Detecting communities helps identify groups of researchers who frequently collaborate. This provides insights into the natural divisions and collaborative efforts within a research field.

For the General Relativity network, we were able to make several observations about community detection using Walktrap algorithm. Walktrap detected 814 communities for General Relativity. This shows that there are many clusters within the network as expected rather than one large cluster. These different clusters could represent different universities, however additional features would be necessary to confirm this. Walktrap modularity is 0.7823643 which is relatively indicates a stronger division of the network into well-defined communities.

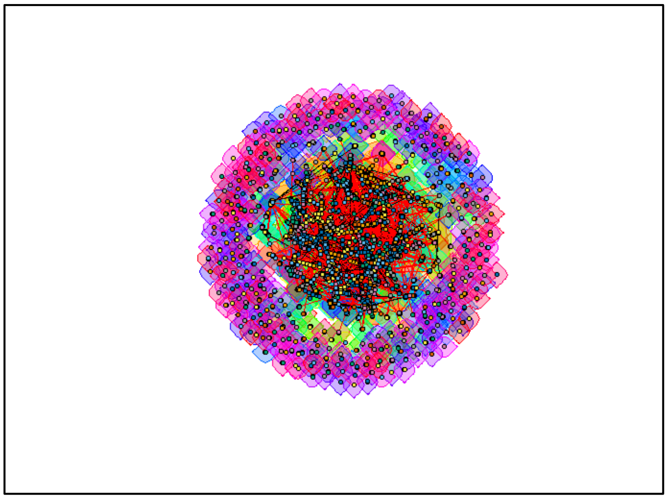


Figure 6: This figure shows the visualization for Walktrap algorithm on the General Relativity network.

To better visualize the patterns emerging from the General Relativity network, the Walktrap algorithm was used only with super nodes, which were nodes with a degree greater than the 75th percentile.

The total number of nodes in the network decreased to 1141 nodes and we observed 102 communities. Even with the super node data, we still see the same pattern of a large number of subgroups within the network. The walk trap modularity of 0.7776915 is relatively indicates a stronger division of the network into well-defined communities and is very similar to the walk trap modularity calculated with all researchers in the network. Please refer to Figure __ below.

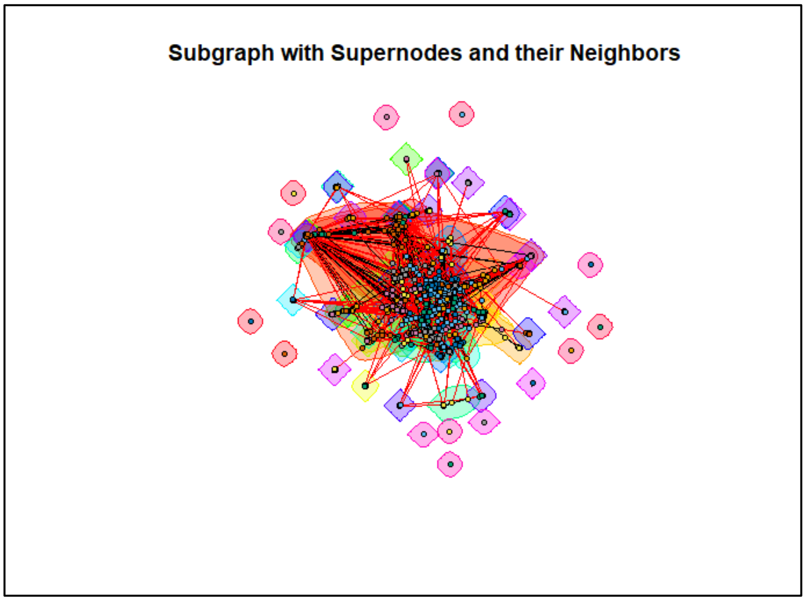


Figure 7: This figure shows the visualization of the Walktrap done using super nodes on the General Relativity network.

For the High energy Physics data, we get similar results with a Walk-trap modularity of 0.6631, and 1295 communities detected. For a total number of nodes 9875, it suggests that on average ea ch of those communities have 7.62 nodes on average. This is higher than the average 6.43 nodes for each community in the General Relativity data. Although higher, it again alludes to multiple s mall clusters. These could be interpreted as multiple universities with 6 to 8 researchers on avera ge working in the fields of General relativity and High Energy Physics.

7. Future Work: Topology Inference

To continue future research into this topic and continue to test the hypotheses outlined, topology inference would be useful. Topology inference aims to predict the structure and connections of a network based on observed data or partial information. When applied to a network of researchers who are nodes and share an edge if they published together in an undirected network, topology inference can provide insights into various aspects like predicting potential collaborations between researchers who have not yet published together but might in the future. This would be useful in testing the hypothesis if author i and author j work together, and author j and author k work together, the odds of author i working with author k are high. We can leverage existing patterns in the collaboration data to help identify potential collaborations.

Topology inference can also help us improve community detection by identifying tightly knit research groups or clusters that often collaborate. These communities can reveal how scientific collaboration is structured and may reveal patterns. Inferred network topologies can help identify critical researchers that are crucial for maintaining connectivity. Understanding this can provide substantial support to the other network descriptive analysis and network simulation conducted to prove or disprove the proposed hypotheses.

8. Practical Implications

The practical implications of this project are to observe the way academic networks are structured and see if there are areas which could be improved in order to make academic ideas and collaboration between researchers more easily accessible. By identifying superstars and observing the way clusters are structured, we may be able to identify hierarchies and determine key leaders in the respective fields. We may also be able to identify researchers who are highly cross-collaborative and may be connected to multiple clusters or cliques. By identifying these key connector nodes, researchers who may be looking to break into different groups will know which other researchers with whom it would be most important to form relationships in order to have greater access to other groups and researchers. This may help researchers with many ideas who have few connections be able to identify other researchers with whom they can connect. It would also be interesting to observe these patterns across different fields of research to identify which fields show a higher level of clustering and which fields may be more evenly connected rather than having a few superstars or key leaders. Perhaps emerging fields would be more highly collaborative than established fields, but this could be explored with further analysis and research.

9. Role Division: All team members collaborated to write the proposal and code.