Mini Project 3 Part 1 *Kevin Koljonen*

The question:

I want to find out how connected/similar english speaking twitch streamers are and if there are separated clusters of streamers. I would also like to check how popular these clusters are in terms of views.

Importance:

Twitch viewers and streamers alike might find the information visualized to help with making contact with certain communities and can show who is connected to who. Essentially this can be a map of contacts and can compare how high viewer streamers affect smaller ones.

Approach:

Using google colab to check the network density and the average path length for all possible node pairs. I will attempt to answer the question using NetworkX and by following the examples made by Sepinoud in her lectures. Using NetworkX I will visualize the network of the dataset. This shows both the average path length and density of groups. The average path length can just be checked by adding all the path lengths together and dividing it by the amount of edges. The visualization of the paths will also show how similar certain “communities” are by looking at how close knit the nodes are. Network density can also be calculated using the networkX density function. Using the datasets features and displaying each node with the views they’ve accumulated I should be able to see which streamers are more popular. The network could be sorted by each feature too to see even more how similar twitch streamers are. Like how long they’ve been streaming for, if they’re partnered or not etc.

Expected outcome:

My expectation for this project is that a majority of twitch streamers will be largely similar to one another. Although there may be outliers I expect the majority to have a dense network. I expect there to be many separate dense clusters of nodes. Some clusters will be more popular than others and many of them will probably be less popular rather than more popular.

Dataset used:

I'm using a Twitch Social Networks dataset which for the english dataset includes over 7,000 nodes and over 35,000 edges.

https://snap.stanford.edu/data/twitch-social-networks.html

Mini Project 3 Part 2

Methodology:

The dataset comes in two csv files, one with the target data and the other with the edges of the network. At the start of the project I've loaded the csv's into a networkX graph. First the 'new_id' is set as the nodes because the edgelist uses the 'new_id' attribute to connect nodes. Then the rest of the attributes, such as 'maturity rating', 'twitch partner', 'views' and 'days streaming' are then assigned to dictionaries and those are attributed to the nodes in the network. Network density, average shortest path length and degree of connectivity is checked which say how connected and dense the entire community is. Then visualization is done and combined with EDA the dataset is narrowed down to nodes with over 100000 views. The visualizations answer the main question being posed.

Results:

The network density of the dataset is ~0.0014

```
Network density: 0.0013914550620165345
```

This means that the network is not very dense.

The average shortest path length on the entire dataset is ~3.68 which means that to get from node to node it takes an average of over 3 steps.

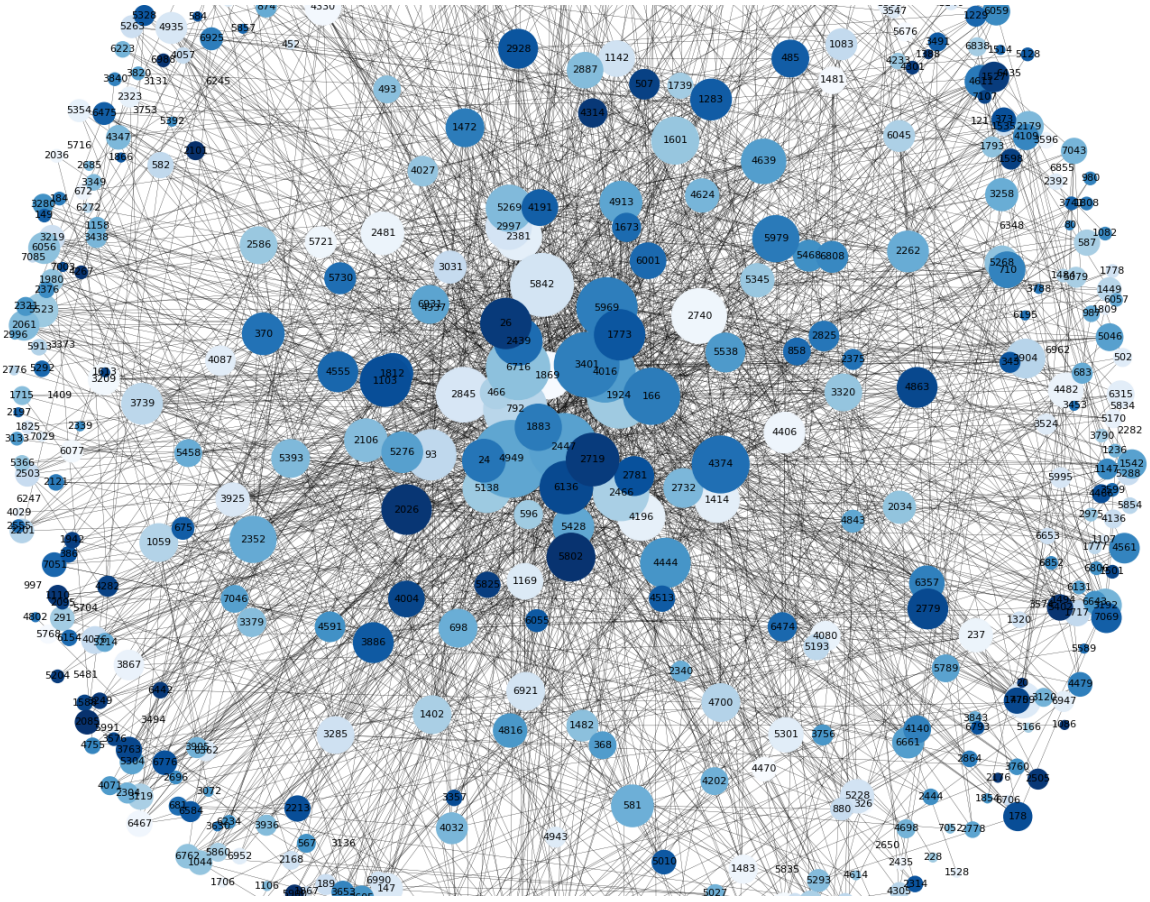
```
[53] nx.average_shortest_path_length(G)

     3.6776157289097005
```

I also check the degree of connectivity between nodes and the values outputted are on average around 50 degrees in the full dataset

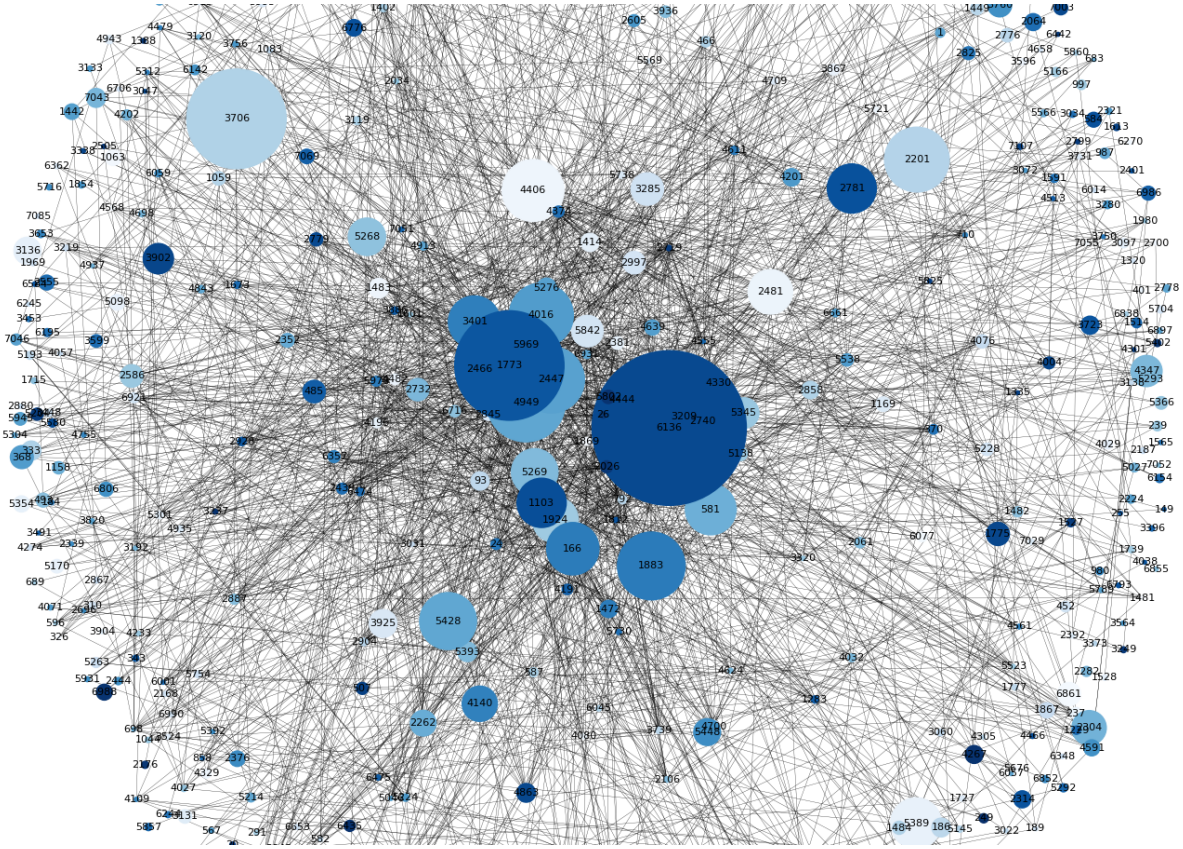
```
{7: 78.6602725896012,
 19: 52.851851851851855,
 9: 75.31616161616162,
 3: 98.41655716162944,
 2: 96.51634320735444,
 14: 66.76112412177986,
```

The visualization of the entire dataset proved to be unreadable because of the amount of nodes and edges so EDA was done and first I removed nodes that had under 100000 views and then assigned node size based on the degree of connectivity, and then the view count

Degree of connectivity:



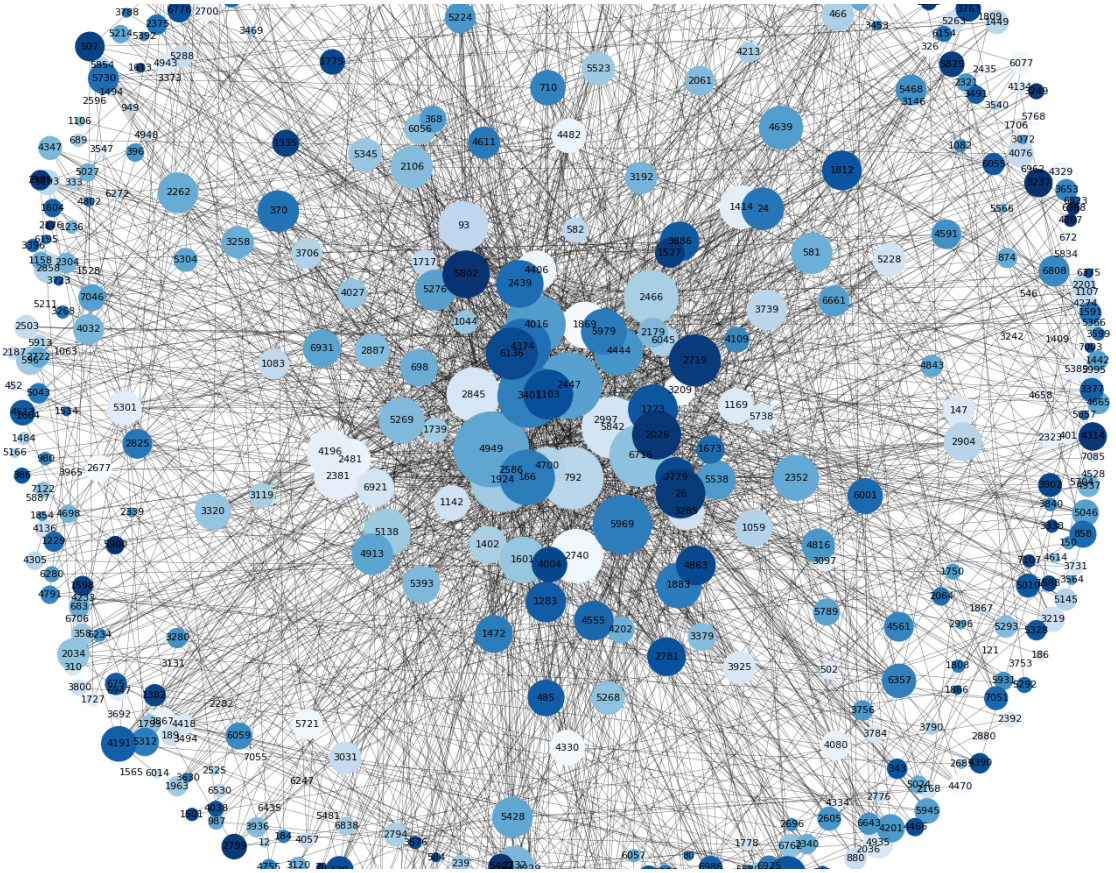
View count:



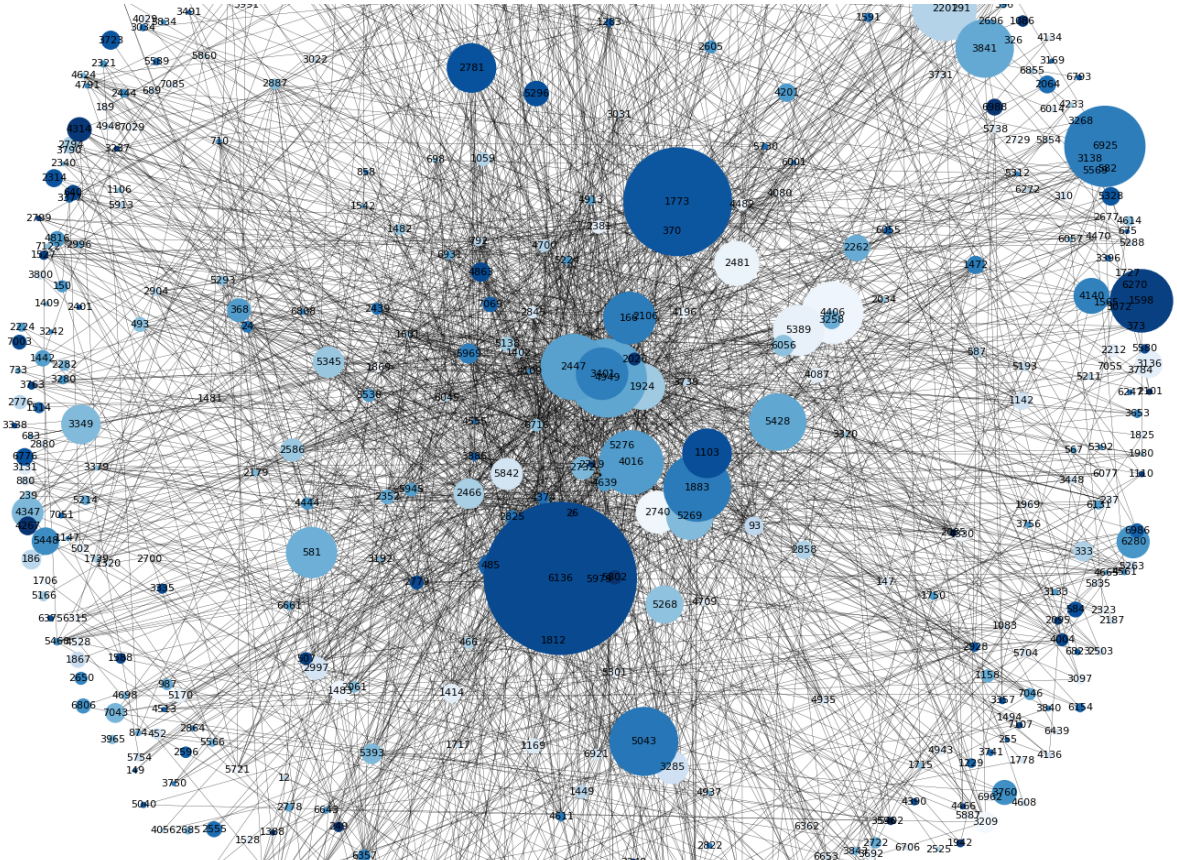
The graphs are still a bit unreadable but now you can see some clustering of nodes on the edge of the network and also in the middle where the high connectivity and high view count nodes are.

To make the graph a bit more readable, nodes with only 1 degree of connectivity have been removed and then the same visualization has been done again.

Degree of connectivity:



View count:



These results also show clustering of nodes with bigger and smaller view counts on both the edges of the graph and the middle of the graph. Although the K value of the spring layout in networkX makes the graph into a circle, the clustering of the nodes on the edges should still be correct.

The network density and average degree of connectivity is also checked again after all the filtering and the results now look like this

```
Network density: 0.016012972534711664
```

```
{7: 20.529761904761905,
 5: 20.823809523809523,
 28: 23.133928571428573,
 4: 21.96875,
 9: 20.625,
 11: 19.335664335664337,
 3: 21.094202898550726,
```

Now the network density has increased to ~0.016 which is higher than in the entire network, but still a relatively low density. The average degree of connectivity has decreased from before and now the average is around 20 degrees

Conclusion:

I think that the results show that English speaking streamers are relatively spread out and some of them form smaller clusters of communities. This, I think, supports my original hypothesis. The easiest part of this project was loading the dataset into networkX and the most difficult part of this project was trying to get the visualizations to look good and have the node size set according to an attribute of the node itself. I think the project could still be improved by still narrowing down the graph size or finding another way to visualize the entire dataset in a comprehensive way.