Analysing popularity through resubmitted content on reddit.com

COMP0123: Complex Networks and Web - Coursework

Mo Afsharmoqaddam

January 6, 2019

Abstract

During the last few years Reddit has become one of the most popular internet destinations for users in United States[1] and in the world. Reddit is an American platform that provides social news aggregation, web content rating as well a discussions website which allows users to submit content [2]. The goal of this study is to use a network analysis approach to identify the best strategies in order for a user or a user's post to become popular through resubmitted content, specifically images. We will discuss what is meant to be popular and how popularity is measured as well as decomposing the data set to understand the underlying network structure. Additionally, the report will showcase key findings of network metrics such as the rich club structure, mixing patterns, PageRank and how they correlate towards popularity. The results indicate that although popularity can be a subjective matter, the network follows a scale free and rich club properties, and the best way to be connected is to have connectivity with the richest users of the network. The study also highlights that one of the popular users identified within the network is indeed the CEO of Reddit.

Contents

1 Introduction

1.1 Popularity

Throughout centuries popularity of a person or even an idea has been defined in terms of liking, attraction, dominance and superiority. Specifically, I will be focusing on sociometric popularity which can be defined as how liked an individual is [5]. Especially with the age of social media the urge of becoming popular has increased significantly. Humans devote approximately 30-40% of all speech to talking about themselves [3] and this increases significantly to 80% within social media. Additionally, sharing has become easier and 68% of people who share is because they want to give others a better sense of who they are and what they care about [4]. However, one of the biggest reasons of why we share is that 78% of people want to stay connected to others. Why someone would want to become popular will have many answers and many people have different intentions, but what is clear is that popularity is a trait that many strive for especially in a status-obsessed world.

1.2 The Problem

Therefore, in this paper we will be investigating the question of what is the best strategy for someone to get popular on Reddit through re-submissions, and whether originality matters? This question as a whole will encompass sub question such as is it better to take a good image and post it or whether its more important linking with popular users. Essentially, is it better to post first or re-post content by popular users? We will also be presenting the top *"popular"* users and their features within the network structure. The analysis presented in this paper will yield interesting results as well as some novel approaches in regards to how popularity is measured.

1.3 Data, Algorithms and Tools

To kick start this project I have used the Stanford Large Network Data-set Collection, which contains 132,308 reddit.com image re-submissions [6]. Algorithms are used to compute the network metrics as well my own measuring system of popularity which will be discussed later in this paper. This study is mainly focused on metrics such as the degree distribution, in-degree, out-degree, rich club coefficient and structure, mixing patterns, network diameter, betweenness centrality, authority, hubs and PageRank. A combination of statistical and graphical tools/libraries were used to perform such analysis. The graph construction was done in Python using networkX [7] for advanced network analysis as well as using Gephi [8] for visualisation purposes and the computation of simple network properties. Microsoft Excel was also used for data handling.

1.4 Achievements and Key Results

REVIEW THIS One of the key achievements were uncovering the underlying structure of the data set and to identify relevant patterns within it. Some of the key results were that the network followed a rich club structure as well as having mixing pattern that was close to neutral. The social graph between between users who had high number of re-submissions as well as in degrees outlined a degree distribution where a few users were considered very popular and most popular users were within the rich club structure. Additionally, custom popularity metrics highlighted different network properties using Reddit's scoring system which will be discussed in more detail later on in the paper. However, we will also discuss other popularity factors through PageRank, HITS and inbetweenness which potentially can be beneficial to other users but perhaps not the best way of gaining popularity.

2 Background

2.1 Reddit

Reddit is an American social news aggregation, web content rating and discussion website. People around the world can register accounts and submit content in all forms of media such as links, images as well as text posts. This study focuses on content specifically that are of images that have been resubmitted multiple times by different users. Such posts are orgnaised by subjects called "subrreddits" which cover a variety of different topics.

Additionally, every post submitted to the website can be given upvotes and downvotes by the other users. Submissions with large amount of upvotes appear towards the top of their respective subreddit and if enough upvotes are received, they can be displayed even on the front page (homepage) of Reddit. With such a scoring system the goal of every user is to gain as much upvotes as possible to gain popularity and recognition by posting interesting and engaging content. Furthermore, a personal scoring system is also devised which simply corresponds to a users "karma" score that can be viewed by other users: $upvotes - downvotes$.

Since February 2018 Reddit has had 542 million monthly visitors, ranking 6th in the world [1]. The website is most popular in the United States, followed by United Kingdom and Canada. Across 2018, there were 153 million submissions, 1.2 billion comments and 27 billion upvotes [9].

2.2 Network Analysis

Networks are everywhere in our everyday lives and behind each complex system, there are networks that define interactions between different elements of the system. We can perform analysis on networks to find out different structures and structural properties through empirical reasoning, mathematical modelling and different algorithms for analysing graphs. By doing so we hope to uncover patterns and statistical properties of the network data. Additionally, we would like to be able to design principles and models as well as understand and

predict behaviours of networked systems. Below are some of the key network properties and algorithms used in this study.

2.2.1 Degree Distribution, In-degree & Out-degree

When studying networks the degree of a node is a representation of the number of links it has with other nodes within the network. The distribution of this is the probability of such degrees over the whole network [10]. Similarly, the in-degree simply means the number of inbound links and the out-degree states the number of outbound links. These are considered first order properties which showcases the node's connectivity as well as evaluating and predicting the structure of the network.

2.2.2 Betweenness centrality

Centralities are second order properties which indicate which node is the most central or important and highlight the connectivity between nodes. In this report I will be focusing on betweenness centrality. Betweenness centrality measures how many pairs of nodes would have to go through a particular node in order to reach one another in the minimum number of hops (shortest path). Nodes with a high betweenness centrality are important for transport.

2.2.3 K-Core and Modularity

Other set of tools analyse communities and clusters within a network. The concept of modularity measures how well a network is partitioned into communities, essentially, internal structures of networks can be discovered as it describes how the network is compartmentalised [11]. Furthermore, K-Core decomposition highlights the hierarchical structure of large networks as it removes the least connected nodes.

2.2.4 Rich Club and Mixing Patterns

Other second order properties include the rich club coefficient and mixing patterns. Mixing patterns, at the link level, show the relation between degrees of the two ends of a node. At the network level they can show to have a positive correlation (assortative mixing), where nodes tend to connect with nodes of similar degree. Negative correlation (disassortative mixing), where high-degree nodes connect with low degree nodes or neutral mixing [12]. Many networks also contain small number of nodes with high degrees, also referred to as rich nodes. Such nodes play a dominant role in a network's structure and function and are part of the "Rich Club"[13].

2.2.5 PageRank and HITS

Link analysis approaches are also essential to compute how important a node is within a network. PageRank and HITS (Hubs and Authorities) algorithms determine influential nodes as well as hubs within a network. Both algorithms try to solve the same problems but differ slightly. PageRank performs by counting the number and quality of links to a page whereas HITS depends on the links that comes from a source node [14][15].

3 Related Work - Literature Survey

3.1 From Popularity Prediction to Ranking Online News

This paper explores the fact that news articles have become an engaging type of line of content which many Internet users interact with on a daily basis, especially with the growth of social media. Most news articles have been absorbed through mobile phones since they are short lived and low cost, hence, there is an increased interest in discovering articles that will become popular amongst users.

Alexandru Tatar et al. [16] addresses the problem of predicting the popularity of news articles based on user comments, using data from two key news sites in France and Netherlands. Being able to predict accurately the popularity of online content is incredibly valuable for different types of stakeholders. Some benefits are: better advertisement campaigns as well as profitable monetization strategies.

The performance and quality of their prediction showed a moderate performance. A way to increase accuracy for better ranking results were to include more features within the model. However, just including more features is not sufficient enough and other sources of information should also be included. For example, an additional source of information could be information about the user profiles that comment on the news articles, since popular users could formulate user communities around certain topics.

3.2 Identifying Social Roles in reddit Using Network Structure

As social networks and content creation increases and grows in size day by day, this network research study analyses the importance of understanding how users interact and produce content. By evaluating different dynamics within communities it is possible to measure content trust, group based recommendations and growth. Cody B et al. [18] explores user posting behaviour on reddit and answers one of the relevant question towards my case study, which is "whether users participate significantly in multiple communities?"

The results suggest that users avoid large interactions across multiple distinct communities. Approximately only 3% of users were identified in multiple communities. Such results is consistent with the paper's intuition of only 1% of the users performing significant participation across multiple communities. However, the paper does not demonstrate whether users that do cross boundaries behave consistently across the communities.

This paper will be very interesting to compare with high performing users conducted in my research as we can evaluate whether "popular" users who resubmit content, post in multiple communities and if so, are the communities similar in content. Since this will highlight the different relationships and dynamics that are created between users in different communities and whether it is beneficial to resubmit content in multiple communities to increase your own popularity.

3.3 The Rich Club Phenomenon in the Classroom

Luis M. Vaquero et al. [17] paper evolves around the evolution of online interactions performed by college students. The results are based on the relationships between social structure and the student's performance. A record of college student interactions (social interaction data) was compared with their academic scores.

The results showed that the higher number of social interactions (more frequent and intense) was usually a good indicator of a higher score for students engaging in them. Such interactions between the high performing students is encompassed within a "Rich Club", while the low performing student barely interact. The "Rich Club" is usually created within the first few weeks of college, which contains the high performing students that have showcased high persistent interactions. Interestingly, once students begin to be more persistence in their activities they do it more often until a maximum saturation point is met. They also show more willingness and collaboration as they *initiate* persistence in comparison to low performing students. Furthermore, low performing students begin to join the "Rich Club" mid way through the course, however, just for it to decrease again towards the end of the course. This shows that low performing students failed to join the "Rich Club" even after showcasing higher persistence since the "Rich Club" had already been formed.

Additionally, the content that was shared between low performing students were merely exchange of documents in trivial manner. However, higher performing students had more complex and a highly organised network of other peers with similar characteristic patterns. High performing students usually exchanged information in a chained manner, whereas low performing students shared information to many and were not included in the chains of high performing students. Generally, within the mid point of the course low performing students reduce their number of interactions which could be an indication of lack of motivation, however, many other external factors could cause less interactions and the lack of data does not allow for conclusion of whether lack of interaction, external factors or both lead to reduced performance.

4 Methodology

4.1 Data Extraction and Cleaning

To begin the research for Reddit re-submission content I used the SNAP Web data: Reddit submissions [6], this is a collection of 132,308 reddit.com submissions. Each submission is of an image, which has been submitted to Reddit multiple times. The data includes following features:

- Image ID - where submission with the same id are of the same image
- Time of submissions and local time
- Submission title

- Total votes - number of upvotes + number of downvotes
- Reddit ID - unique id of the submission on Reddit
- Individual number of upvotes and downvotes
- Subreddit - The community that the image was posted on
- Score - also known as "karma", number of upvotes - number of downvotes
- Number of comments the submission received
- Username

After extracting the data, the data consisted of many blank fields within the usernames. All blank fields were removed which reduced the dataset to 112,046 submissions.

4.2 Graph Generation and Popularity Methodology

After cleaning the data, I was than able to create the intended network. Using Python an edge list was created where each node represents a user and an edge between users are the re-submission of the same image posted in chronological order and converted into a CSV file. Due to having access to the time period of each submission, the graph generated will be a directed. The graphs were created with the use of networkX library.

Additionally, a nodes CSV table was generated which aggregated all of the important features for every user. The results show 63327 unique users and for every user the following features are calculated:

- Count of Reddit ID, i.e total number of re-submissions of content
- Sum of total votes
- Total number of upvotes
- Total number of downvotes
- Sum of score
- Total number of comments
- Total number of out degree
- Total number of in degree
- Popularity of image
- Popularity of person (user)

To measure the network properties outside of aggregated statistical data, a custom popularity definition was created. For each user to measure the popularity of an image they posted the following formula was used:

$$ImagePopularity = \sum(numberOfComments + score)$$

The idea is that for an image to be inherently popular, a high number of comments on a post would indicate a high amount of engagement, where users are highly interested in the content. Additionally a high **score** number also indicates a popular image which many people have enjoyed and liked.

Similarly the formula used to measure a users's popularity is as follows:

$$UsersPopularity = \sum(TotalVotes + InDegree)$$

The idea behind a user's popularity is around receiving the total number of votes, regardless of the votes being positive or negative. Since this shows the total attraction a person's received, as a person can be popular but not well liked. Additionally, the in degree of a user indicates how many users are being influenced by the node. The idea is to take in-links as votes. For example if a link from user A to user B indicates that user A is recommending, endorsing or voting for user B (similar idea to PageRank).

4.3 Algorithms

4.3.1 Rich Club Coefficient

As explained earlier many networks contain a small number of nodes with high degree, also known as "rich" nodes. We can calculate the rich club coefficient which is a quantitative measure of the density of inter-connectivity among the group of the riches nodes in a network. In this study I proceed to calculate rich-club coefficient as a function of node degree, where the group of rich nodes are those with degrees bigger than k:

$$\phi(k) = \frac{2E_k}{N_k(N_k - 1)}$$

where:

- N_k is the number of nodes with degree larger than k
- $\frac{N_k(N_k-1)}{2}$ is the maximum number of links among N_k nodes
- E_k is the actual number of link among the N_k nodes

4.3.2 Mixing Patterns

Mixing patterns show the relation between degrees of two end nodes of a link and whether if there is an positive, negative or neutral mixing. Mixing pattern measured by assortative

coefficient α is as follow:

$$\alpha = \frac{L^{-1}\sum_i K_i K_i' - \left[\frac{1}{2}L^{-1}\sum_i (K_i + K_i')\right]^2}{\frac{1}{2}L^{-1}\sum_i (K_i^2 + K_i'^2) - \left[\frac{1}{2}L^{-1}\sum_i (K_i + K_i')\right]^2}$$

where:

- L is the number of links
- K_i and K_i' are degree of two end nodes of link i
- $\alpha = 0$, neutral mixing, $\alpha > 0$, assortative mixing and $\alpha < 0$, disaassortative mixing

4.3.3 Betweenness Centrality

Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v:

$$C_B(v) = \frac{\sum_{j<k} g_{jk}(v)}{g_{jk}}$$

where:

- g_{jk} is the number of shortest paths between nodes j and k
- $g_{jk}(i)$ is the number of shortest paths between j and k passing through node v and $i, j, k = 1, 2, 3, ..., N$.

4.4 Methods and Tools for Network Analysis

For the network analysis stage and computation of above and other algorithms, different statistical and graphical tool-sets were used. Gephi was the main platform for network visualisation as well as performing up to third order network properties and link analysis such as: degree distribution, Betweenness centrality, modularity, PageRank, HITS and network diameter. Additionally, networkX was used to compute the rich club coefficient, assortativity as well as in degrees and out degrees. Furthermore, custom python scripts were written to calculate average neighbour degree distribution as well as the log based degree distribution for the network. The graphs and other analysis were created using other python libraries such as numpy, pandas and matplotlib.

5 Results

5.1 The Whole Picture

Before diving in and showcasing some of the results and the filtering process, I visualised the network as a whole to gain an understanding of its structure by just viewing how it looks, by doing so it gave me the overall outlook of the network and what relevant metrics I can begin to experiment with.

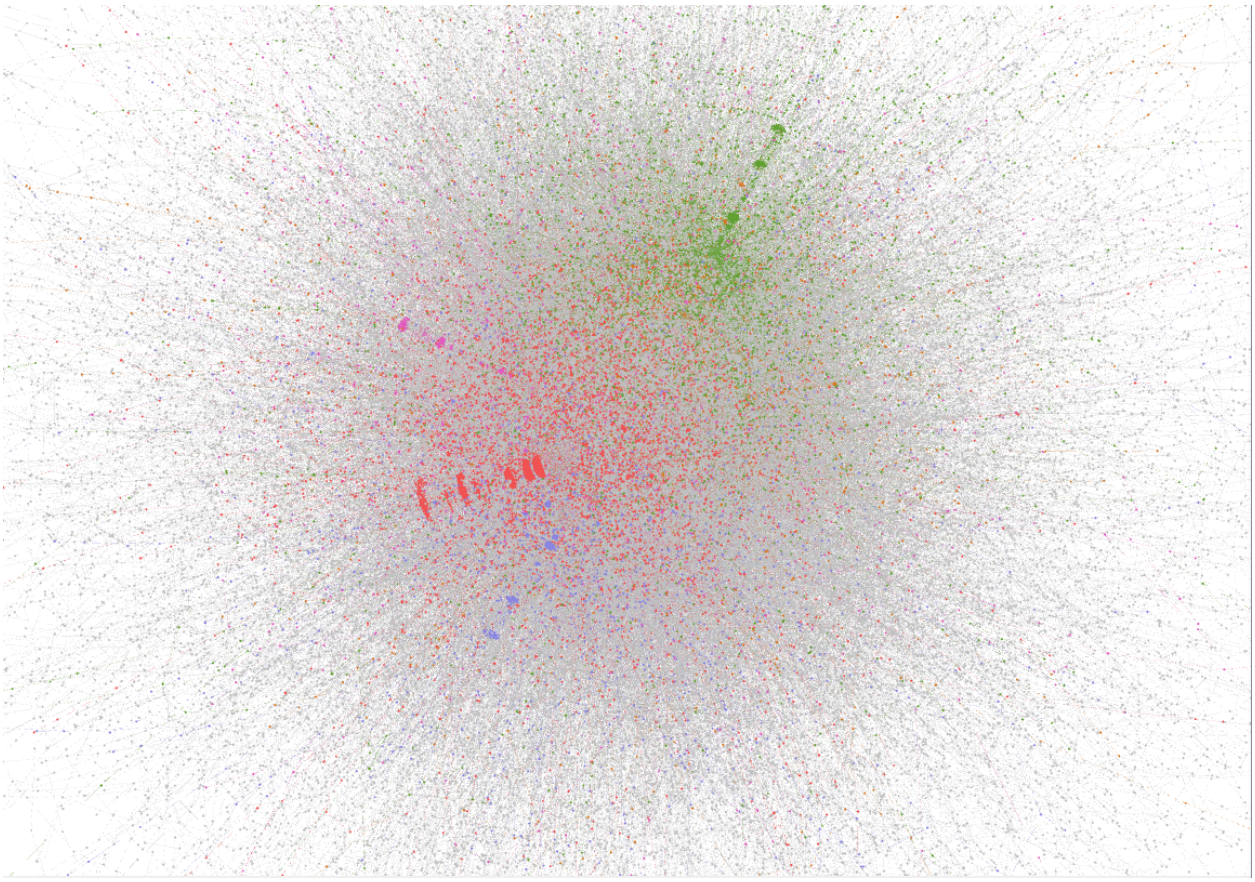


Figure 1: Showcasing the whole dataset of around sixty thousand users and their re-submissions. The colour is created by the modularity module which shows the most prominent communities

Table 1: Basic Features of Unfiltered data-set

Number of Nodes	Number of Edges	Average Degree	Network Diameter	Number of communities
63.329	91.040	1.438	65	1.486

As we can see from the image above, the decomposition of this dataset resulted in many different components, the number of Weakly Connected Components resulted to 1345 and the number of Strongly Connected Components resulted to 28639 as well as a large network diameter. However, due to these properties, and looking at the image above we can also view this graph as one giant component. Additionally, it is also clear that a few communities play a dominant role in the network.

5.2 Filtering

To get a closer indication of popularity, as the average degree is only 1.438, I will be performing a degree filtering methodology to the dataset to get rid of all the nodes that have less than 6 degree. This will remove "inactive" nodes (users) that have only resubmitted up to 5 images in the last 5 years. By performing this filtering, the core structure of the

network has remained the same (as seen in figure 2), and it will be easier to perform the ranking methodologies to explore which users are indeed the most popular.

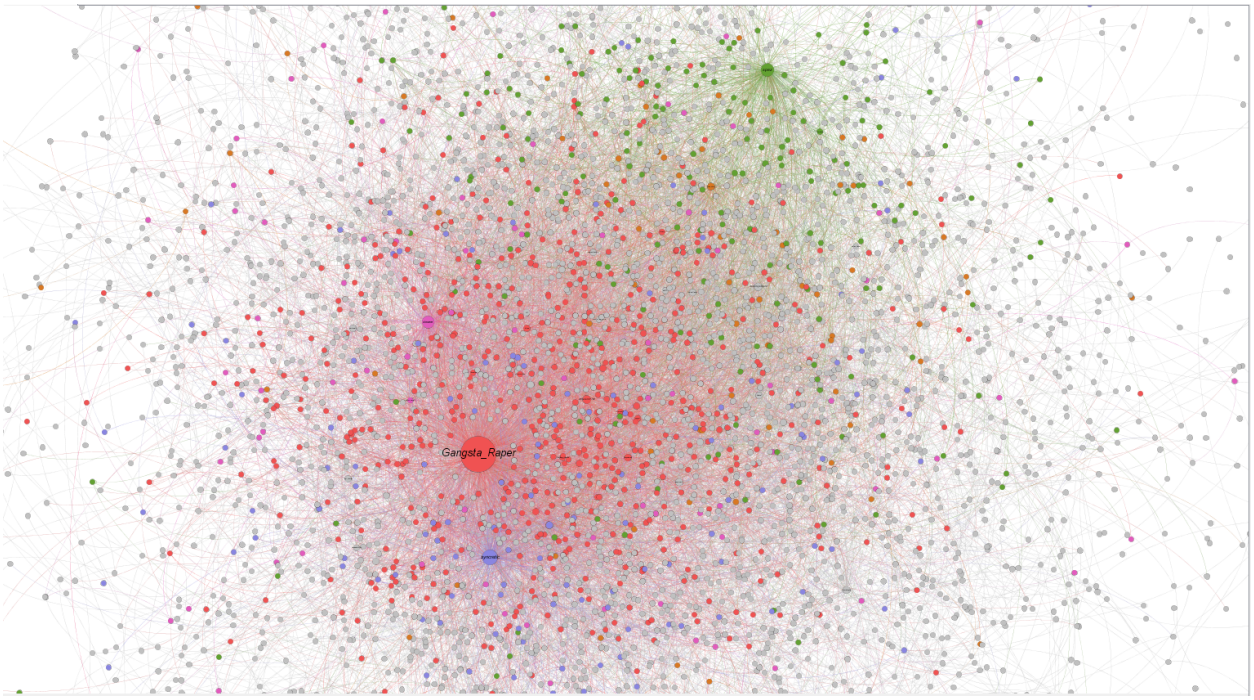


Figure 2: Showcasing the filtered dataset (degree range: 6 - 4278) of 4,656 users and their re-submissions. The colour is created by the modularity module which shows the most prominent communities

Table 2: Basic Features of filtered Data-set

Number of Nodes	Number of Edges	Average Degree	Network Diameter	Number of communities
4.656	11.776	2.529	17	282

5.3 Scale Free

Figure 3 showcases the degree distribution following a power-law, that is the fraction of $P(k)$ of nodes in the network having k connections to other nodes goes for large values of k as:

$$P(k) \sim k^{-\gamma}$$

This shows that, our network is indeed a scale free network. However due to having a very small average clustering coefficient of 0.074 and a high diameter of 17, it does not possess the small world phenomena. Additionally, figure 4 show cases the in-degree and out-degree of the network which also follows a power-law distribution. Generally, power-law can give us information about a networks resilience, where, real-world power-law networks are more resilient to random errors.

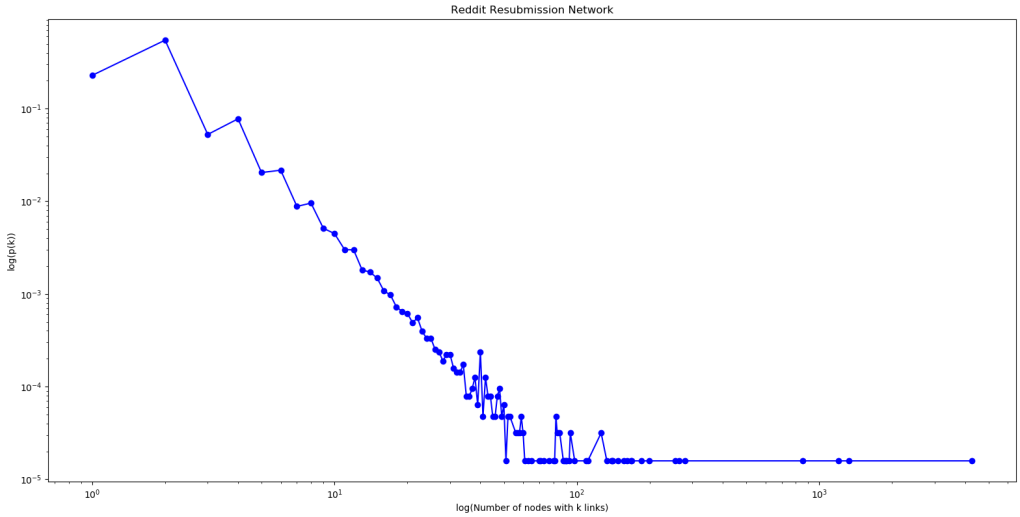


Figure 3: Loglog scale of the degree distribution of the reddit resubmission network following a power-law structure

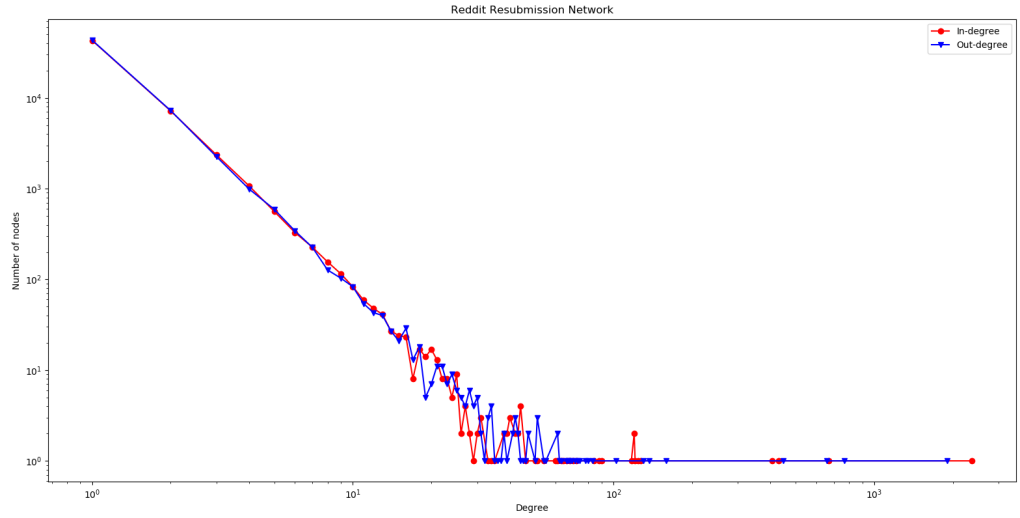


Figure 4: Loglog scale of the in-degree and out-degree distribution of the reddit resubmission network following a power-law structure

5.4 Rich Club Property and Mixing Pattern

The rich club property is an interesting observation to check within our network. This is because if a rich club exists, i.e popular users in the network are connected to each other, this posses a great strategy in gaining popularity. As seen in figure 5, a strong rich club property is observed which indicates a good strategy towards popularity would be to link with a rich/popular users.

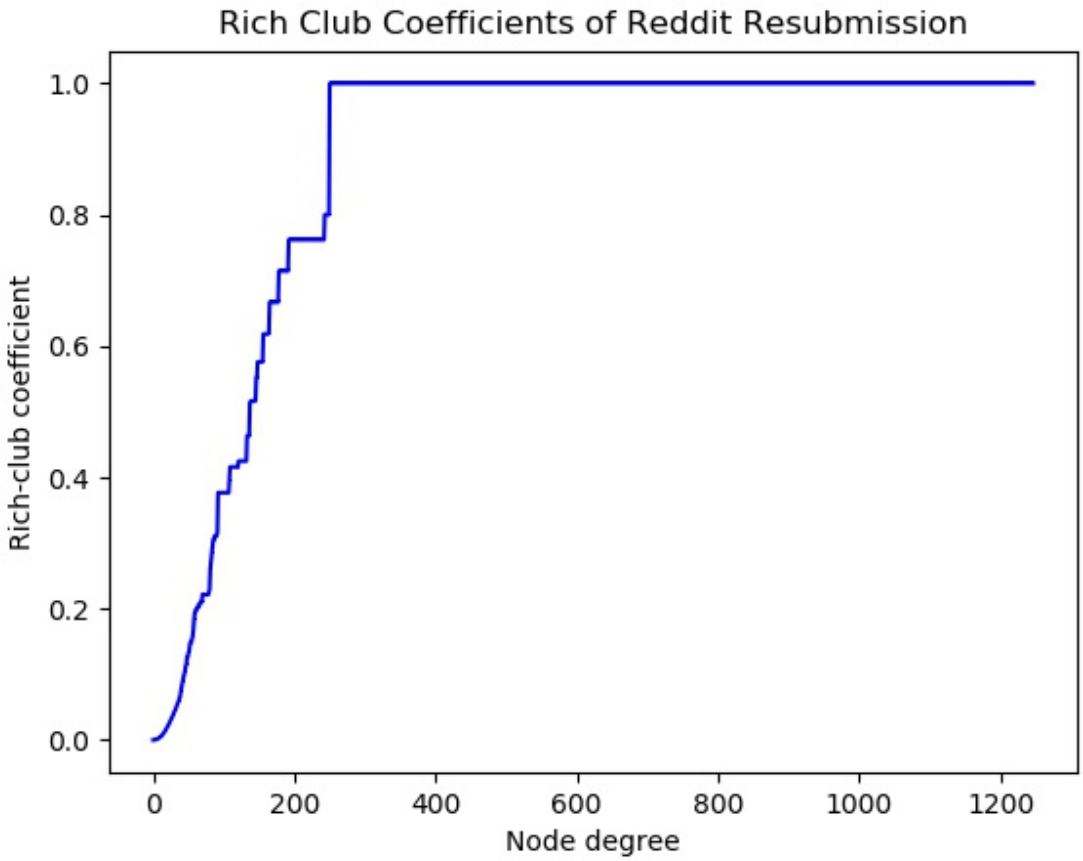


Figure 5: A strong Rich Club Property observed within the network

Additionally, a normalized network is show below in figure 6 (normalize using randomized network) [19]:

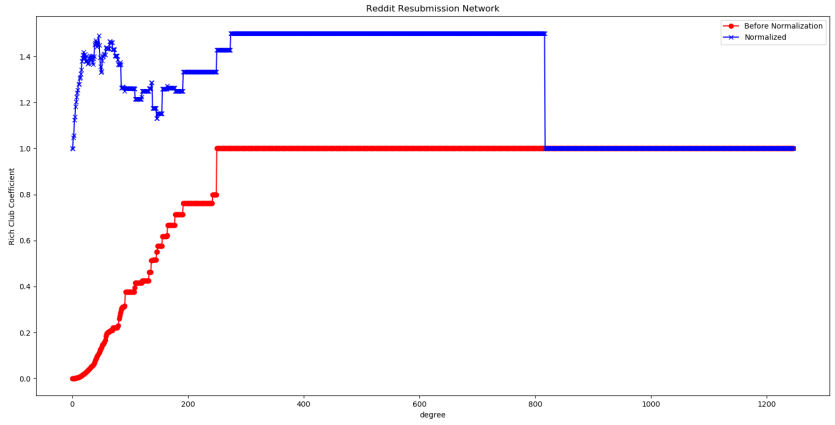


Figure 6: Normalized using a randomised network [19] and the un-ormalized version

The 7 core decomposition of the network also highlights the most influential nodes/users in the graph and their connectivity with each other:

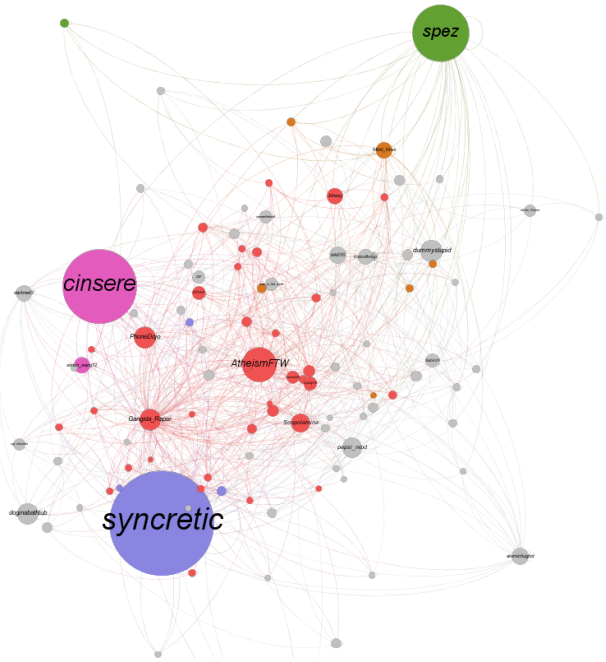


Figure 7: 7 Core decomposition of the network, highlights the most popular nodes are connected to each other, showing the "Rich Club"

Furthermore, the assortative coefficient of this network resulted in:

$$\alpha = -0.0326$$

This result is closer to 0, hence does not show a dissortative mixing but it is close to a **neutral** mixing. This is also showed by the nearest neighbours average degree in figure 8.

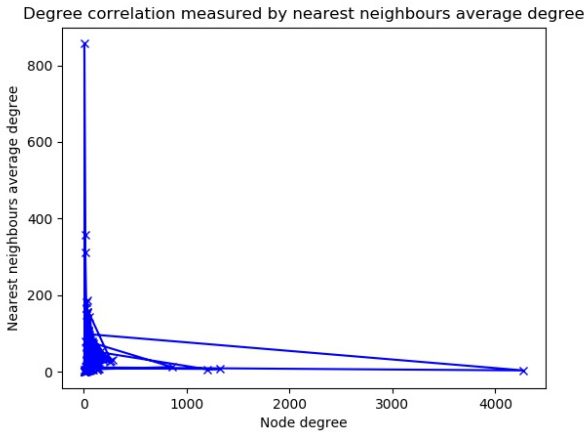


Figure 8: Nearest neighbours average degree

5.5 Visualisation of ranking methods of measuring popularity

In this section I will be presenting visualisations of different ranking methods employed to measure popularity in the network.One measure of popularity is through PageRank and HITS algorithms. Interestingly, both algorithms created similar networks that can be observed in figures 9-12. Out of the four thousand users, 4 users seem to be the most popular using these algorithms.

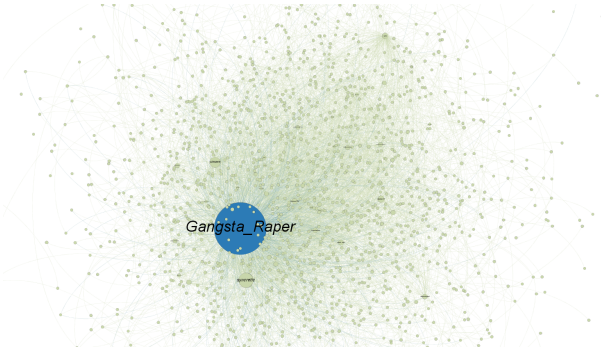


Figure 9: PageRank

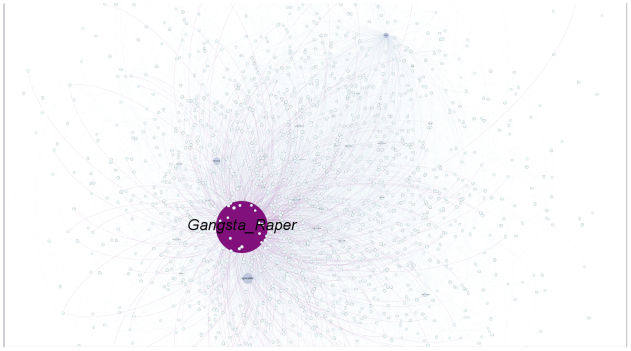


Figure 10: Authority

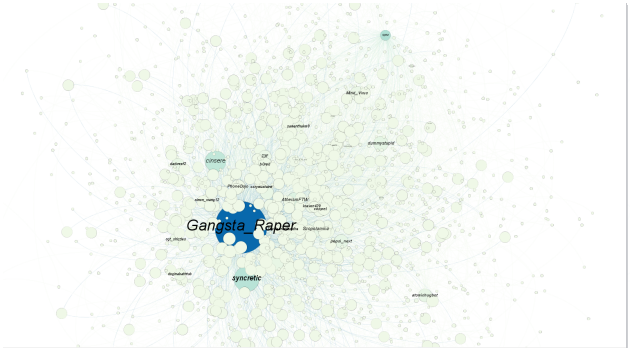


Figure 11: Hubs

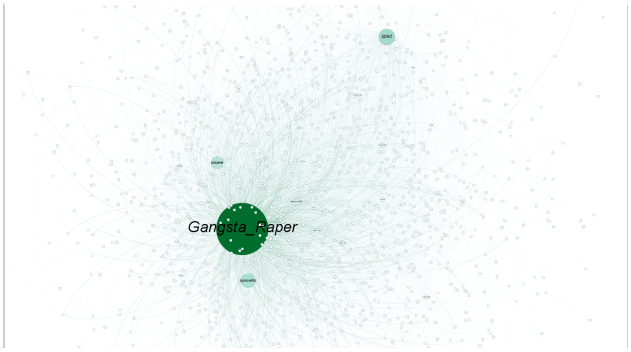
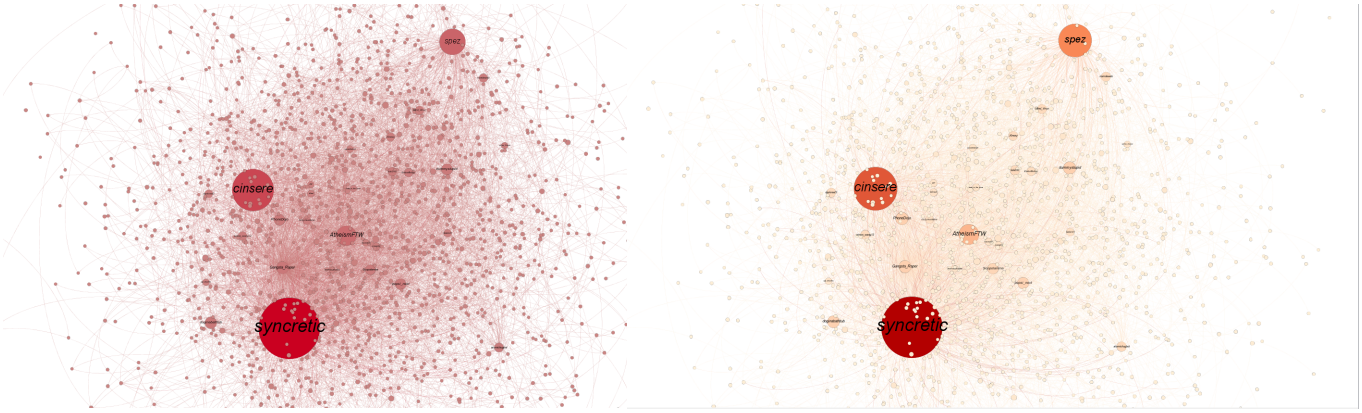
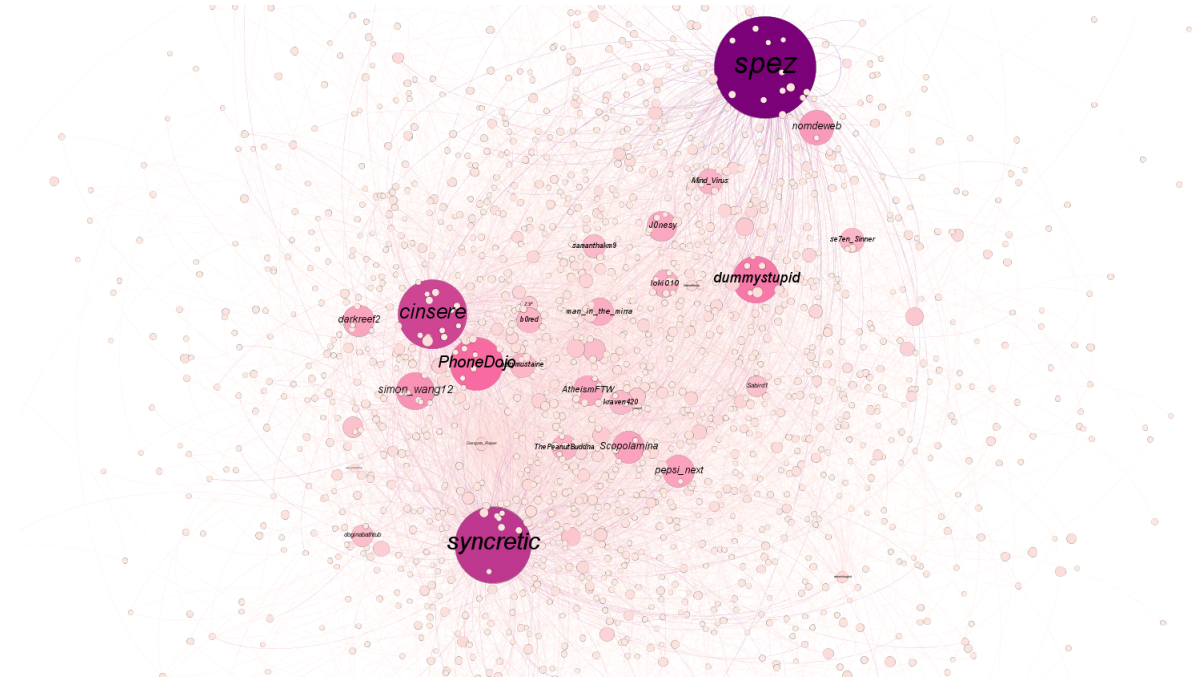


Figure 12: Betweenness Centrality

However, looking at the custom made popularity metrics as explained in the methodology section, such networks bring slightly different results. it is interesting to note that the user spez is the CEO of Reddit and one of the most popular nodes in the entire network. Three visualisations are reviewed, the popularity of a user (c) , popularity of an image for each user (b) as well as the "karma" score of users (a). An interesting results yields that one of the most influential nodes identified by the previous algorithms "Gangsta_Raper" is much smaller than before. The reason behind this could be that popularity in Reddit is very much so correlated with how much each user likes other user, and that popularity heavily relies on Reddit's scoring system.



(a) **(b)**



(c)

Figure 13: Custom popularity metrics based on Reddit's scoring system, popular users indicated here also appear in other ranking algorithms, however, they are more prominent in this ranking. (a) = karma score, (b) = image popularity, (c) = user's popularity

User	Resubmissions	Total Votes	Score	Comments	Out-degree	In-degree	Image Popularity	User's Popularity	Different subrredits
syncretic	931	1434251	560241	26763	658	669	587004	1434920	4
cinsere	545	1290796	372672	32619	448	407	405291	1291203	2
spez (CEO)	858	1950878	221044	81867	770	429	302911	1951307	3
AtheismFTW	176	501866	162568	8997	137	127	171565	501993	1
dummystupid	169	842077	78247	16262	159	120	94509	842197	3
PhoneDojo	48	960555	77135	17271	30	9	94406	960564	2
Gangsta_Raper	5608	151457	87431	4897	1903	2373	92328	153830	1
doginabathtub	79	334854	84190	6314	55	39	90504	334893	1
pepsi_next	144	538494	75694	7651	68	117	83345	538611	5
Scopolamina	173	537814	66954	9929	78	121	76883	537935	5

Table 3: Top 10 popular users and their features

6 Discussion

During the last few years Reddit has become one of the most popular internet destinations for users in United States and around the world. The amount of content shared within social media and online has increased significantly as it is much easier and low cost to consume than ever before. With this said, the urge of becoming popular has increased significantly. In this study we focused on identifying what attributes are needed in order to get popular, through image re-submissions on reddit.com.

Using network analysis techniques I discovered that our network holds a rich club property as well being close to a neutral mixing. The top 10 communities in the network accounts for 68% of the nodes as seen in figure 14. The network is a scale-free networking as the degree distribution as well as the in-degree and out-degree distributions follow a power-law property.

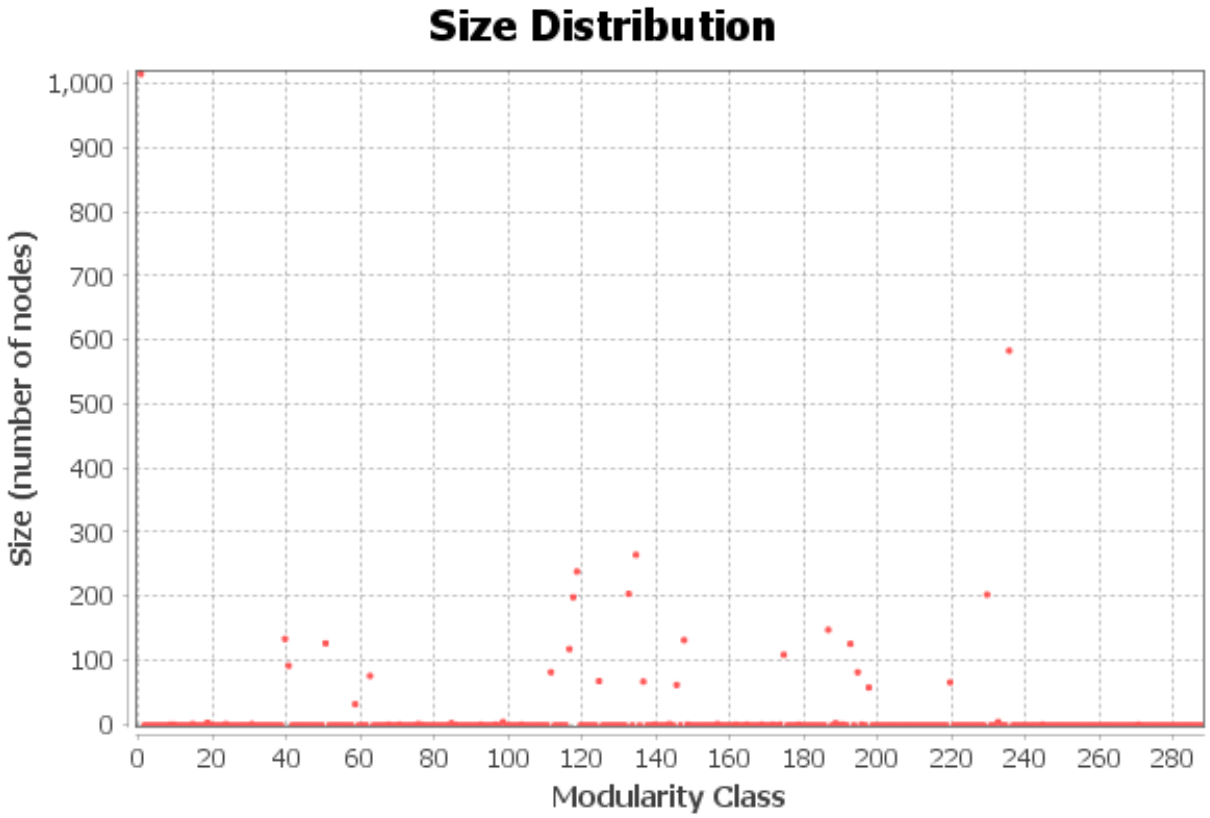


Figure 14: Modularity and Community Detection of the network

Different ranking algorithms were used in order to understand a user's popularity. Link analysis techniques such as PageRank and HITS as well as the betweenness centrality highlighted very few users to be popular. The popularity aspects of these algorithms did not consider the scoring system of Reddit itself, whereas, it used different network properties.

When taking the scoring system of Reddit into account and using other features such the number of comments displayed and total votes for a user, a similar but different network was highlighted. It's interesting that a very popular user that was ranked highly by PageRank had a much lower popularity score when compared with the custom ranking system. Having said that all the popular users were in the rich club as well as having moderate ranking's in both methodologies.

It is also interesting to note that majority of popular users had over 100 re-submission (high degrees) of content as well as high number of votes in their posts. It is also interesting that most users were also involved in multiple subreddits which increased their visibility to different communities (although some users transitioned between similar communities). However, we can see that staying committed to only a few subrreddits is also a viable strategy as you are able to repeatedly produce high scoring content which in return increases your visibility and popularity.

My conclusion is that it is very hard to precisely define popularity. To become popular through re-submissions it is clear that you will have a bigger advantage if you are part of rich club within the network. Additionally, users need to be active and engage through many content creations even though the content itself does not need to be original. As well as that if users are well liked (high karma score), this will help much more than having just a high re-submission score. This is because if a user is liked there is potential higher level of trust in the close tight nit communities, so a user will be much more influential. It also helps if you have connection with the CEO of Reddit.

One of the main limitations of this study was evaluating popularity through on image re-submission. Many content on reddit, does not involve images, therefore, some user groups maybe have been disregarded in that sense. Also I have been analysing a fairly small dataset, so in the future a larger dataset could prove to be more meaningful. Further, future work could be to enhance the popularity metric into more sophisticated formulas and ratios that take into consideration more features and patterns, to rank a user's popularity. Additionally, other networks could be created with different datasets that are not just re-submission of images to observe popularity of users. Since many subrredits involve discussions through only text. Additionally, users receive votes when they post a comment, taking into account users that have high scores through commenting on a content could also give an interesting insight in how that user is perceived and whether their opinion is validated or important.

References

[1] *Reddit.com Site Info. Alexa Internet.* Retrieved Jan 5th, 2019.

[2] Ohlheiser, Abby. *"Reddit will limit the reach of a pro-Trump board and crack down on its 'most toxic users'"*. Washington Post. Retrieved 4th Jan 2019.

[3] *Social Media Taps Into Our Most Primal Urge: Talking About Ourselves*, available at https://www.forbes.com/sites/alicegwalton/2012/06/29/facebooks-share-button-taps-into-the-wiring-of-our-brain/. Retrieved Jan 1st, 2019

[4] *The Psychology of Sharing: A Persona-based Approach to Sharing on Social Media*, available at https://engage.social/blog/social-share/the-psychology-of-sharing-a-persona-based-approach-to-sharing-on-social-media/. Retrieved Jan 1st, 2019

[5] Lansu, T. M., & Cillessen, A. N. (2012). *Peer status in emerging adulthood: Associations of popularity and preference with social roles and behavior.* Journal of Adolescent Research, 27(1), 132-150 Retrieved December 10th, 2019

[6] *Web Reddit Data of resubmitted images*, available at http://snap.stanford.edu/data/web-Reddit.html. Retrieved Jan 1st, 2019

[7] *NetworkX Python Library*, available at https://networkx.github.io/documentation/latest/index.html. Retrieved Dec 8th, 2018

[8] *Gephi - Visualisation tool for network systems*, available at https://gephi.org/. Retrieved Dec 8th, 2018

[9] *Reddit's Year In Review: 2018*. Reddit. December 10, 2018. Retrieved Jan 1st, 2019.

[10] Albert, R.; Barabasi, A.-L. (2002). *Statistical mechanics of complex networks*. Reviews of Modern Physics Retrieved Dec 15th, 2018

[11] *Fast unfolding of communities in large networks* Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre Retrieved Dec 18th, 2018

[12] *Mixing patterns in networks* M. E. J. Newman Phys. Rev. E 67, 026126 Published 27 February 2003 Retrieved Dec 20th, 2018

[13] *The rich-club phenomenon in the Internet topology* Shi Zhou ; R.J. Mondragon Retrieved Dec 20th, 2018

[14] Cutts, Matt. *Algorithms Rank Relevant Results Higher*. www.google.com. Archived from the original on July 2, 2013. Retrieved 5th Jan 2019.

[15] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schtze (2008). *Introduction to Information Retrieval*. Cambridge University Press. Retrieved Jan 1st 2019.

[16] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, Serge Fdida.*From Popularity Prediction to Ranking Online News*. Social Network Analysis and Mining, Springer, 2014 Retrieved Dec 25th, 2018

[17] *The rich club phenomenon in the classroom* Luis M. Vaquero1 & Manuel Cebrian2,3 (2013) Retrieved Dec 30th, 2018

[18] *Identifying Social Roles in reddit Using Network Structure* Cody Buntain Jennifer Golbeck Retrieved Jan 3rd, 2019

[19] Julian J. McAuley, Luciano da Fontoura Costa, and Tibrio S. Caetano, *The rich-club phenomenon across complex network hierarchies*, Applied Physics Letters Vol 91 Issue 8, August 2007 Retrieved Jan 5th, 2019