Raymond Mead 4/19/15

Analysis of Reddit Communities

Project Description:
Reddit is one of the most popular entertainment, social networking, and new sites. It is the 29th most popular website according to Alexa Rankings. Reddit has a unique structure, since it has many smaller communities, called subreddits, that make up Reddit as a whole. Anyone can create a subreddit community, and anyone can decide whether or not to be part of the community by subscribing to a subreddit. Each subreddit community is centered around posts made by the members of the community. Users make posts, which can be anything from text, to images, to links to anywhere on the internet, and other users vote the posts up or down, and possibly comment on the posts. The entirety of Reddit is driven by user behavior, and there is little involvement from the people who operate it, which makes it an interesting source of user behavior.

I wanted to look at how the Reddit community as a whole is structured. Finding, which communities are similar, which are different, and how communities tend to interact with each other. To do so, I wanted to look at the activity of each of these communities and how it affects the activity of similar communities. For example if the activity in the Game of Thrones community went up, how would that affect similar communities? Also, what are the behaviors of the users within Reddit as a whole? Do people tend to stick to small tightly knit groups, or is each user more independent and a part of many different communities?

I gathered two datasets about the communities. One was a general overall look at the communities and their similarities, where I gathered the top 1000 posts of all time. This allows an overall picture of the relationship of communities. I also gathered a second dataset, where I got the top 250 posts for each month in 2014. From that dataset I want to look at the activity and relationships of subreddits over time.

The reason I structured the project in this way was because it takes more time to get the information for subreddits over time, so the initial overview is there for overall information, while the second is if there is enough time to go into deeper analysis.

Data Gathering:
The process I used to gathering data involved getting a list of all the subreddit communities as a whole, and an overview of the information about each of the communities. Reddit has their own API which I used to gather data.

For each community I gathered:

- Rank (Where did it fit in the overall popularity)
- Name and URL
- Size (Number of subscribers to the subreddit)

I gathered information between all of the communities with over 50,000 subscribers, aside from the top communities. The top communities are those which frequently reach the top posts of Reddit as a whole. There was a large gap of 300,000 subscribers between the top communities, and the next popular one. I used this large divide as a cutoff for the top communities.

Afterwards, I searched through the top pages for each subreddit, where each page contained 25 posts. I gathered the links to the top 1000 posts of all time for each of the communities I considered. I then saved these links, and later went to each post to gather more detailed information from each post.

Once I had the links to the top posts of a community, I gathered information for each post. The information I gathered is as follows:

- Post URL (link to the post)
- Link (what the post links to)
- Title
- Submitter (User that posted the link)
- Time of the post
- Score (Related to whether people voted the post up or down)
- Number of Comments
- Who commented on the post, and how many times they commented

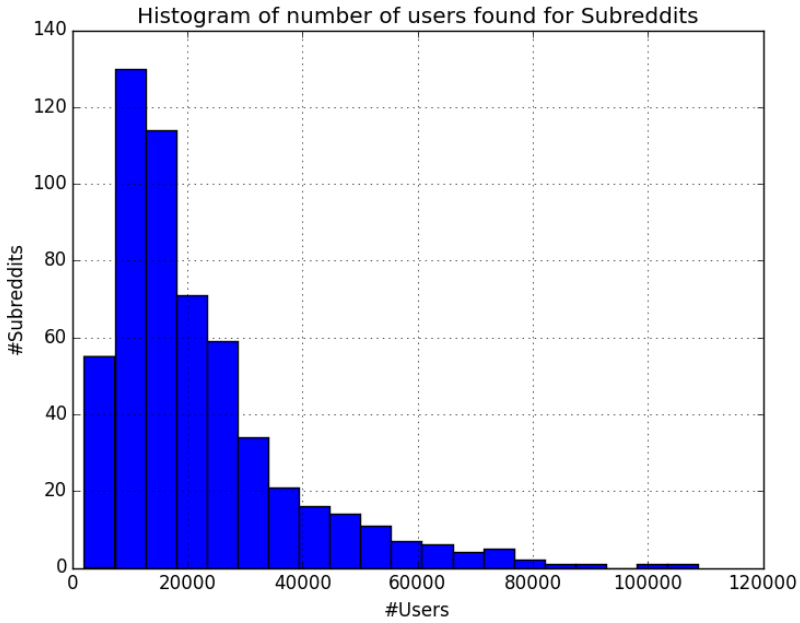
For the second dataset, I gathered the top 250 posts for each month in 2014, from all of the subreddits which have over 50,000 subscribers. For each of the posts I gathered the same type of information.

The users for each subreddit community are found from those who commented on posts. There is not another way tell whether a user is part of a community or not. Unfortunately this means that my user set for each community is only a small sample of the size of the community as a whole. It also means that I do not now all the communities a user is a part of. Despite this, the user sample is a good measure of the relative similarities between communities, but it makes analyzing the user base, for the most part, an unapproachable task, since the user data is very sparse.

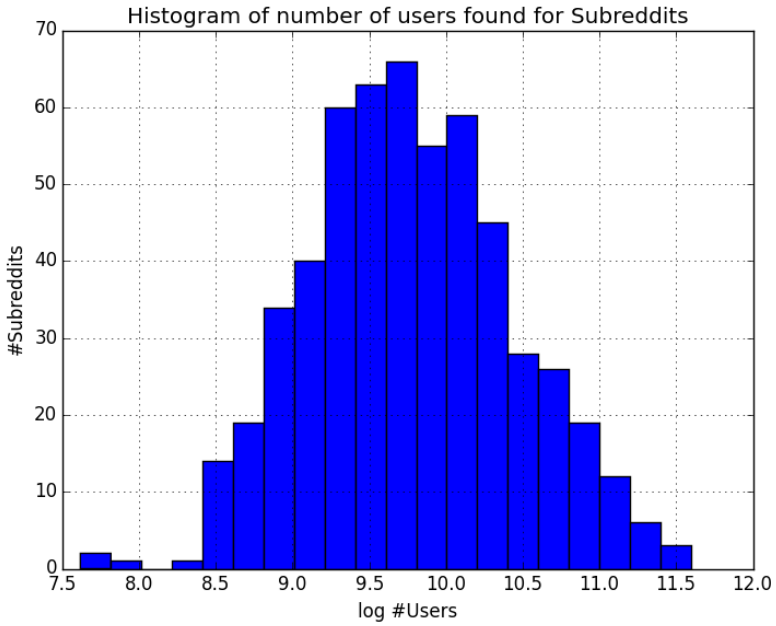
General Statistics:
Number of subreddits analyzed = 553
Total Number of users gathered = 2,585,741
Total Number of posts scraped $\approx$ 553,000

Graphs that look at a general overview of the information

Histogram of number of users gathered from subreddits:



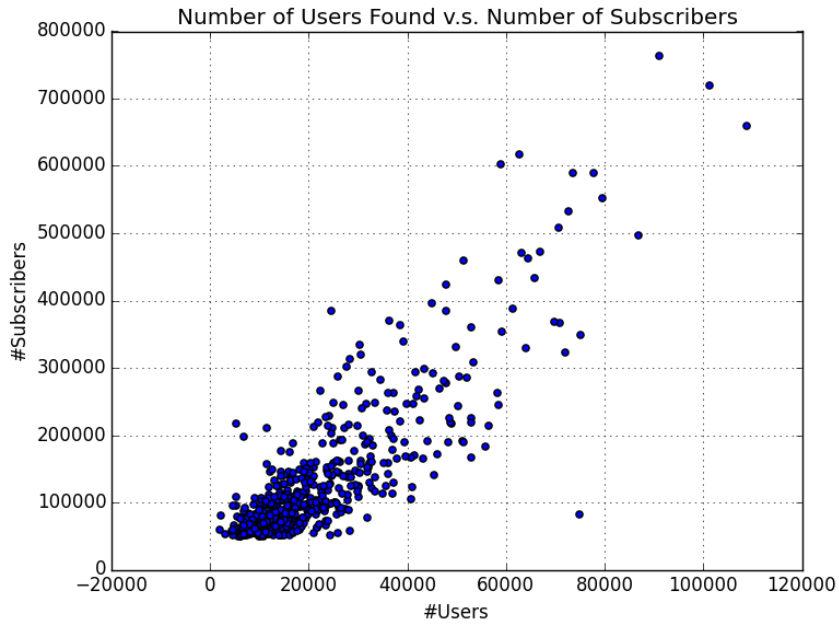
This plot looks like a lognormal distribution.
I plotted the histogram of the log of the number of users gathered to confirm that suspicion.
Histogram of log(number of users gathered from subreddits).



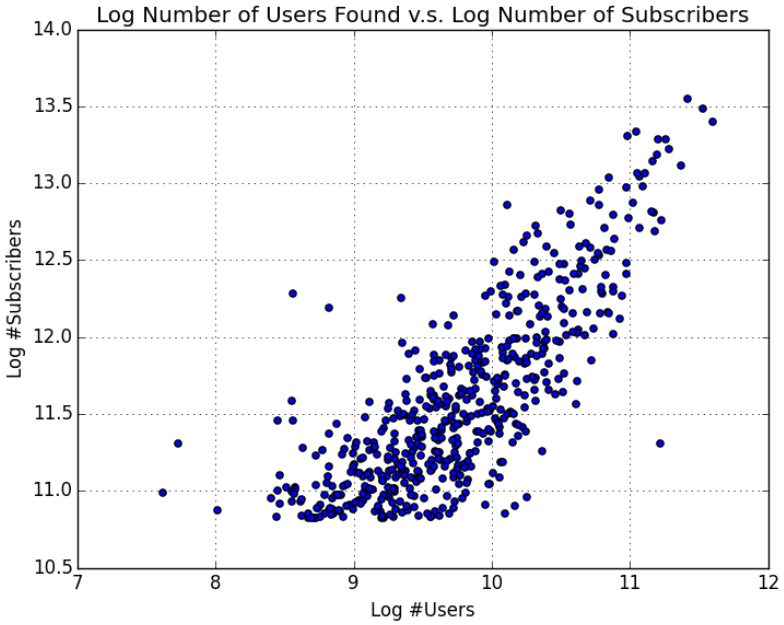
This shows that the dataset of users I have is lognormal. Though this should not be looked at too deeply since I employed an arbitrary cutoff for which communities I scraped.

I wanted to see how my sample of users compared to the true size of the subreddit, which is measured by the number of subscribers for a subreddit.

Gathered user size vs number of subscribers:

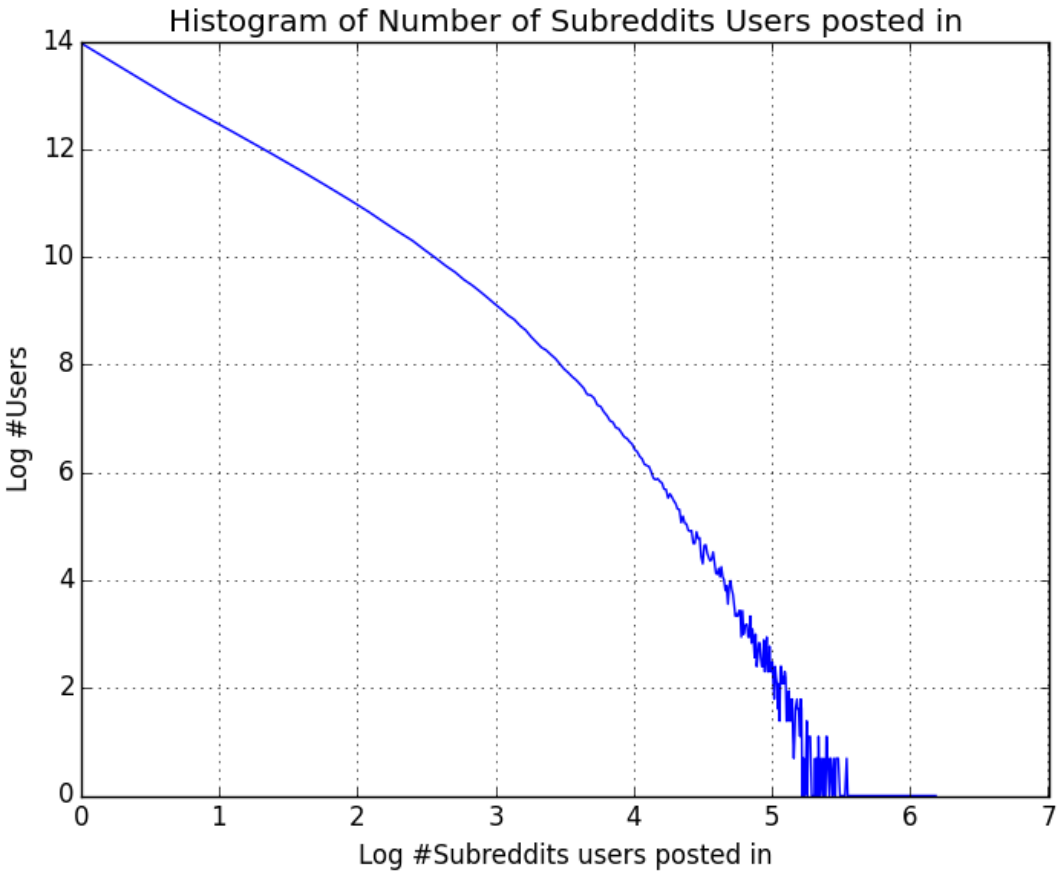


Log-Log plot of Gathered user size vs number of subscribers:



You can see that for the most part there is a proportional correlation between the two, with some outliers. But this shows that the sample of users is not that bad, and the relative variation is more apparent with the smaller subreddits. It should also be taken into consideration that certain subreddit communities do not have as many people who comment as a similarly sized community, due to the nature of the content in the community.

Log-Log Histogram of Number of Subreddits Users posted in:



We would expect to see a large bias towards the lower end of the scale, since I only gathered a sample of posts. Therefore it would be much less likely to find someone in multiple subreddits even though they may post in both, and also we would not find every subreddit a user posted in at the upper end of the scale as well. However, even with that consideration, the plot follows a power-law distribution for a while until it gets to the larger. So there is some evidence that the number of communities someone is in would follow a power-law distribution, but there are considerations that prevent us from asserting that completely.

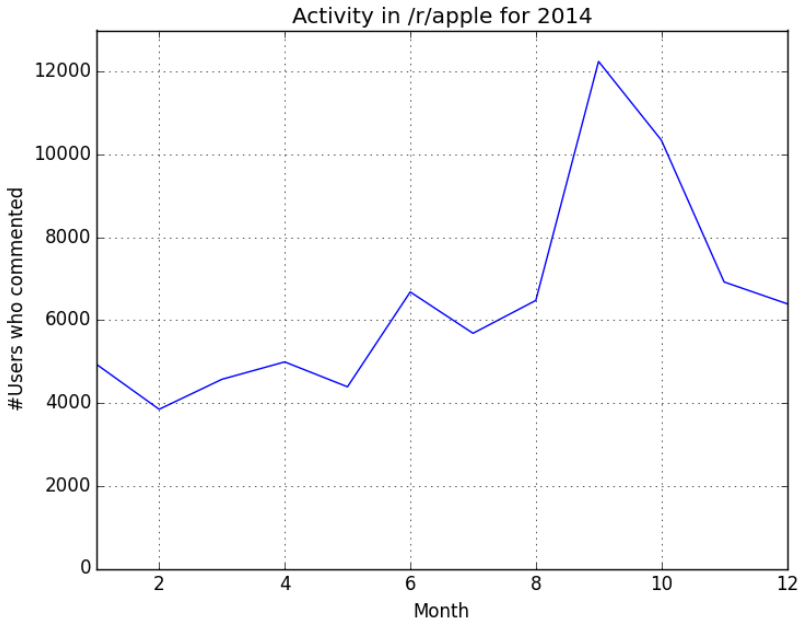
Most Related Subreddit Communities according to Jaccard Similarity of the set of users who posted in each community.

Top 10 Jaccard Similarities:

Subreddit	Subreddit	Jaccard Similarity	Comments
britishproblems	unitedkingdom	0.195191	Overlap between two British subreddits
darksouls	DarkSouls2	0.193226	Game and its Sequel
AskMen	AskWomen	0.191231	It's very funny that those who comment in AskMen also do so in AskWomen, showing that you may not be asking men or women exclusively
nintendo	wiiu	0.183381	Two Nintendo related communities
Autos	cars	0.178064	Two Car related communities
CFB	CollegeBasketball	0.174071	Overlap between college sports: Football(CFB) and Basketball
asoiaf	gameofthrones	0.171421	Two Game of thrones communities: asoiaf(A Song of Ice and Fire) is the name of the series of books that George R.R. Martin wrote, and gameofthrones for the corresponding T.V. Show
webdev	web_design	0.168647	Web Design and Development
business	Economics	0.163653	Those interested in Economics also interested in business
3DS	wiiu	0.160669	Again Two Nintendo related communities, they must be a very tight cluster

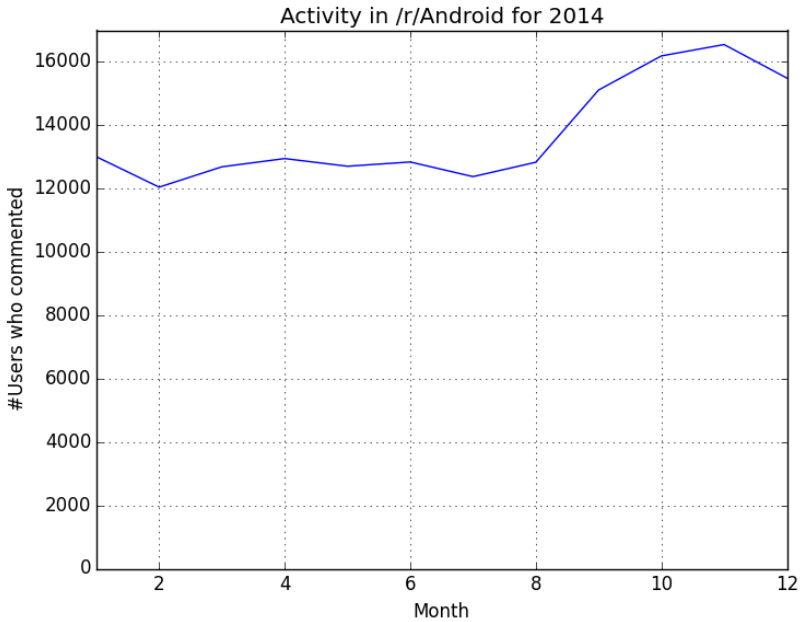
Some Initial Plots of Activity of Various Communities

Apple:



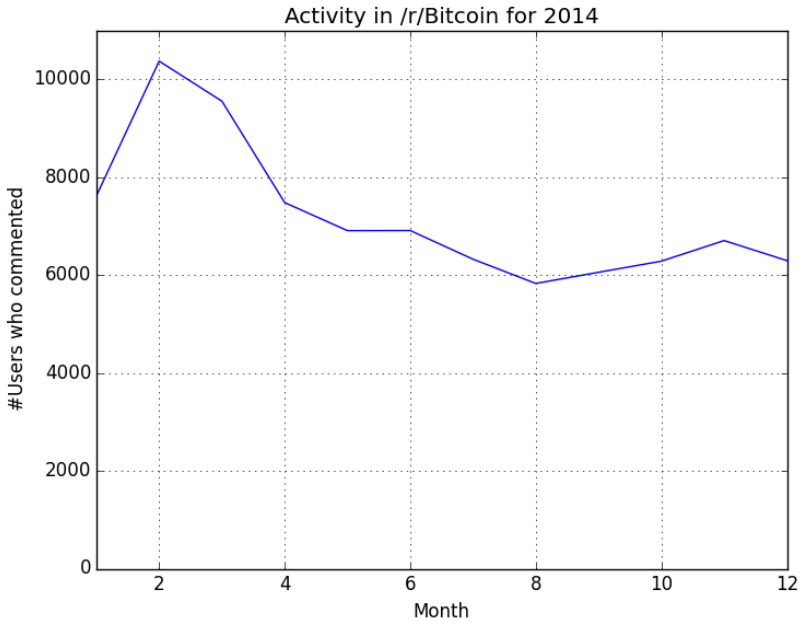
The spike in the graph is around the release time of the iPhone 6. There is also a small spike in June when Apple had their Worldwide Developer Conference, where they unveiled iOS8 and other new software developments.

Android:



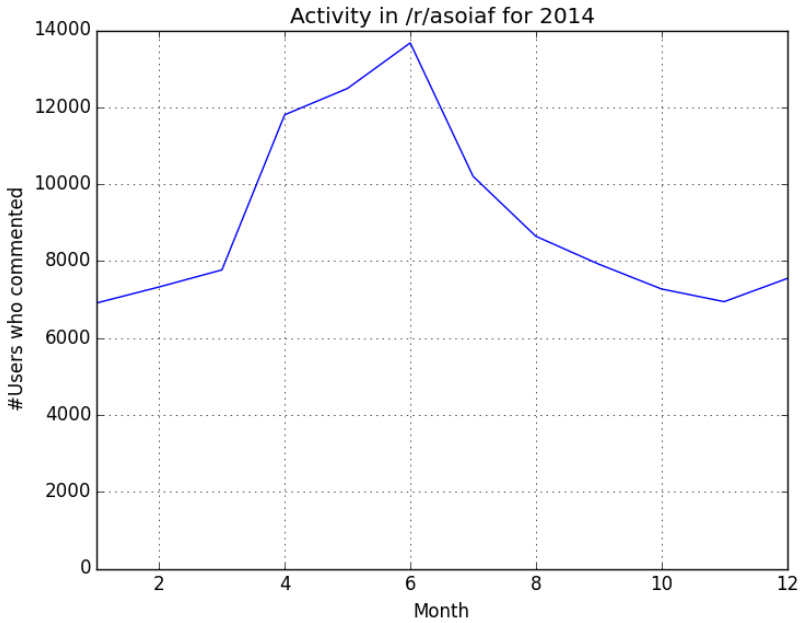
The increase at the end of the year was when Google was rolling out Android 5.0. The spike here is much less pronounced than the one for Apple's new release, possibly showing the different mentalities towards each company.

Bitcoin:



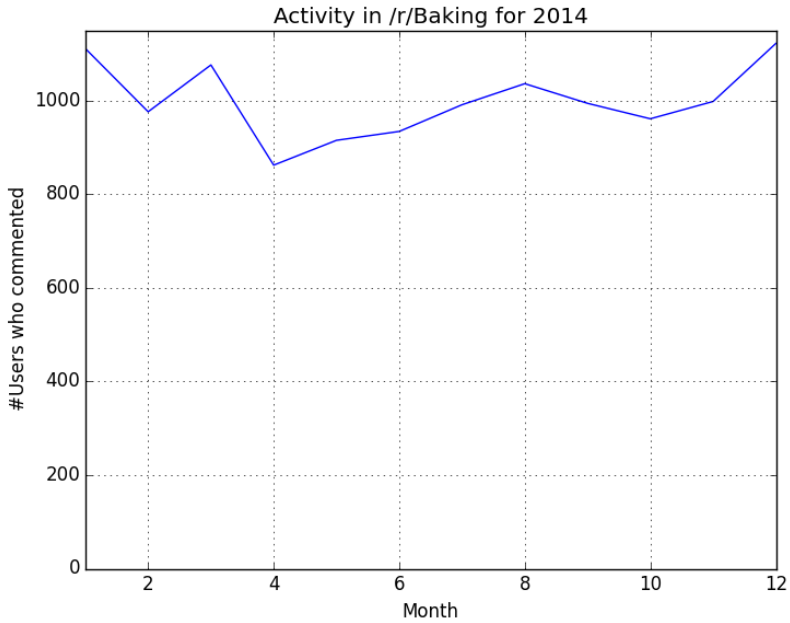
In February there was a collapse of the largest Bitcoin exchange, Mt. Gox, and 850,000 Bitcoins were lost, with a value of around $500 million. The Mt. Gox CEO had to declare bankruptcy due to the liability of such a loss.

Asoiaf(A Song of Ice and Fire – Name of Game of Thrones Series of Books):



The activity in the community reached almost twice its base amount during the time that HBO was showing the Game of Thrones T.V. show. Even though the subreddit is not officially about the T.V. show, since it is supposed to be dedicated towards the books, there is still a rise in activity.

Baking:



This is an example of a community that is not related to events throughout the year. Because of this there is not much variation in the amount of activity over the course of the year, which is expected.

Further Plans

This is just an initial examination of the data. I am still in the process of gathering the historical data for some of the subreddit communities. Once I have all of the data there are a few things that I hope to look for.

First of all, I would like to do a clustering of the subreddits based on their Jaccard distance to see how the communities are clustered. In addition to that I would want to project the relationship of subreddits onto a 2D graph, so that you could see a "map" of the Reddit communities as a whole. I would also want to look at a PageRank of the subreddits, to see where most people would aggregate if they were doing a random walk through the various communities. The transition probabilities in this case would be proportional to the size of the intersection of two communities.

In addition, I want to look at the relationship between Jaccard Similarity of two subreddit communities, compared to the covariance of their activity. I expect to find a relationship, but this would show how much of a relationship there is between the overlap of users and how much activity would change in similar community.

Also, I want to look at the growth of all of the communities throughout the year, and how much the similarity of subreddits changes throughout the year. Do they mostly stay the same? Or is there a large fluctuation depending on the different times of the year?

You could also look for various spikes in the activity of various subreddit communities, which would reflect major events related to those communities. This would shed light on what things affect people's activity the most throughout the year.

With this data, there is a lot of information to examine networks of communities and how they interact. Does activity in a community positively or negatively affect other communities? How does growth spread throughout a network? Since Reddit activity reflects of the actions of people and groups of people, there are probably patterns in the data set that reflect our behavior as human beings in complex networks.