Subreddit Similar Interactions Network Analysis

Dipak Subramaniam[1]

[1]Southern Illinois University Edwardsville
April 18, 2023

Abstract

This paper presents a network analysis of interactions between popular subreddits on the Reddit platform. The data was extracted using web scraping techniques, and formed a directed network graph using NetworkX in Python. The graph was analyzed and visualized using a combination of Gephi and NetworkX. The results show that subreddits with similar topics tend to be very heavily connected, whereas subreddits with dissimilar topics have loose connections that represent the wide-ranging interests of the population. Certain subreddits act as important hubs that facilitate interactions between different communities. The network graph's global characteristics, important nodes, distribution graphs, resilience against targeted/random attacks, and implications for spreading models are discussed. It can be conclusively determined that the Reddit platform's structure is scale-free, and subreddits play an essential role in shaping communication and information flow across the platform.

Keywords– network science, Reddit, subreddits, similar interactions, user overlap, networkx, Gephi

1 Introduction

Reddit is a social media platform consisting of subreddits dedicated to specific topics. It has over 430 million active users and more than 130,000 active subreddits as of 2023 [2]. The website is ranked 18th in the world with over 1.7 billion unique visitors per month. The most popular subreddits have millions of subscribers, with the largest subreddit being r/announcements, which has over 45 million subscribers. The top categories for subreddits include news, gaming, technology, sports, and entertainment. The platform is particularly popular among younger users, with 64 percent of its users aged 18-29. The majority of Reddit users identify as male, with 69 percent of users identifying as male and 29 percent identifying as female. The average user spends over 16 minutes per day on the platform, with the most active time on the platform being between 9am and 5pm on weekdays. Over 54 percent of Reddit users are from the United States, followed by the United Kingdom, Canada, and Australia. However, its use is slowly expanding in other areas as well. Overall, these statistics highlight the significant user base and engagement on Reddit, with a wide range of subreddits catering to various interests and topics. The platform's popularity among younger users and its high traffic make it an attractive destination for marketers and advertisers, as well as network science students.

The interactions between these subreddits can be analyzed using network science techniques. In this study, the aim is to construct a network graph that captures the interactions between subreddits and analyze the structure of the graph to gain insights into the Reddit platform. The paper will disseminate detailed information on the network graph, including its visualizations, global characteristics, important nodes, distribution graphs, resilience against targeted/random attacks, and implications for spreading models. By studying the structure and behavior of this network, there will be a better understanding of the dynamics of online communities and the impact of different factors on their evolution.

2 Methods

2.1 Network Setup

Web scraping techniques were leveraged to extract data on the interactions between subreddits. The data was obtained (as shown in Figure 1) from the Subreddit Stats website using Selenium and Python [3]. Given the immense number of subreddits and fluctuating / inconsistent statistics on the top fifty to hundred subreddits, the network was limited to the top thirty-one subscribed subreddits, gathered from the OneUpApp.io site [1].

```
# node list
top31names = []
for elm in top31:
    top31names.append(elm["name"])

# adding nodes
first = 0
for elm in top31:
    sub_weight = elm["subs"]/1000000
    print(elm["name"])
    print(sub_weight)
    G.add_node(elm["name"], weight=sub_weight)

    base_url = "https://subredditstats.com/subreddit-user-overlaps/"
    stats_url = "https://subredditstats.com/r/"
    driver.get(base_url + elm["name"])
    time.sleep(4)
    pre = driver.find_element_by_xpath("//*[@id='outputEl']")
    divs = pre.find_elements_by_tag_name("div")

    similar = []
    ite = 0
    if first != 0:
        for ch in divs:
            col = ch.text.split(" ", 1)
            col[1] = "".join(col[1].split())
            if col[1] in top31names:
                G.add_edge(elm["name"], col[1], weight=float(col[0]))
    first = first + 1

    nx.write_gml(G, "top31subreddits.gml")
```

Figure 1: Code snippet using Selenium to scrape each subreddit's page for statistics

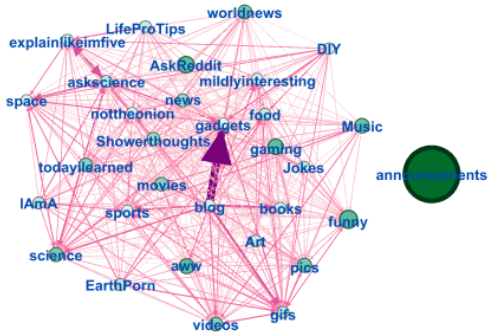

Figure 2: Initial network, with 31 nodes and 580 directed links

A directed network graph was constructed using NetworkX, where each subreddit was represented as a node, and the directed edges between the nodes represented the user overlap interactions between subreddits. The edge weights were calculated using a similarity score that measures the likelihood that being a part of one subreddit will lead to posting or commenting on another subreddit [3] For example, if a subreddit (A) has a similarity score of four with another (B), it would imply that users in A are four times as likely to post/comment on subreddit B. Figure 2 shows the initial network - note that the largest node is a singleton as it was closed to user posts/comments and is aptly restricted to moderator announcements. Aside from this node, every node has links to every other node, since there are similarity scores between these thirty subreddits. After browsing the data, it became prudent to limit the edges to only those with edge weights greater than 3; this signalled a significant enough level of overlap.

2.2 Modifications

The modified network shown in figure 3 excludes any edge that has an insignificant weight (similarity score less than 3). This produced two singletons (circled in blue): r/announcements and r/AskReddit. Givent that r/announcements and r/AskReddit do not have significant enough overlaps with the rest of the nodes, they were excluded from the network analysis. This leaves 29 nodes with 337 links. The network diameter of three is indicative of the ultra-small nature of Reddit. These nodes are all within a single giant component, with an average degree of 11.621 (weighted degree of 29.137). The average path length is 1.4. The network is moderately sparse, with a density

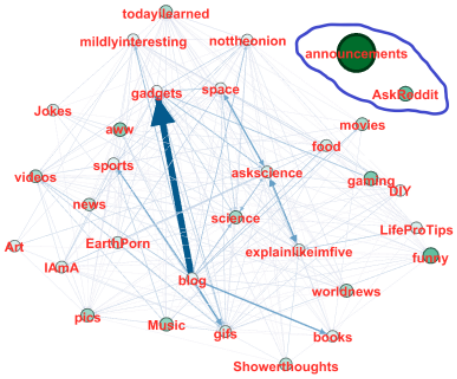

Figure 3: Modified network excluding edges with weight less than 3

of 0.415. The modularity is very low at 0.171, giving a preview that clustering this network will be fairly difficult.

2.3 Important Nodes

The analysis of the network revealed several important nodes that serve as hubs within the network. These hubs are characterized by having high out-degree or in-degree, indicating their influence within the network. The nodes with the highest out-degree were found to be r/blog, r/gifs, and r/videos, suggesting that these subreddits are popular origin points and may act as influencers within the network. Closer inspection of these subreddits shows that they contain a multitude of links to other subreddits in the posts or comments, which explains how users visit many other communities from these high out-degree nodes. On the other hand, the nodes with the highest in-degree were found to be r/gifs, r/gadgets, and r/nottheonion, indicating that these subreddits are popular destinations for users to visit. This is reinforced by the fact that these communities often have crossposts to other subreddits, where users who see the post in the smaller subreddits may click on the original post and arrive at the high in-degree node.

One edge that stands out above all others is the directed link from r/blog to r/gadgets, whereas users in r/blog were more than twenty-seven times as likely to post or comment in r/gadgets, which is staggering. It implies that r/blog is the most highly influential node, and r/gadgets is the most primary benefactor of the discussions taking place in the former node. Other relationships of note include the other nodes which r/blog heavily influences, such as r/sports, r/books, and r/gifs. There is also the double relationship between r/askscience and r/explainlikeimfive, which are equal in similarity to each other. These two nodes have a perfect give-and-take harmony of user traffic. The r/space community has this sort of relationship to r/askscience, albeit in a lesser sense.

Another important measure of node centrality is betweenness centrality, which identifies nodes that act as critical intermediaries in the network. The subreddits with the highest betweenness centrality were found to be r/gifs, r/gadgets, and r/pics, suggesting that these subreddits may play an important role in connecting different parts of the network and facilitating the flow of information. They play relatively equal parts in acting as influencers and serving as popular destinations. Overall, these findings shed light on the structure and behavior of the network, highlighting the important role of certain nodes in shaping its dynamics.

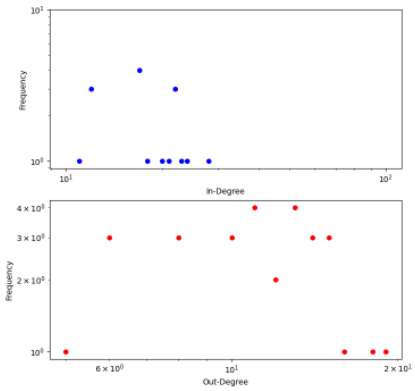

Figure 4: In/Out-Degree Frequency Distribution (log-log)

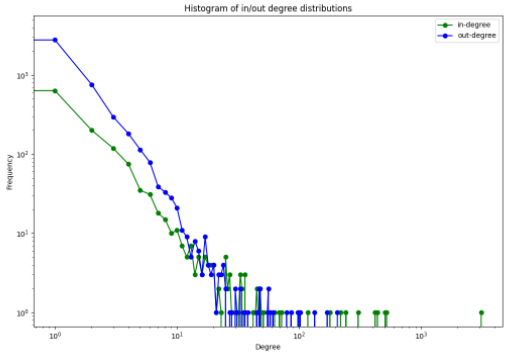

Figure 5: In/Out-Degree Histogram

3 Results

3.1 Degree Distributions

The shapes of the in-degree and out-degree distributions in Figure 4 are quite dissimilar. The in-degree distribution is concentrated on the lower degree nodes, whereas the out-degree distribution is slightly more evenly spread, with a bias for higher out-degree nodes on average. Given that the in-degree distribution is right-skewed, while the out-degree distribution is left-skewed, this could indicate that nodes are more likely to receive incoming connections from a larger number of less-connected nodes, but are also more likely to make outgoing connections to a few highly connected nodes. Figure 5 shows largely overlapping degree distributions - that out-degree has a greater slope is indicative of the above. This overall linear trend

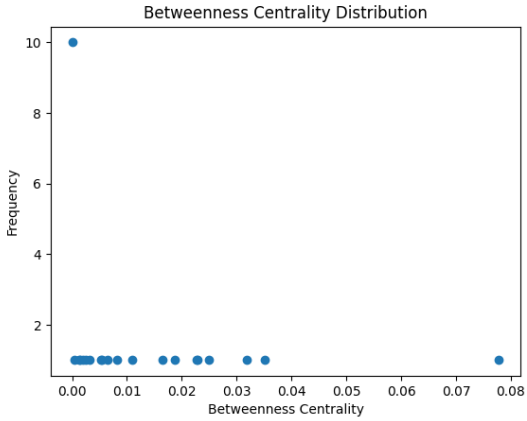

Figure 6: Betweenness Centrality Distribution

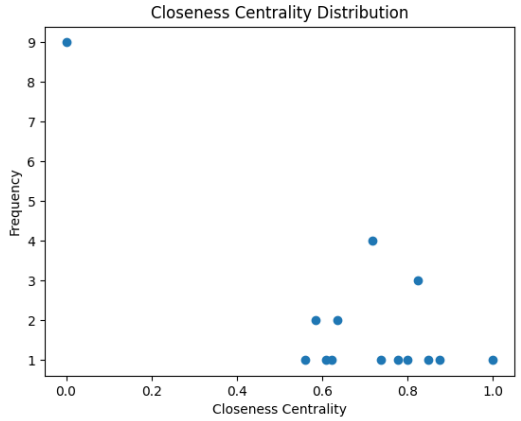

Figure 7: Closeness Centrality Distribution

combined with the a pooling of high degree nodes is most telling of this network's scale-free behavior.

3.2 Centrality Distributions

Betweenness centrality measures the importance of a node in the flow of information through the network, while closeness centrality measures how easily a node can reach all other nodes in the network. The distributions for these measures are shown in Figures 6 and 7. For the top thirty one subreddits, the betweenness centrality is very low, with most of the nodes' individual betweenness values being concentrated on the lower side. Only one node stands as an outlier in betweenness, represented by r/gifs (due to its pivotal bottleneck position). However, the closeness distribution suggests that the network is skewed left, with the high closeness centrality average of the nodes showing a more tightly connected network. This is consistent with the high degree and large number of

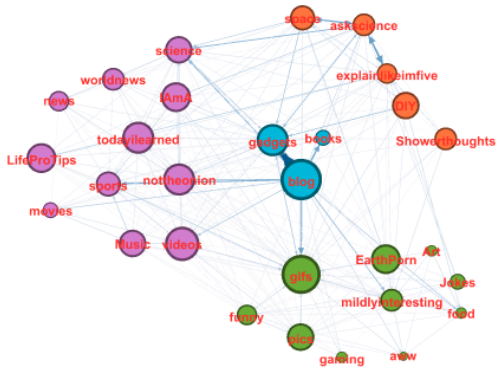

Figure 8: Four discernible clusters

edges throughout the network. It's also indicative of the Reddit platform, since information can be readily shared and crossposted onto other subreddits.

3.3 Clustering

The clustering coefficient measures the degree to which nodes in a network tend to cluster together, indicating the presence of tightly connected groups or communities. The average clustering coefficient for this network is 0.604, which suggests that the network is fairly highly clustered, with many nodes forming tighter connected groups or communities. It's likely this high because of the greater number of triangles present as a result of relative node groups. This can have implications for the way information flows through the network, whereas nodes within a community are more likely to share and exchange information with each other than with nodes outside of the community. A high clustering coefficient can also inform about the presence of small-world characteristics in the network, where most nodes are not directly connected to each other but can be reached through a small number of intermediaries. This can have implications for the efficiency of information flow through the network, as it suggests that information can be transmitted quickly and efficiently even across long distances within the network - this is true for most social networks that share information.

Girvan-Newman clustering did not work well for this network and always yielded only two communities, one of which was a single node (almost always r/gifs). However, the Louvain clustering based on optimizing modularity produced 4 distinct communities (shown in Figure 8) which are interpreted as follows. The cluster of pink nodes

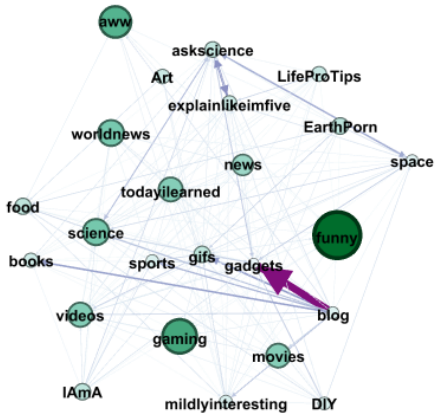

Figure 9: Random 5 nodes removed

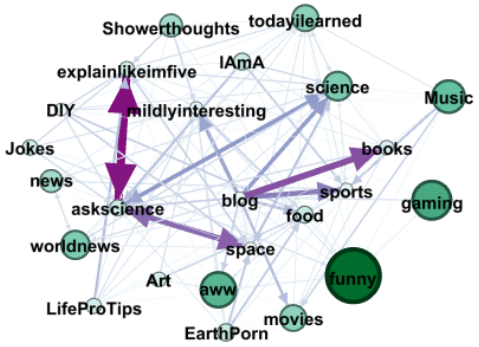

Figure 10: Targeted 5 highest betweenness nodes removed

represents subreddits such as r/worldnews, r/music and r/todayilearned, that share news, recent discoveries and the latest pop culture. The orange nodes such as r/space, r/explainlikeimfive, r/DIY and r/askscience are firmly in the cluster which covers science and concise explanations of topics. The green cluster disseminates all varietes of humor and any media content that is highly entertaining and engaging. The r/funny, r/gifs, r/food and r/gaming communities are good examples of this. There is also a central community which represents subreddits partaking in discussions and general information spreading. These forum-esque nodes often feature a high amount of comments when compared to the other communities.

3.4 Resilience

This network was reproduced and attacked in two ways: a random strategy which selected five random nodes to be removed from the network, and a targeted strategy, which removed the five highest betweenness nodes. Figures 9 and 10 show the resultant networks when the two strategies have been carried out; node the shift in node weights and edge weights. Table 1 shows how the random attack did not alter average degree, through it did increase the relative edge weights of the network/ The network became more dense as a result of eliminating nodes on random, which would likely only remove some sparse nodes. The modularity, clustering coefficient and average path length remained the same. In the visualization of Figure 9, the network looks more or less the same as the baseline network However, when considering a targeted attack on high betweenness nodes, the network performed much more poorly, which a significant reduction in edges. Thought the diameter of 3 held, the average degree plummeted, as did the density of the network. Modularity went up slightly as a result but the overall clustering coefficient was reduced severely. Perhaps it was slightly easier to identify communities with modularity but the number of triangles dropped greatly, whereas the path length increased. Overall, the network was more negatively impacted by this targeted attack. In Figure 10, the targeted attack visualization shows a completely changed network with different relative weight shadings for the edges. The relationships between r/mildlyinteresting and r/askscience, along with the aforementioned r/blog influences, are highly reflected in this network.

Table 1: Network properties across attacks

Metrics	Baseline	Random	Targeted
Nodes	29	24	24
Edges	337	286	160
Diameter	3	3	3
Avg. Degree	11.621	11.917	6.667
Avg. W. Deg.	29.137	78.428	34.483
Density	0.415	0.518	0.29
Modularity	0.171	0.130	0.215
Clust. Coeff.	0.604	0.635	0.489
Avg. Path Len.	1.4	1.314	1.670

3.5 Network Nature

Figures 11-15 show the randomly generated set of different types of networks: Barabasi-Albert, Erdos Renyi, and random networks with varying edge probabilities. Random networks are a type of network in which nodes are connected randomly and uniformly, without any preferential attachment or bias. As a result, random networks tend to have a Poisson degree distribution, which means that the degree of nodes in the network follows a normal distribution. Erdos-Renyi (ER) networks

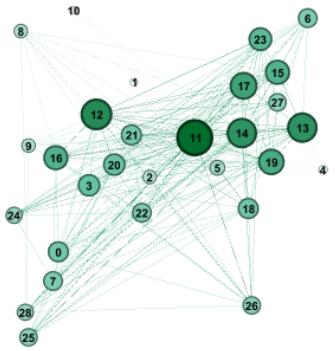

Figure 11: Randomly generated Barabasi-Albert network

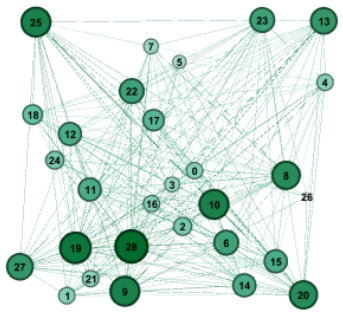

Figure 12: Randomly generated Erdos-Renyi network

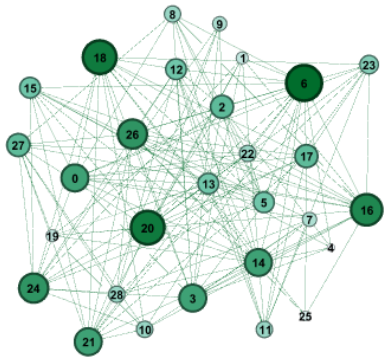

Figure 13: Randomly generated random network with p=0.3

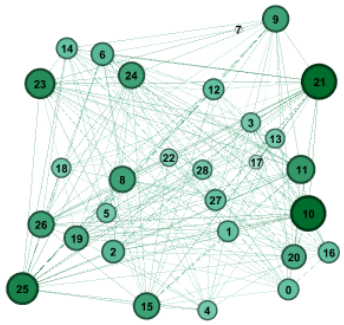

Figure 14: Randomly generated random network with p=0.4

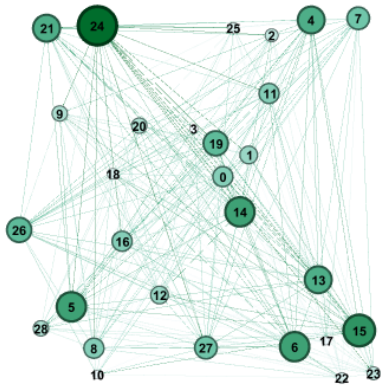

Figure 15: Randomly generated random network with p=0.5

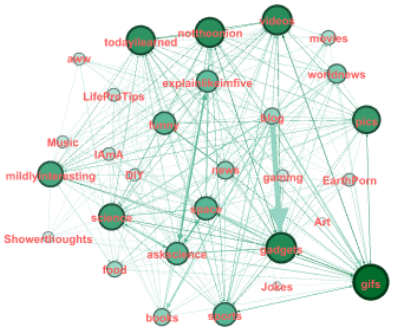

Figure 16: Original network

are a specific type of random network that is generated by connecting pairs of nodes with a certain probability. In an ER network, each edge has the same probability of existing, regardless of the degree of the nodes it connects. As a result, ER networks tend to have a Poisson degree distribution, similar to random networks. Barabasi-Albert (BA) networks, on the other hand, are generated by a preferential attachment mechanism, in which new nodes are more likely to connect to nodes that already have a high degree. This leads to a power-law degree distribution, in which a small number of nodes have a very high degree, while the majority of nodes have a low degree. BA networks tend to have a highly clustered structure, with a small number of highly connected hubs and many disconnected nodes. Comparing the baseline network in Figure 16 to these derivations reveals that, though the network may look deceptively random due to high connectivity (especially in the unmodified network in Figure 2), it most resembles the BA network, further indicating its nature of being scale-free. This is logical considering the heavy presence of preferential attachment across Reddit.

3.6 Spreading

This subreddit network would probably fit an SIR model - whereas some communities may become toxic in their discourse (S), they may eventually have a chance of becoming restricted or outright banned (I), which may last for quite some time depending on the severity of the toxicity. Eventually, some communities may come out of this ban, albeit with more rules and scrutiny (R). In such a model, r/gifs and r/nottheonion may be the most susceptible to being "infected", since the general nature of these bottleneck nodes receiving and sending information could increase chances for toxicity. When considering an actual disease, perhaps in the form of a virus or a network outage, then most of the network would be greatly affected due to the high connectivity and high closeness of the network. However, the application of this spreading model deserves its own study.

4 Conclusions

In conclusion, this study presents a network analysis of interactions between subreddits on the Reddit platform. The network graph's visualizations and global characteristics reveal that the platform's structure is scale-free, and certain subreddits act as important hubs that facilitate communication and information flow. The network is fairly dense and well-connected, with high influencer and destination hub types. The clustering reveals four general subdivisions based on genre. The vulnerability of the network to targeted attacks suggests that measures should be taken to ehance the nature of discourse critical subreddits from toxicity to prevent network failure. The resilience of the network to random failures suggests that the platform can tolerate the loss of low-degree subreddits without significant impact on the network's structure. Finally, the implications for spreading models suggest that the network's structure can greatly influence the spread of information on the platform, and measures should be taken to prevent the spread of misinformation from critical subreddits. In the future, it would be prudent to scrape data on a much larger scale to include many more top subreddits, as well as implementing the spreading model and seeing its implications on the network.

References

[1] Baer, D. (2022, May 4). The 31 biggest subreddits (2022 update). OneUp Blog. Retrieved April 4, 2023, from https://blog.oneupapp.io/biggest-subreddits/

[2] Ruby, D. (2023, March 11). 101+ reddit statistics for 2023 (Users amp; Traffic Data). Demand Sage. Retrieved April 15, 2023, from https://www.demandsage.com/reddit-statistics/

[3] Subreddit Stats. (n.d.). Subreddit user-overlap. Related Subreddits By User/Redditor Overlap. Retrieved April 10, 2023, from https://subredditstats.com/subreddit-user-overlaps