Social Network Analysis: Subreddit Interactions

Farhad Bayrami, Artificial Intelligence, 1073949
Mahmut Kaan Molla, Artificial Intelligence, 1067839

August 2024

1 Introduction

The study of online communities, particularly on platforms such as Reddit, provides insights into the dynamics of digital social networks and community interactions. Reddit, a social media platform organized into communities known as subreddits, offers a unique environment to analyze how communities connect, interact, and sometimes conflict with one another. The subreddit hyperlink network, which represents directed connections between subreddits, serves as a valuable dataset for exploring these interactions.

This project leverages a dataset extracted from publicly available Reddit data spanning from January 2014 to April 2017, capturing the directed hyperlinks between subreddits. The dataset includes detailed information on the nature of these hyperlinks, such as timestamps, sentiment, and textual attributes, offering a rich source for analyzing community behavior and interaction dynamics. By applying Social Network Analysis (SNA) techniques, this study aims to uncover patterns and factors that influence community engagement, cooperation, and conflict within Reddit.

The network analyzed in this project is a monomodal directed network, where all nodes represent subreddits. Edges in the network represent hyperlinks originating from posts in one subreddit (the source) and pointing to posts in another subreddit (the target). Since both the source and target nodes are of the same type (subreddits), the network does not exhibit a bipartite structure. Instead, it is characterized by a single mode of nodes connected by directed edges, reflecting the nature of subreddit interactions through hyperlinks.

2 Problem and Motivation

The primary problem addressed by this project is understanding the dynamics of subreddit interactions and the underlying factors that drive community engagement and conflict. As digital communities continue to grow and evolve, understanding the mechanisms of interaction between subreddits becomes increasingly important. This knowledge can provide insights into broader social phenomena such as online polarization, community influence, and information dissemination.

The motivation for this study stems from the need to analyze and interpret the complex relationships within the subreddit hyperlink network. By identifying patterns of positive and negative interactions, the project seeks to contribute to the broader understanding of online community dynamics and the factors that lead to conflicts and alliances. The findings can

inform the design of algorithms and policies that promote healthier online environments and foster constructive community interactions.

3 Datasets

The dataset [2] utilized in this study is derived from publicly available Reddit data collected over 2.5 years, from January 2014 to April 2017. It comprises two primary components: the Subreddit Hyperlink Network and Subreddit Embeddings.

Subreddit Hyperlink Network: This directed network captures the hyperlinks between subreddits, where each hyperlink originates from a source community post and points to a target community post. The network includes 55,863 nodes (subreddits) and 858,490 edges (hyperlinks). Each hyperlink is annotated with several attributes, including the timestamp of creation, the sentiment (-1 for negative or +1 for positive) of the source post towards the target post, and text property vectors representing the source post's content. The network is characterized by its directed, signed, temporal, and attributed nature, allowing for a comprehensive analysis of subreddit interactions.

Subreddit Embeddings: In addition to the hyperlink network, the dataset provides embedding vectors for each subreddit, offering a numerical representation of the subreddit characteristics based on its interactions. The embedding file contains 51,278 embeddings, although not all subreddits could be represented due to data limitations.

This dataset is an invaluable resource for studying community interactions on Reddit and provides the necessary data for conducting a thorough social network analysis of subreddit relationships. The dataset was used in a research project [1] investigating how subreddits attack one another.

3.1 Dataset Statistics

Statistic	Value
Number of nodes (subreddits)	55,863
Number of edges (hyperlink between subreddits)	858,490
Edge weights (label of hyperlink)	-1 or +1
Edge attributes	Text property vectors
Timespan	Jan 2014 - April 2017

Table 1: Statistics of the subreddit hyperlink network dataset.

3.2 Dataset Columns

The dataset consists of several columns that describe the properties of each hyperlink between subreddits:

- **SOURCE_SUBREDDIT**: The subreddit where the hyperlink originates.
- **TARGET_SUBREDDIT**: The subreddit where the hyperlink ends.
- **POST_ID**: The ID of the post in the source subreddit that initiates the hyperlink.

- **TIMESTAMP**: The time when the post was created.
- **LINK_SENTIMENT**: A label indicating if the source post is explicitly negative (-1) or neutral/positive (1) towards the target post, determined using crowd-sourcing and a trained classifier.

3.3 Preprocessing

During preprocessing, the title and body hyperlink datasets were joined on the `POST_ID` column. This resulted in a new dataset with the following columns: `SOURCE_SUBREDDIT`, `TARGET_SUBREDDIT_title`, `TARGET_SUBREDDIT_body`, `POST_ID`, `TIMESTAMP` and `LINK_SENTIMENT` This join allows for the comparison of the target subreddits in both the title and body hyperlinks, revealing potential differences in the target subreddit based on where the hyperlink appears in the post.

The `POST_PROPERTIES` attribute, which contains detailed text properties of the source post, was removed as it is not directly relevant to the project's focus. This step streamlined the dataset to concentrate on the interaction dynamics between subreddits rather than the intricacies of post content.

3.4 Focus on the Largest Strongly Connected Component (Computational concerns)

In our analysis of the subreddit hyperlink network, we have chosen to focus on the largest strongly connected component (SCC) of the network. This decision is based on the characteristics and significance of the SCC in understanding the core structure and interactions within the network.

3.4.1 Rationale for Focusing on the Largest SCC

The largest SCC represents the core of the network, where interactions are dense and every node can reach every other node. By focusing on this component, we aim to analyze the most central and structurally significant part of the network. Below, we outline the key advantages and disadvantages of this decision.

3.4.2 Advantages of Focusing on the Largest SCC

- **Cohesiveness and Interconnectedness**: The largest SCC captures the most interconnected subgraph in the network, where all nodes are reachable from one another. This allows for a detailed analysis of the core interactions within the network, which is crucial for understanding the central dynamics and key communities.
- **Simplification of the Network**: Focusing on the largest SCC reduces the size of the graph, making the computation of complex network measures (such as centrality and community detection) more feasible. By limiting the scope to the most connected part of the network, we can perform deeper analyses with reduced computational overhead.
- **Core Dynamics**: The largest SCC often represents the most active and influential portion of the network. Studying this component allows us to concentrate on the interactions that are most likely driving the overall behavior of the network.

3.4.3 Disadvantages of Focusing on the Largest SCC

- **Loss of Generality**: By restricting our analysis to the largest SCC, we exclude nodes that are not part of this component, including peripheral and isolated subreddits. This exclusion may result in the loss of important insights regarding the overall network structure and dynamics, particularly for those nodes that do not belong to the core but still play a role in the broader network.
- **Exclusion of Smaller Communities**: Small communities or subgraphs that are not strongly connected may be overlooked. These smaller groups could represent niche subreddits or specialized topics that, while not part of the core network, contribute to the diversity and richness of the overall structure.

The final dataset in hand has 11564 nodes and 98166 edges and its average degree is 17.

4 Validity and Reliability

4.1 Validity

The dataset used in this study offers a robust representation of subreddit interactions, capturing both positive and negative sentiment and including temporal data. The directed nature of the hyperlinks between subreddits provides a framework for analyzing online community dynamics, including cooperation, conflict, and information dissemination.

4.1.1 Data Representation and Sentiment Classification

The sentiment labels in the dataset are derived from crowd-sourced input and a trained classifier, improving the accuracy of sentiment detection compared to simple sentiment analysis. This enhances the dataset's alignment with real-world interactions on Reddit, making it more valid for studying genuine community behavior.

4.1.2 Dataset Integration

The integration of title and body hyperlink datasets on `POST_ID` enriches the dataset by capturing potential differences in subreddit linkages. Retaining both `TARGET_SUBREDDIT_title` and `TARGET_SUBREDDIT_body` provides a more comprehensive view of subreddit interactions.

4.1.3 Limitations and Focus on the Largest SCC

Focusing on the largest strongly connected component (SCC) allows us to analyze the most interconnected part of the network, where all nodes are mutually reachable. This decision simplifies the analysis but comes with trade-offs, as it excludes peripheral nodes, potentially limiting the generality of the findings.

4.2 Reliability

4.2.1 Data Sources and Preprocessing

The dataset is sourced from publicly available Reddit data, ensuring accessibility for replication. Preprocessing steps, such as the removal of `POST_PROPERTIES` and the joining of datasets

on POST_ID, are well-documented, allowing for consistent application across different analyses.

4.2.2 Use of Established Methods

The study employs established Social Network Analysis (SNA) techniques using tools like Gephi and NetworkX, contributing to the reliability of the results. These methods ensure that other researchers can reproduce the findings using the same dataset and analysis pipeline.

4.2.3 Impact of Focusing on the Largest SCC

Focusing on the largest SCC increases the reliability of the analysis by reducing network complexity and enabling more consistent calculations. However, this focus may reduce the generality of the findings, as interactions outside the largest SCC are not considered.

4.2.4 Reproducibility

The use of explicit sentiment labels and temporal data, along with clear preprocessing pipelines, ensures a high degree of reproducibility. This allows for independent validation of the findings and ensures that the study's results can be replicated in future research.

5 Measures and Results

5.1 Community Detection

To understand the structure of the subreddit hyperlink network, we applied the Louvain community detection algorithm, which optimizes modularity to identify densely connected subgroups within the network. This method is particularly effective for large networks like ours, where we aim to detect clusters of subreddits that are more interconnected internally than with the rest of the network.

5.1.1 Community Statistics

The analysis identified a total of **23 communities** within the network. These communities exhibit significant variation in size, ranging from small groups of just a few subreddits to larger clusters containing thousands of subreddits. The largest community contains **2754 nodes**, indicating that a substantial portion of the network is concentrated within this single cluster, whereas the smallest communities consist of **3 nodes**.

Key statistics of the detected communities are summarized below:

Metric	Value
Total Number of Communities	23
Largest Community Size	2754 nodes
Smallest Community Size	3 nodes
Average Community Size	~503 nodes
Median Community Size	162 nodes

Table 2: Community Statistics

5.1.2 Distribution of Community Sizes

The distribution of community sizes reveals that most communities are relatively small, with only a few large clusters dominating the network. This suggests a hierarchical structure where a few large communities play central roles, potentially acting as hubs for information flow and interaction within the network. On the other hand, smaller communities may represent more specialized or niche groups of subreddits.

To visualize this, a plot of the distribution of community sizes (Figure 1) highlights the disparity in community sizes, emphasizing the dominance of a few large communities alongside a long tail of smaller ones.

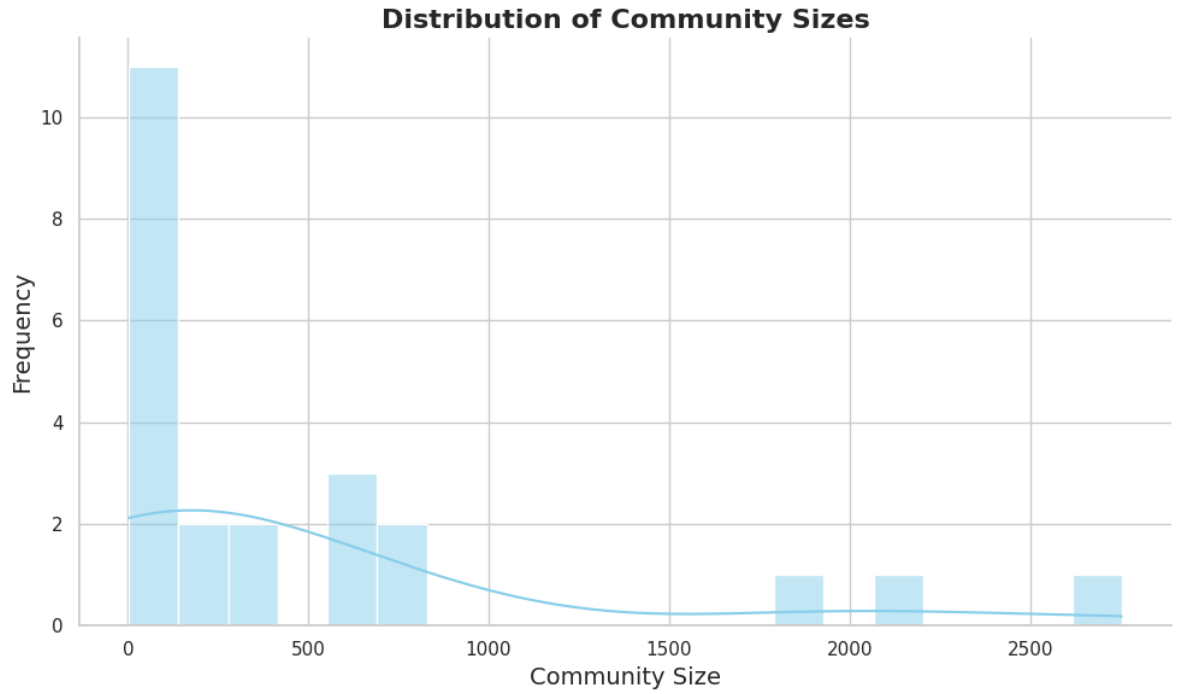


Figure 1: Distribution of Community Sizes in the Subreddit Hyperlink Network.

5.2 Degree Centrality

5.2.1 Definition and Relevance

Degree centrality measures the number of edges connected to a node. In a directed network like ours, this can be split into in-degree (incoming edges) and out-degree (outgoing edges). For simplicity, we focus on the total degree, which captures the overall connectivity of each node. Nodes with high degree centrality are highly connected, indicating that they play a central role in linking different subreddits.

5.2.2 Application to the Study

Degree centrality was applied to identify the subreddits with the most connections, which can be seen as hubs in the network. These hubs are essential for maintaining the cohesion of the network and facilitating the flow of information between communities. In our dataset, the top nodes by degree centrality include `askreddit`, `iama`, and `subredditdrama`, which are highly

interactive subreddits that serve as central hubs for user engagement across Reddit. The top 10 nodes by degree centrality are depicted in tha Table 3.

Node	Degree Centrality
askreddit	0.1622
iama	0.1436
subredditdrama	0.1346
outoftheloop	0.0791
pics	0.0662
videos	0.0619
gaming	0.0599
writingprompts	0.0581
conspiracy	0.0560
legaladvice	0.0558

Table 3: Top 10 Nodes by Degree Centrality

These subreddits have the highest number of connections, indicating that they are key nodes for communication and content sharing within the network.

5.3 Betweenness Centrality

5.3.1 Definition and Relevance

Betweenness centrality measures how often a node lies on the shortest path between other nodes. Nodes with high betweenness centrality act as bridges in the network, facilitating the flow of information between different parts of the network. These nodes are crucial for connecting otherwise distant communities.

5.3.2 Application to the Study

Betweenness centrality was used to identify subreddits that serve as intermediaries between different communities. These nodes are important for ensuring that information can spread throughout the network, even between communities that are not directly connected. In our analysis, `subredditdrama`, `iama`, and `askreddit` emerged as key nodes with high betweenness centrality, indicating their role in connecting various parts of the Reddit community. The top 10 nodes by betweenness centrality are depicted in the Table 4.

These subreddits serve as key intermediaries, facilitating information flow across different communities.

5.4 Eigenvector Centrality

5.4.1 Definition and Relevance

Eigenvector centrality measures a node's influence based not only on the number of connections it has but also on the importance of its neighbors. A node with high eigenvector centrality is connected to other well-connected nodes, making it an influential part of the network. This measure captures the idea of "influence by association."

Node	Betweenness centrality
subredditdrama	0.1257
iama	0.1225
askreddit	0.1158
outoftheloop	0.0699
gaming	0.0457
writingprompts	0.0377
anime	0.0345
legaladvice	0.0314
conspiracy	0.0283
dota2	0.0275

Table 4: Top 10 Nodes by Betweenness Centrality Value

5.4.2 Application to the Study

Eigenvector centrality was applied to identify subreddits that are not only well-connected but are also connected to other influential subreddits. This helps in understanding the hierarchical structure of influence within the Reddit network. The results indicate that subreddits like, iama, and askreddit are not only central but also well-positioned within clusters of other central nodes, reflecting their significant influence within the network. The top 10 nodes by eigenvector centrality are depicted in the Table 5.

Node	Eigenvector Centrality Value
askreddit	0.2763
iama	0.2641
videos	0.1877
pics	0.1812
worldnews	0.1498
funny	0.1445
news	0.1423
gaming	0.1407
technology	0.1301
explainlikeimfive	0.1291

Table 5: Top 10 Nodes by Eigenvector Centrality Value

These subreddits have strong connections to other influential nodes, highlighting their importance in the broader network.

5.5 PageRank

5.5.1 Definition and Relevance

PageRank is a variant of eigenvector centrality originally developed by Google to rank web pages. It assigns higher scores to nodes that are connected to other high-ranking nodes, with the added factor of damping, which prevents overemphasis on certain parts of the network. In social networks, PageRank helps identify influential nodes that are well-linked across the network.

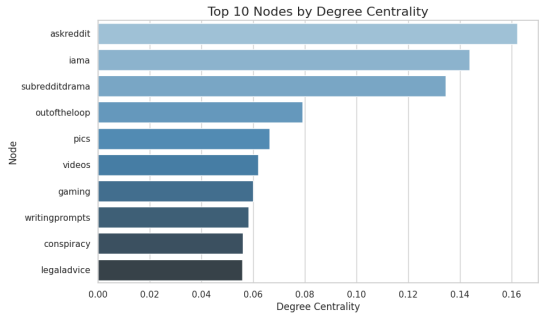
5.5.2 Application to the Study

PageRank was used to identify influential subreddits based on their connectivity to other influential nodes. This measure is particularly useful in large-scale networks like Reddit, where influence is not just about the number of connections but the quality of those connections. The results align with the other centrality measures, confirming the importance of subreddits like `iama`, and `askreddit`. The top 10 nodes by PageRank are depicted in the Table 6:

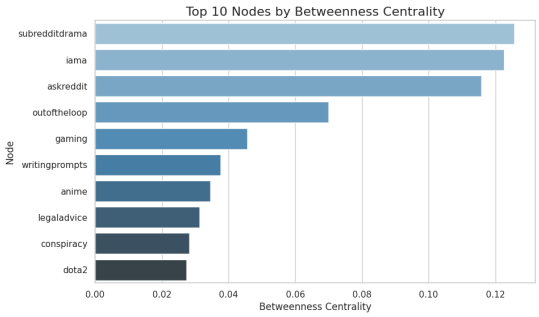
Node	PageRank Value
`iama`	0.0135
`askreddit`	0.0134
`videos`	0.0070
`pics`	0.0068
`videos_discussion`	0.0060
`outoftheloop`	0.0053
`gaming`	0.0043
`funny`	0.0040
`leagueoflegends`	0.0039
`worldnews`	0.0038

Table 6: Top 10 Nodes by PageRank Value

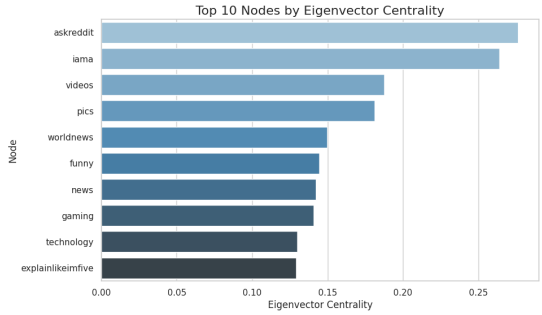
PageRank identifies these subreddits as influential based on their connections to other highly ranked nodes in the network.



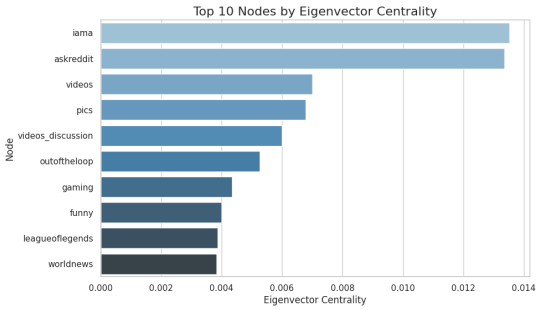
(a) Degree Centrality



(b) Betweenness Centrality



(c) Eigenvector Centrality



(d) PageRank

Figure 2: Centrality Measures for Subreddit Network

5.6 Clustering Coefficient Analysis

Clustering coefficients provide insight into the tendency of nodes in the network to form tightly-knit groups, which is a key indicator of the presence of community structures within the network. Two measures were calculated: the global clustering coefficient and the local clustering coefficient.

5.6.1 Global Clustering Coefficient

The global clustering coefficient, also known as transitivity, measures the overall probability that two neighbors of a node are also neighbors themselves, forming a triangle. For the largest strongly connected component (SCC) of our subreddit network, the global clustering coefficient was found to be **0.0711**. This relatively low value suggests that, on average, subreddits do not form many tightly-knit triads, indicating that while there are clusters within the network, these are not particularly dense or pervasive throughout.

5.6.2 Local Clustering Coefficient

The local clustering coefficient measures the tendency of individual nodes to form triangles with their neighbors, providing a more granular view of clustering behavior across the network. The distribution of local clustering coefficients is visualized in Figure 3, which shows four different perspectives: a histogram, a box plot, a density plot, and a violin plot.

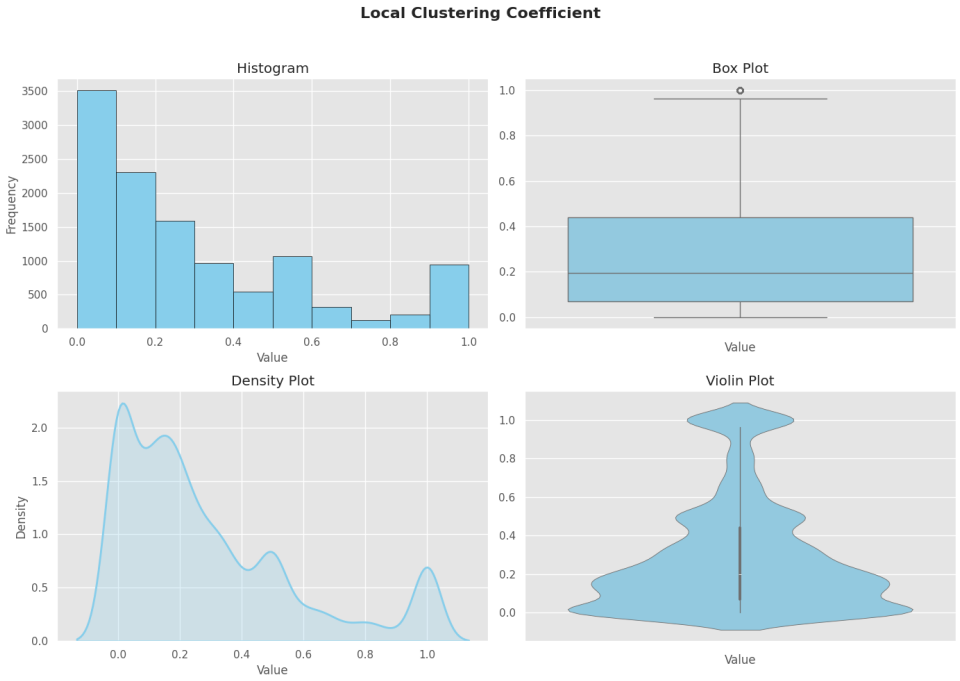


Figure 3: Distribution of Local Clustering Coefficients for the Largest SCC

Overall, these plots indicate that while there are some subreddits with higher local clustering coefficients, the majority of subreddits do not form strong clustering patterns, aligning with the low global clustering coefficient of the network. This suggests that, while there are pockets of closely connected subreddits, the network as a whole is not highly clustered.

5.7 Edge Redundancy and Edge Betweenness

In social network analysis, edge redundancy and edge betweenness are critical for understanding the flow of information and the robustness of the network. **Edge redundancy** refers to the existence of alternative paths between nodes, ensuring that the removal of a single edge does not disrupt communication. High redundancy can indicate robustness but might also point to inefficiency if there are too many unnecessary connections.

5.7.1 Edge Betweenness

Edge betweenness is a measure of the importance of an edge in terms of the number of shortest paths passing through it. Edges with high betweenness are critical for maintaining connections between different parts of the network. Conversely, edges with low betweenness are part of multiple alternative paths, suggesting that the network can function effectively even if these edges are removed.

In the context of our subreddit network, low edge betweenness is desirable, as it indicates that the network is not overly reliant on specific connections. The presence of alternative paths implies that the network is robust, with no single edge acting as a bottleneck for information flow. In this study, the maximum edge betweenness observed was 9.014×10^{-5}, which is exceptionally low. This suggests that the network has high redundancy, with multiple pathways for interactions between subreddits.

6 Conclusion

This study analyzed the subreddit hyperlink network by applying various centrality measures, clustering coefficients, and community detection techniques to the largest strongly connected component (SCC). The quantitative analysis revealed key subreddits that play central roles in the network, such as `askreddit`, `iama`, and `subredditdrama`, which emerged as influential nodes across different centrality metrics. The clustering coefficient analysis showed that, while some subreddits exhibit strong community behavior, the network as a whole is not highly clustered, reflecting a more dispersed interaction structure.

The low edge betweenness values observed across the network suggest that the network is highly redundant, with no single edge disproportionately influencing the overall communication between subreddits. This indicates that the network is robust, capable of withstanding disruptions in specific connections without significantly affecting the flow of information.

Overall, the qualitative analysis of these quantitative findings suggests that the subreddit network, while containing influential nodes, is characterized by a relatively low level of clustering and high redundancy. These features contribute to the network's resilience and its ability to facilitate widespread interaction across various subreddits.

7 Critique

While this study successfully applied Social Network Analysis (SNA) techniques to identify key characteristics of the subreddit network, several limitations affected the scope and depth of the analysis. Due to memory and CPU limitations, we were unable to perform more computationally intensive analyses, such as calculating small-worldness and conducting a triad census.

These measures could have provided further insights into the global structure of the network and the prevalence of specific interaction patterns, respectively.

Additionally, the focus on the largest SCC means that our analysis did not cover the full network. As a result, peripheral nodes and smaller components, which may have had significant roles in the broader Reddit ecosystem, were excluded from the study. This limits the generality of the findings, as the behavior and influence of these peripheral subreddits remain unexplored.

Furthermore, future work should incorporate the use of edge weights, specifically the `LINK_SENTIMENT` and `TIMESTAMP` attributes, which were not fully utilized in this study. Analyzing how sentiment evolves over time and how temporal dynamics affect subreddit interactions could provide a richer understanding of the network's behavior. The integration of these temporal and qualitative dimensions would allow for a more comprehensive analysis of the network, moving beyond static snapshots to capture the dynamic evolution of interactions over time.

In summary, while this study provides a solid foundation for understanding the core dynamics of the subreddit hyperlink network, future work should address the limitations mentioned above. By incorporating additional measures, utilizing the full network, and considering the temporal and sentiment-driven aspects of subreddit interactions, a more complete picture of the network's structure and behavior can be achieved.

References

[1] Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 933–943). International World Wide Web Conferences Steering Committee.

[2] `https://snap.stanford.edu/data/soc-RedditHyperlinks.html`