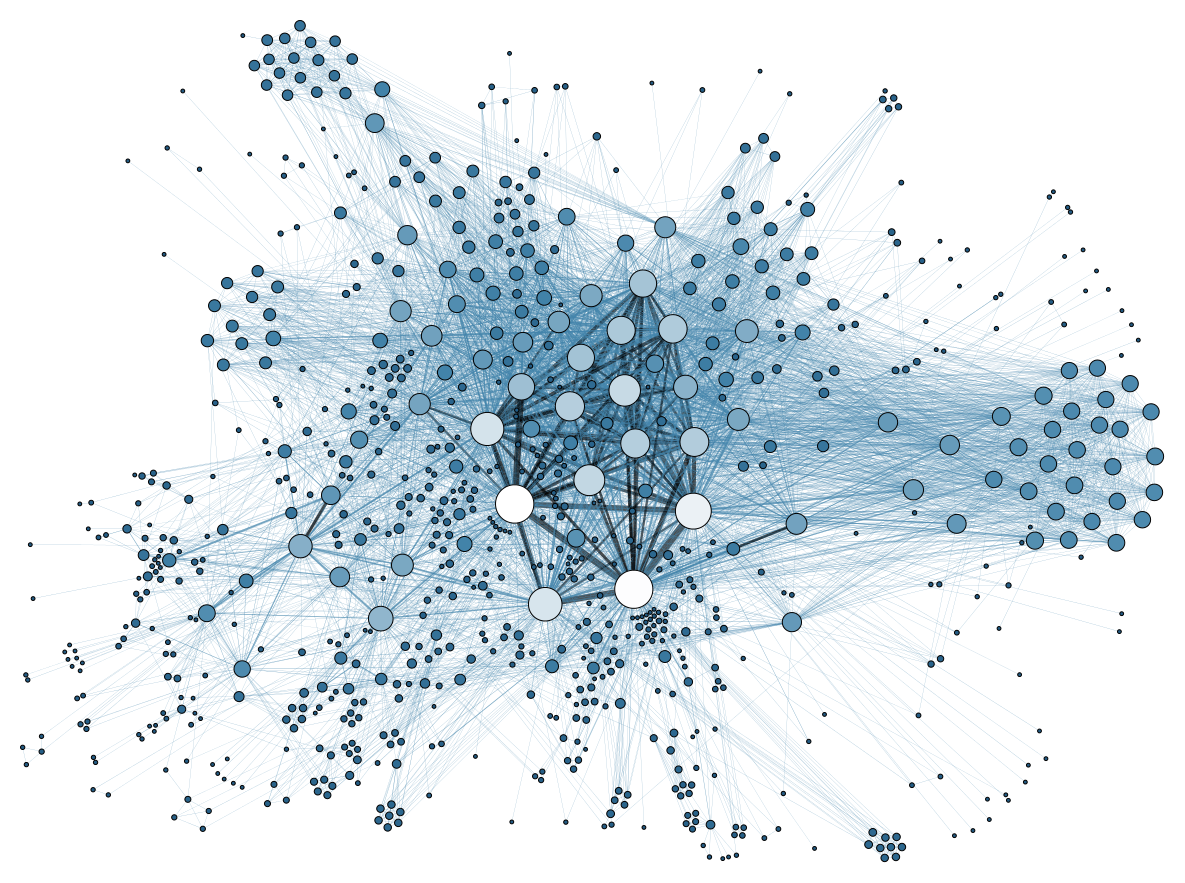
****

**IDS 564 – Social Network Analysis**

**Final Project Report**

**An Analysis of General Relativity and Quantum Cosmology Author Collaborative Network**

**Group members: Leyla Elnaggar, Danielle Strejc**

**Outline**

**Introduction**

**1. Problem Statement**

**2. Review of relevant prior work**

**3. Data Collection**

**3.1 Dataset Basic statistics**

**3.2 Degree Distribution**

**3.3 Cliques and Clusters**

**3.4 Community Detection**

**3.5 Hubs in the Network**

**4. Conclusions**

**References**

**Appendix**

**Introduction**

Social networking is one of the important parts of our daily life because it enables us to communicate, exchange information and collaborate with others. Social Network Analysis is a set of methods used to visualize networks, describe specific characteristics of overall network structure, and build mathematical and statistical models of network structures and dynamics (Luke 2015). Social networks are formally defined as a set of nodes that are tied by one or more types of relations. Nodes are most commonly persons or organizations, but in principle any units that can be connected to other units can be studied as nodes.

Scientific collaboration networks are a symbol of contemporary academic research. Researchers are no longer independent players, but members of teams that bring together complementary skills and multidisciplinary approaches around common goals. Social network analysis and co-authorship networks are increasingly used as powerful tools to assess collaboration trends and to identify leading scientists and organizations.

**1. Problem Statement**

In our project we will be studying General Relativity and Quantum Cosmology collaboration network. And we will be performing a variety of analyses and visualizations on this collaboration network. It is a **scale-free network** whose node degrees follow a power-law distribution showcasing a skewed distribution of links. The size of connected components also follow a power-law, and the giant component contains many authors.

**2. Review of relevant prior work**

In our project, we try to analyze this collaboration network of scientists working on general relativity papers. Although we study it in a static manner, we have found many research papers that try to analyze the dynamics of this kind of network throughout time, or study this kind of network in an innovative matter as described in this paper, “We constructed a multilayer network, in which each layer represents a different kind of collaboration. After having analyzed the evolution over time of specific parameters of the co-authorship network, we investigated the effects of adding one type of collaboration edge at a time, in a cumulative fashion, on the values of these parameters and on the topology of the collaboration network through time, including rapid shifts in the dynamic evolution of the largest component.” (Roberto Lalli, Riaz Howey & Dirk Wintergrün)

**3. Data Collection**

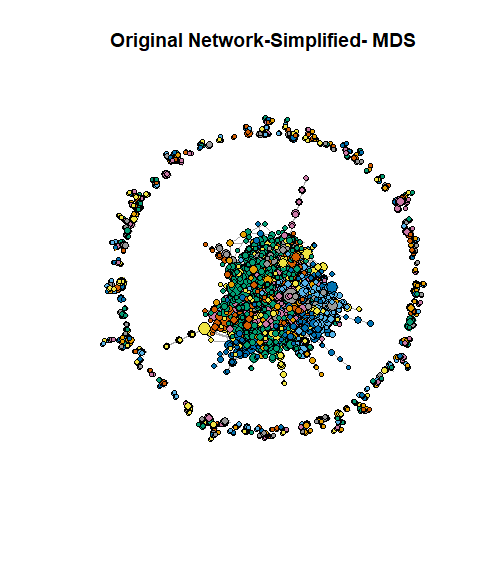
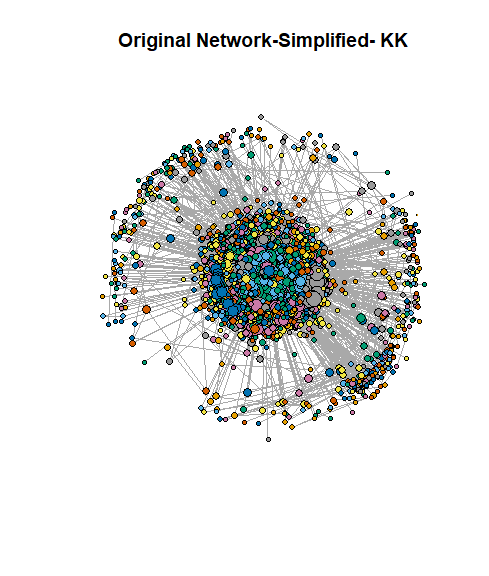
We found this dataset on a repository for Stanford university with many collections of datasets. The data covers papers published in the period from January 1993 to April 2003 (124 months).

It is an undirected and unweighted collaboration network. Each author is represented as a node. If an author ‘i’ co-authored a paper with author ‘j’, then the graph contains an 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.

**3.1 Dataset Basic statistics:**

|  |  |
| --- | --- |
| **Table 1. Dataset Statistics** |  |
| **Nodes** | **5242** |
| **Edges** | **28980** |
| **Nodes in largest SCC** | **4158 (0.793)** |
| **Edges in largest SCC** | **13428 (0.926)** |
| **Average clustering coefficient** | **0.5296** |
| **Number of triangles** | **48260** |
| **Fraction of closed triangles** | **0.3619** |
| **Nodes mean degrees(unsimplified)** | **11.05685** |
| **Avg path length** | **0.629** |
| **Transitivity** | **0.6298** |
| **Diameter (longest shortest path)** | **17** |

**Building the Network visuals**

******Below we show two plots of our simplified Network:**

The MDS plot uses metric multidimensional scaling for generating the coordinates, and the other uses the KK layout generator. The vertices are weighted by graph strength. The medoids plot allows us to see two groups split, the outer ring and inner circle are hardly connected with each other, which could be something interesting to explore in the future. Both plots help us view and understand the network better.

**3.2 Degree Distribution**

One tool we have for analyzing our network is the degree distribution. In our network each node represents an author, the degree would represent the number of authors this specific author is collaborating with, it may also be an indicative of highly active researcher who is publishing many papers and collaborating within the literature community.

We have plotted three different plots: Normalized degree distribution, degree frequency as well as log-log degree distribution, to show the degree distribution we can see that most of the nodes have less than a degree of 16. Our mean degree is 5.52 which means on average an author collaborates with 5.52 other authors.

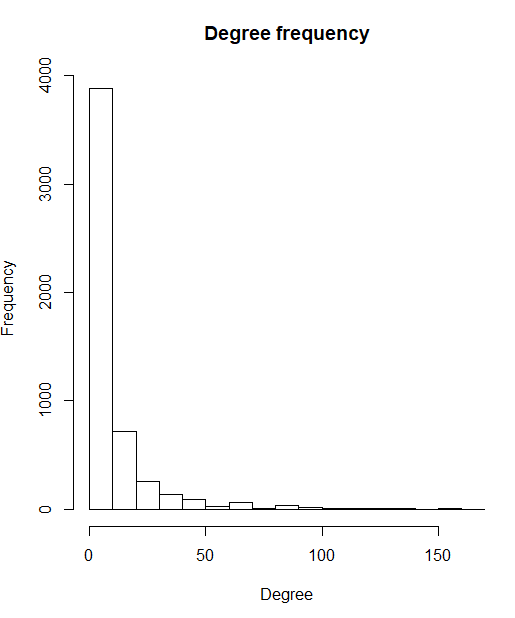
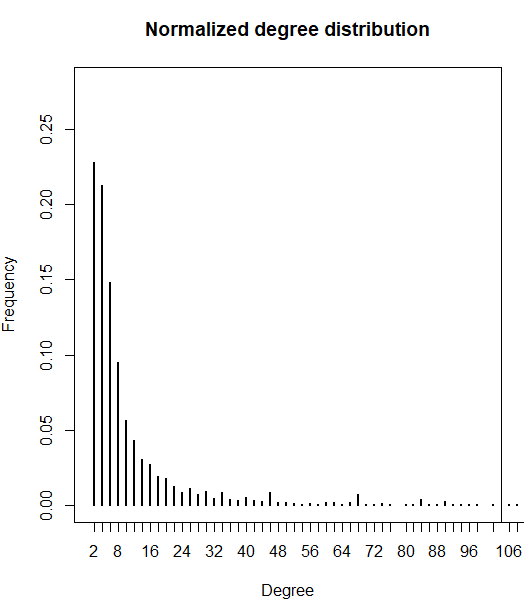
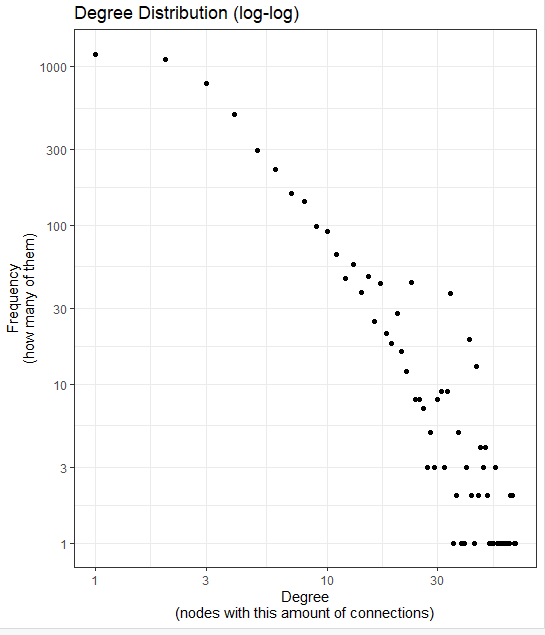


Figure1 (a) Figure1 (b)

 Figure1(c)

To visualize and detect a power law behavior, we plotted CCDF for the network, and the distribution has a long tail shape. This means that the author collaboration is low for most of the authors, but there are a few authors with very high levels of interactions with other authors. In addition, the curve shows a clear upward curvature, a sign that it does not match the power law model on the entire domain.

Recall that the CCDF follows a power law with exponent α – 1, The distributions have a long tail shape.

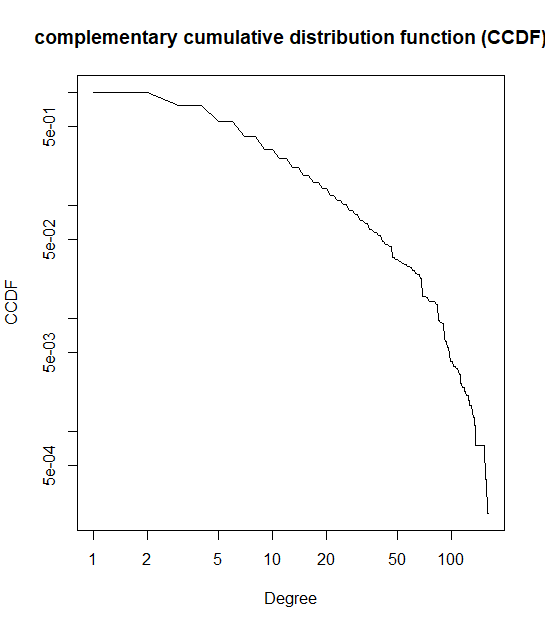
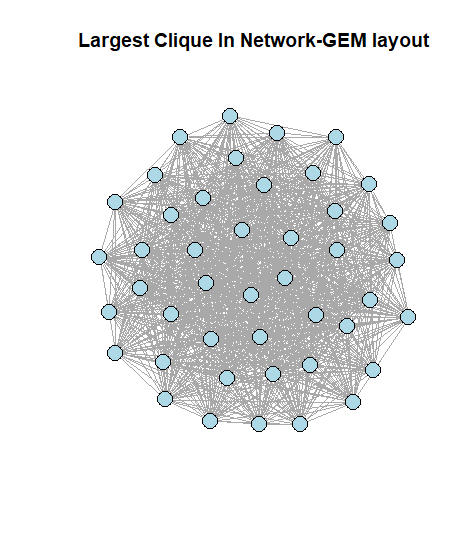


Figure2: CCDF

We next briefly examine the correlation between four important network measures: degree centrality, edge betweenness, and node betweenness. As degree centrality increases, so does node betweenness. This makes sense since more paths going through a vertex (node betweenness) would likely mean that the vertex has a higher number of adjacent edges (degree centrality). There is not a strong relationship between edge betweenness and degree centrality or node betweenness, but it is positive in both cases which tells us that as edge betweenness (number of the shortest paths that go through an edge in a graph) increases, a vertex will have more adjacent edges and paths through it.

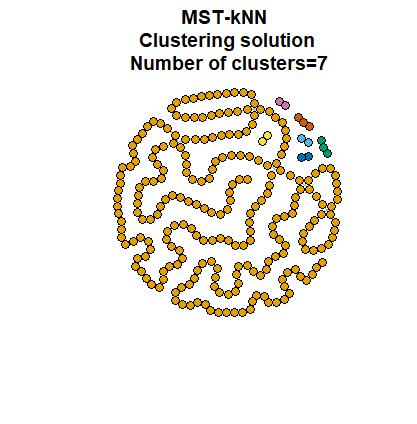
|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2. Correlation Between Measures in the original simplified network** | | | |
|  | **Degree Cent.** | **Edge Between.** | **Node between.** |
| **Degree Cent.** | 1 | .010 | .516 |
| **Edge Between.** | .010 | 1 | .018 |
| **Node between.** | .516 | .018 | 1 |

**3.3 Cliques and Clusters**

A **clique** is defined as a maximal complete subgraph of a given graph—i.e., a group of people where everybody is connected directly to everyone else. There are 355 total clusters in the network, ranging from size 1-177. The largest clique contains 44 nodes, which is quite large even for a network of this size. If this network was labeled, it would allow us to know who those 44 people are, and probably influential & high profile researchers who are making advances in their field, in addition they were all collaborating with each other at some point or another in the 10 year timeline that we look at. Please see **Table 8** in the appendix for a further breakdown of strongly connected clusters in the simplified network.

From **Table 3**, which shows number of vertices in maximal cliques (cliques that cannot be enlarged), we see that 90% of nodes are of size 7 or less , which is close to the size of average degree we have of 5.5.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3:** Maximal cliques in simplified graph | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 1606 | 1394 | 594 | 182 | 52 | 28 | 13 | 9 | 4 |



This next plot uses the R package “MST- KNN”. This is a clustering algorithm within a graph-based approach (Inostroza-Ponta 2008). It uses two proximity graphs: minimum spanning tree (MST) and k nearest neighbor (kNN).

This graph is what we get if we look at the nodes involved in clusters of size 2, 3, and 4. We include this to just show how much of the network is connected when we break it down a little further. By looking at this smaller subgroup, we get a clear understanding of how most of the vertices here are connected in a neat way and this plot shows that. Refer to Table 7 for graph density of this subgroup as it compares to the original network.

**3.4 Community Detection**

When we look at the overall simplified network with community detection through the walk trap algorithm, modularity is 0.78236, which tells us that the network can be relatively well split up into groups. We try two different methods of community detection- walk trap community and fast greedy. They are both bottom up approaches to community building. Fast greedy tries to optimize modularity in a “greedy” manner where nodes are in separate communities to start, and gradually vertices are merged so that each merge is locally optimal. Walk trap runs short random walks and uses these random walks to merge separate communities.

Modularity looks at the strength of groups within a network and its communities. The relatively high score for the original network is not surprising, since a high modularity score indicates sparse connections across communities, but also strong ones within the communities. The original network and the largest clusters subgroup have the highest modularity scores. We suspect there are some densely connected communities in this network, particularly among the most influential authors.

The top 15 hubs are all fully connected, so the modularity score is 0. There will be more on top hub subgroups in the next section.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 4. Modularity in several networks** | | | | |
|  | **Original Simplified** | **Top 15 Hubs** | **Random 15 node subset** | **Top 60 Hubs** |
| **Walk trap** | 0.78236 | 0 | 0 | 0.044684 |
| **Fast Greedy** | 0.81978 | 0 | 0 | 0.082729 |

Top 15 Hubs and the random 15 subset have the same modularity scores, but they are opposite in terms of connectivity. The top 15 hubs are all connected whereas the random subset has only one 2-way connection, so it is difficult to split the network into any communities. It is surprising that the original network has a higher modularity score than the top 60 hubs. This could be because the top 60 hubs are connected in a way that makes splitting them up very difficult.

**3.5 Hubs in the Network**

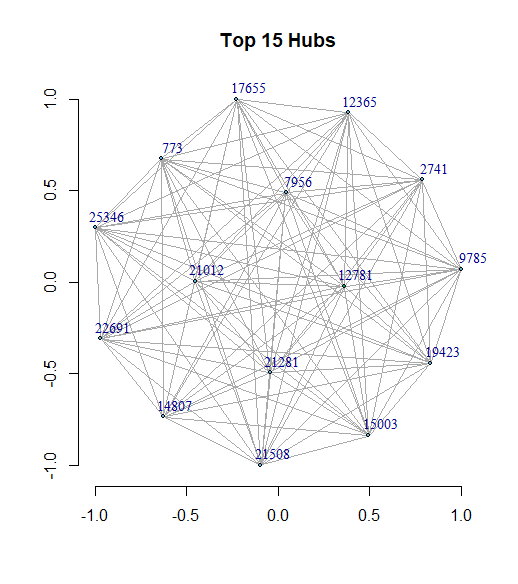
Hubs in our network are the authors that tend to collaborate the most with other authors, meaning they are likely very influential people. There is no one node that has a significantly higher score than the others, so these top 6 hubs are probably close to each other in importance in the industry.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 5. Top “Hubs” in the Network** | | | | | |
| **21012** | **2741** | **12365** | **21508** | **9785** | **15003** |
| 0.1555625 | 0.1535750 | 0.1530727 | 0.1511946 | 0.1509039 | 0.1504084 |

**Further Exploration of Top Hubs**

We were still curious about the top hubs and understanding those nodes a little more. **Table** **6** below looks at the number of neighbors that each of the hubs has. Notice how the relationship is not linear, where neighbors would decrease as hub score decreases. This tells us that hub 2741 has more connections with influential authors than 12365. Node 12365 has 12 more neighbors, but it has a lower hub score since its connections are not as important in the network as the connections of 2741.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 6. Number of Neighbors of Top “Hubs” in the Network** | | | | | |
| **21012** | **2741** | **12365** | **21508** | **9785** | **15003** |
| 81 | 65 | 77 | 67 | 68 | 62 |

Next, we wanted to see a graphical interpretation of the top hubs. We made a subgroup with just the top 15 hubs. This looks very much like the largest clique graph. As it turns out, a many of the top hubs are in that top clique subgroup which confirms our prior interpretation. In fact, upon further analysis we saw that 43/44 top hubs are represented in that largest clique which we find to be significant since it reinforces the strength of connections in this citation network. See the **Plot 2** in the appendix for a side-by-side look at the largest clique and the top 44 hubs.

We next take a random subset with 15 nodes since that size yielded the best graph density score from our random subset analysis. With that in mind we compare this graph density value to the top 15 top hubs to see if there is an actual significant difference in network density between the two subgroups. Refer to the appendix for **Plot 1** and an explanation of the code used.

The graph density of a random optimally sized subgroup is .015. When we compare that to a subset with the top 15 hubs, the graph density is 1 which tells us that the top hubs are much more densely connected than if we just compare random nodes in the network. Simply put, authors who are collaborating with many other authors are more likely to continue to collaborate with other authors who are also influential.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 7. Graph Density in several networks** | | | | |
|  | **Original Simplified** | **Top 15 Hubs** | **Random 15 node subset** | **Top 60 Hubs** |
| **Graph Density** | 0.00105 | 1.0 | 0.015 | 0.65254 |

We want to look at the graph density to understand the connectedness of the network more. Once again, we note that the top hub subgroup is much more connected than a subgroup of random nodes (whose size is determined by the best possible graph density), which confirms our earlier findings.

**4. Conclusions**

We have applied various methods learned during the course to analyze and visualize this collaboration network.  Our results suggested that prolific authors with higher degrees tend to collaborate more.

The network is somewhat well connected and contains significant community structure, so we tried to analyze those communities and gain some insight to the network. This is a scale-free network and we can easily see the high degree of connectivity among the most influential authors, which tells us that those who are at the top of their field are working with others who are also very important. If an author wishes to become more prominent in this field, they should try and get in with the “top hubs” and hopefully in time get their name out there.

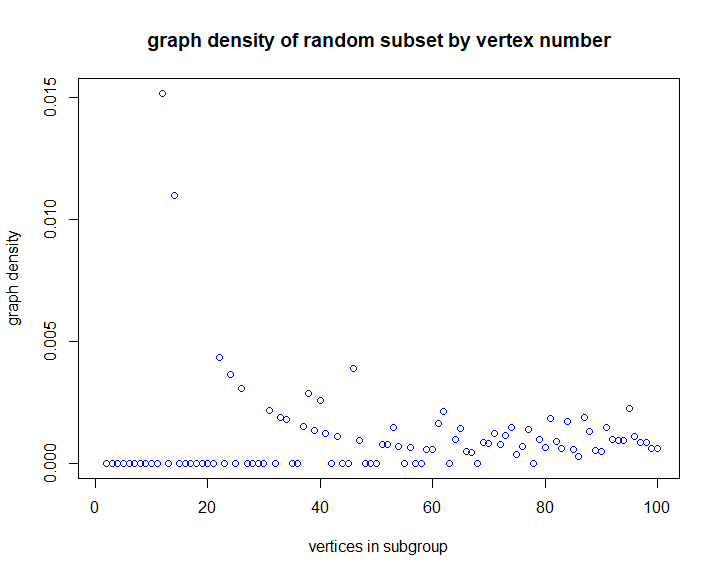
**Future work:** We could consider studying how the graph evolved over years and find whether the network was becoming denser and each author had more collaborators on average as the network changed over time.

**References:**

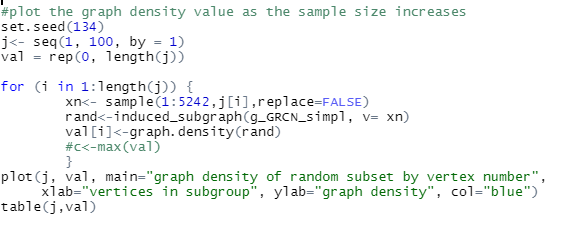
* Data source : <http://snap.stanford.edu/data/ca-GrQc.html>.
* Bishop , Michael. “What Are the Differences between Community Detection Algorithms in Igraph?” *Stack Overflow*, Nov. 2013, [stackoverflow.com/questions/9471906/what-are-the-differences-between-community-detection-algorithms-in-igraph](file:///C:\Users\dtstr\Downloads\stackoverflow.com\questions\9471906\what-are-the-differences-between-community-detection-algorithms-in-igraph).
* Luke, Douglas A. 2015. A User’s Guide to Network Analysis in R.
* The dynamics of collaboration networks and the history of general relativity, 1925–1970. Roberto Lalli, Riaz Howey & Dirk Wintergrün.
* J. Leskovec, J. Kleinberg and C. Faloutsos. [Graph Evolution: Densification and Shrinking Diameters](http://www.cs.cmu.edu/~jure/pubs/powergrowth-tkdd.pdf). ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.
* *Degree Distribution*, University of Chieti-Pescara, [www.sci.unich.it/~francesc/teaching/network/distribution.html](http://www.sci.unich.it/~francesc/teaching/network/distribution.html).
* Parraga-Alava, Jorge, and Mario Inostroza-Ponta. *A Quick Guide of Mstknnclust Package*, CRAN, 17 Sept. 2020, [cran.r-project.org/web/packages/mstknnclust/vignettes/guide.html](file:///C:\Users\dtstr\Downloads\cran.r-project.org\web\packages\mstknnclust\vignettes\guide.html).

**Appendix**

**Plot 1**

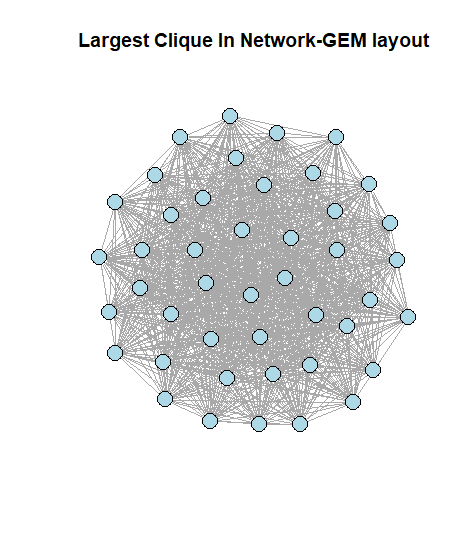
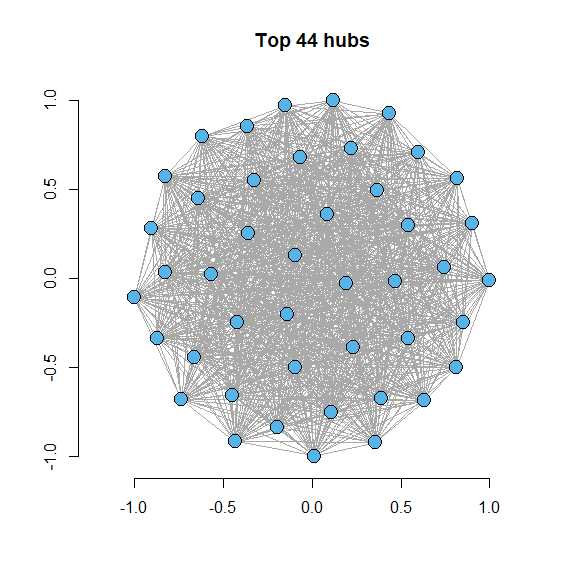
****

This plot helped us determine the best number of vertices to include in a random subgroup to compare to the top hubs subgroup. We wanted to see if the difference in graph density was attributed to the top hubs, or just because the top hubs took a subset of the network and that in itself was bound to lead to an increase in graph density. We cut off at 100 since after that results were close to zero. Below is the code sample of how we got this plot. We get a random subgroup of 1-100 nodes and plot the graph densities for each one to see which subgroup size is optimal. We want the best one so that we can compare the best subgroup from the original to the top hubs to really see if there is an important difference in connectivity between these subgroups. Below is the code used:



**Plot 2**

These are plots comparing the largest clique to a top hub subgroup of the same size. They are nearly identical, and both have very high graph densities since every node is connected with nearly every other node.

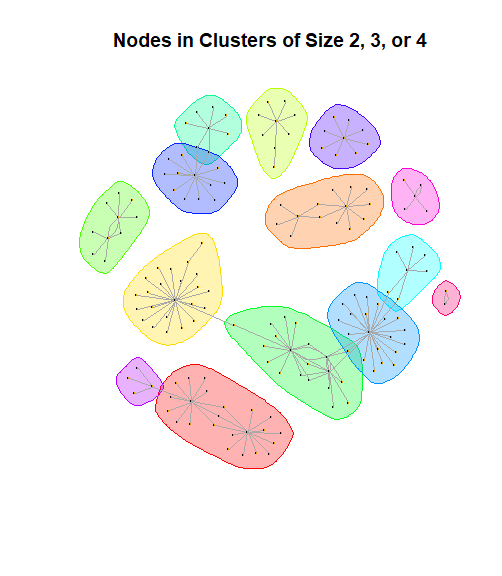


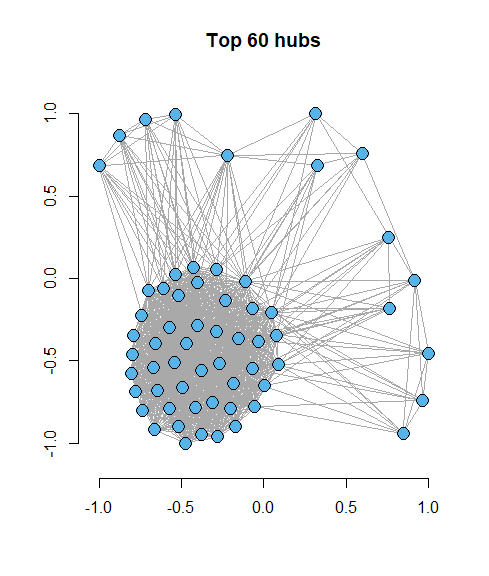
As mentioned in the main section, 43/44 of the vertices that they have overlap.

This table reads as: 177 clusters of size 2, 98 or size 3, etc. There are clusters that exist within larger clusters as well.

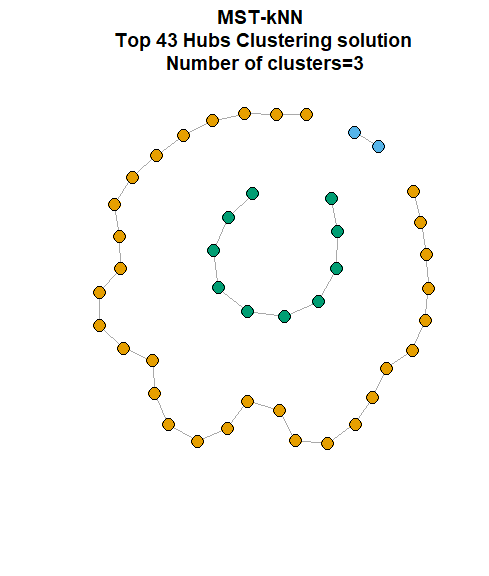
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 8. Clusters in the Network** | | | | | | | | | | | | | |
|  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **12** | **14** | **4158** |
| **Number of Clusters of Size** | 1 | 177 | 98 | 30 | 17 | 12 | 8 | 6 | 2 | 1 | 1 | 1 | 1 |

**Plot 3**

This plot shows how easy it is to identify groups in the network once we split it up by only including the nodes involved in the three most commonly sized clusters (sizes 2, 3, and 4). There are 166 unique nodes in this subgroup.

**Plot 3**

Plot of the Top 60 Hubs- once again we see a large proportion of nodes in their center densely connected network, and a few on the outskirts.

**Plot 4**

**Additional Plot with MST-KNN**- shows the clusters made with these 43 top hubs and how connected a majority are.