Reddit Data Analysis

Aditya Sharma A20516668
Raj J. Shah A20524266

1 Abstract

The purpose of this project was to gather data from social media and compile it into a dataset. We were also required to search for, examine, and visualize the data using graphs. After that, we had to provide the retrieved data for a new analysis. Reddit was our social media platform of choice for obtaining the necessary data.

2 Introduction

It is necessary to construct a user's subreddit network, which may be seen as a graph with users as nodes and relationships between users' subreddits as edges. We intend to examine the data using metrics like Degree Distribution, Clustering Coefficient, Betweenness, Closeness, etc. after the graph representation.

3 Project Outline

- Data Collection
- Data Visualization
- Network Measures Calculation
- Result

4 Data Collection

For this project, we have utilized **Reddit** for social media analysis. The Reddit platform is a content-sharing, forum-based platform aggregated and organized into communities known as 'subreddits'. The platform provides the option to write posts, links, images, and video-based content, and similar to other social media has the option to 'upvote' and comment on the posts. Reddit has the option to surf anonymously and requires credentials for posting and commenting.

Reddit offers a variety of developer tools and services including a platform as well as data API for accessing and utilizing content on the platform. Besides the data API, Reddit offers Ads as well as Embeds API.

The Reddit Data API[6] allows the developers to access and modify Reddit Data programmatically. The platform requires OAuth as a purpose of authentication. For accessing the Reddit API, we need to **register** on the platform providing the purpose as well as the application name. The client also needs a *User-Agent* which is unique across the platform and needed when accessing the API. The Reddit API can be used for research purposes academically but allows the limitation of re-distributing the data across other platforms and based on Reddit Data.

It's important to note that while academic research is permitted, redistribution of data across other platforms based on Reddit data is strictly prohibited. To facilitate interaction with the Reddit Data API,

we utilized the popular Python Reddit API Wrapper (PRAW)[3], documented within Reddit's official resources. PRAW simplifies data access and manipulation, acting as a convenient layer built upon the core Reddit APIs.

To access Reddit API, we require three client-based credentials which are **clientId**, **clientSecret** as well, and **userAgent**, provided by the Reddit Developer and submitting a request for the application.

Some of the challenges we faced when accessing/manipulating Reddit data:

- To optimize data collection efficiency, we initially leveraged the active user community within select (Top 20) popular subreddits. However, due to a significant user influx, we implemented temporary measures to limit the number of submissions from new users and the inclusion of additional subreddits, effectively capping the data volume at a manageable level (Top 20 submissions).
- Due to the ever-evolving nature of the Reddit community, characterized by high user engagement and frequent content updates, we observed a corresponding increase in user activity and time spent accessing user-related subreddits. To ensure compliance with Reddit's rate limits and manage data volume efficiently, we restricted our analysis to the top 100 submissions per user and subreddit.
- With the advent of data and the number of users increasing, we parsed the data and stored it temporarily in a JSON file with the format
  {
  "user1": ["list of subreddits"],
  "user2": ["list of subreddits"],
  .
  .
  .
  }
  to be able to manipulate and use for further analysis.
- The Reddit Data API implements tiered access for certain subreddit-level data, including upvote/downvote counts. This paywalled data restricts the comprehensiveness of freely available information and limits the analysis of user engagement within subreddits.

4.1 User privacy policy[7]

- Reddit collects minimal information from users to identify. You can browse without any account.
- Information such as interests is asked to design the home feed. It is up to the user to provide information like bio, gender, age, location, or profile picture.
- The content we submit such as messages, chats, mail including saved drafts, or audio is collected via Services.
- If purchases are made on the Reddit platform, then certain information including name, address, email, phone number, and information about the product you are purchasing. It uses payment processor platforms like Stripe.
- Reddit also tracks information related to interactions within the platform such as voting, saving, etc.
- Reddit automatically keeps track of information when you access such as IP address, browser type, mobile carrier, pages visited, and links clicked. Out of these, information is collected from cookies as well.
- Reddit receives information about the use of any third-party service that uses a Reddit account. In general, Reddit does not control how third-party services collect data when they serve their content directly via these embeds (used for displaying content from third parties).
- The Reddit Data is used as per policy :

  - Personalize services, content, and features to match the user's activity and preferences.
  - Provide, optimize, target, and measure the effectiveness of ads shown as well as Research and develop new services;
  - Communicating about products, services, offers, promotions, and events that might 'interest' a user.

- Information collected from services is public and accessible to everyone. Any visitors to and users of services will be able to access the content.
- The scope of data collection and processing, when users share content from Reddit on other media, is under the privacy policies of the other services.
- Reddit as per policy doesn't sell any information without your consent or any third-party service. Data can be shared if there is any legal regulation/process or a bodily emergency to a person.

4.2 Data Usage Policy[5]

- You must be of legal age to use the data and bind to the US government's terms and conditions or other countries.
- To use Data API, you must provide identification(e.g. contacts).
- We are not allowed to modify user content except if it is to be used for displaying in your application/service.
- May receive notices or requests to remove user content under Reddit's privacy policy.
- You will only access (or attempt to access) Data APIs using the Access Info described in the Developer Documentation for the Data APIs. You must use the Access Info we provided you (e.g., the OAuth token) when accessing the Data APIs, and you will not misrepresent or mask either the user agent or OAuth identity when using the Data APIs.
- If we utilize any third-party libraries, extensions, and wrappers, we need to comply with any limitations or restrictions imposed by the applicable third party and Reddit.
- We are not allowed to interfere, modify, disrupt, or disable any functionalities of Data API.
- We cannot sell, lease, or sublicense the Data APIs or access thereto or derive revenues from the use or provision of the Data APIs, whether for direct commercial or monetary gain unless there is express written approval from Reddit.
- We are not permitted to use the Reddit Trademarks in, or as part of the name of your App, or any logos used to promote or identify our App, unless expressly authorized in writing by Reddit or in the Data API Terms.
- We can not make any statement regarding our use of the Data APIs.
- We can use Confidential Information only as necessary in exercising our rights granted under the Data API Terms and we can not disclose it to any third party without Reddit's prior written consent.
- We should protect the data that we are using, from unauthorized access, use, or disclosure as if it is our data.
  For more information on Reddit's Privacy and Data Usage Policy, you can refer to the link given below.

5 Data Visualization

Our social media graph analysis focused on user relationships based on their shared subreddit memberships. This "friendship network" approach aimed to understand user similarities and potential connections through common interests.

Data was initially parsed and stored as a JSON file, then converted into a dictionary format with users as keys and the number of shared subreddits (weight) as values.

NetworkX[8], a powerful library for network creation and manipulation, was chosen to construct the graph. It facilitates analysis and visualization of complex social networks, providing insights into user behavior, information flow, and network dynamics. Additionally, NetworkX offers valuable centrality measures like degree, betweenness, and eigenvector centrality, which aid in identifying influential users and key network connectors.

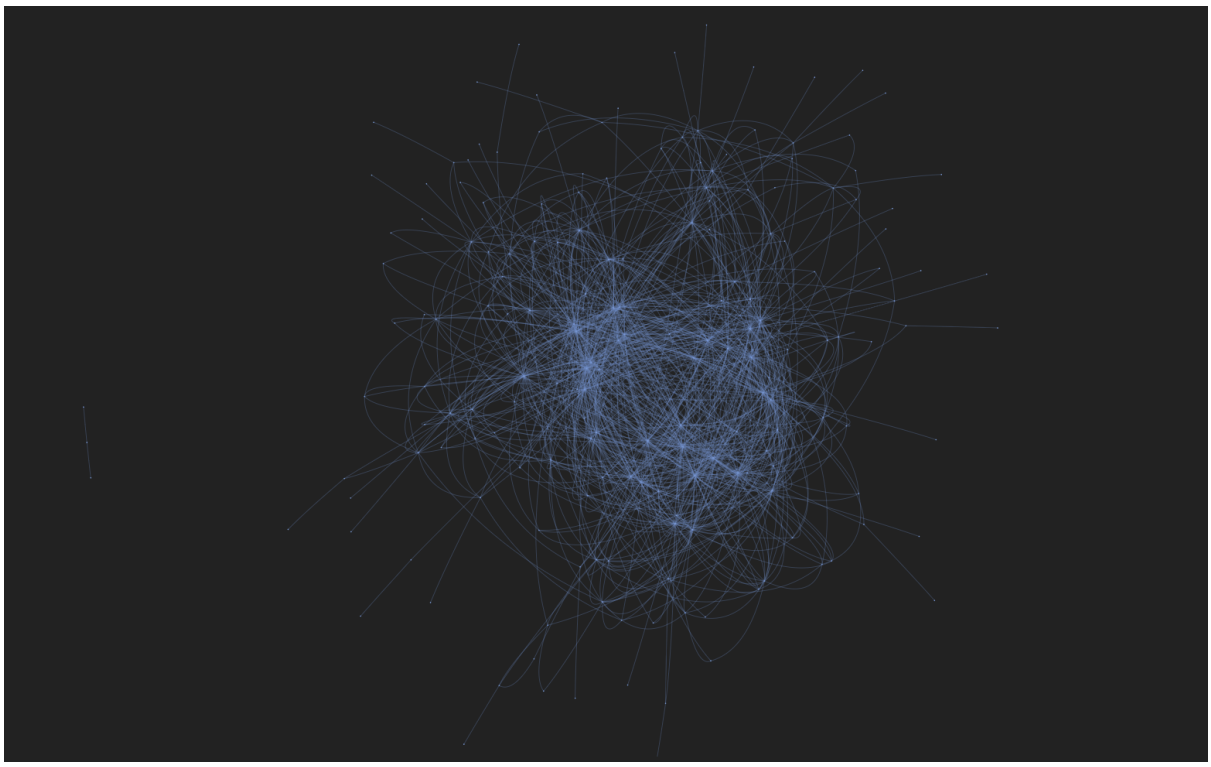
Furthermore, NetworkX seamlessly integrates with libraries like **Pyvis**[4], enabling the creation of interactive and visually appealing network visualizations. Pyvis allows customization of node size, color, shape, and edge style to represent various user attributes or interaction types, enhancing comprehension of relationships and interactions. This combination of NetworkX for analysis and Pyvis for visualization forms a powerful workflow for comprehensive social media analysis.

The graph structure utilizes an *adjacency list* represented as a dictionary, where keys represent users and values represent the number of common subreddits (weight). The network is defined as a weighted undirected graph, with weight signifying the degree of shared subreddit membership. To normalize the weights, a logarithmic scaling factor is applied.
The adjacency list is created in the following format
{
"user1": {"user2": 1}, {"user3": 2},...
"user2": {"user1": 1}, {"user3": 4},...
.
.
.
}

The Graph is represented as:

- **Nodes**: The Reddit Users (Active in Popular Subreddits)
- **Edges**: If there are 'subreddits' subscribed by both users
- **Weight**: Number of subreddits subscribed to by both users

Figure 1: The Social Network Graph depicted with nodes as 'users' and edges representing the 'commonality'



6 Network Measures

Since we have leveraged the *NetworkX* library for our social media graph construction grants us access to its built-in functionalities for graph analysis and measurement.

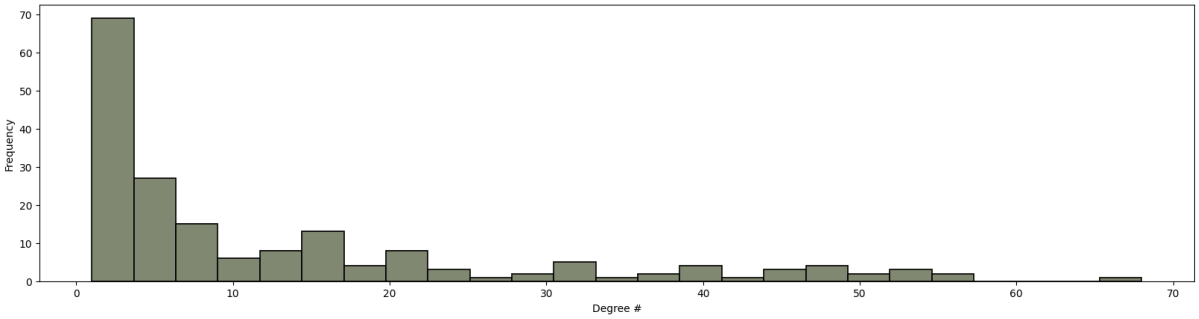
We analyzed a sample of real-time social media data using a graph where connections between nodes (users) are weighted based on their common interests (intersection) and only considered if the weight is greater than 5.

The number of nodes is **184**, with the number of edges being around **1192**.

We have focused on 3 such measures regarding the social graph which are :

- Degree Distribution: Degree is defined as the *number of edges* connected to a node. It is an essential measure of robustness as well as connectivity within the graph. As part of social media analysis, the degree distribution can help us identify the influential nodes within the network and isolated components within the graph. We can analyze the information flow and diffusion within the network if needed.

Figure 2: Degree Distribution for the nodes represented as a histogram



We have used the 'graph.degree()' method for finding the number of degrees for each node. Using the degree measure above, we observe the highest degree within the network being **68** and the minimum being **1**. Also, the graph is right-skewed, having more number of nodes with low degrees.

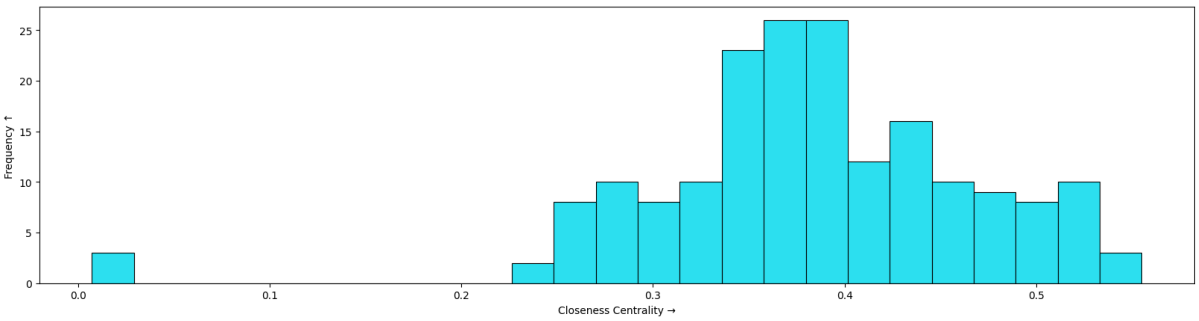
- Closeness Centrality: Centrality within a network is a measure of a node based on its network position. The closeness centrality defines the centrality in the network calculated as the reciprocal of the sum of lengths of shortest paths between the target node and other nodes in a graph. The higher value indicates higher closeness as well as centrality.

  As per *Networkx*, the closeness centrality is defined as the :

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$$

  where n-1 is the number of nodes reachable from node u and $d(v,u)$ is the shortest path from node 'v' to node 'u'

Figure 3: Closenness Centrality for the nodes represented as a histogram

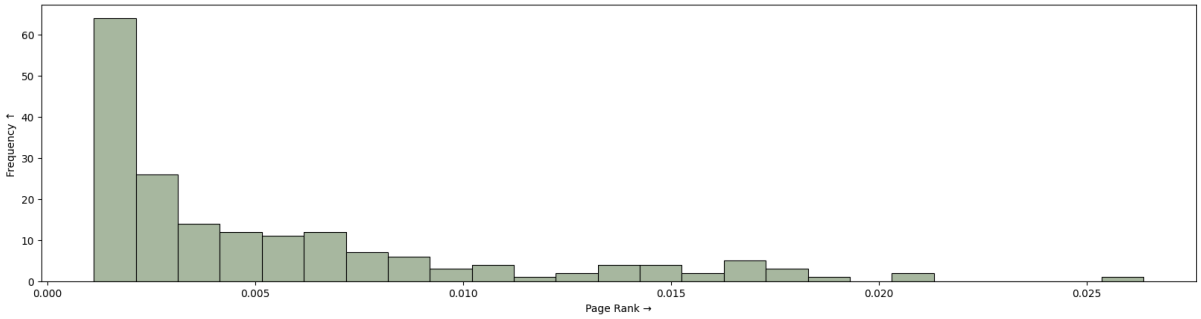


We have used the 'nx.closeness_centrality()' method for finding the closeness centrality for each node. Using the centrality measure above, we observe the node with maximum closeness centrality within the network being **0.55** and the node with minimum closeness centrality being **0.0073**.

- Pagerank: The Pagerank[2] algorithm similar to centrality is used to provide a 'ranking' to a node which can loosely translate to priority derived from the network structure and the incoming links. The idea is taken from the original 'PageRank' algorithm[9] which was used to rank web pages for recommendations. PageRank assigns higher scores to nodes with more incoming links (followers) from other important nodes. This helps identify users who are likely to have a wider reach and potentially influence others. PageRank can predict how information or trends might spread through the network. Users with higher PageRank scores are more likely to be starting points for information diffusion, as their content reaches a broader audience.

As per *Networkx*, the PageRank algorithm was designed for directed graphs but the algorithm does not check if the input graph is directed and will execute on undirected graphs by converting each edge in the directed graph to two edges.

Figure 4: Pagerank values for the nodes represented as a histogram



We have used the 'nx.pagerank()' method for finding the page rank for each node with default values (damping factor: 0.85, max_iter: 100). Using the centrality measure above, we observe the node with maximum page rank within the network being **0.0264** and the node with minimum page rank being **0.0011**.

7 Discussion of Result

7.1 Insights from Results

The analysis of the Reddit social network graph has provided several insights into user interactions and network dynamics within the platform. By leveraging NetworkX for graph construction and measurement, we were able to explore key metrics such as Degree Distribution, Closeness Centrality, and PageRank.

Degree Distribution: The presence of a large number of zero or nearly zero weights, indicating minimal or no shared subreddits between certain users, contributed to the complexity of the network by increasing the number of edges. The weights ranged from 2 to 170, requiring eventual scaling down and normalization.
The right-skewed distribution suggests that while there are a few highly connected users (nodes) with a high number of subreddits (edges), the majority of users have lower degrees. This indicates a hierarchical structure within the network where a small number of users potentially hold significant influence or act as connectors.

Closeness Centrality: Nodes with higher closeness centrality scores are more central within the network, indicating their proximity to other nodes. This suggests that certain users may have a higher potential to disseminate information or influence others due to their central position in the network.

PageRank: Users with higher PageRank scores are likely to have a wider reach and influence within the network. These users may serve as important starting points for information diffusion and are crucial in predicting how trends or information spread throughout the community.

7.2 Further Questions

While the analysis provides valuable insights, it also raises several further questions that warrant exploration:

Community Detection: Are there distinct communities or clusters within the Reddit network? Understanding community structure can provide deeper insights into user behavior and interactions.

Temporal Analysis: How do dynamics evolve? Analyzing changes in user interactions, influential nodes, and community formation over different periods can reveal trends and patterns in user engagement.

Content Analysis: How do user interactions correlate with the content shared on subreddits? Examining the relationship between network metrics and the nature of discussions or posts can shed light on factors driving user engagement and network growth.

7.3 Next Steps for Investigation

Community Detection Algorithms: Implementing algorithms such as modularity optimization or spectral clustering to identify communities within the network and analyze their characteristics.

Longitudinal Analysis: Collecting data over multiple time points and conducting longitudinal analysis to track changes in network structure, user behavior, and content popularity over time.

Natural Language Processing (NLP): Integrating NLP techniques to analyze the content of posts and comments, and investigate how language dynamics influence network interactions and community formation.

References

[1] Matplotlib. URL: https://matplotlib.org/stable/plot_types/index.html.

[2] Page rank - social media analysis. URL: https://www.andreaperlato.com/graphpost/page-rank-in-network-analysis/.

[3] Python reddit api wrapper docs. URL: https://praw.readthedocs.io/en/stable/.

[4] Pyvis. URL: https://pyvis.readthedocs.io/en/latest/documentation.html.

[5] Reddit data policy. URL: https://www.redditinc.com/policies/data-api-terms.

[6] Reddit dev api. URL: https://www.reddit.com/dev/api/.

[7] Reddit privacy policy. URL: https://www.reddit.com/policies/privacy-policy.

[8] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008. URL: https://networkx.org/documentation/stable/index.html.

[9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking : Bringing order to the web. In *The Web Conference*, 1999. URL: https://api.semanticscholar.org/CorpusID:1508503.