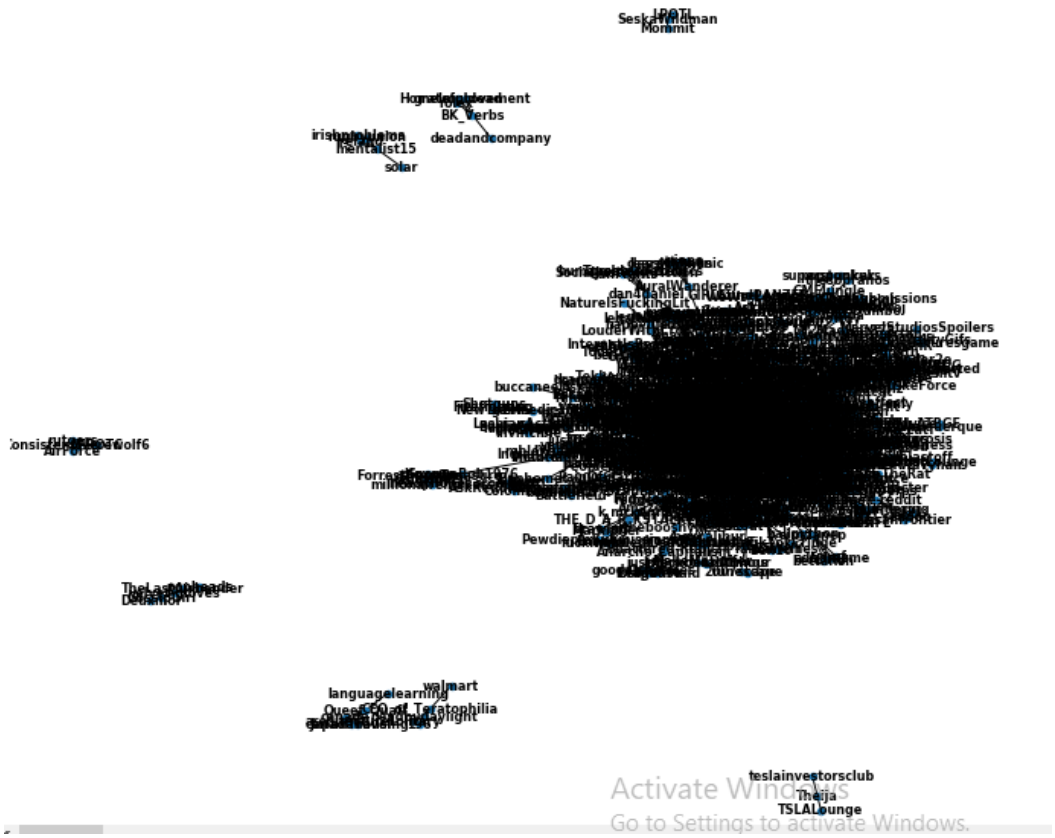
Project 1 Report

Jackson Wakefield - jswakefi - 1216414420

1) **Data-Mining**
    a) In this project, I had to find a way to connect users of Reddit to one another through sets of nodes. This was not as easy a task for Reddit as it would have been for most other social media platforms, as Reddit does not have explicit connections between "friends" like most platforms have. Instead, all I had to work with were users' comments, frequently visited subreddits, etc.
    b) In order to connect the users to one another, I decided to acquire a list of 250 users and their most recent 50 comments on Reddit. Then, I poured through each of the 50 comments and started a tally for which subreddit the comment was made on. Through this, I was able to get a 4500+ line json file containing thousands of individual subreddits that people have recently commented on (to be used as connecting nodes).
2) **Network Graph**
    a) Nodes represent: User and Subreddit - Users can be connected through Subreddits. If two users have both commented in a similar subreddit, for example r/aww, they will be connected through intermediary node r/aww.
    b) When beginning to create my network graph, I found that the sheer volume of my graph was too high to not weigh the number of times a person comments in a particular subreddit, so I raised the cap for inclusion into the network up to anywhere between 3-8 comments to create an edge between a user and a subreddit. The lower I raised the bar, the more I found that the main cluster of nodes increased in magnitude, and the outlying clusters began disappearing. Setting the bar to 20 or more comments in a subreddit erases the cluster entirely, and replaces it with tens to hundreds of outliers and smaller clusters.
    c) The following are the graphs i get with different values of my minimum comments:

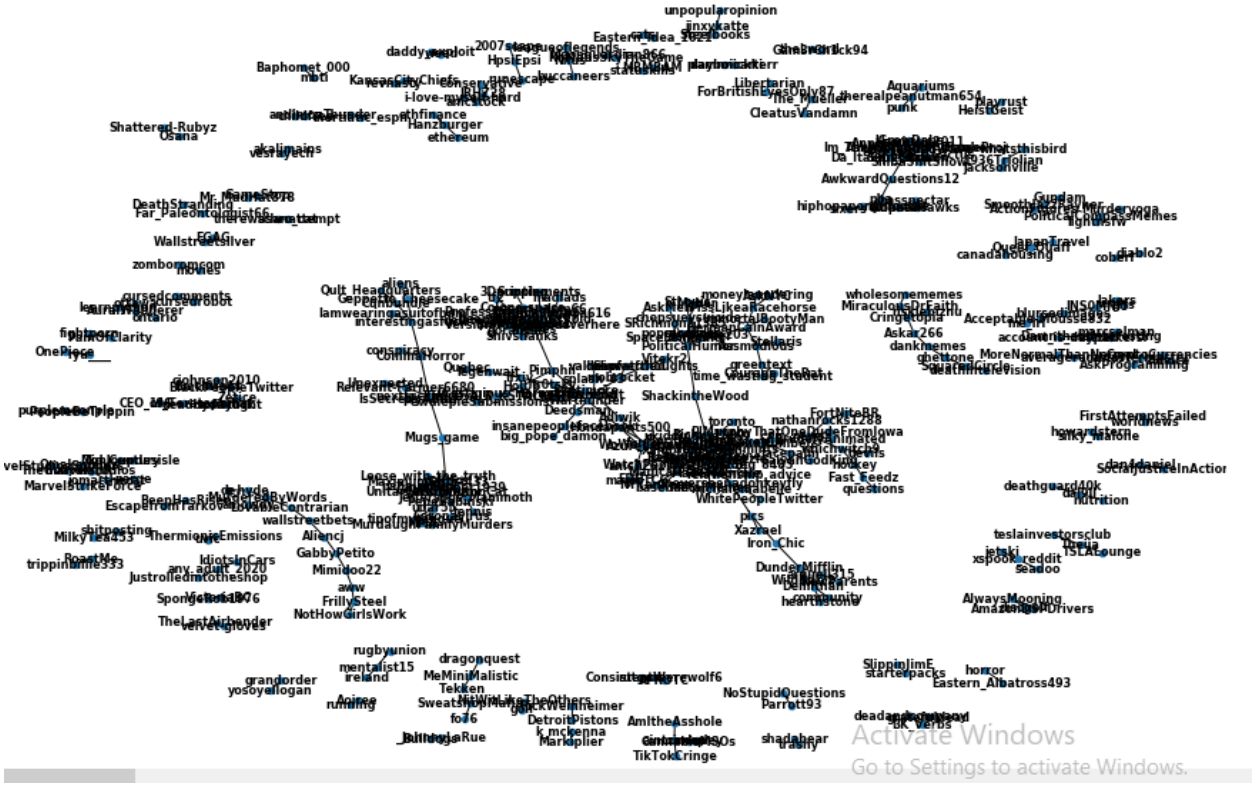
i) Minimum comments: 3



(1) Few outside clusters or outliers, massive inner cluster

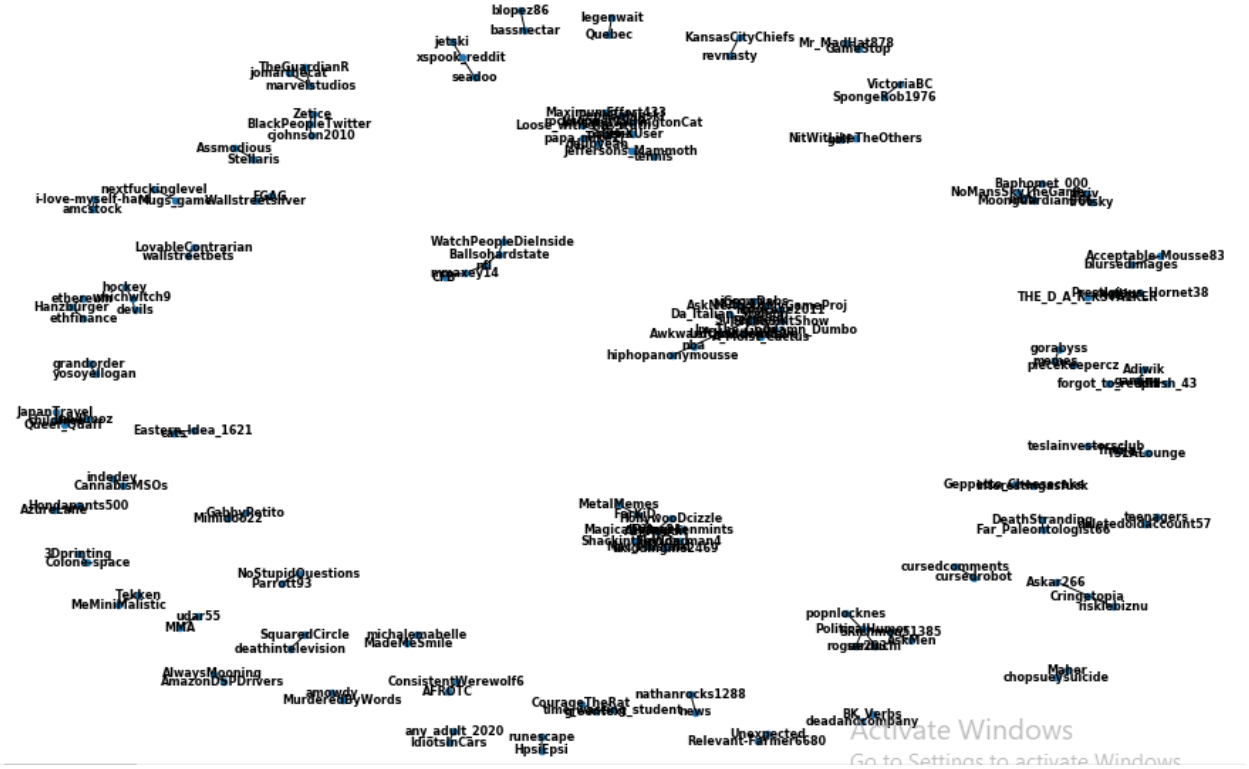
ii)

iii) Minimum comments: 8



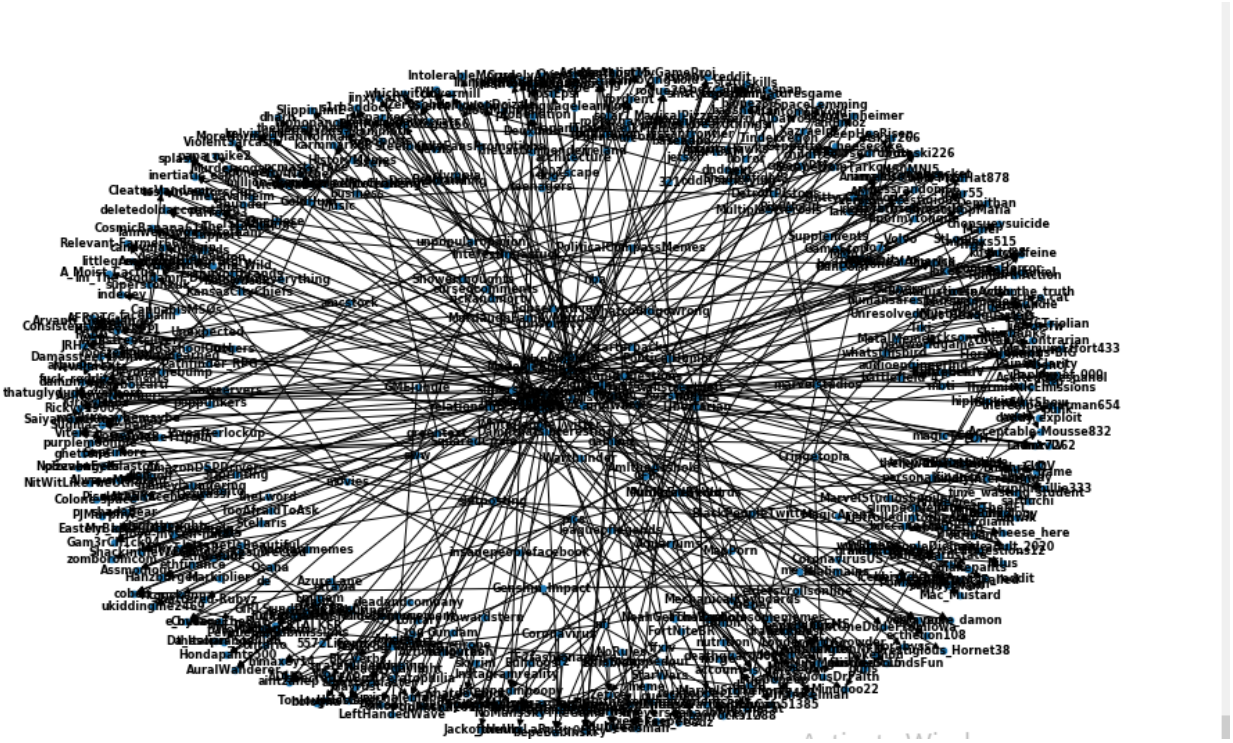
(1) Emergence of the mega-cluster, several small single-linked clusters and many outliers

iv) Minimum comments: 15



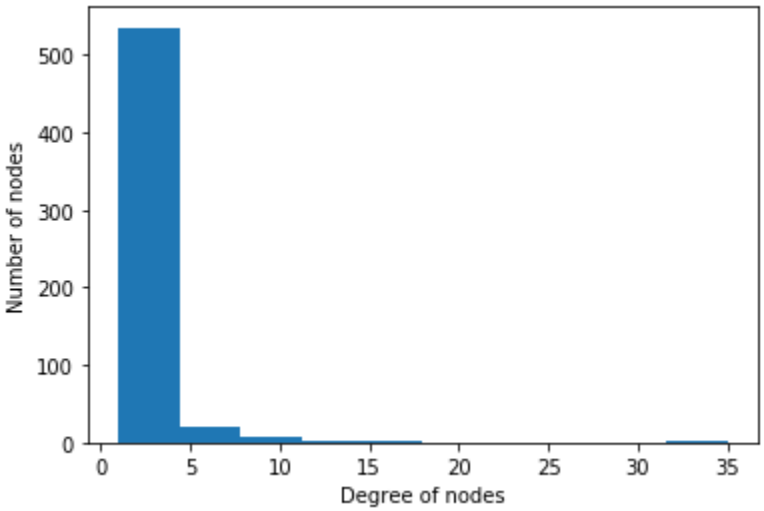
(1) Few small clusters, mostly consists of outliers or tiny-clusters

v) Minimum comments: 5 as a directed graph from user -> subreddit



vi) Minimum

3) Degree Distribution



a) What this graph tells me is that despite this network graph requiring a minimum of 5 comments to be added as an edge, 5 comments generating a very large

cluster, the overwhelming majority of nodes still have a degree of under 5. This means that people do not post more than 5 comments in very many subreddits. Since the volume of the large cluster is so high, however, it is safe to assume that very few subreddits comprise most of people's commenting habits. It is a widely used saying that the top 1% of subreddits dominate the site, and this histogram supports that notion.

4) **Pagerank**
    a) The pagerank was certainly one of the more interesting metrics I attempted for this graph. The pageranks are all very small of course, since there are hundreds of nodes in the graph and they aren't connected well enough to have a high pagerank, but these are the largest pagerank nodes in the set *pardon the french, can't stop reddit from being reddit*:

```
interestingasfuck: 0.005882182317226995
HolUp: 0.005141380273710755
WhitePeopleTwitter: 0.005178420375886566
AskReddit: 0.016290451028630182
nextfuckinglevel: 0.006900785127061827
memes: 0.009363951921753329
Superstonk: 0.00954915243263239
politics: 0.010827035957697906
PoliticalHumor: 0.005511781295468875
```

        i) These metrics did not surprise me whatsoever. All pageranks for individuals were low (as to be expected, I only traversed the last 50 comments, so it would be impossible to have more than 10 nodes attached to any single user), and generally, the pageranks for subreddits were much higher. Continuing on the notion that Reddit is dominated by a scarce few subreddits, these communities rank among the highest subscribed communities in the app, so it is no wonder that "AskReddit," one of the apps top 3 most trafficked communities, dwarfs most of its competitors in pagerank.

5) **Betweenness Centrality**
    a) I was not planning on using both Betweenness Centrality and Pagerank, but I decided to give it a shot and came up with some great data. The following is the highest scoring nodes in Betweenness Centrality:

```
Damnthatsinteresting: 0.06836619295991164
AskReddit: 0.3149008316089728
CosmicBanana616: 0.07170990513013129
memes: 0.11684778570153156
bruteski226: 0.105301308432536
ufc: 0.0929717822485925
nba: 0.09168961176268822
Superstonk: 0.07525225804013452
funny: 0.05664397316855923
politics: 0.08091745313761294
hiphopanonymousse: 0.09164190753430299
jackatman: 0.06301761268337018
AwkwardQuestions12: 0.07539397604021011
```

b) Here we see some of the same subreddits that dominated Pagerank. The interesting part about Betweenness in particular is the inclusion of individual users as well as subreddits. This did not make any particular sense to me, as a subreddit like AskReddit or Politics should have many more than 10 nodes, and therefore is connected to many more individuals. However, when I looked in my JSON for the names of the specific people who are challenging subreddits in Betweenness metrics, I found the source of their large showings. It turns out that each of the users that have high betweenness ratings have nodes pointing to most if not all of the subreddits that scored high in Pagerank and Betweenness. This means that using betweenness metrics, I accidentally tricked the Betweenness algorithm into thinking that "CosmicBanana616", who has made more than 5 comments to 4 of the subreddits listed above, is actually connected closely to the other users of these massive subreddits. Basically, these users are piggybacking off of the success of multiple of the most Between nodes in the graph by linking themselves to some or all of them. By linking himself to r/AskReddit, r/Memes, and r/Funny, "CosmicBanana616" has scored himself a higher Betweenness rating than r/Funny itself, which just so happens to be another one of the top 3 most trafficked subreddits. Similarly, "AwkwardQuestions12" has linked himself to r/NBA and r/Superstonk, and guaranteed himself in the top 20 most Between nodes in the graph. I found this absolutely fascinating as it highlights the main correlation problem when using Betweenness over Pagerank in directed network graphs.