Introduction

We had several ideas at the beginning, but the one that stuck out at us was analyzing Reddit. Reddit is a forum like site based on different sub communities called subreddits. These subreddits all have a topic and the topic relates to their name. For example /r/funny is a place to post anything funny. Our idea was to get a better idea of how these different communities interact with each other. Many subreddits share similar topics, and Reddit has a convention to share this content between subreddits, it is called a cross post or x-post. If a post contains x-post then that content was originally from another subreddit. Using each subreddit as a node and x-posts as edges, we knew we could form a network.

Once we had this network formed in our minds, we had some initial questions that we wanted to answer. Questions like what will be the hubs of our network, and are there subreddits that have a disproportionate degree compared to the number of subscribers? Questions aside, we also had to ask ourselves why this was important. Reddit is a site with over 200 million unique visitors a month, and the people that visit reddit have a wide variety of interests. We knew that gathering data on a site like this could lead us to some unique insights, and could lead us to see something that no one has investigated previously.

Only one project we found came anywhere close to what we wanted to do. So this provided us with an interesting opportunity to do something totally unique and draw insights into data that no one has seen before. The next step for us was to create our scripts to gather the data and then clean the data making sure all of the data is relevant and fits our criteria.

Data Description

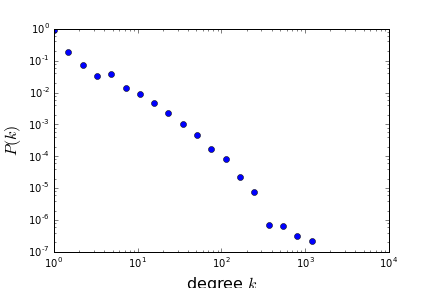
Our data was collected by first identifying the subreddits we wanted to search for cross posts. We agreed that we wanted to use all subreddits that had greater than 1,000 subscribers which amounted to the top ~18,500. We queried the website <http://redditmetrics.com/> to create our list of these top subreddits and used the list this in creation of our queries.

We created a reddit scraper that searched each of these top subreddits for all the the posts with the keyword ‘x-post’ within the last year. Reddit’s API would return pages of JSON which we then parsed with a python function to extract the title, content and subreddit. After Initially running our script, we noticed that 8 subreddits had hit the maximum amount of records that reddit would return to us per query. For these 8 subreddits we created custom queries based on the timestamp of the last post before it maxed out and the timestamp of a full year from when we ran the query. After gathering the data we needed from the overflow subreddits, we began to work on the data.

From the title and content, we used another custom function with a hierarchical structure of try-except blocks to find the best candidate for the subreddit that the content was taken from. After identifying the linked subreddit, we cleaned our data with another function, eliminating self-linked records and formatting everything to import to Gephi.

Before starting our project we extensively searched Reddit for anything similar to our project as we wanted it to be totally unique. We wanted to gather our own data and make insights that no one had done before. After searching different keywords in general and specifically within subreddits, the only similar project we could find was burried within r/dataisbeautiful : <https://redditstuff.github.io/sna/vizit/> . This project was not to the caliber or scale that we wished to examine subreddits and did not think that it would be an issue for our goal.

Empirical Results

We ran several tests on our data, and what the statistics we found most relevant were the Clustering Coefficient, Average Degree, and Average Path length. We also made a degree distribution graph shown here.

What we see here is a scale free network, the graph shows a power law tail showing this. The average degree is 7.77, which is means it is pretty highly connected. Along with the path length which is 4.5, shows that this network has the small world phenomena. The last statistic we decided was interesting was the clustering coefficient, which is .100. Many nodes on our visualization are pushed to the outer edge being only connected to one node. Compared to the data we used on our visualization this is low, our visualizations clustering coefficient is .373. The explanation is simple, in our full data there are many more subreddits that are small and most likely only have a degree of one or two, but in our visualization we downsized it to some of the largest subreddits. What would be interesting is to see a visualization of our full data, and compare it to the one we made.

Insights

Our network has afforded us some powerful insights into Reddit and the committees formed by subreddits. We observed two large interconnected groups within our network with many smaller communities that we also analyzed. The largest group is comprised of the three largest nodes in the network, r/funny, r/pics and r/videos, which are strongly connected to each other with many links to smaller subreddits as well. These nodes mainly comprise the entertainment/humor section of Reddit and are within the top 10 largest subreddits as well.

The other large group that we see is comprised of r/gaming and r/pcmasterrace. From this gaming/technology section of Reddit we see that many intermediately sized subreddits such as specific games or pc topics linked to these larger hubs. We theorize that large subreddits such as gaming might be unfairly biased towards the most popular games that have more influence and link more content than being supportive towards all types of gaming. We also see a very large link between r/pcmasterrace and r/buildapc showing that most of the content cross posted to r/buildapc is from r/pcmasterrace.

Looking at a smaller community focused on sports subreddits, we see some interesting characteristics as well. Looking at the nodes it is comprised of, we see one of the large nodes linked to r/soccer being r/gunners, the specialized subreddit for one of England's premier league teams. While not particularly famous or good, we can deduce that they have a very large fan base and active community from our network. We also see that in the area of our sports community are the subreddits for golf, bicycling, and fishing as parts of other communities, having not much in common with r/sports at all. From this and other smaller network, we are able to see how similar topics may not necessarily be linked and how they may have more in common with other communities elsewhere.

Conclusion

Looking back at our goals, and what we wanted to accomplish with this project we were very successful. We were able to answer the questions we initially had and found answers to questions that could only be thought up looking at our visualization. Our initial questions were basic what are the hubs going to be, and what subreddits are disproportionately represented. The first was easy to guess, but the second could only be seen from the data. Subreddits like /r/techsupport had a small number of subscribers, but a large degree. These types of results were interesting to us.

We have went over some of the insights drawn from looking at our visualization. Seeing the bias that appears in communities based on the communities around it, was something we never thought we would see. We believe more insights can be drawn from this data because it is so extensive. One limiting factor is not seeing being able to visual all of our data. Most likely we would see more small communities that could be interesting to see their interactions with the communities around them.

Looking at the skills we have learned doing this project, we both have strengthened our programming skills through using python for mass data collection and cleaning data. Obviously, we learned new analytical techniques using network science to analyze large data sets. Networkx provided an easy gateway into the techniques we used on our dataset. Finally, we feel that there could be more work done, and more insights drawn with a little more work.

Future work

For our work to continue and become much more accurate, we hope that in the future, Reddit allows users to input the subreddit that they are taking content from in the same way that they enter where they are posting it to. Most of the challenges we had with the project were identifying that source subreddit and we will be able to produce even more powerful insights if out data was perfectly accurate.

More conclusions about the interconnectivity of reddit can be drawn from analysis of the smaller highly connected nodes within our network. Specifically, r/techsupport was ranked 585th in size but had the 35th highest degree, showing that much of the content was cross posted and it was highly connected despite its size. Looking at the most disproportionate nodes like this would be our focus for analysis of smaller communities.

We would also look at the links to the largest subreddits individually or the large communities individually to provide us insight into their structure. Looking at these individually would allow us to see outliers and identify what really comprises each community. We would be able to see more examples of subreddits being disconnected and separate from their overall category (golf and sports) and how some pairs can be extremely interconnected (buildapc and pcmasterrace).

Overall, this project was extremely fun and we enjoyed every minute of it. Taking our concept from the initial idea through creating our scripts, gathering our data, cleaning, modeling, analysing, and presenting was rewarding and exciting for both of us. We may not have the opportunity to enact all these plans for future work, but these ideas reflect our drive and passion for the product that we have produced.