## CS 579: Online Social Network Analysis

## Project 1-Social Media Data Analysis

Pratik Jadhav (A20543728)

Shreenidhi R (A20513287)

## Introduction:

In the digital age, online social networks have become a prominent platform for individuals to communicate, share information, and express opinions. Among these platforms, Reddit stands out as a popular forum-based website where users can create communities (subreddits) based on common interests. Understanding the dynamics of interactions within such online social networks is crucial for various applications, including targeted advertising, content recommendation, and community detection. In this research, we focus on analyzing the social network structure of Reddit, with nodes representing authors of posts and authors of comments on these posts. By examining the relationships between these nodes, we aim to gain insights into the underlying patterns of user interactions, information flow, and community formation within the Reddit ecosystem.

**Methodology**

**Data Collection:**

We have created a Reddit developer account and obtained the API key, which we named as Project1\_OSNA. We use PRAW, a Python wrapper for the Reddit API, to establish connections and crawl data from Reddit.

A screenshot of a computer

Description automatically generated

The following process was followed for data collection and cleaning:

1. PRAW was utilized to establish a connection between the Python client and the Reddit API.

2. A request was made to retrieve the top 25 posts from the subreddit

3. A recursive call was implemented to retrieve comments for each post, with a maximum limit of 10 comments per post.

4. In the project, a data frame was constructed where the post author was designated as the node and the comment author as the destination.

5. The post author and the comment author were depicted as nodes, with the edge symbolizing the relationship between the comment author and the post author.

6. During a recursive call, a CSV file is generated to store data under the columns "Source" and "Target." In the CSV file, the columns represent the nodes, and each row signifies an edge between two nodes.

7. To avoid duplicating authors as nodes, we created a list of previously added nodes. During each subsequent recursion, this list is checked to prevent the addition of multiple nodes for the same author.

###

8.In the final step of the data collection phase, we saved the account IDs in the data frame’s 'Source' and 'Target' columns. This data was then stored in a file named 'nodedate.CSV'.

**Data Visualization:**

The data was extracted from the CSV file, and each column of the data frame was stored in individual variables. The data was then displayed in labeled formats. As the number of nodes increases, the display becomes more congested. Nodes with higher degrees are represented with larger sizes based on the visualization settings, indicating their significance relative to their degree.

In this project, we utilized the following Python packages: Networkx for graph manipulation and analysis, Pandas for data manipulation and storage, and Matplotlib for data visualization.

**Network Graph:**

In the graph, nodes represent both the authors of posts and comments. The edges between these nodes signify the connections between comment authors and the authors of the posts they comment on. The graph consists of 188 nodes and 237 edges. A loop represents an edge that starts and ends on the same node, essentially indicating a self-referential relationship or a node connected to itself which means the author of the post has also commented on the post.

A network diagram of blue dots and lines

Description automatically generated

**Network Measures Calculation:**

**Degree Distribution:**

Degree distribution in refers to the probability distribution of the degrees of nodes in a network. The degree of a node is the number of connections it has to other nodes. Degree distribution is important because it provides insights into the structure of the network, such as whether it is scale-free (with a few highly connected nodes and many nodes with few connections) or random (with a more uniform distribution of connections). Understanding degree distribution helps in analyzing the robustness, information flow, and other characteristics of online social networks.

Since the network has high number of nodes that have a low degree (few connections), it could mean that there is

Low engagement: Authors are not actively engaging with each other's content, leading to a sparse network where interactions are limited

Content Consumption: Some authors may be more focused on consuming content rather than actively engaging and creating their own content, leading to lower degrees as they have fewer interactions.

A graph of a person

Description automatically generated

**Closeness Centrality:**

Closeness centrality is a measure used in network analysis to quantify how close a node is to all other nodes in the network. It is calculated as the reciprocal of the sum of the shortest path distances between a node and all other nodes in the network. Nodes with high closeness centrality are considered to be more central within the network, as they can reach other nodes more quickly than nodes with lower closeness centrality. Closeness centrality is often used to identify nodes that are important for information dissemination or that act as bridges between different parts of the network.

A graph of a number of blue bars

Description automatically generated

**Degree Centrality:**

Degree centrality is a measure used in network analysis to quantify the number of connections a node has in a network. It is calculated as the number of edges incident to a node, normalized by the maximum possible degree in the network. Nodes with high degree centrality are more central within the network, as they are directly connected to a larger number of other nodes. Degree centrality is often used to identify nodes that are influential or popular within a network, as they have the potential to reach a larger audience or have a greater impact on information flow.

**A graph of a person with a beard

Description automatically generated with medium confidence**

**Betweenness Centrality:**

Betweenness centrality is a measure used in network analysis to quantify the extent to which a node lies on the shortest paths between other nodes in the network. It is calculated as the number of shortest paths that pass through a node, normalized by the total number of shortest paths in the network. Nodes with high betweenness centrality are important for maintaining the network's connectivity, as they act as bridges between different parts of the network. They also play a crucial role in controlling the flow of information through the network, as they lie on many of the shortest paths between other nodes.

A white rectangular object with blue and black text

Description automatically generated with medium confidence

**Katz Centrality:**

Katz centrality is a measure used in network analysis to quantify the influence of a node in a network, considering both the number of its immediate neighbors and the nodes that can be reached through multiple paths. It is calculated as the sum of the weights of all paths between a node and other nodes in the network, with the weights of longer paths being discounted. Katz centrality assigns higher scores to nodes that are connected to other nodes with high centrality, reflecting their potential to control or influence the flow of information in the network.

A graph of a person with a blue bar graph

Description automatically generated with medium confidence

**PageRank Centrality:**

PageRank centrality is a measure used in network analysis to quantify the importance or centrality of a node in a network. It was originally developed by Google to rank web pages in search engine results. PageRank centrality is based on the idea that a node is important if it is pointed to by other important nodes.

A network of blue dots and lines

Description automatically generated

**Reference:**

1.Software Used To run python code : Visual Studio

2. Matplot lib, https://matplotlib.org/

3. Pandas, https://pandas.pydata.org/

4. NetworkX Documentation :

https://networkx.org/documentation/networkx-1.9/

5.Praw : https://praw.readthedocs.io/en/stable/

**Contribution:**

**Shreenidhi R**: Developed code for collecting and processing, Visualization data for project requirements, and report formation.

**Pratik Jadhav**: Developed code for network measures and calculations and report formation.