Jose Enrique R. Lopez 15 June 2021

CMSC 191 Final Project

Prominence and Clique-Formation in the Reddit Hyperlink Network

I. Introduction

Reddit is a social media platform based around communities rather than people. Unlike Facebook and Twitter which primarily connect individuals through online friendships, Reddit links 'subreddits'—communities which are devoted to a particular interest or cause. There are over 100, 000 active subreddits which span various academic disciplines, hobbies, ideologies, and any other topic of discussion that bind users of similar interest together. Subreddits enable members to post relevant content which gets upvoted or downvoted by other members, and posts with enough upvotes are promoted in the home page. Branding itself as 'the front page of the internet', Reddit enjoys the 18th rank among most-visited websites according to Alexa (2020).

As a social network, Reddit readily lends itself as an arena for testing homophily, the idea that entities with certain similarities are more likely to associate with each other (McPherson, Smith-Lovin, Cook, 2001). Specifically, it is of interest whether such an idea, applied more so to individuals, extends to larger communities as well. For example, are subreddits with greater thematic similarity more likely to hyperlink each other? Another reason why studying Reddit is worthwhile is strategic content moderation. Like other social media platforms, Reddit is confronted with the spread of objectionable user content; for example, those that actively promote violence and transphobia. Occasionally, Reddit bans or quarantines such controversial subreddits for these reasons. But aside from the formal measures imposed top-down by Reddit, it is also important to focus on informal sanctions exercised by fellow subreddits. Therefore, it is useful to determine which subreddits are most influential (depending on how one defines influence) and are in a position to act as gatekeepers for the wider Reddit community. Essentially, studying Reddit as a social network is both of theoretical and practical importance.

II. Methodology

We conduct exploratory social network analysis on the subreddit hyperlink network using Pajek 5.12. We use the dataset generated by Kumar, Hamilton, Leskovec, and Jurafsky (2018) which tracked the hyperlink ties between any two subreddits from January 2014 – April 2017. For this initial analysis, we used the network file for the bodies of posts only, excluding the titles. The Excel2Pajek software was used to convert the tab-separated network data into a workable Pajek file. This resulted to a directed graph with 12, 822 vertices, 37, 287 arcs and a network density of 0.04%.

The level of influence of each subreddit was determined using three metrics. First, we measure the **degree prestige** of each subreddit based on incoming arcs. This tells us by how much a particular subreddit has been hyperlinked by other subreddits. Subreddits with the highest degree prestige are most popular and are in a better position to influence other subreddits directly and help informally moderate their content. Second, we measure the **proximity prestige** of each subreddit based on incoming arcs. This summarizes the influence domain of each subreddit and the number of subreddits which could hyperlink them by association with those that have. Proximity prestige tells us the number of potential communities which a particular subreddit can influence, not just those that they already influence. Subreddits with the highest proximity prestige have the highest *potential* to be hyperlinked by other subreddits. Lastly, we measure the **betweenness centrality** of each subreddit on a symmetrized version of the network. This metric tells us by how much a particular subreddit acts as a bridge between other subreddits in the Reddit community. This is useful for establishing the gatekeepers between communities especially those that link controversial and non-controversial communities. Taken together, these measures provide a high-dimensional account of centrality and prestige in the Reddit community. All these measures were found through built-in options in the vector tab of Pajek.

We also identify cliques in the network in relation to influence. Cliques are subgraphs where each vertex connects to every other vertex and where the inclusion of any outside vertex does not result to an altogether complete graph. We perform clique-finding on the symmetrized network. Fragments were used to search for 3-cliques, 4-cliques, and 5-cliques in the network. Such information is useful for exploring homophily and whether subreddits with closely aligned

topics are more likely to form larger subreddit groups. Such larger subreddit groups potentially exert greater influence on other subreddits than if these subreddits acted alone.

Throughout the analysis, the social networks were visually represented through graph drawings, some of which are presented in this paper. To reduce edge occlusion, only relevant subnetworks were visually represented (e.g., top 10% scorers in the metrics). The networks were also energized through Kamada-Kawai, featuring the most prominent nodes in the center.

III. Results and Discussion

A. Prominence

The simplest measure of prominence is degree prestige which in this case pertains to how much a subreddit has been hyperlinked by other subreddits. The table below shows the ranking of the top 10 subreddits based on input degree prestige as generated by Pajek through an input degree vector. It also includes a cluster column which indicates the number of times a subreddit has been hyperlinked by another subreddit.

Table 1. Top 10 Subreddits by Input Degree Prestige

Rank	Vertex	Cluster	Id
1	35	2246	askreddit
2	23	1336	iama
3	7742	1045	pics
4	9697	799	todayilearned
5	9704	783	videos
6	738	776	funny
7	1386	642	worldnews
8	13	624	dogecoin
9	9700	571	adviceanimals
10	1	518	leagueoflegends

AskReddit and I AmA occupy the top 2 spots; in fact, they do so consistently for the remaining measures. This means that they are most recognized by other subreddits, and may be most influential in helping informally moderate other subreddits' content. Observe that many of the above subreddits are general information communities such as "pics", "videos", and "world news". This may explain why they are popular: generality implies inclusivity. Conversely, specialized subreddits such as those devoted to theoretical computer science, Iglesia ni Cristo, and rugby may have less degree prestige simply because fewer people recognize these topics.

The drawing below shows the top 10% of subreddits based on input degree prestige. It has been partitioned according to the cluster values of incoming arcs so that vertices of identical colors have the same number of incoming arcs. The network has also been energized through Kamada-Kawai, featuring the most prestigious nodes in the center.

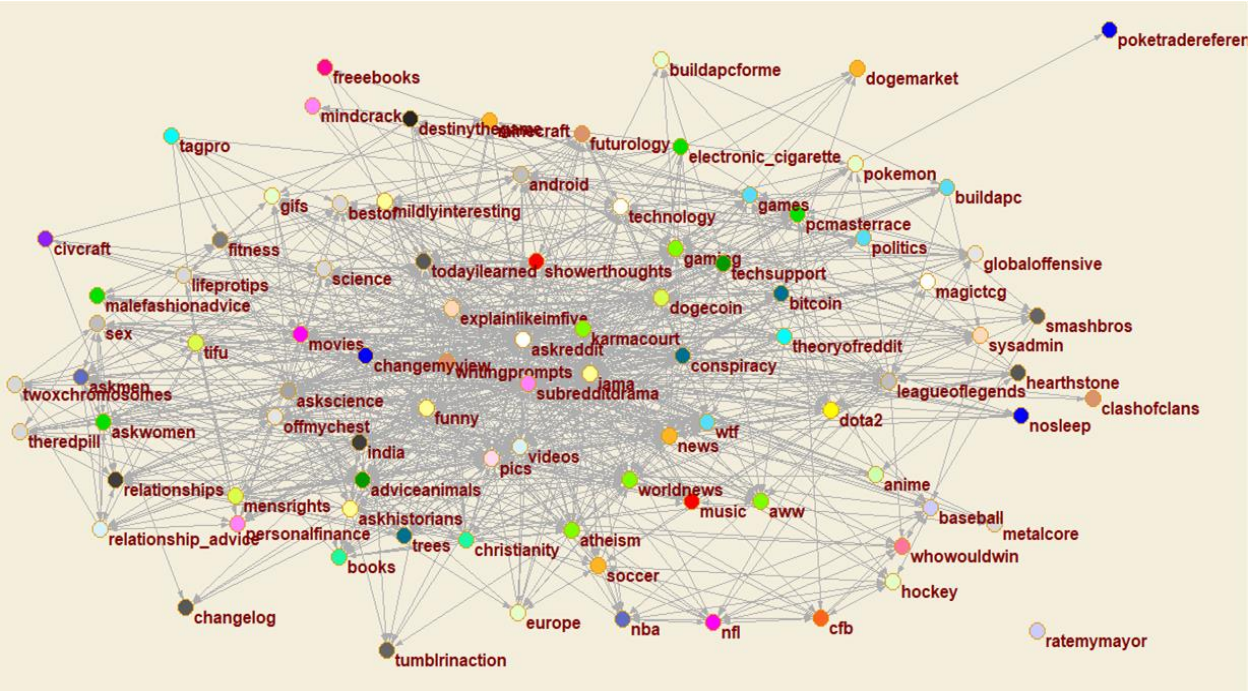


Figure 1. Energized Social Network of Top 10% Subreddits According to Input Degree

What does the generality trend imply about content moderation? One may be tempted to think that general subreddits like “pics” and “videos” are too politically neutral to be able to take a stand against objectionable content generated by other subreddits. However, if we examine these subreddits closely, we may conclude that this is not the case. For example, I AmA, which allows users to introduce themselves and their personal projects, has its top post endorsement for the Black Lives Matter Movement. So, generality does not imply political neutrality, and these top subreddits may be able to help in content moderation due to their influence.

The results for proximity prestige are similar to that of the former, and are shown below:

Table 2. Top 10 Subreddits by Proximity Prestige

Rank	Vertex	Value	Id
1	35	0.2853	askreddit
2	23	0.2798	iama
3	7742	0.2567	pics
4	9704	0.2561	videos
5	738	0.2482	funny
6	9697	0.2481	todayilearned
7	563	0.2464	gaming
8	1386	0.2411	worldnews
9	6495	0.2406	technology
10	9703	0.2397	wtf

Proximity prestige values range from 0 to 1; the higher the value, the greater the number of direct and indirect nominations of a particular subreddit. Again, we see askReddit and I AmA occupying the top spots, but this time including indirect influence as well. Compared to the former, this table features even more general subreddits, including todayilearned and technology, indicating that general topic subreddits do tend to be more popular.

Lastly, the scores for betweenness centrality are somewhat different compared to the first two measures. The relevant table is shown below:

Table 3. Top 10 Subreddits by Betweenness Centrality

Rank	Vertex	Value	Id
1	35	0.1758	askreddit
2	23	0.1287	iama
3	71	0.0873	subredditdrama
4	13	0.0476	dogecoin
5	7742	0.0375	pics
6	43	0.0313	writingprompts
7	563	0.0310	gaming
8	46	0.0301	hailcorporate
9	9704	0.0280	videos
10	72	0.0257	bitcoin

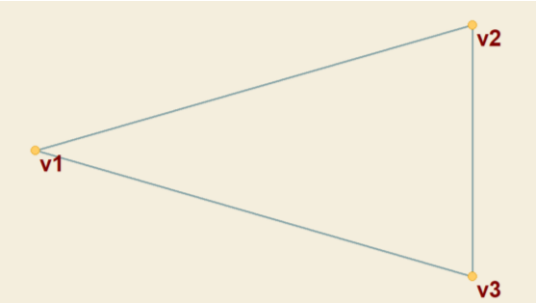
While askReddit and I AmA are top 2, notice that it is the first time for some subreddits to make it to the top 10 such as subredditdrama, hailcorporate, and bitcoin. Subredditdrama is particularly noteworthy as a community devoted exclusively to discussing fights, flame wars, and

rants within and among subreddits. One may consider it as a court of public opinion within the Reddit community. It makes sense therefore that it acts a bridge, or a common ground between and among subreddits which may not be in good terms with each other. It is well-placed as a special community which exercises informal conflict resolution and social control within the Reddit community. The other subreddits in the top 10 are also well-poised to assume these role of conflict resolution and social controls especially in relation to issues that affect the subreddit's interests (e.g., good economic practices for hail corporate and freedom of speech for writing prompts).

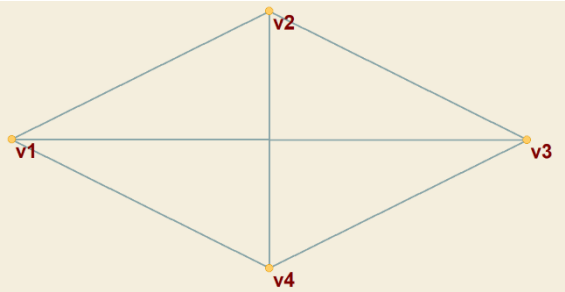
B. Clique Formation

Cliques signify tight-knit relations among entities as these are the most dense a network can be. Tight-knit relations most often exist among entities with similar values and interests. This is the principle of homophily (McPherson et al., 2001), which is more so associated with individuals rather than communities. This is understandable as homophily among large-scale groups is logistically harder to track. This makes Reddit a convenient case for studying homophily among online communities.

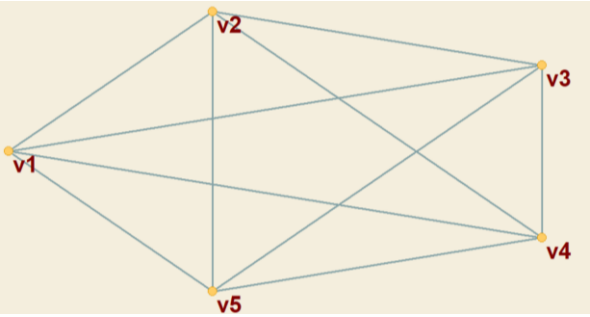
For this initial study, the n-cliques we consider are 3-cliques, 4-cliques, and 5-cliques. Finding cliques in the Reddit community was less straightforward since there is no corresponding built-in Pajek function. Several steps were needed to be performed for this purpose. First, fragments were created as motifs for finding the respective cliques, as follows:



A triad



A quartet



A quintet

Figure 2. Respective Fragments Used for Hunting 3-Cliques, 4-Cliques, and 5-Cliques (clockwise)

Second, each of these fragments were located in the symmetrized Reddit network and three resulting networks containing the fragments were generated. Third, each network was partitioned into clusters based on the degree of each node. For instance, if UPLBmemories subreddit was part of Cluster 2 among all 3-cliques, then UPLBmemories has degree 2 (only 2 subreddits have hyperlinked it) and is part of one (and only one) 3-clique. Lastly, we extract all n-cliques that belonged to Cluster n-1. For 5-cliques, for example, we extract the nodes that had degree 4 only. This is an important step—without which, we may be able to extract complete graphs but not *maximally* complete graphs (not cliques!). Hence, we may not be able to get all n-cliques using this method but at least we are sure that we are strictly getting cliques.

Let's first consider the 3-cliques of the Reddit network, some of which are shown below.

Figure 3. Some 3-Cliques from the Reddit Hyperlink Network

It would seem that the subreddit cliques converge around similar topics. For instance, on the upper left corner, we see a clique involved in game theory and its applications to society. On the upper right hand corner, we see cities in San Francisco. This may indicate that homophily applies to communities as well.

Following are the resulting graphs for 4-cliques and 5-cliques:

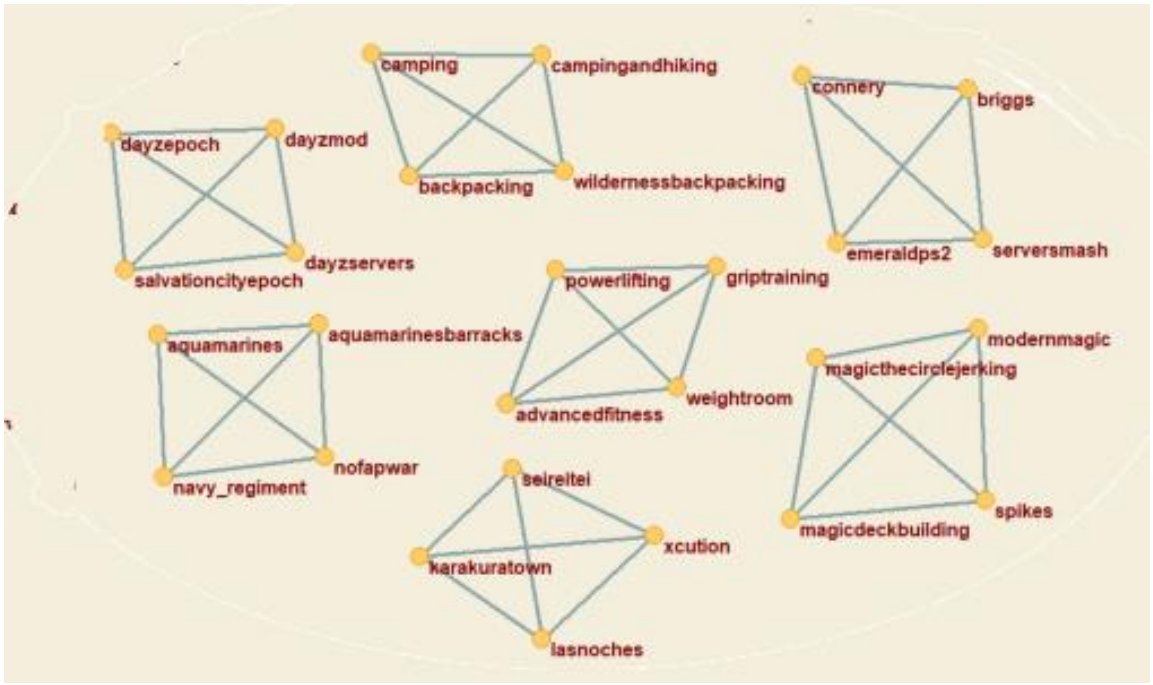


Figure 4. Some 4-Cliques from the Reddit Hyperlink Network

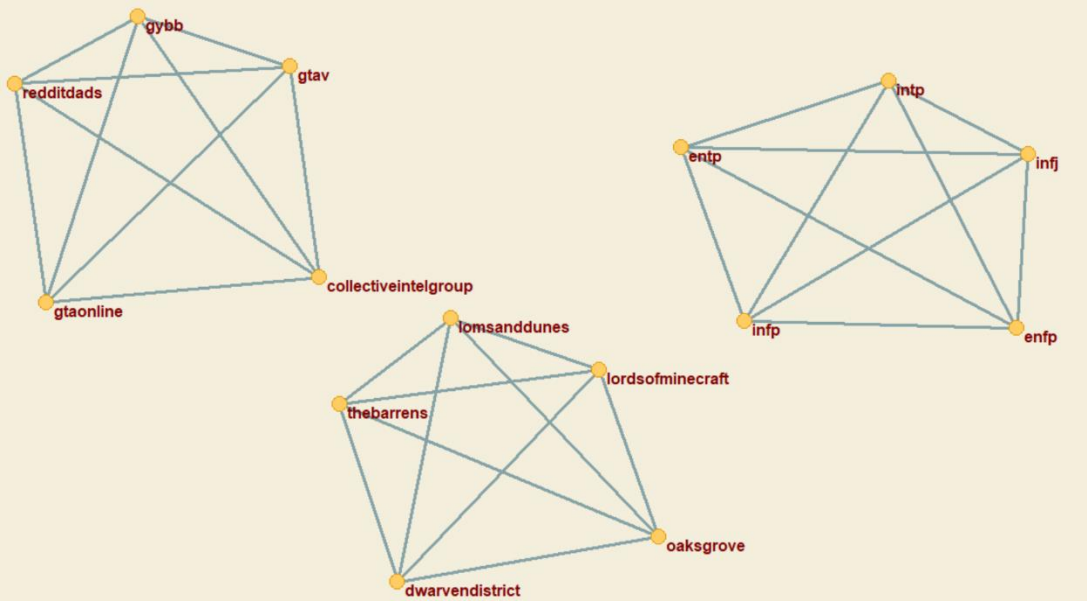


Figure 5. Some 5-Cliques from the Reddit Hyperlink Network

Indeed, the subreddits of 3-cliques, 4-cliques, and 5-cliques converge around thematically similar topics, which isn't surprising. The implication is, however, is that some of these cliques may become so cloistered to the point that they become echo chambers. This isn't a real danger unless the subreddits actively promote violence or transphobia. As it is, the above cliques feature

topics that are mundane such as camping and gaming. As n gets larger, we may expect to see more diversity among topics and even subreddits which are in conflict with each other, which may be the subject of a more rigorous statistical study.

IV. Summary

This exploratory study analyzed prominence and clique formation within the Reddit community. The most prominent subreddits are those that are of general interest such as world news and sports, which is a reason why they are most popular. Their generality, however, does not imply neutrality and they are well-placed in helping moderate the content of other subreddits. Other less prominent subreddits such as subredditdrama which scored high in betweenness centrality also assume important roles in gatekeeping among controversial and non-controversial subreddits. On the other hand, clique formation was consistent with homophily as subreddits in cliques revolved around common themes. Hence, this study also serves as conceptual foundation for more rigorous statistical studies testing homophily among online communities. In particular, we are interested in how homophily persists or changes among n-cliques as n gets larger with more prominent nodes being included. This may help in challenging echo chambers and pave the way for strategic content moderation in Reddit and other social media platforms.

References

Alexa Internet. (2020). Reddit Competitive Analysis, Marketing Mix and Traffic". http://www.alexa.com/siteinfo/reddit.com

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). *Community Interaction and Conflict on the Web. Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18.* doi:10.1145/3178876.3186141

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). *Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology, 27(1), 415–444.* doi:10.1146/annurev.soc.27.1.415