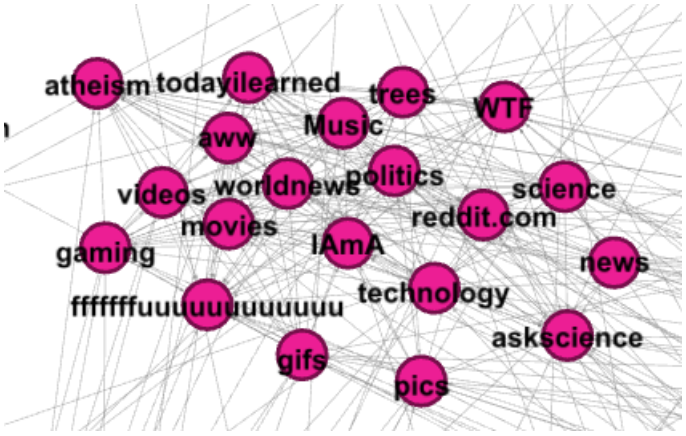
Robert Johns
December 5, 2014

CMSC 491-3: Complex Networks
Category Diffusion in the Reddit Network
Project Final Report

1. *Introduction*

My project deals with a social network analysis of reddit. Reddit consists of thousands of communities, called subreddits, where users post information relative to a topic. The most recent data I could find indicates that there nearly 450,000 total subreddits, around 10,000 are active (being consistently posted to and commented upon). Each subreddit has, in addition to a list of the top user posts (which can be links or "self posts", which don't link to an external website but are prompts for discussion), a sidebar containing descriptions of the subreddit's rules, links to related external websites, and, of the most interest to this project, links to other subreddits with similar content. My project focuses on the sidebar links to other subreddits, treating each hyperlink as a directed edge from one subreddit node to another. Using this as a model, many interesting dynamics of this "community of communities" can be studied.

At first, the main focus of my project was merely the data acquisition. After I obtained the data and studied it, though, I developed a hypothesis about the community structure of reddit and a better algorithm to discern it. The main issue I noted has to do with the modularity algorithm and its applicability to discerning qualitatively related nodes in a network. Because the modularity algorithm can make its determinations based only on edges and nodes, and while generally very good at determining community structure based on those factors, cannot determine that two nodes in the same modularity class are totally unrelated, or that nodes in different classes are similar in their qualitative category. For example, take the modularity algorithm's results for reddits's default subreddits (those users are subscribed to after making accounts):



Because a large chunk of the user base is subscribed to these subreddits, they will be densely connected to each other in the graph and thus will be in the same modularity class. However, in terms of their content just about all of these subreddits are completely unrelated. The goal of this project is to develop an algorithm to fix that.

2. *Data Acquisition*

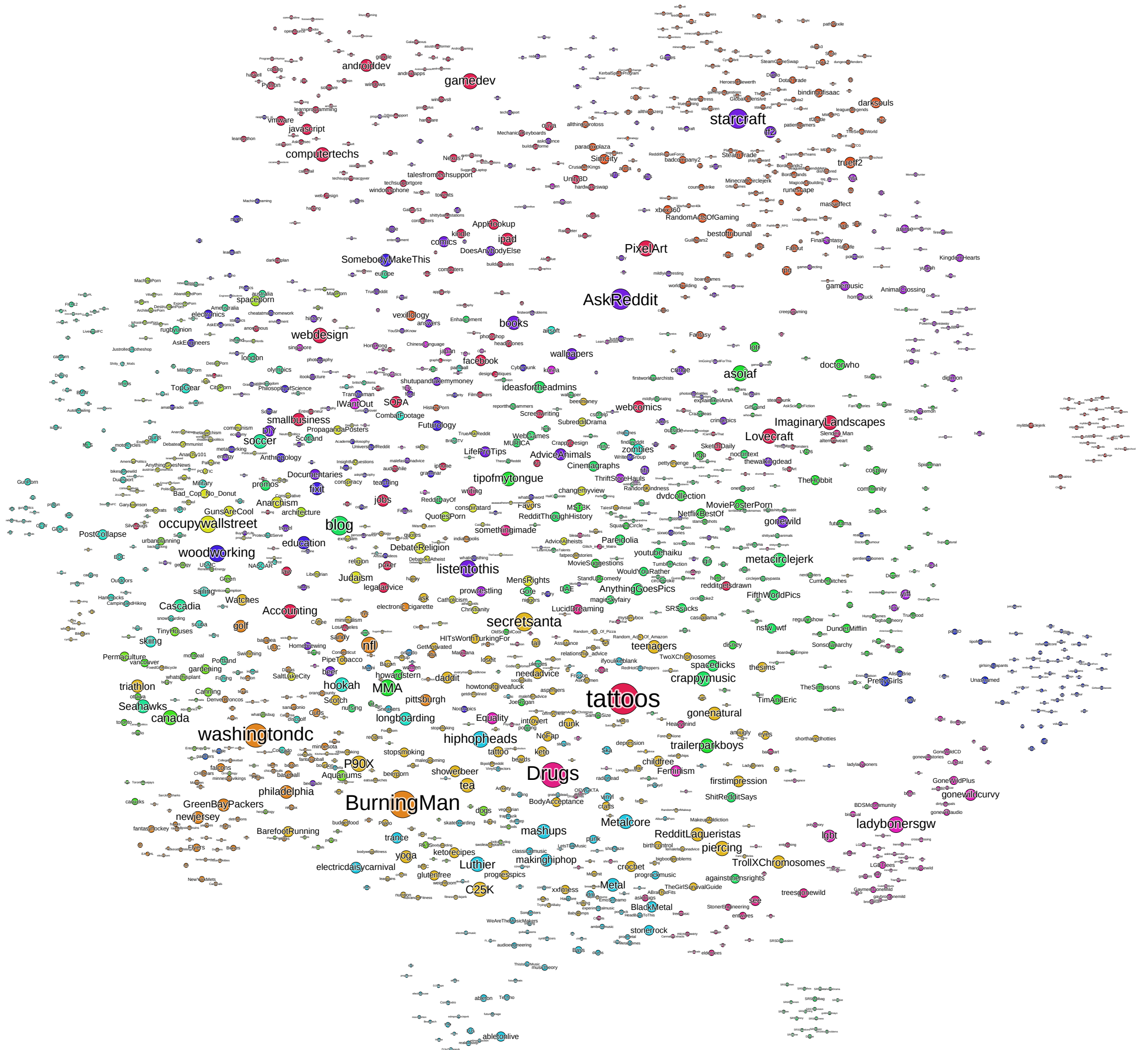
My original plan was to crawl reddit for data and collect two different types of connections: connections from the sidebar between subreddits and connections from user posts to external websites. My plan was to incorporate the type of website linked in the top posts in the classification algorithm. To this end, I crawled the website redditlist.com for a list of the top 5000 subreddits, and put them in a text file. I then wrote a script to pull the links from any given subreddit page and tested it on a small set of the top 5000 subreddits, and set it to work. After a short period of time, however, I was dismayed to discover that the reddit servers limit the API calls to around two per minute. While I could easily work around this if given enough time, due dates prevented me from allowing my script to run for days on end. I then went to google hoping that someone without time constraints had collected similar data, and was delighted to find that a reddit user had a github repository at http://rhiever.github.io/redditviz/ containing the data I needed, minus the connections to external websites, and had written an interesting academic paper on the network. As the subreddit connections was the real key to the data, this new dataset was more than

satisfactory.

The aforementioned dataset contains information about over 22,400 connections between more than 5,288 subreddits. The data obtained was in the form of a .json file, which gephi produces as output but cannot take as input. I wrote a python script to transform the data into a .csv file containing the edges information which gephi can take as input, and examined the results.

In case it is of interest, I will also include the code I used for my original crawler, which would work well if I had two weeks for it to run.

On the next page is a graph of the collected data with node size determined by betweenness centrality and color determined by modularity class.


androiddev
gamedev
starcraft
darksouls
computertechs
javascript
vmware
PixelArt
AskReddit
SomebodyMakeThis
webdesign
books
facebook
wallpapers
ideasfortheadmins
asoiaf
doctorwho
webcomics
ImaginaryLandscapes
Lovecraft
smallbusiness
soccer
IWantOut
tipofmytongue
AdviceAnimals
occupywallstreet
blog
woodworking
education
listentothis
metacirclejerk
secretsanta
Accounting
Cascadia
tattoos
washingtondc
canada
Seahawks
triathlon
hiphopheads
Drugs
BurningMan
MMA
hookah
nfl
P90X
longboarding
pittsburgh
trailerparkboys
crappymusic
teenagers
spacedicks
gonenatural
ladybonersgw
Metalcore
mashups
RedditLaqueristas
piercing
TrollXChromosomes
Luthier
C25K
Metal
makinghiphop
GreenBayPackers
newjersey
philadelphia
yoga
electricdaisycarnival
BarefootRunning
firstimpression
ShitRedditSays

3. *Algorithm*

   The algorithm I developed was inspired by our section on information diffusion in the class. The man concept was this: instead of information diffusing through a network, usually in some binary on-off way, why not have the diffused information be a category of the nodes, and let each node have a number representing a degree of belonging to each category.

   To start, I manually created a seed set of 403 nodes I categorized myself. I then created a $|V| \times 23$ array, with each column (aside from the subreddit name) corresponding to the subreddit's degree of belonging to that category. The categories were: movies, TV, music, art, food, books/literature, food, games, technology, science, meta (reddit), sports, local, life, sex, LGBT, news, politics, history, religion, comedy, animals, nature. I then started diffusing the information through the network, the idea being that if a subreddit is connected to nodes that don't share its category, that category will not spread too far, but if the others do share its category, the degrees of belonging will reinforce each other and the surrounding nodes will increase their degree of belonging to that category. I percolated using the following parameters: for a degree of belonging ranging from 0-9, I then uploaded wrote the data to a .gdf file and examined the results in gephi.

4. *Results*

   To measure the success of the percolation, I examined the "primary category" for each node; that is, the category with the highest degree of belonging for each node. The results are listed in the following table, along with a percentage of nodes with any degree of belonging to that topic:

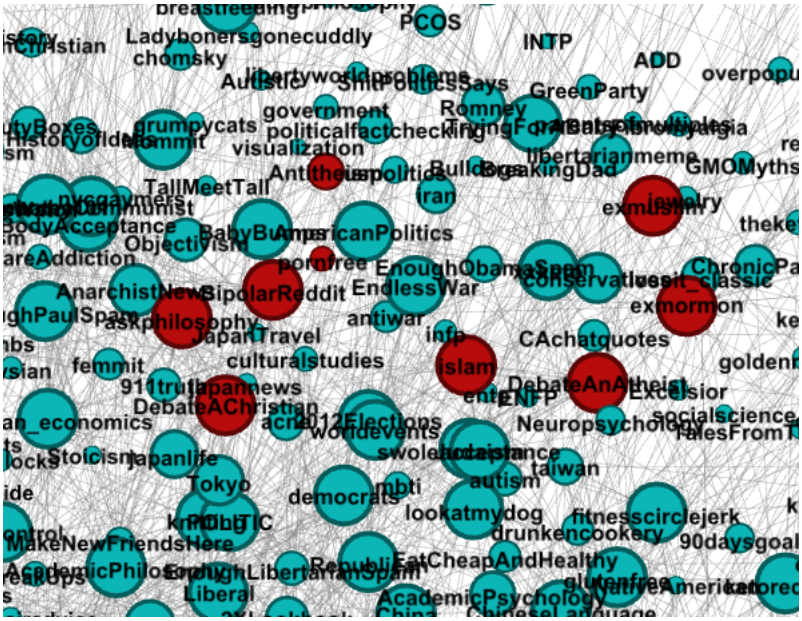
topic	%primary	&true
none	12.78	12.78
movies	1.3	12.62
tv	4.21	20.37
music	2.79	13.62
art	0.13	2.05
books	0.94	9.41
food	3.18	22.13
games	9.58	31.18

topic	%primary	&true
tech	5.57	23.81
science	2.79	23.78
meta	0.77	12.72
sports	6.5	25.31
local	9.47	32.48
life	12.74	39.88
sex	7.54	25.31

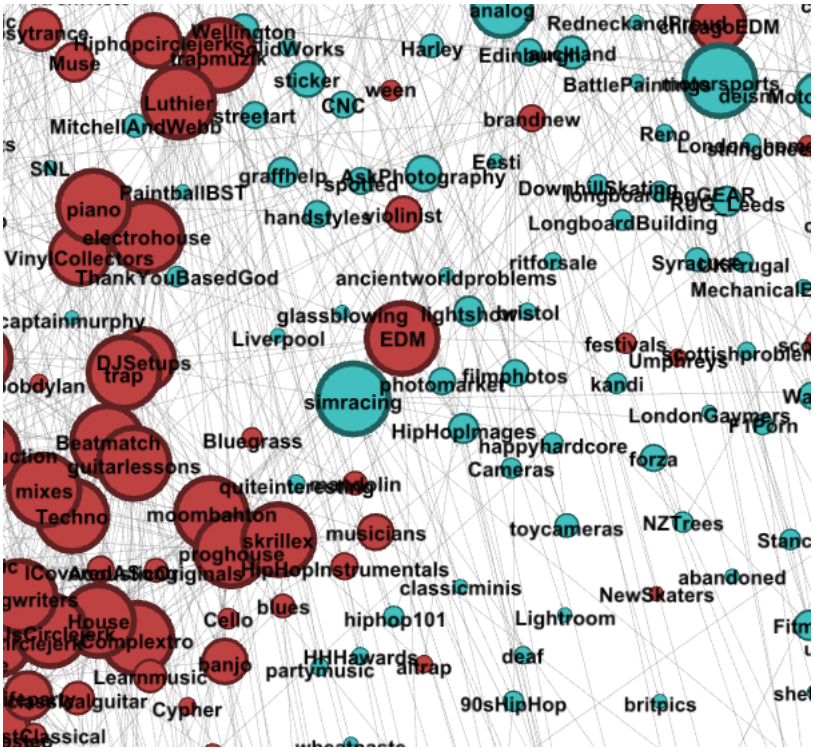
topic	%primary	&true
LGBT	1.34	3.42
news	4.71	37.2
politics	1.55	29.86
history	0.50	11.17
religion	0.1	2.79
comedy	6.2	27.7
animals	2.91	16.2
nature	2.43	19.41

I also performed some eyeball analysis on the results, and saw that the category diffusion gave overall a more satisfying representation of the subreddit network than the modularity algorithm, though the graph is not as clean-cut. In the following snapshots, I show how the diffusion algorithm can pick out correct subreddit categories in dense communities:

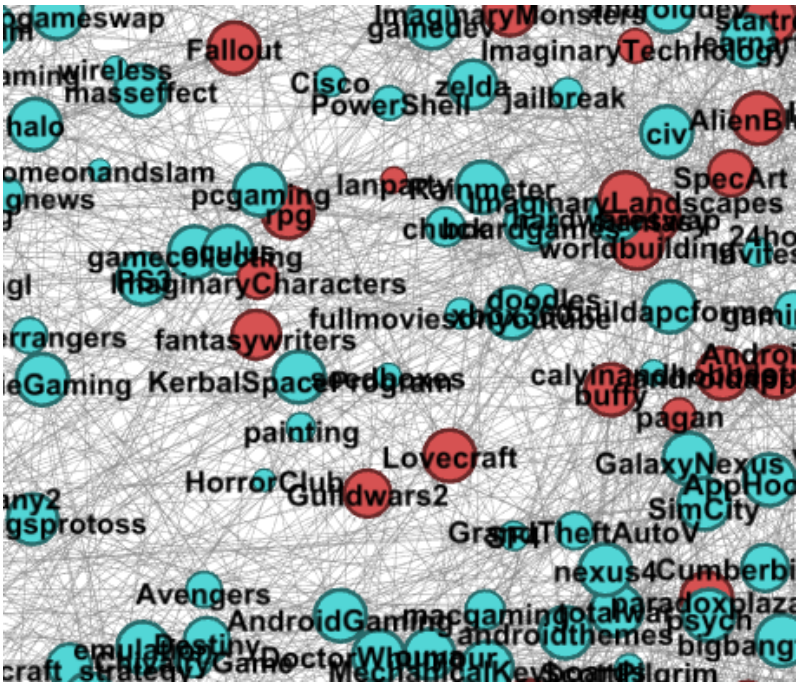
Religion:



Music:



Books/literature:



5. *Problems, Improvements & Future Research*

   I would like to work with a better seed set. The one I started with selected subreddits randomly from the list, and I would like to pick optimal ones with large degree centrality. I also am a flawed judge of subreddit topics, as there are many I don't know anything about.

   I would like to improve the categories. Well into the creation of the seed set, I realized that I had left out several common topics, like pictures, videos, discussion/online community, philosophy, etc. If I had another crack at it I'd do a better job, but it took several hours to make the initial seed set and I'm not keen to try again anytime soon.

   I would like to optimize the diffusion parameters. I made up the existing ones arbitrarily, and I bet there are optimal ones that would give the best results. In addition, I would like to improve the diffusion algorithm to give cleaner results.

   If possible, a dictionary/semantic similarity algorithm could automate the categorization process for the seed set, and would allow for the algorithm to be extended to other applications, but such an algorithm is difficult to design and beyond the scope of this project.

6. *Conclusion*

   I am very pleased with the results of my project. While flawed, my algorithm accurately describes the categories of most subreddits starting from a small seed set. I am very excited to work on this process in the future and see what can be made out of it.

7. *References*

   Olson, Randal S. & Neal, Zachary P. "Navigating the massive world of reddit: Using backbone networks to map user interests in social media." December 12, 2013