CSC 555 - Social Computing and Decentralized AI
Project Report (R3)
Team: Adithya Raghu Ganesh (araghug), Meghana Ravindra Vasist(mravind)

SOCIAL NETWORK ANALYSIS FOR YOUTUBE

ABSTRACT
Nowadays, people choose to interact with each other over social media rather than face to face. Facebook and Youtube are such platforms where people express their opinions and interests. In recent times, due to the Covid-19 pandemic, these platforms have developed a greater role in public interactions. As part of this project, we propose to analyze and find trends in the way these interactions happen and to find out how awareness can be spread through these media.

HYPOTHESES
Hypothesis 1: The number of subscribers in a community plays a significant role in the influence of the community across the overall network.

Evaluation of hypothesis: In order to calculate the influence of nodes in a network, we propose to use **Pagerank** as a metric. This measure can identify influence of nodes over the whole network by ranking each node in the network based on the structure of the nodes.

Hypothesis 2: Social entities who are part of the same communities are more likely to be socially related to each other.

Evaluation of hypothesis: To validate the hypothesis, the **cohesion** of the nodes are considered and checked to see if they follow the same communities. Cohesion gives a clear idea of how the extent of connectedness and togetherness among actors within a network is measured.

JUSTIFICATION
Hypothesis 1: Pagerank is a good all round social network analysis score because it assigns relative scores to all nodes in the network based on the concept that the connections to high scoring nodes have a higher weight in comparison to the connections to the low scoring nodes. It has a damping factor to penalize the nodes which are distant from the node under consideration. Since the hypothesis is to measure the influence of a node, Pagerank provides us with the data to validate or invalidate the hypothesis.

Hypothesis 2: This hypothesis proposes to find a relation between common bond and common identity groups by stating that people with similar interests are more likely to connect with each other not only on a group level but also on a personal level. Since cohesion gives us a measure of how tightly coupled two individuals are in the network, this measure can be used to validate or invalidate the hypothesis.

DESCRIPTION OF DATA (Youtube Social Network: Undirected graph)
We are comparing the graph structures within Youtube. Given below are the details of the dataset.

The data contains two columns in which the first column indicates the source node and the second column indicates the target node.
Contains 1.1M nodes and 2.9M edges

Network statistics	
Nodes	1134890
Edges	2987624
Nodes in largest WCC	1134890 (1.000)
Edges in largest WCC	2987624 (1.000)
Nodes in largest SCC	1134890 (1.000)
Edges in largest SCC	2987624 (1.000)
Average clustering coefficient	0.0808
Number of triangles	3056386
Fraction of closed triangles	0.002081
Diameter (longest shortest path)	20
90-percentile effective diameter	6.5
Community statistics	
Number of communities	8,385
Average community size	13.50
Average membership size	0.10

Ref: https://snap.stanford.edu/data/com-Youtube.html

The dataset also provides information about the different communities that the users are part of. Here, each row of the dataset contains all the nodes that are part of the community (which is represented as the row number). This helps us analyze and understand the common identity groups in the network.

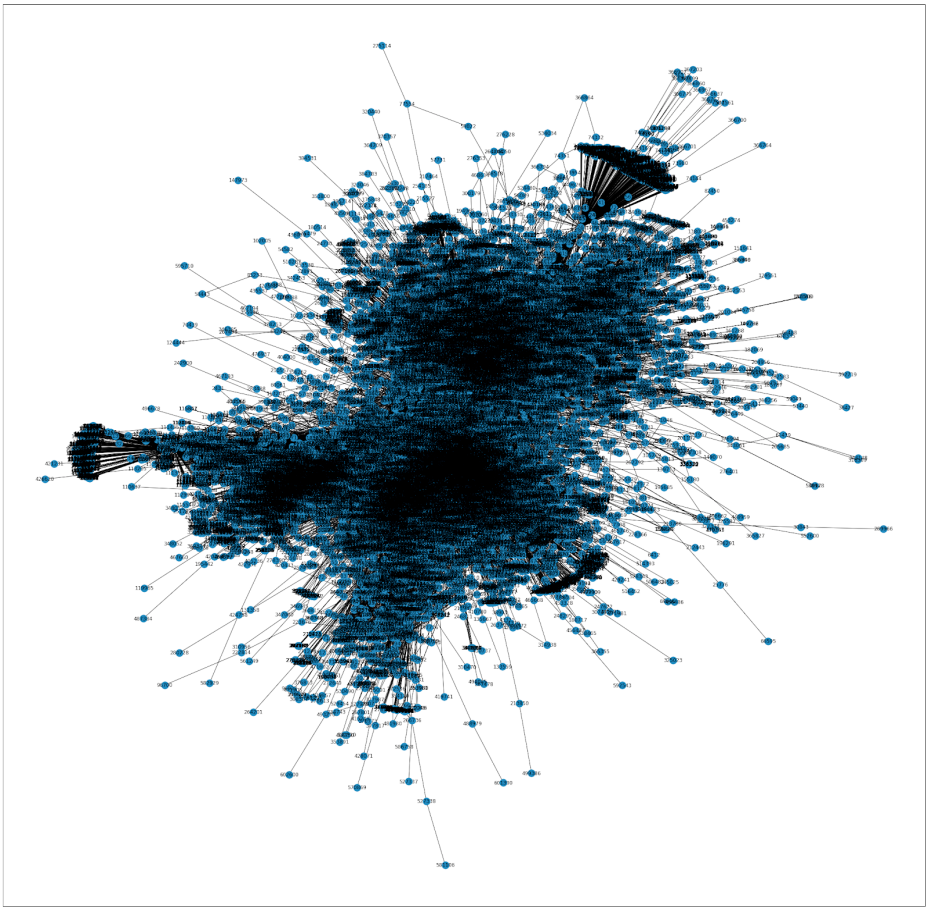
OBJECTIVE
There are many small scale influencers that are trying to spread goodwill and awareness amongst people in the society. But since these influencers do not know who the right audience is for the spread of information, the needed spread of important values and information are cut short. We are trying to find, analyze and identify stronger social ties that these individuals can work with for a better societal and communal spread.
Furthermore, in the current unprecedented times, there are fewer in-person interactions and such virtual campaigns are all the more essential for the betterment of the society. It is also true that social media campaigns can reach a wider audience than just a speech in a park.

DESIGN

We used the connections between individuals in youtube to analyse and understand their roles in the overall youtube network and also the interactions in the communities they are in.

We leveraged the networkx library to build the initial graph to understand how a sample graph would look.

An Overview of the Youtube Network



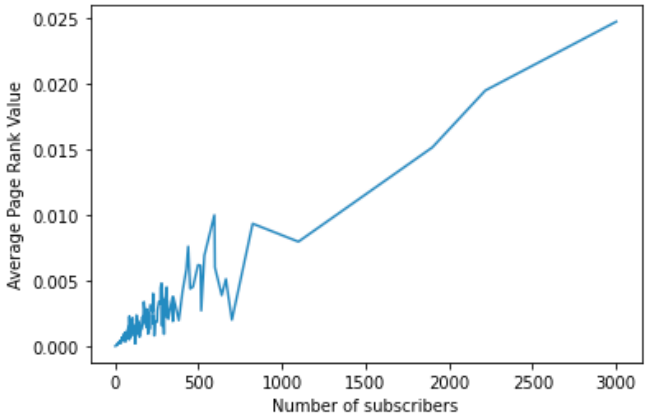
Below is the design for validating/invalidating the two hypotheses:

Hypothesis 1:

The given files for both youtube overall network and youtube communities were in textual format. We processed each file to retrieve nodes and edges and created the overall graph using the networkx library. We iterated through the communities and calculated the page rank of each node and assigned an average page rank for each community.

Upon generation of the average page rank for each community, we plotted a graph that related the page rank value to the size of the community.

Given below is a graphical representation of how the hypothesis about the size of the community and influence holds true.



Looking at the above graph, we can see that as the number of subscribers increases, the influence, on an average, increases. The graph produces a linearly increasing regression line even though the average pagerank value for consecutive subscribers is not very consistently increasing. This is due to the possibility that some nodes in some communities have greater influence. This will significantly increase the average pagerank of that community. But the overall graph seems to be non decreasing.

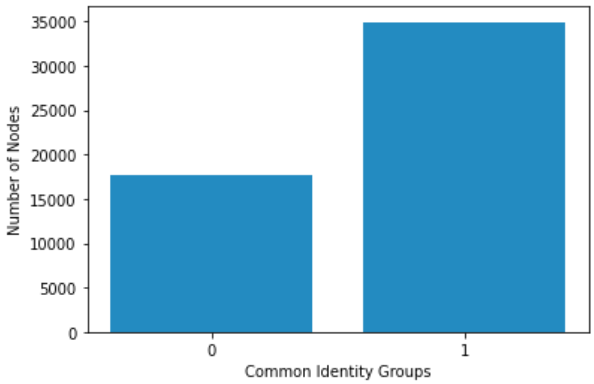
Examining the results, we can also conclude that the influence of the nodes also plays a role in the influence of the community along with the number of subscribers. In conclusion, we can say that the **hypothesis validates to be true**. Hence, we can say that the number of subscribers in a community plays a significant role in the influence of the community across the overall network.

***Hypothesis 2*:**

In order to validate/invalidate hypothesis 2, let us understand what common bond and common identity is: common bond relation is the individual relations between two individuals. This type of relation measures the closeness of two entities on a personal level and we cannot deduce anything with respect to a superficial group they may or may not belong to. Common identity relations between two entities describes the relation on a group level. Here, the two entities are connected to a common entity.

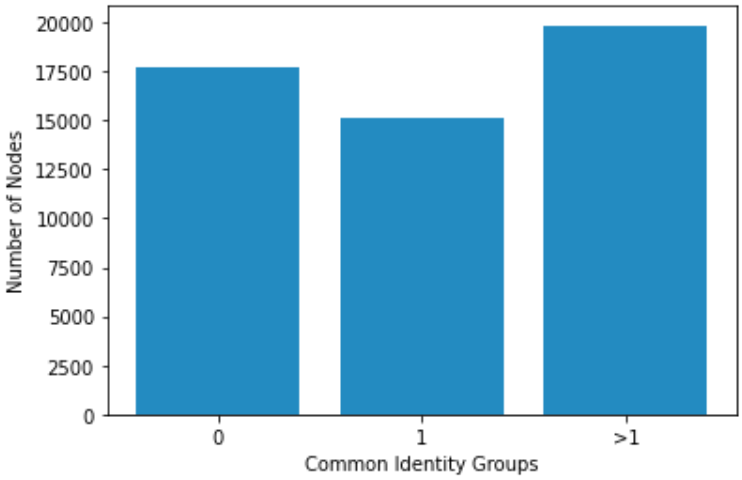
Coming to the proof for the hypothesis, we need to check if two entities are in the same community and then determine if two social entities are socially related to each other in the overall network or vice versa. For each node, we take the communities the neighboring nodes are in and check if there are common identity groups. This is done for all the nodes in the youtube network that have at least one community. We then identify all the pairs of entities that have a common identity relationship as well as a common bond between them.

Upon analyzing the results, we see a clear 2:1 ratio on the users that have common bonds to be part of common communities (common identities).



There are 52675 nodes and open analysis of the same, we realise that people with similar interests are more likely to be connected with each other as 34943 individuals or accounts had common identities in comparison to 17732 which is a clear minority. In the above graph, 0 refers to the number of people who are linked to each other in the overall network but are not in the same community and 1 indicates the number of people who are in the same communities and are also connected to each other in the overall network.

We wanted to bolster the hypothesis further by identifying the number of common identity groups present between the individuals that share common bonds to see the strength of the bonds they form. We notice that from the generated analysis that, the groups of individuals that share both common bond and identity do in fact have more common groups between them when there exists at least a single common identity group allowing us to conclude that when individuals share a common bond, there is a probability that they may indeed have around 8 common groups between them.



IMPLEMENTATION

The project is developed in Python due to the simplicity of usage of the libraries. The entire network is represented by a networkx graph. Networkx provides a wide range of algorithms to perform social network analysis. The following were used:

- Python 3.6 - An interpreter
- Networkx - To represent the social network and perform social network analysis
- Statistics - To calculate the mean, median and mode
- Collections - To construct defaultdict
- Pandas - To read the dataset from the file and write to a dataframe
- Matplotlib - To plot the graphs and visually represent the
- Google Colab - For collaborative development

Data: The data was in a text file which contained the links between nodes. This was first converted to a csv file which was read to a dataframe object in the code using pandas.
In order to access the communities, each row of data in the text file was read into a dictionary with key as the community number and value as the list of nodes that belong to the community.

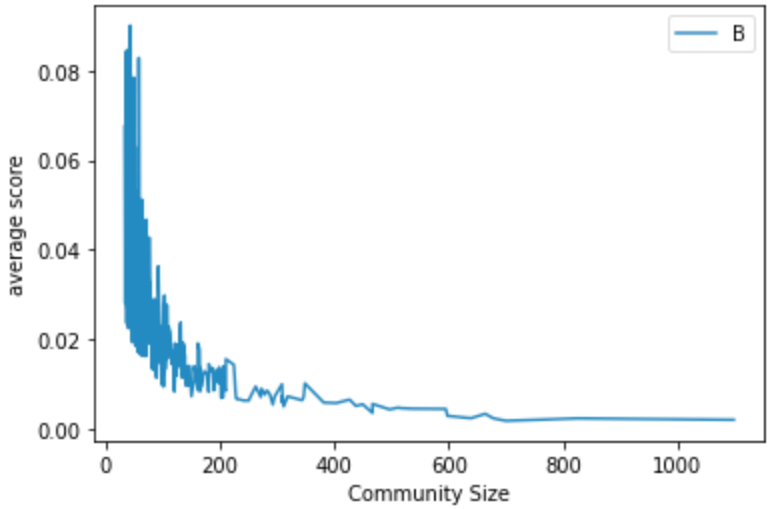
Collaboration between team members: The team members collaborated on Zoom. We divided the work equally and developed the project in a Google Colab notebook. After performing the SNA, we plotted the results using matplotlib. The results are shown under the Results section.

CHALLENGES FACED

- Katz centrality - In our first design, we said we would validate/invalidate Hypothesis 1 using Katz centrality instead of pagerank. We changed to pagerank because it is faster, more efficient and economical.
- Size of dataset - The dataset contains about 1.1 million nodes which makes the computation expensive and highly time consuming.
- Small world coefficient (omega)- There are about 16000 communities in the dataset. To calculate the omega for all the communities taken as subgraphs takes more than a day. Hence, we could calculate omega for one of the larger communities in the network which has a size of 136 nodes. This took about 2 hours to run and give the output.
- Processing limitations - The systems that we used for development had restricted resources. Hence, we started using Google Colab but those resources were also not sufficient to run all the algorithms for all the nodes. Hence, we had to run the algorithms on subgraphs of the overall graph.
- Read/Write graphs: We tried to read and write the networkx graph but the data was not in the correct format that was required by the different read/write methods given by the networkx library.

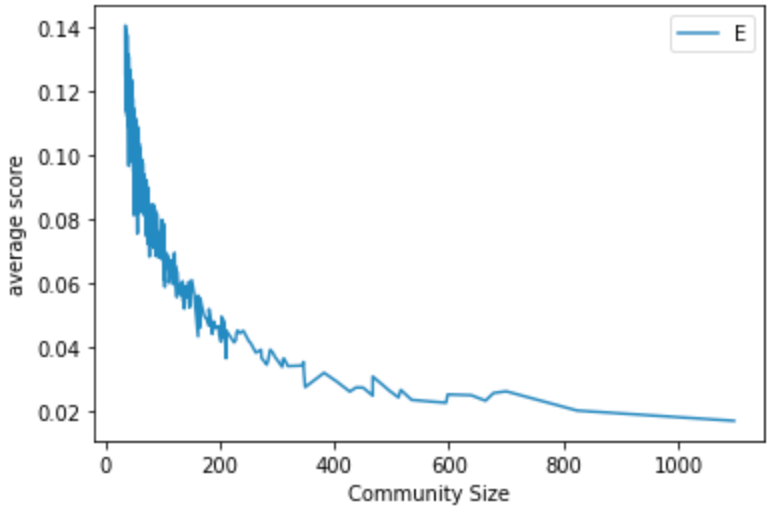
RESULTS AND INTERESTING FINDINGS

- **Betweenness centrality:** The betweenness centrality for all the communities in the overall network were calculated and the following result was obtained:



Here, the x-axis is the size of the community and the y-axis is the average betweenness centrality of all the communities of that size. The above graph is plotted only for the top 500 communities in order to better depict the trend. It can be seen that the communities with larger sizes have lesser betweenness centrality than the smaller ones. This shows that, in larger graphs the connections between the nodes doesn't imply that the whole network is connected. For example, if a graph contains 3 nodes and two of them are linked to each other, then the betweenness centrality of the graph is high. If a graph contains 100 nodes and 20 of them are connected, the betweenness still remains lesser than the first graph due to the high number of nodes.

- **Eigen centrality**: The eigen centrality for all the communities in the overall network was calculated and the following result was obtained:

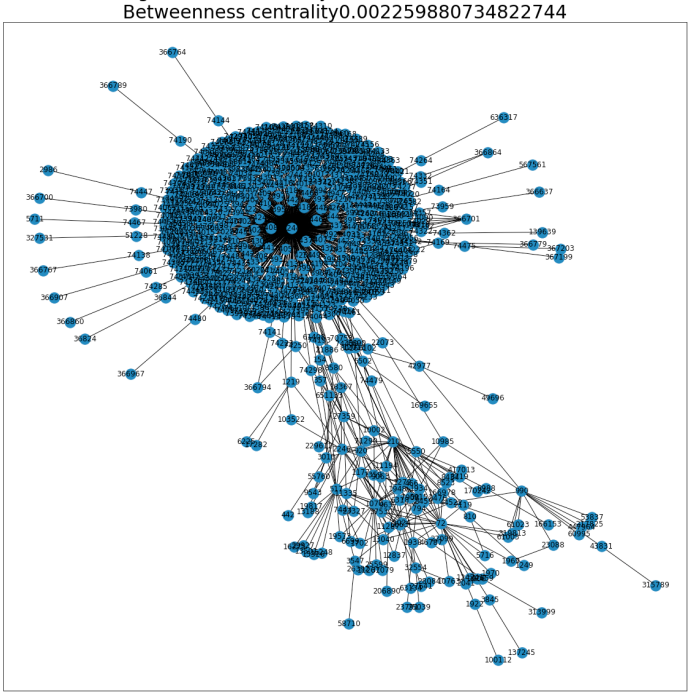


In the above graph x-axis is the size of the community and y-axis is the average eigen centrality of all the communities of that size. The above graph is plotted for the top 500 communities in order to better depict the trend. It is seen that the graph follows the same trend as before.

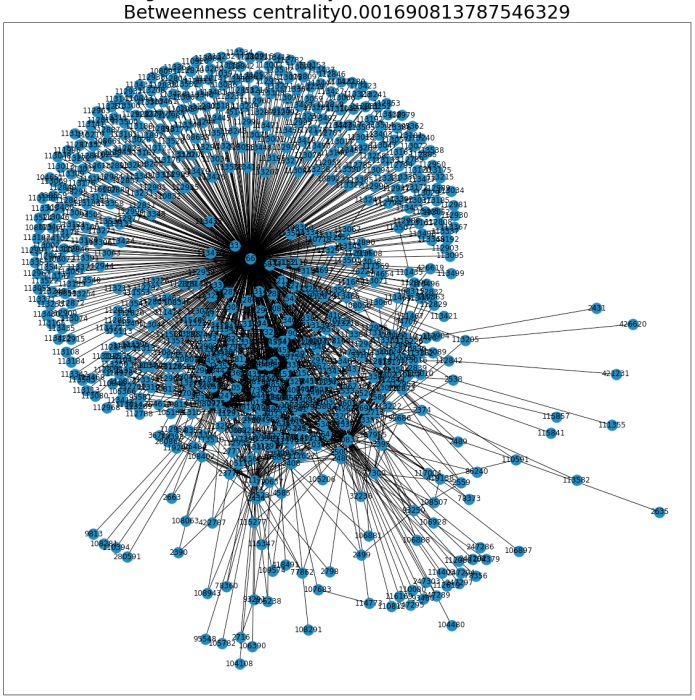
- **Different subgraphs of the overall graph:** We plotted the overall network which is shown in the design section of the report. From that graph, we can see some spots that are densely populated but nothing more can be inferred from it other than them belonging to the common communities.

  Hence, we decided to identify popular communities and generate a subgraph to see how the interactions between users would be. We noticed several patterns in these communities.

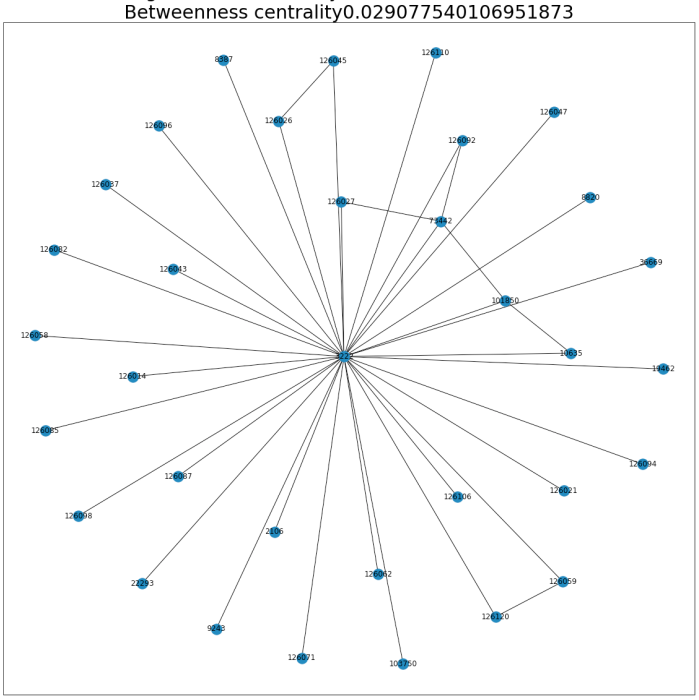
Eigen vector centrality: 0.025768536740373767
Betweenness centrality0.002259880734822744



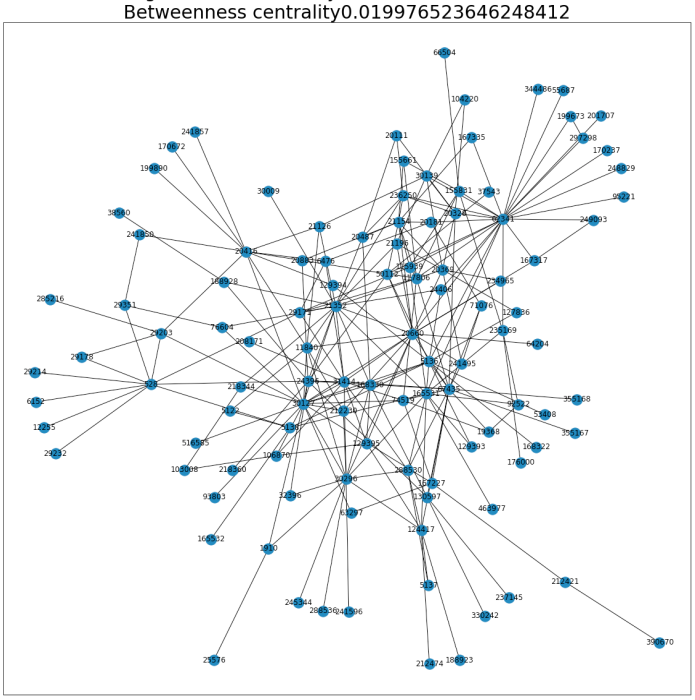
Eigen vector centrality: 0.02630796405982042
Betweenness centrality0.001690813787546329



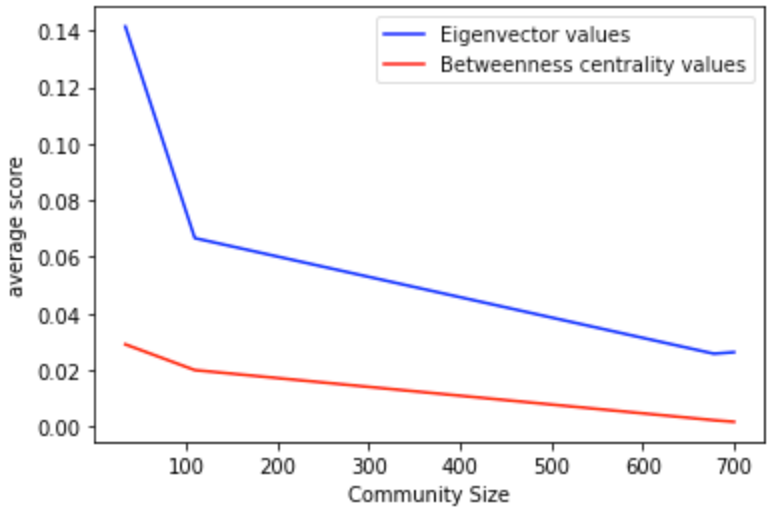
Eigen vector centrality: 0.14143534689526416
Betweenness centrality0.029077540106951873



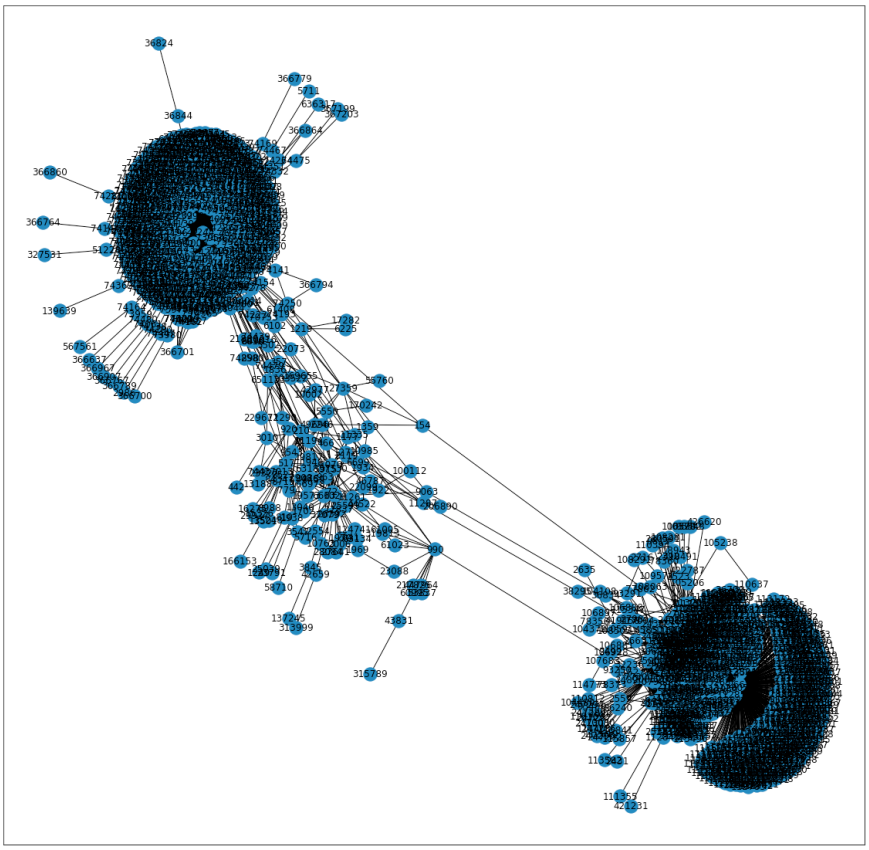
Eigen vector centrality: 0.06662796317358215
Betweenness centrality0.019976523646248412



As we can see, the values computed for the different communities are also plotted. This follows the same trends as the graphs previously displayed with the value decreasing as the community gets larger.



- **Common bonds in the network:** In order to study the influence of one type of community over another, we plotted the graph for two communities with predominantly different reach:

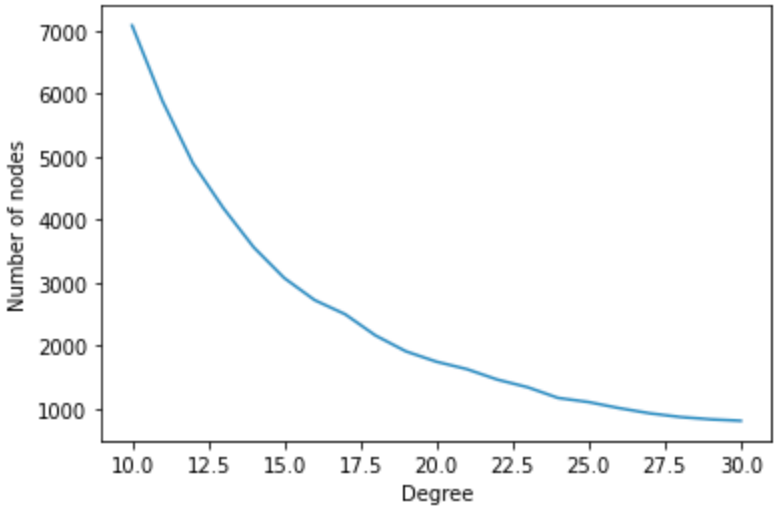


In the above graph, the dense cluster on the left is the for example, a tech community and the dense cluster on the right, a non-tech cluster. Due to some common bonds between a portion of the users in the communities, there is a connection between these two communities.

- **Clustering coefficient:** We calculated the average clustering coefficient of the overall network. The value obtained was `0.10718`. The clustering coefficient is low which means that the nodes in the network are not very well connected but are connected to each other to an extent.

- **Small world coefficient (Omega):** Normally, one would think that the youtube network is a small world network. In order to verify that, we chose a relatively large community of size 136 and calculated the omega value of this subgraph in the overall network. We found that the omega value was `0.61617`. Since 0.6 is close to 1, this subgraph is a random graph and not a small world graph.

- **Relation between degree and number of nodes:** We represented the relation between the degree of a node and the number of nodes that have the same degree associated with them.



Here, x-axis represents the degree and y-axis represents the number of nodes. It is surprising to see that the number of nodes that have a lower degree are higher and as the degree increases, the number of nodes that have that degree decreases.

CONCLUSION

From the analysis shown above, it can be concluded that the Youtube social network graph constructed using the dataset is a slightly random graph with an omega value of 0.6. Based on the measures shown above, we can see that there are certain aspects which might seem one thing to the naked eye but when statistical measures are used to evaluate these aspects, it results in concluding a completely different thing altogether. This tells us the importance of the way the ordering and depiction is done is very important in determining the characteristics of the network.

We can also formulate various hypotheses on the network and try to address this based on statistical measures such as centrality measures, pagerank algorithm, small world network analysis etc, which will help prove/disprove the hypotheses. This gives a better understanding of the network on a more concrete level rather than just basing your conclusions on mere observations.