A Network Representation of Reddit's Communities

Erik Strauss, Josh Shaffer, and Matt Quintana

Abstract— **Reddit has many users that upload content into separate communities that exist on the site. These communities all tend to follow a specific them; some are very general / broad, and others are very specific and niche. The vast variety of communities present on the popular site Reddit creates an interesting network showing how Reddit's communities interact with each other. This document is taking the top posts of all time on Reddit as well as currently hot posts as of April, 13, 2018. We created six different data sets and started creating network graph representations of the interconnectedness of Reddit's most popular communities. We have found that subreddits that are broad such as r/pics and r/funny are more likely to have posts makes it to the front page.**

I. INTRODUCTION

Reddit is a popular website where people from all over the world can come together and share content that they enjoy or find interesting. This content can range from anything like news article to videos showcasing a cool clip from a video game. With such a wide variety of content present on the site, it only makes sense that people would come together to create special communities that revolve around certain kinds of content. These communities are known as ”subreddits” and there are thousands of subreddits that exist on the site. The best posts of these subreddits are often displayed on the front page of Reddit, along with the name of the subreddit that the post had originated from.

Using Reddit's front page, we want to gather the top posts of all time and use the data from that to create a network representation of Reddit's communities. We believe that this will show the interconnectedness of Reddit's communities and identify which subreddits are the most influential on the site.

II. MOTIVATION

Network graphs of social networks are very interesting because it shows how connected people are and different cliques that exist in the community. Now, Reddit is not considered a social network, at least not to the extent of sites like FaceBook or Twitter, but there are social aspects on the site. The subreddits are all run by community members and they each fit certain themes that different community members find interesting. We think that we could essentially model a network of Reddit that could be reminiscent of a social network.

There have been a couple of other people who have modeled some networks of Reddit before. One such network is analyzing comments on posts and figuring out which key words are most used in different subreddits. Other graphs try to accomplish what we are doing here as well. However, there do not appear to be many, if any, recent network representations of Reddit's front page. Reddit is constantly changing and new subreddits are created constantly, for example in the past year alone, r/prequelmemes has became one of the most popular subreddits and reaches the front page frequently. We want to show a network of Reddit's front page that is more representative of this year's current trends. Currently there does not seem to be any other network analysis out there that are doing that.

III. DATA

The first thing we asked ourselves was ”how are we going to pull the information we need off of Reddit?” We wanted our data sets to be fairly large, so doing it by hand would be too time consuming and tedious. Not only that, but we wanted to gather a bunch of different information about each individual post, not just their respective subreddits. Fortunately, we quickly discovered that the Reddit API is available for free to anyone, but only a set amount of data can be accessed per day (we did not need to worry about that since the data we wanted was not enough to reach the limit). We then needed to find a way to pull data out of the API automatically and not manually.

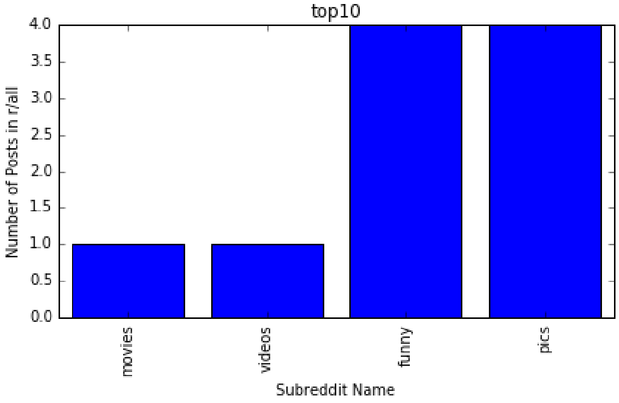
Luckily for us, Reddit makes it really easy to gather information from the site. The Python Reddit API Rapper (PRAW) can easily scan through the front page and other subreddits to gather information on the various posts. Using the PRAW, we created a script for a bot that we used to scan through the front page of Reddit and give us back post titles, post authors, post date, post subreddit, and of course the post karma (obligatory points that are given to a post to rate its popularity). Using this information, we were able to create a few interesting bar graphs showing the overall popularity of the different subreddits that reach the front page. We also used Python's networkx library to create a few network graphs with our data. We then ran a few analyses on those graphs to gather some more information about the overall network structure of Reddit's communities. If you are interested in reproducing our analysis be sure to check out our GitHub at: https://github.com/jls865/499NetworksProject

IV. Analysis

After performing the techniques and methods outlined in the previous section, we successfully gathered multiple data sets representing the contents of Reddit's front page. We used the data sets to generate three different bar graphs showing which subreddits have more top posts of all time that reach the front page. We also generated three other bar graphs showing the subreddits total posts that were currently hot at our time of retrieving data on, April 13, 2018. For the purpose of this experiment we used a fresh account to do the analysis so its front page has the same content that every basic user does.

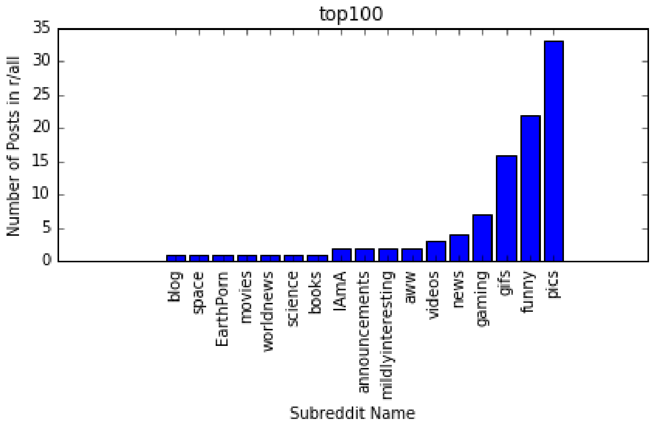
Top Bar Graphs:

First we will start with a data set of ten posts from the top posts of all time:



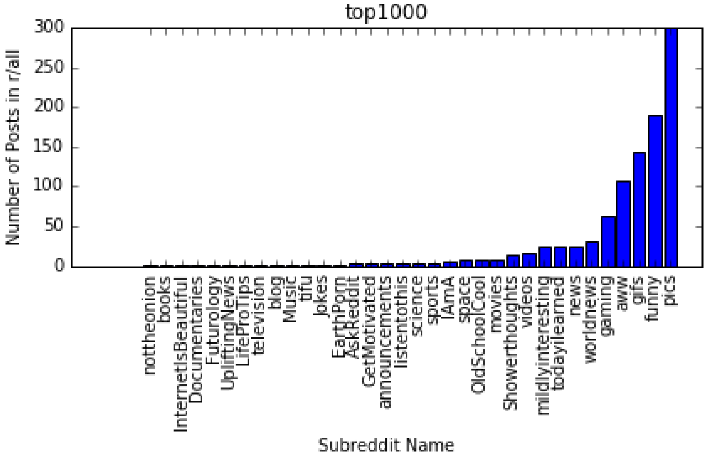
From the top ten posts, there were only four subreddits that appeared on the front page. Already, we can start to see that some subreddits tend to have more content appear on r/all because of their broad themes. We can also start to see how these subreddits act as an almost bigger community of subreddits with cross posts happening frequently.

The next graph we created was with a data set of the top 100 posts on Reddit:



Now we can really start to see that the subreddits are getting skewed. According to this graph, the most popular subreddit is r/pics, and if we think about it that does make sense. The theme of the subreddit is anything that is a picture, so it naturally has a wide variety of content on it. The wide variety of a subreddit seems to be what leads to reaching the front page more often. The subreddits with the smallest amount of posts on the front page are a little more specific and don't apply to as wide a variety as subreddits like r/pics or r/funny.

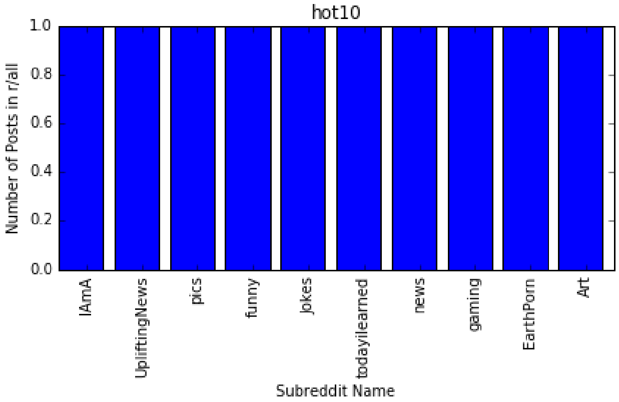
A bar graph using our largest data set, the top 1000 posts on Reddit:



Even with ten times more posts, r/pics still remains at the top, so it would seem that it is the most popular subreddit on Reddit. As we work our way down the data, the subreddits start to lose the broad and general feel and starts to hone in on more specialized content that appeal to a smaller amount of people. So according to this data alone, it would seem that we can say that a subreddit is more popular if it holds a wide variety of content instead of more specialized content.

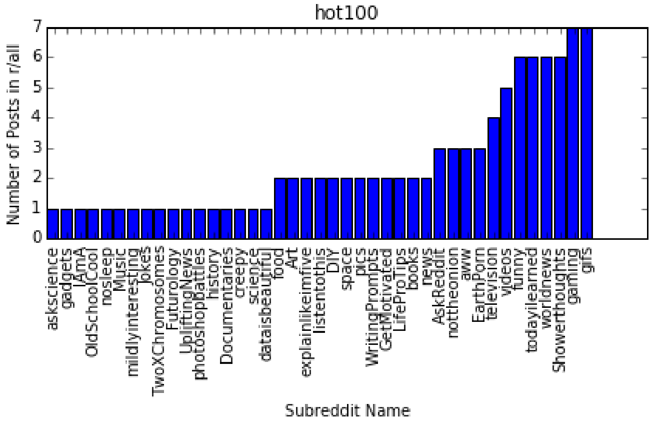
Hot Bar Graphs:

Now we will look at ”hot” posts on Reddit. As opposed to the top of all time, hot represents what is currently trending at a given time on Reddit. This is more representative of a single day of Reddit. Below is the bar graph for top ten hot posts from the front page based on subreddit:



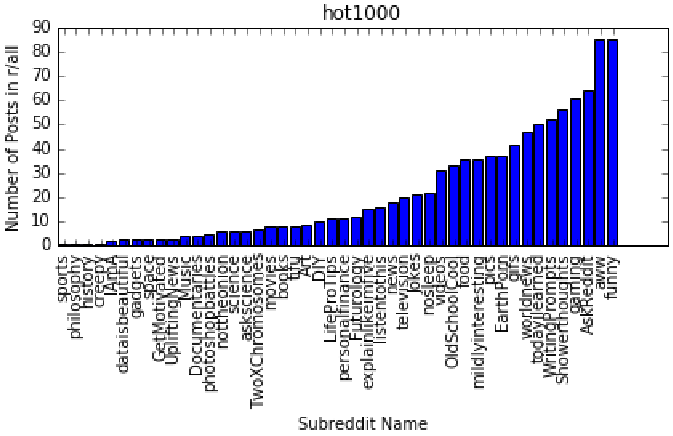
This graph shows that there is a lot more variety of subreddits that make it to the hot front page. In fact it shows an even distribution between the ten currently hot posts coming from ten subreddits. This was pretty interesting to see especially since the top all time was so skewed.

Below is the bar graph for top hundred hot posts from the front page based on subreddit:

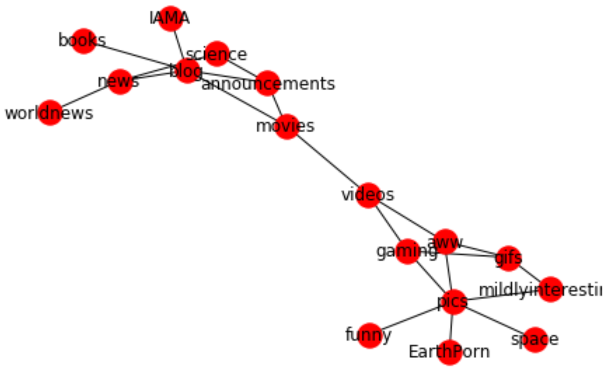


Now with a lot more posts in the data set we can see that the graph is starting to skew like the top post bar graphs. Though the graph is not skewed nearly as much as it was in the latter. There is a difference in results regarding subreddit types that have the most posts. For example, some of the more specific or niche subreddits have a decent number of posts on the front page. The broad ranged subreddits still seem to have the majority of posts but there is a noticeable mix of niche and broad.

Below is the bar graph for top thousand hot posts from the front page based on subreddit:

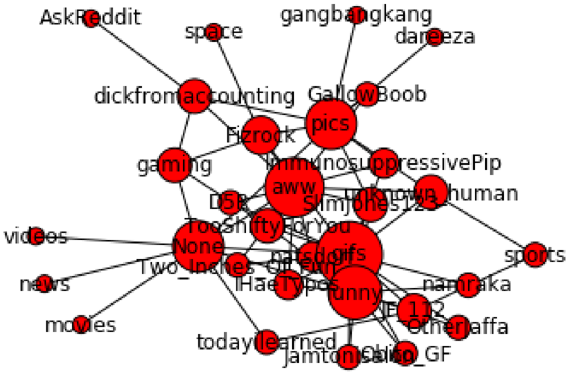


These results are interesting because for a thousand posts there are still only about fifty subreddits that they originate from. There are countless subreddits in existence but it seems that only about fifty subreddits have the most influence on the front page.



Here we took the subreddits present in the ”Top 100” data set and created a network graph based on the similarities in post content. Nodes are connected if the content between them is very similar in format or in types of posts. We characterize similarities between posts by looking at whether the content of it was an image, a text post, or a hyperlink to an external location. We can see a clear distinction between news and blog subreddits compared to picture and video subreddits. We can also see that r/pics is at the center of all of the picture subreddits. It appears too that r/movies is the real link between the picture subreddits and the more text-based and article-based subreddits.

Graph displaying relationship between subreddits and users:



For this part of the analysis, we decided to look at the relation between the different subreddits that appear on the front page and their authors in order to see if specific authors reached the front page more than others. In this graphic each node represents either a subreddit or an author and edges connect authors to subreddits that they have made a submission to that ended up making the front page. Similar to the previous results, many of the authors created submissions to the most popular subreddits of pics, aww, and gifs.

V. Results and Discussion

From our data, a large proportion of front page posts come from only a few different subreddits. For example, although we used data from 1000 different top submissions, less than 50 actually are relevant, as hundreds of front page posts are attributed to the subreddits of r/pics, r/funny, r/gifs and r/aww, subreddits that are related to more general topics that allow for a wider range of content to be submitted. We also found that when looking at the top posts of all time the majority of posts are more likely to originate from these broad subreddits rather than niche subreddits. One of the things that was most interesting is that if Reddit is the front page of the internet then the king of the that page is r/pics by far having the most top posts of all time. When comparing top all time posts with hot posts there seems to be a noticeable difference between the number and types of subreddits. For example, r/funny and r/aww are much more popular than r/pics, there seems to be a lot more subreddit variance in hot than in top.

VI. Conclusion

We have found that the front page of Reddit usually only features a select number of subreddits. This could be due to the fact that usually every users front page is different depending on what subreddits they are subscribed to. For the purpose of this experiment we had used a fresh account to do the analysis so its front page has the bare minimum of what everyone else's does. This could be the reason for there only being the broad subreddits and not niche ones because the account had not subscribed to any other subreddits. But this account is a good representation of the average user and what there front page typically consists of. Our results shows that there is power in the hands of few with specific subreddits and even users controlling the front page posts. In the future we could possibly gather data from hot periodically throughout the year and compare how the subreddit popularity fluctuates depending on the time of year. Another thing that would be interesting to do is to increase the number of posts to create massive data sets of top all time one hundred thousand and run the analysis on those to see how it effects the results. The last thing we would like to do in the future is to create a network graph that takes the top posts and groups them according to their submission date. This analysis would also be able to show the popular trends of each year or month in the past.

References

[1] https://praw.readthedocs.io/en/latest/