CS 579: Online Social Network Analysis

Project I - Social Media Data Analysis

Group 45
Vivekanand Reddy Malipatel (A20524871)
Mohammed Shoaib (A20512491)

We have chosen **Reddit** Social Media platform to crawl the data.

In this project, we crawled data from Reddit using the PRAW API and created a social network based on the comments made by users on a selected subreddit. We then analysed the network using the NetworkX package in Python and calculated several network measures, including the degree distribution, clustering coefficient, and betweenness centrality.

Data Collection

To crawl data from Reddit and create the social network, we will use the Python Reddit API Wrapper (PRAW) library. PRAW is a Python library that makes it easy to access the Reddit API and retrieve information about Reddit posts, comments, and users.
Here's an outline of the steps we followed to crawl Reddit data and create a social network:

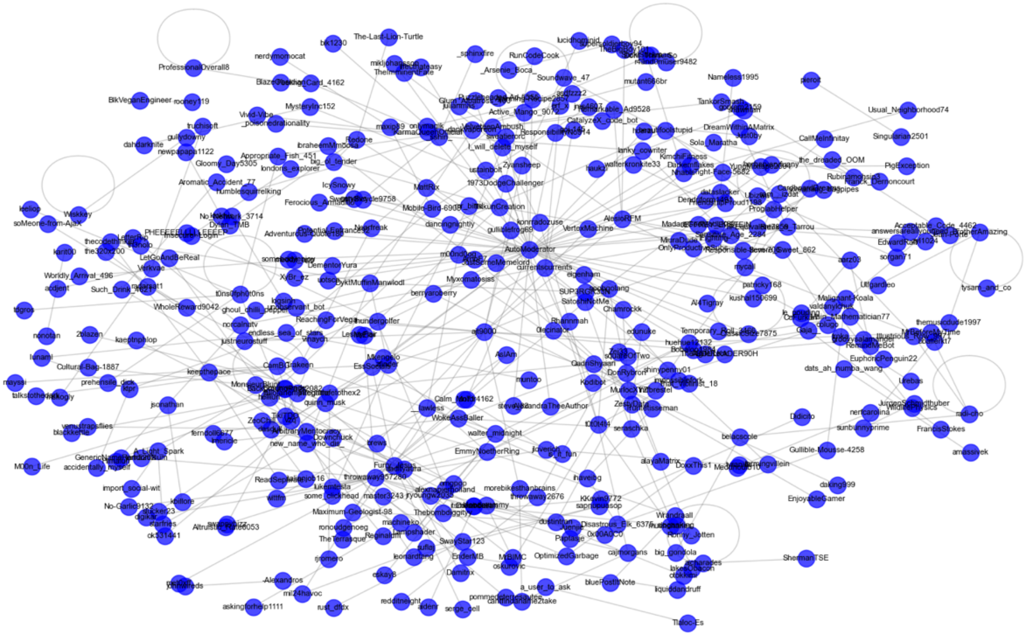
- Get the Reddit API credentials: Before we can start using the Reddit API, we need to obtain the Reddit API credentials, including the client ID and client secret.

- Authenticate with the Reddit API: We will use the Reddit API credentials to authenticate our PRAW client and start interacting with the Reddit API.

- For creating a social network with 100-500 nodes, we can extract the data for a specific subreddit and store the information about the users who submitted the posts as nodes and the relationships between them as edges. We store the data into csv files namely, nodes.csv and edges.csv.

Data Visualization

We used the NetworkX package to create a social network graph based on the comments made by users on the selected subreddit. We visualized the graph using the `matplotlib` library.

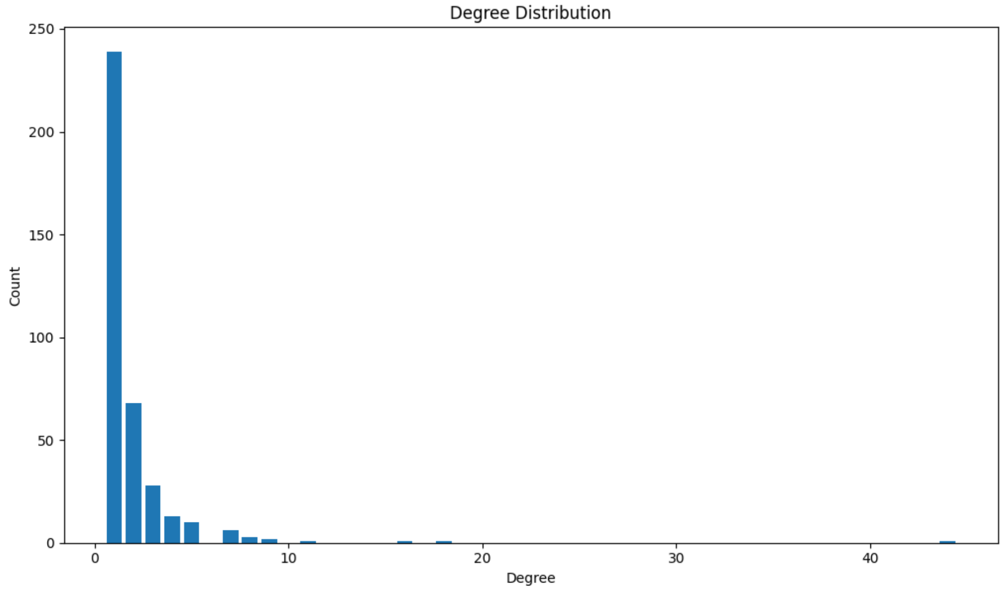


The above image shows the social network graph we created. Each node in the graph represents a user, and an edge between two nodes indicates that one user replied to another user's comment.

Network Measures Calculation

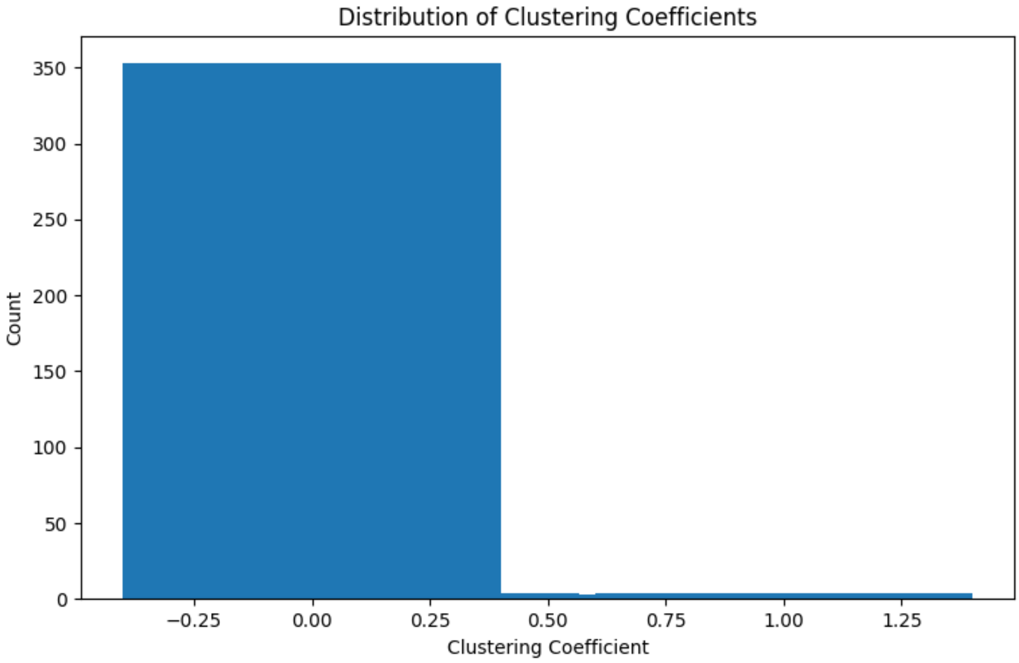
we first load the graph from the edge.csv file and then calculate the clustering coefficient and betweenness centrality for each node in the graph. We store the results in clustering_coeffs and betweenness, which are both dictionaries where the keys are node IDs and the values are the respective measures.

We calculated several network measures using the NetworkX package. First, we calculated the degree distribution of the network, which shows the number of nodes with a given degree.



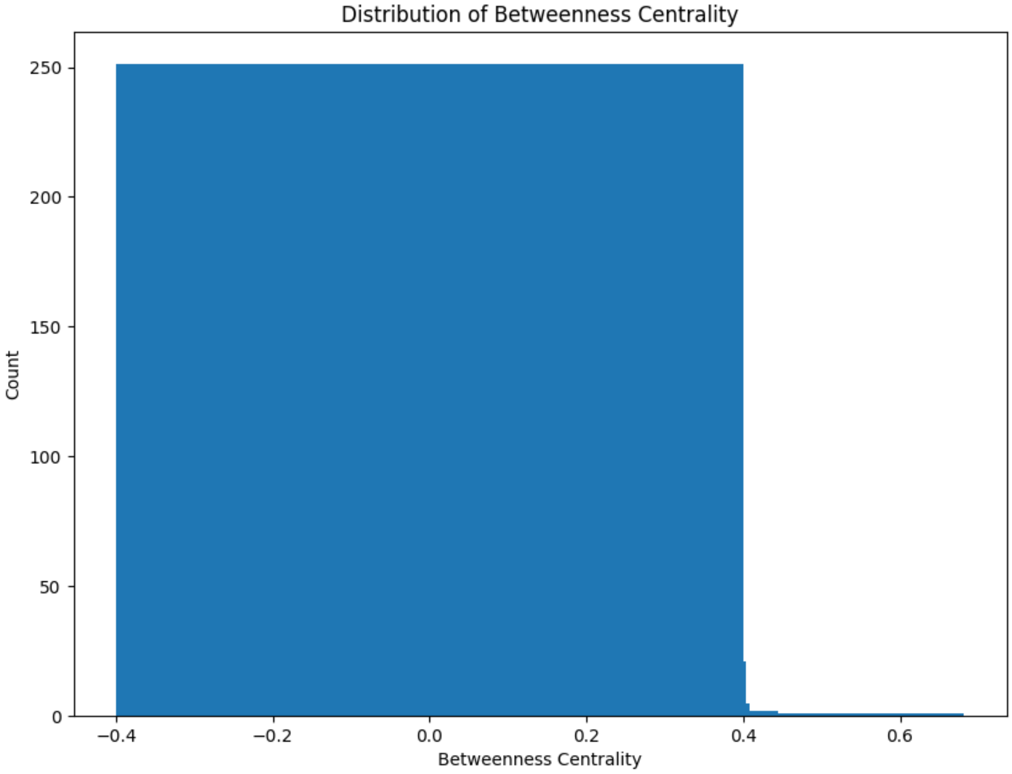
The above image shows the degree distribution of the network. As we can see, the majority of users have a low degree (less than 5), but there are a few users with a very high degree (over 10).

We also calculated the clustering coefficient of the network, which measures the degree to which nodes in the network tend to cluster together. The clustering coefficient of a node is the proportion of its neighbours that are also neighbours of each other.



The above image shows the distribution of clustering coefficients for the nodes in the network. We can see that most nodes have a low clustering coefficient (less than 0.5), indicating that they are not well-connected to their neighbours and few have over 0.5 to 1.25 indicating that they are well connected.

Finally, we calculated the betweenness centrality of the nodes in the network, which measures the extent to which a node lies on paths between other nodes. Nodes with high betweenness centrality are important "bridges" between different parts of the network.



The above image shows the distribution of betweenness centrality for the nodes in the network. We can see that most nodes have a low betweenness centrality.

Conclusion

In this project, we crawled data from Reddit, created a social network based on the comments made by users, and analyzed the network using several network measures. The results of the analysis showed that the network has a few users with a very high degree, but most users are not well-connected to their neighbours. There are also a few nodes that are very central to the network, indicating that they may be important "bridges" between different parts of the network.

References

Packages: Pandas, Networkx, Praw, matplot.lib

Software: Jupyter Notebook, VS Code, git

Web Reference:
https://towardsdatascience.com/scraping-reddit-data-1c0af3040768
https://www.reddit.com/prefs/apps