# SC 435: Introduction to Complex Networks
# Course Project
# Reddit Hyperlink Network Analysis

Kritika Gupta (201601129)
*DA-IICT, Gandhingar*
(Dated: November 17, 2019)

## INTRODUCTION

Reddit[1] is an American social news aggregation, web content rating, and discussion website. Users of the website can submit content in the form of posts, which may contain images, text, links, etc. Posts on Reddit are organized into user-created areas of interest called "subreddits". Subreddits have come to represent communities of people whose interests align with area of the subreddit.

**Subreddit Hyperlink Network**[2]: 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. Every post has label indicating if the post is explicitly negative towards the target subreddit.

## I. NETWORK PROPERTIES

Each node of the network represents a subreddit. The network is a directed multigraph. Each edge from node A to node B represents a post by subreddit A which consists of a hyperlink to subreddit B. The sentiment of the post that creates the hyperlink can be negative or positive/neutral. Negative posts are assigned an edge weight of $-1$ and positive/neutral are assigned an edge weight of $+1$.
The network consists of :

1. Number of subreddits (nodes) = 67180
2. Number of posts (edges) = 858488
3. Number of pairs of connected subreddits = 339643

### A. Density

The network density equals a value of 0.00019023 and hence, we can say that it is a sparse network, i.e., the number of edges are significantly less than the possible number of edges. This is usually true for social networks, and in this case, implies that subreddits do not post posts with links to other subreddits very often.

### B. Degree Distribution

Fig 1 shows that the degree distribution of the network follows a power law, i.e., most subreddits have very few links to other subreddits and only a few subreddits have links to a lot of other subreddits. Since the degree distribution follows a power law, the network is scale free.[3]

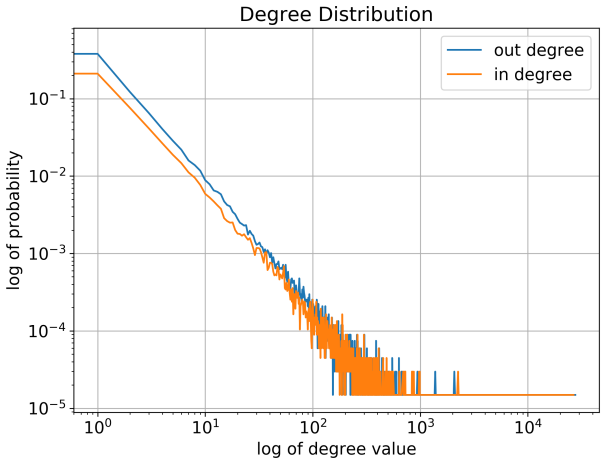


FIG. 1: Log-log plot of in and out degrees of the nodes of the network.

### C. Components

Strongly connected components : 45564
The giant strongly connected component consists of 31% of the nodes and the remaining consist of less than 0.007% of the nodes.
Weakly connected components : 712
The giant weakly connected component consists of 97% of the nodes and the remaining consist of less than 0.01% of the nodes. Hence, most of the subreddits are connected to each other via a hyperlink.

## II. PAGERANK CENTRALITY

To understand which subreddits are the most important in this hyperlink network, we calculate the PageRank centrality of the graph.

| Subreddit | PageRank Centrality |
|---|---|
| askreddit | 0.021 |
| iama | 0.018 |
| pics | 0.011 |
| funny | 0.009 |
| videos | 0.009 |
| todayilearned | 0.006 |
| gaming | 0.006 |
| worldnews | 0.005 |
| science | 0.004 |
| news | 0.004 |

TABLE I: PageRank Centrality

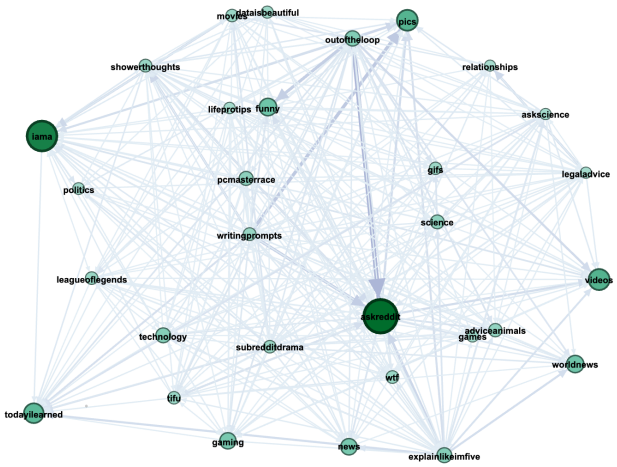


FIG. 2: Nodes with top eigenvector centrality visualized in Gephi[4]. For simplicity, all nodes and edges are not visualized. However, centrality is calculated using the entire network.

## III. RECIPROCITY

A negative link from subreddit A to subreddit B implies that there exists a post in subreddit A that hatefully attacks subreddit B. It is interesting to ask whether subreddit B attacks subreddit A back or not. The reciprocity of a node is defined as the ratio of number of bidirectional edges to the number of total edges it has.

We define *negative reciprocity* as the reciprocity of a node considering only the negative links in the network. Hence, a subreddit with high negative reciprocity suggests that the subreddit is hated by others as much as it spreads hate, and it hates others as much as it is hated. Similarly, *positive reciprocity* determines, out of all the appreciation that a subreddit gives and receives, how much of it is reciprocated back to it.

| Subreddit | Negative Reciprocity |
|---|---|
| transgenderuk | 1.0 |
| mixedasians | 1.0 |
| survivorcirclejerk | 1.0 |
| wellington | 1.0 |
| lcfc | 1.0 |

TABLE II: Subreddits with the highest negative reciprocity

However, it is observed that more than 75% of the nodes have a negative reciprocity value less than 0.1. Specifically, 91% of the nodes have a negative reciprocity value of 0, and 0.6% have a value of 1.

| Subreddit | Positive Reciprocity |
|---|---|
| newwhitehaven | 1.0 |
| mrref | 1.0 |
| dungeonscape | 1.0 |
| kommunitykombat | 1.0 |
| privatepracticeslps | 1.0 |

TABLE III: Subreddits with the highest positive reciprocity

However, it is observed that more than 75% of the nodes have a positive reciprocity value less than 0.3. Specifically, 78% of the nodes have a positive reciprocity value of 0, and 2% have a value of 1.

## IV. NEGATIVITY AMONG SUBREDDITS

In this section we analyze which subreddits attack others the most, and which subreddits get attacked the most.

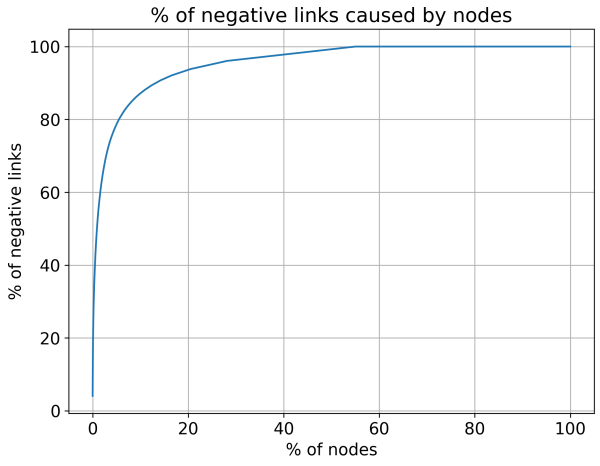


FIG. 3: Fraction of negative links caused by number of nodes.

Fig 3 shows that very few of the subreddits ($\approx 10\%$) cause 80% the negative attacks on Reddit.

| Subreddit | % of attacks caused |
|---|---|
| subredditdrama | 10.932 |
| bestof | 4.214 |
| drama | 3.272 |
| circlebroke2 | 2.152 |
| shitpost | 1.591 |

TABLE IV: Subreddits that cause the most negative links.

| Subreddit | % of attacks received |
|---|---|
| askreddit | 4.094 |
| worldnews | 2.073 |
| news | 1.975 |
| pics | 1.794 |
| todayilearned | 1.714 |

TABLE V: Subreddits that are hit by the most negative links.

## V. NEGATIVE REPUTATIONS OF SUBREDDITS

Reddit posts with hyperlinks to other subreddits can be used to change the perception users have of a subreddit. For instance, let us assume a user opens Reddit and begins browsing, starting at a random subreddit. On encountering a post with a hyperlink to another subredddit, they follow the hyperlink to the target subreddit only when the sentiment of the post is positive or neutral. If the post has a negative sentiment, the user forms a negative impression of the target subreddit and its reputation worsens. The number of links that led to reaching this target subreddit is analogous to the time the user spends on browsing Reddit. This process is repeated by multiple users and would lead to users having negative opinions about different subreddits. In the following section, we examine which subreddits can acquire a negative reputation via such usage of Reddit by users, and also the average browsing time of users required to develop such an opinion. This is done by simulation a random walk on the network using the following rules and assumptions:

### A. Random Walk Assumptions

1. Each simulation corresponds to one browsing session by a user.
2. One step of the random walk corresponds to the user clicking a link in a post by the source subreddit with positive sentiment, and reaching the target subreddit. This comes from the assumption that a user would prefer to read and follow positive/neutral posts and not pursue negativity online.
3. On reaching a negative sentiment post, the user does not continue browsing any further links, and the session ends.
4. If the random walk gets stuck at an absorbing state, the user will never encounter a negative link. Hence, a maximum time threshold ($t_{max}$) is set to break each simulation. This is done under the realistic assumption that if a user does not encounter a negative link, they would not continue to browse infinitely.
5. Whenever any user reaches a negative sentiment post, the reputation of the target subreddit is worsened by an equal amount.
6. For simplicity, only the giant weakly component is considered for the simulation.

### B. Simulation

The process described above is simulated as a random walk on the network. 1000 simulations are done with ($t_{max} = 100$).

| Subreddit | Reputation | Steps per user |
|---|---|---|
| askreddit | -282 | 19.0 |
| pics | -129 | 18.0 |
| worldnews | -120 | 18.0 |
| funny | -117 | 17.0 |
| videos | -113 | 19.0 |
| todayilearned | -102 | 17.0 |
| adviceanimals | -84 | 19.0 |
| wtf | -82 | 18.0 |
| news | -72 | 22.0 |
| iama | -71 | 17.0 |

TABLE VI: Random Walk Simulation Results

From Table VI, we see that if 1000 users browse Reddit for one session, the subreddit 'askreddit' gains the worst reputation, and each user reaches this subreddit after an average of 19 hops via positive/moderate hyperlinks. In other words, if 1000 users browse Reddit, 282 of them will end up at 'askreddit' via a path of positive hyperlinks and average length 19.

[1] "https://www.reddit.com/,".
[2] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," `http://snap.stanford.edu/data` (2014).
[3] M. Newman, *Networks* (Oxford university press, 2018).
[4] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," (2009).