Social Networks Analysis: Project Report

Evolution of communities in Reddit over time

Tapan Chugh , CS12B055
Abhishek Yadav, CS12B032
Pranav P. Nair, CS12B062

9th May 2015

1 Introduction

Reddit is an entertainment, social networking, and news website where registered community members can submit content, such as text posts or direct links. Registered users can then vote submissions "up" or "down" to organize the posts and determine their position on the site's pages. Content entries are organized by areas of interest called "subreddits."

We look at the communities that are formed among users who share the same posts repeatedly by forming a graph between usernames of people who shared similar content. We also try to understand how these communities actually evolve and change over time by sampling our data at monthly intervals for a period of around 5 years.

We have tried multiple approaches to solve this problem. We look at the connected components and their diameters. We also tried some of the community detection techniques like Spectral Clustering, Block Model approximation etc.

2 Dataset

We have used the data for reddit posts available from SNAP (http://snap.stanford.edu/data/web-Reddit.html).

The given data was converted to a bi-partite edge list between post id and user name at every month. Then it was also converted into a list of edges between usernames who had shared a same post during that period.

Number of submissions	132,308
Number of unique images	16,736
Number of users	63,201
Average number of times an image is resubmitted	7.9
Timespan	July 2008 - Jan 2013

3 Approach

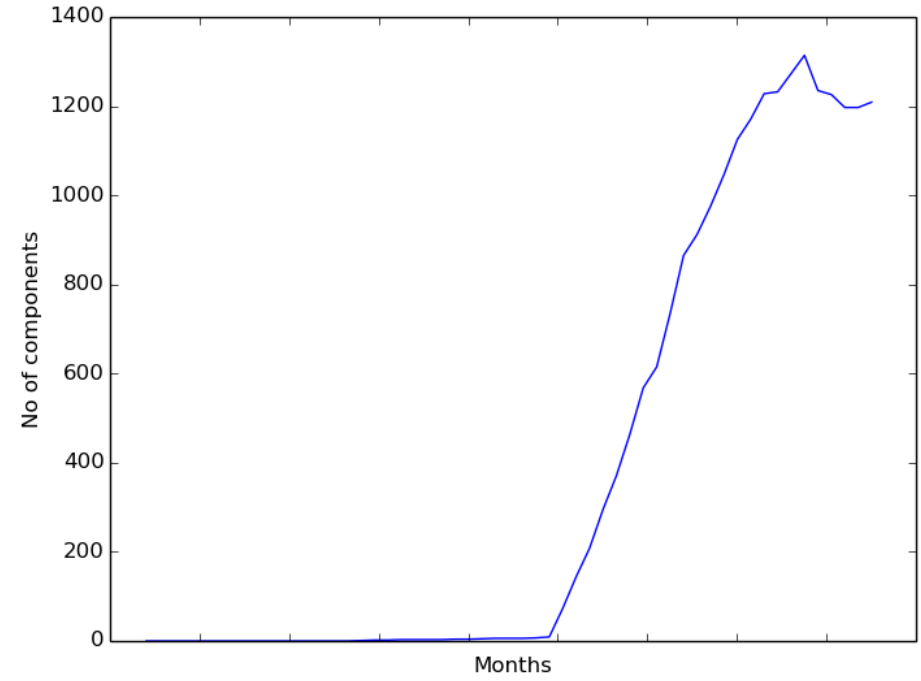
We have two aspects of the problem.

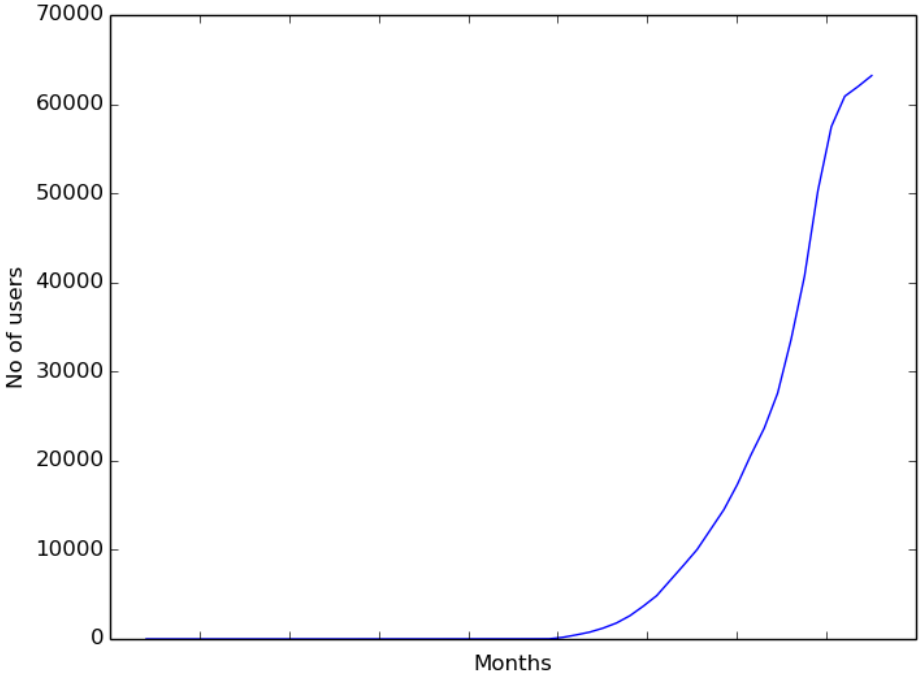
1. To understand the formation of communities. We try to figure out how strong communities are present in the data.

2. To understand the changes with time. We want to look at how these communities actually evolve with time.

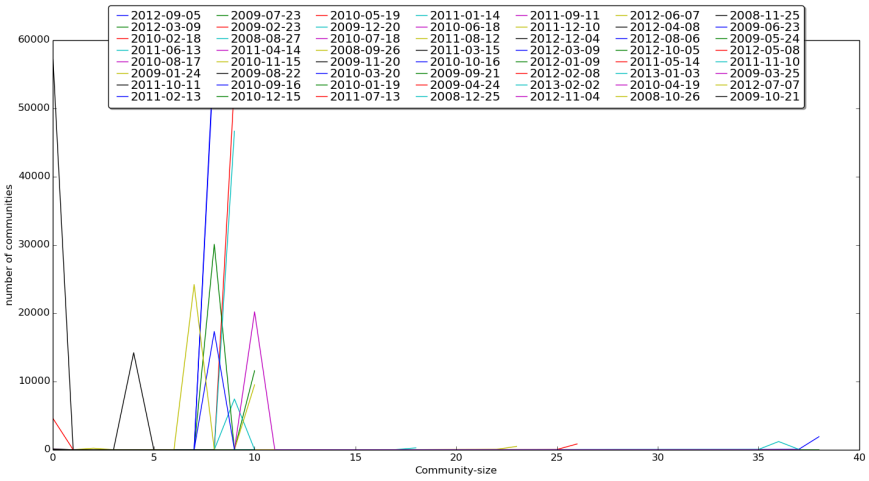
We started by sorting the data based upon the months. We created a seperate file for each month, which had all the data till that month. Then from these bipartite graphs between image ids and usernames, we created the list of edges between usernames who had shared the same post during that period. Then from these data files, we constructed graphs and ran some experiments. We also got some interesting results out of it.

1. **No of connected components in the graph**
   We measured the no of connected components in the graph and plotted a curve. We found that it was almost constant for the first inital months (where there was very little data), but then we found a linear growth in the no of components each month, and decreases a little in the end (explained later). This was accompanied by an exponential increase in the number of nodes. By looking at this data, we can consider each component to be roughly a single community. This suggests, that there are groups of people who have shared each others posts, and there are other groups and there is no intersection between them. However, we are still not sure about the quality of the communities. This does not really explain whether these all these nodes are closely connected or not.

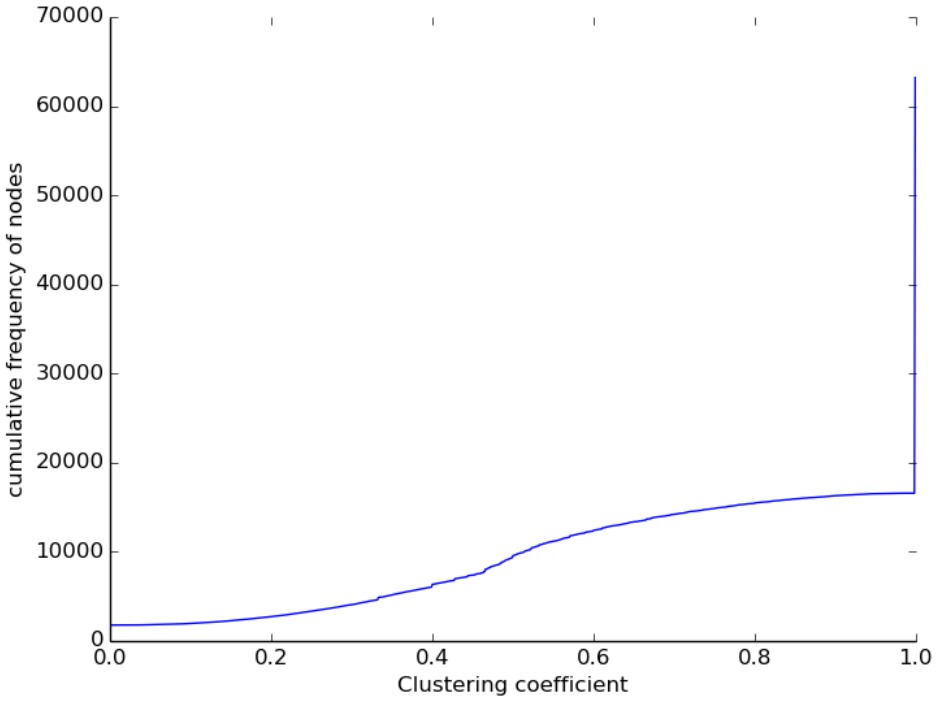

1400
1200
1000
800
600
400
200
0
No of components
Months

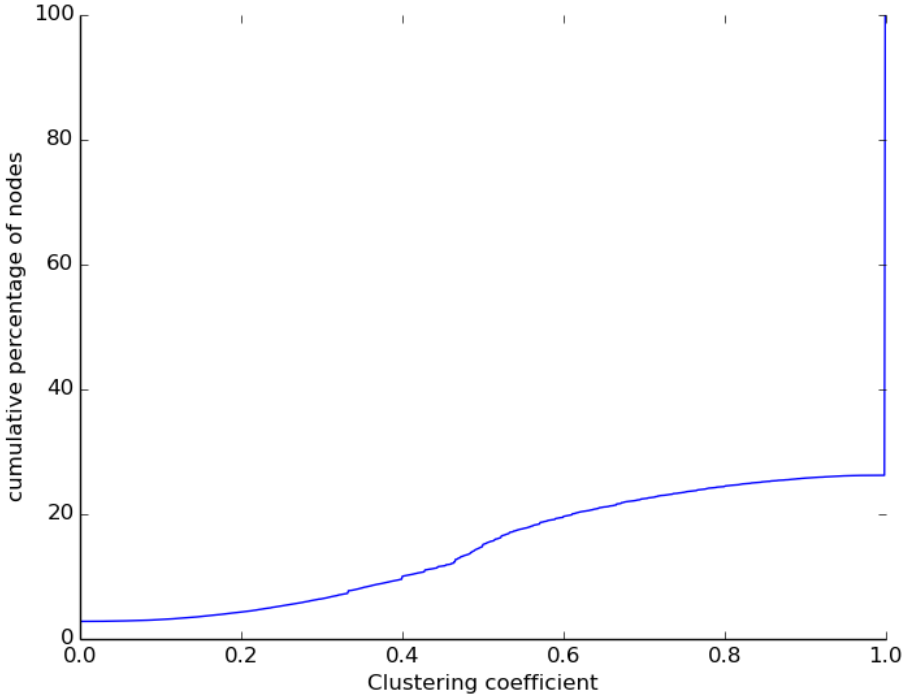

70000
60000
50000
40000
30000
20000
10000
0
No of users
Months

2. **Community Detection over the data every month**
We used the community detection algorithms to detect algorithms in the data, every month. We recorded the size of each community. We plotted a graph of the size of the communities vs the number of communities. We see that majority of the communities are pretty small in size. The graph has a peak, and the majority of the communities are of that size. This peak keeps on moving towards right every month. This suggests that the average size of the communities also increases every month, as the data available increases. There is only a miniscule number of communities that are big in size.

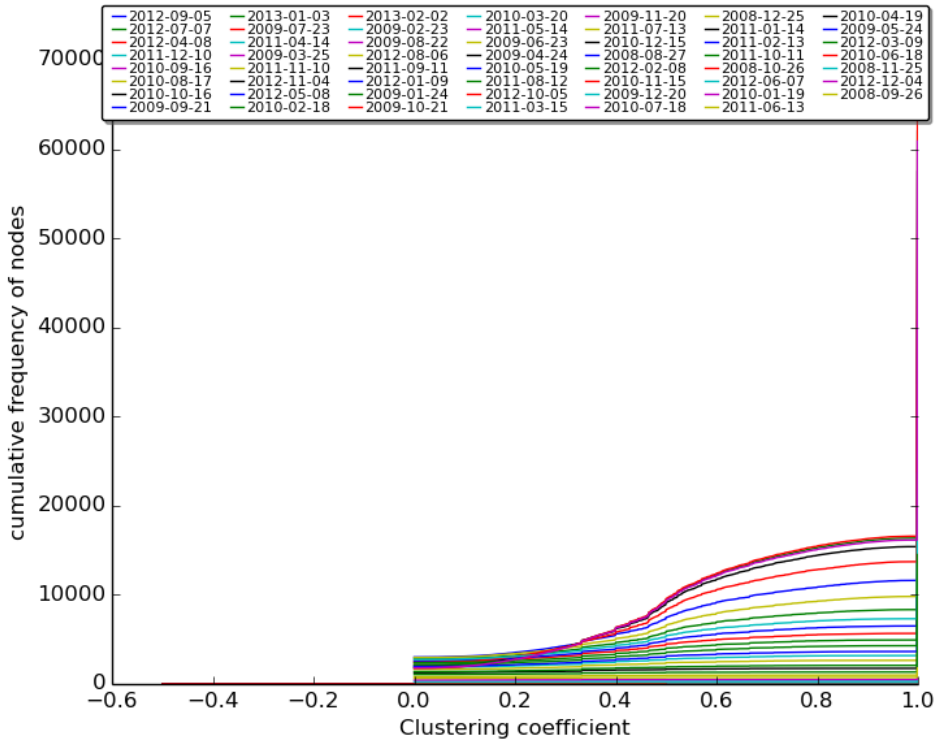


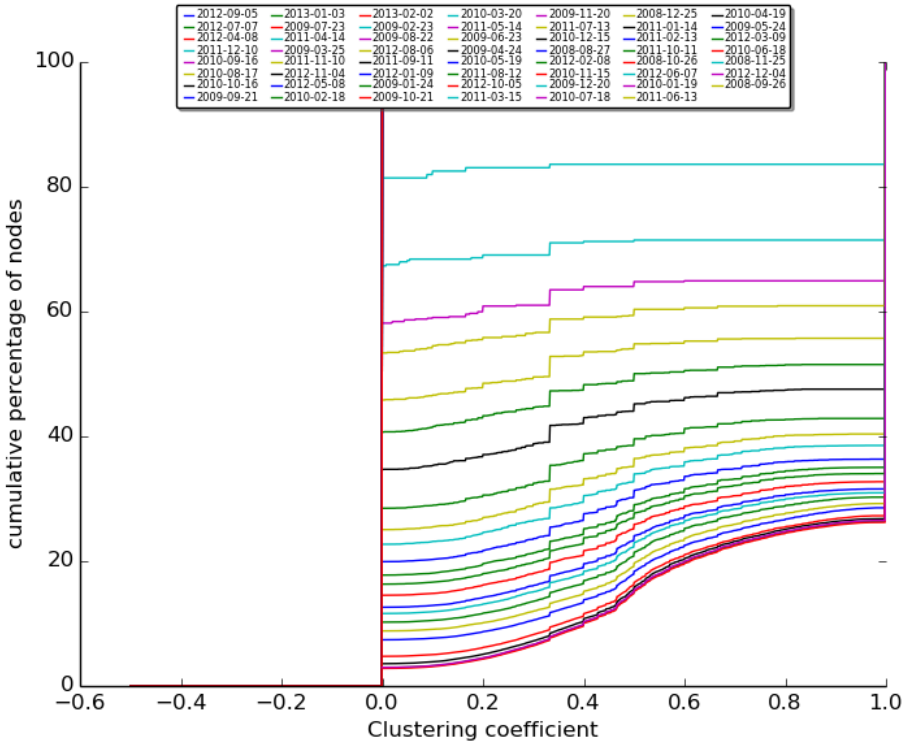
3. **Clustering Coefficient of the data**
We calculated the clustering coefficients of the nodes in the graph. Then, we plotted a graph between the clustering coefficient and the no of nodes having it. We found that majority of the nodes had really high clustering coefficients (between 0.8 - 1). This suggests that the communities are pretty decently well connected. There is a high probability of users having similar interests being connected rather than, because they shared some post which they found interesting randomly. The nodes with the low clustering coefficient can be the bridges, who have more than one type of interests. This development of varied interests also explains the decrease in the number of componenents in the later stages of the experiments.


70000
60000
50000
40000
30000
20000
10000
0
cumulative frequency of nodes
0.0
0.2
0.4
0.6
0.8
1.0
Clustering coefficient


100
80
60
40
20
0
cumulative percentage of nodes
0.0
0.2
0.4
0.6
0.8
1.0
Clustering coefficient

4. **Clustering Coefficient calculated every month**
   To understand the evolution of the communities over the whole period, we calculated the clustering coeffiecient of the data. We then plotted the graph of the clustering coefficient vs the number of nodes having clustering coefficient it. Since the number of nodes changes in the data every month, we calculated a graph with the percentage of number of nodes having a clustering coefficient having it. Initially, we saw that all the nodes had almost zero clustering coefficient. We notice that the graph keeps on shifting towards right as time progresses. We see that the fraction of nodes with low clustering coefficient to be very less compared to those with very high clustering coefficient.





4 Results

1. We saw that the graph (and the intermediate graphs) are disconnected. Each connected component approximately represents a community or an interest area and all people in that community share that interest.

2. We saw that the for an exponential increase in the total people under observation, only a linear increase in the components is increased.

3. We saw that most of these communities are of small sizes. The average size of the communities increases slightly as the number of people under observation increases.

4. We see that the clustering coefficient of these communities increases as a function of time. This suggests that people in the component tend to reshare the content already shared in their interest areas.

5. A large fraction of the nodes tend to have very high clustering coefficients (more than 60% have clustering coefficient close to 1). This suggests that there is a high degree of homophily in the network.