# Analyzing Reddit Hyperlink Network using Advanced network analysis technique with R

Rajesh Baidya, MAT: 3462771

28. September 2021

## 1 Introduction

Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory [2]. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. In this study, we are studying the network between subreddits. So for this case Subreddits are the nodes and the hyperlink between one subreddit to others are edges. Throughout this work, I would try to find different aspects of this dataset. In this process, first I will discuss the networking dataset and its properties that I have selected to focus on. I would try to compute the degree and different kinds of centralities of this network. After that, I would try to plot some subreddits in 3-dimensinal space and find the similarities between the subreddits. In this study, I would try to answer the following questions using several network analysis techniques.

- How to measure the significance of a subreddit in a network?
- How some of the most popular subreddits are connected with each other?
- How the network grows with time?
- Is it possible to generate a meaningful cluster of posts from a subreddit using given subreddit embeddings?

## 2 Dataset

The dataset I am going to use in this work is named 'Social Network: Reddit Hyperlink Network'[1]. Reddit is a popular social network where every community is called a Subreddit. The hyperlink network represents the directed connections between two subreddits. This network dataset is extracted from publicly available Reddit data of 2.5 years from Jan 2014 to April 2017. The subreddit-to-subreddit hyperlink network is extracted from the posts that create hyperlinks from one subreddit to another. A hyperlink originates from a post in the source community and links to a post in the target community. Each hyperlink is annotated with three properties: the timestamp, the sentiment of the source community post towards the target community post, and the text property vector of the source post. The network is directed, signed, temporal, and attributed. The hyperlink can be present in either the title of the post or in the body. So in this dataset, they provided two different network files, one network created from the hyperlinks in the title and the other one from the post. They have also provided embedding vectors (a distributed representation for text) representing each subreddit.

This dataset contains a total 55,863 number of subreddits/nodes and 858,490 hyperlinks between subreddits/number of edges. Every edge is also labeled with positive or negative emotions by -1 or +1. There is also a Text Property Vector or Embedding Vector associated with every source subreddit.

## 3 Preprocessing

This 'Reddit Hyperlink Network' dataset contains two files. One contains the Network of subreddit-to-subreddit hyperlinks extracted from hyperlinks in the body of the post and the other has the hyperlinks extracted from the title of the post. I choose the network file extracted from the titles for further experiments. Only this title network contains 571,928 edges which are very hard to visualize as a network. That's why I have selected only the edges which are activated between 1st January 2014 to 31 December 2014. As a part of the preprocessing steps, I exported the hyperlinks between source-target subreddits including the timestamps and post properties

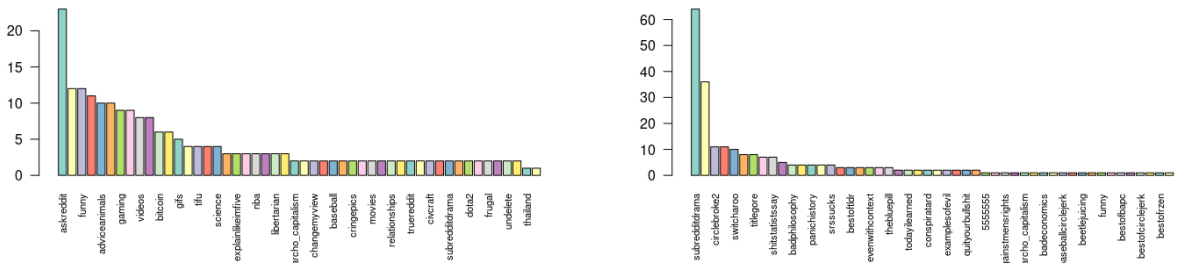


Abbildung 1: Bar plot of subreddits on degree in (left) and degree out (right)

against every source subreddits on separate files. Now both of these files contain 128,003 rows.

While doing the analysis of the network, I have merged all the similar edges indifferent of the timestamp which contain the same source and target subreddits and added another additional column named frequency, where I counted the occurrence number of the edges. After doing this merge, the number of edges narrows down to 63,352. Still, it is too large to be visualized as a network. Then I did further processing and filtered the edges which have a frequency of more than 30. Now I have only 314 edges of the most important subreddits which are ready for further visualization and analyzing.

# 4 Experiments

## 4.1 Descriptive Statistics

First I wanted to calculate the incoming and outgoing hyperlinks between the subreddits on this network. So I calculated the in-degree and out-degree of the network. You can see the visualization of the degree in figure 1. Most of the subreddits have low incoming hyperlinks, where 105 subreddits have 0 in degree and 100 subreddits has only 1 in degree. Only 1 subreddit (askreddit) has the highest number of incoming degrees of 23 and the second highest in-degree is 12 (worldnews). The numbers in outgoing hyperlinks are more contrasting. Only one subreddit named 'subredditdrama' created most of the outgoing hyperlinks where the second-highest subreddit (bestof) has created only 36 outgoing degrees. 132 subreddits have no outgoing hyperlinks

## 4.2 Centralities

Next, I wanted to compute the network centralities other than the Degree centrality. First I computed the Eigen Centrality, where Eigenvector centrality is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes [3]. You can see the Eigen Centralites and Top Eigen Value subreddits on the figure 2.

For a given graph $G := (V, E)$ with $|V|$ number of vertices let $A = (a_{v,t})$ be the adjacency matrix. The relative centrality score of vertex $v$ can be defined as:

$$x_v = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} X_t$$

Later I computed the betweenness centralities of the network. Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes [3]. You can see the betweenness centralities of the network on the figure 3. From the figure we can see that, only subreddits hold the highest betweenness score, rest of the subreddits hold a betweenness score closer to 0.

The betweenness of a vertex $v$ in a graph $G := (V, E)$ with $V$ vertices

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Correlation between the Betweenness Vector and Eigen Vector is 0.6135. From both centralities, we can see that Centrality is much more distributed in Eigen centrality rather than the Betweenness centralities. In

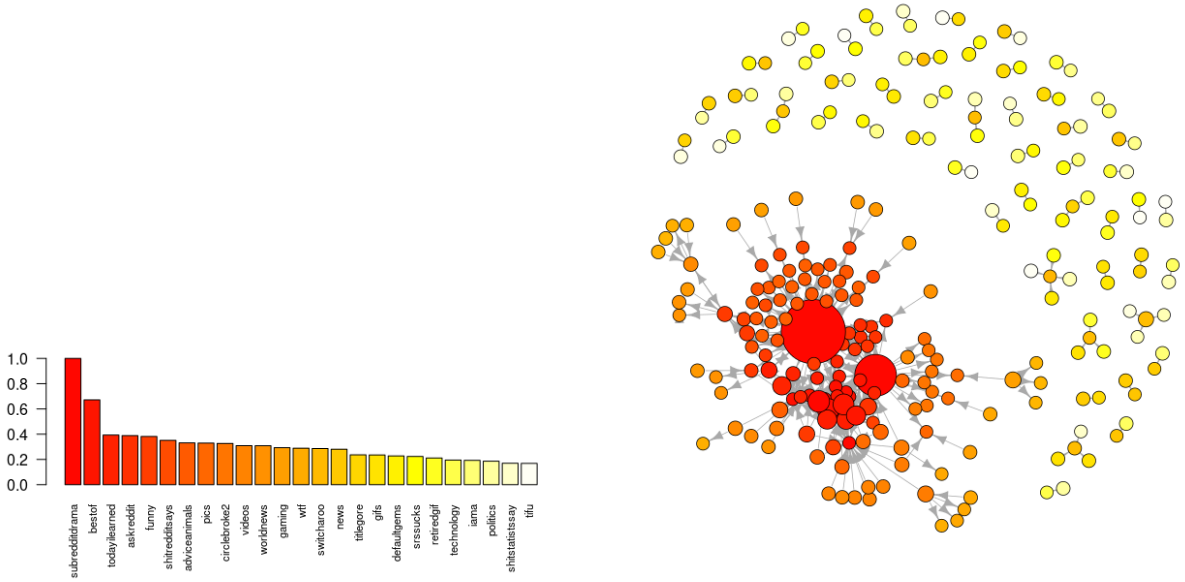


Abbildung 2: Eigen Centralities of the Network between subreddits

Betweenness, only five subreddits hold the betweenness value of more than 10, where the highest one (subredditdrama) holds a value of 158.5. On the other hand, in Eigen centralities, more than 25 subreddits hold a centrality score of more than 0.10 where 'subredditdrama' is still the most central subreddit with the highest score of 1.00.

### 4.3 Temporal Properties

A temporal network is a network whose links are active only at certain points in time. Our network also contains Temporal Properties. Although its number of links only get increased over the course of time. To analyze the network's temporal properties I took the data before its similar edges have been merged to use the original time information. You can see how our network transforms at different times in figure 4. The top-left image has been created with the edges before 2014-01-01, where top-middle and top-right images were created with the edges before 2014-01-05 and 2014-01-10 respectively. The bottom images were created by taking snapshots of data on 2014-03-31, 2014-06-30, 2014-12-31 respectively. Although, It was not possible to get further insight into the network from these temporal properties without more extreme preprocessing.

### 4.4 Subreddit Post Properties

In this part of the experiment, I took a look into the post properties under source subreddits. I picked the 4 most popular subreddits and selected their 86 dimensional post properties for further experiments. After performing Principal Component Analysis, I plotted the posts under the subreddits into a 3-dimensional space to find similarities among the posts from the same subreddit. But no kind of notable similarities were found among the posts after this limited processing. You can see the plot on figure 5.

## 5 Conclusion

From the above analysis, we can see that this subreddit network is quite centric and Subreddit named 'Subredditdrama', 'bestof', 'todayilearned' are the most significant subreddits in various aspects. Most of the subreddits are connected to each other by these three subreddits. As time grows, more and more subreddits get connected to these significant nodes and the network gets more centric. Although calculating the similarities between subreddits and their posts from the post properties was not that straightforward. Just performing the Principal component analysis was not enough to cluster the similar posts together. Further experiments are possible in this direction and also sentiments towards target subreddits from the source subreddits can be examined. For more information about this experiment and implementation, you can visit the GitHub link. `https://github.com/06rajesh/subreddit-hyperlink-analysis`

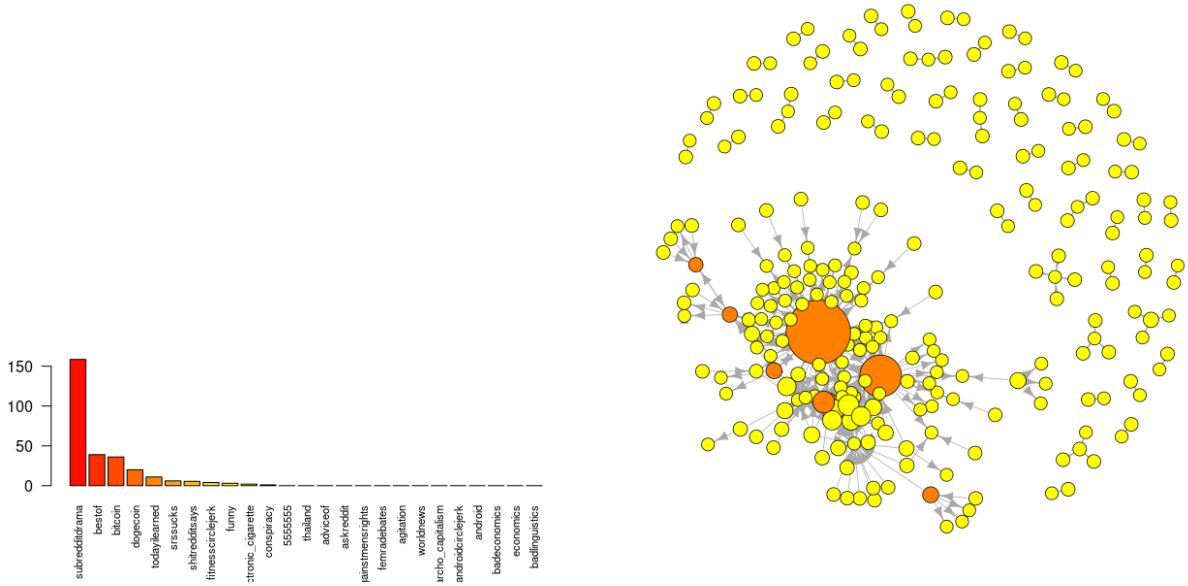


Abbildung 3: Betweenness Centralities of the Network between subreddits

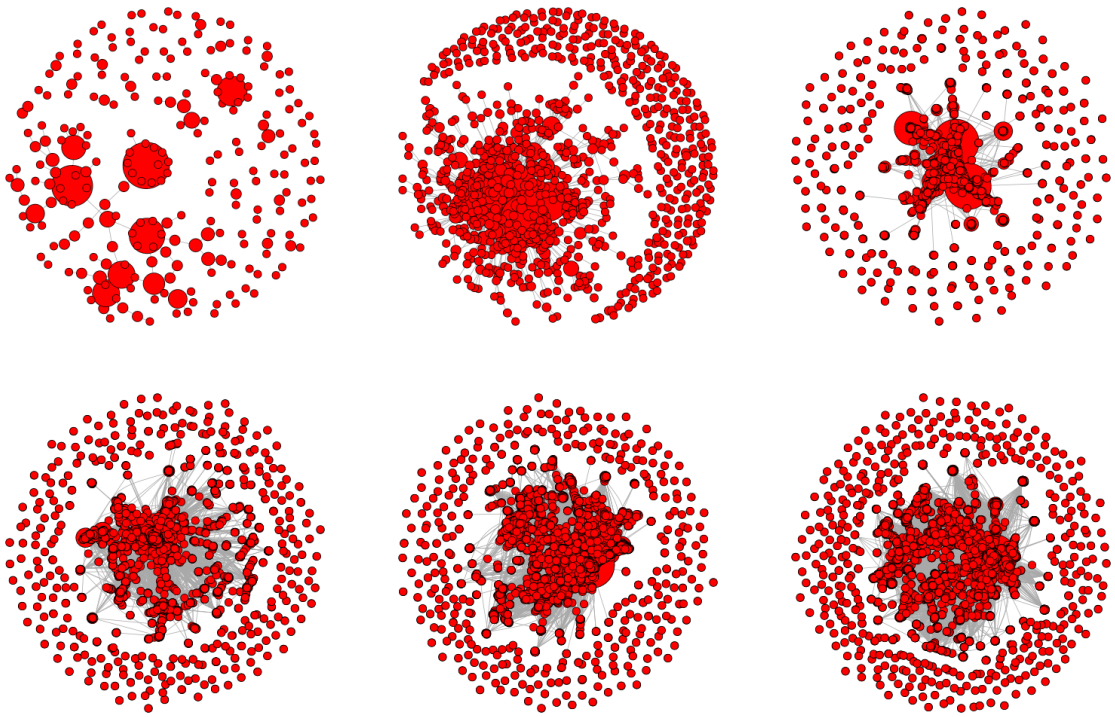

Abbildung 4: Evalutation of the network through time (Left-top to right-bottom)

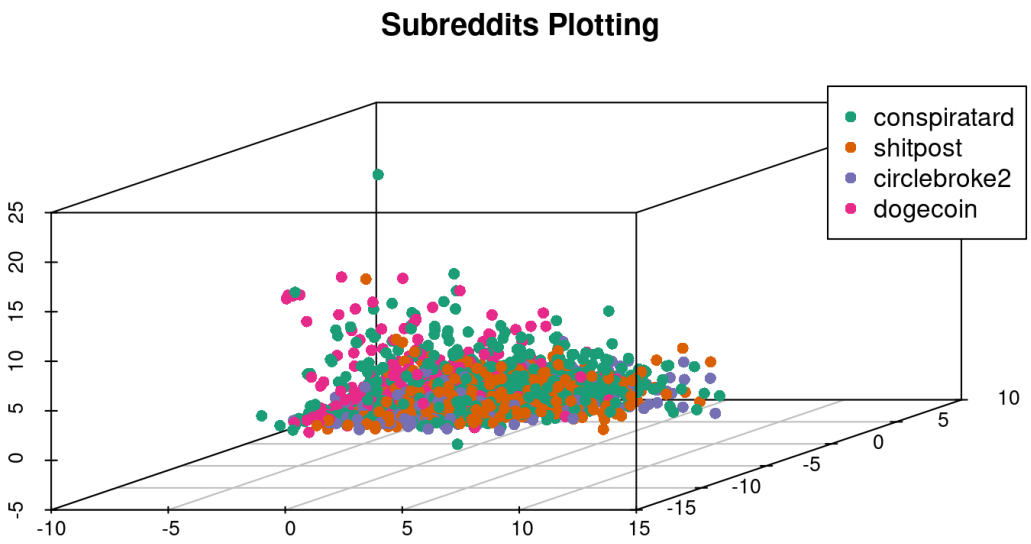


Abbildung 5: Posts from different subreddits are plotted on a 3-dimensional Space

# Literatur

[1] S. Kumar, W.L. Hamilton, J. Leskovec, D. Jurafsky. Community interaction and conflict on the web. `https://snap.stanford.edu/data/soc-RedditHyperlinks.html`, World Wide Web Conference, 2018.

[2] Social Network Analysis. `https://en.wikipedia.org/wiki/Social_network_analysis`

[3] Centrality `https://en.wikipedia.org/wiki/Centrality`