Characterizing and Detecting Quarantined Subreddits

Neel Bedekar, Nishtha Bhatia, Joan Chen

Github Repository

1. Introduction

The advent of the web gave birth to strong, online communities. Anonymity and the free speech movement enabled open discussion and communication among community members online, but it also led to content and communities that occupied hateful and toxic stances. In an effort to combat the empowerment of widespread toxicity, harm, and violence by such online interactions, measures were taken to both regulate and respond to them.[1]

One such regulation instituted by Steve Huffman, the CEO of Reddit, involved a quarantine system, under which technically allowable, yet *generally offensive* subreddits, would only be viewable through explicit opt-in and would be hidden from searches or recommendations.[2] The system dramatically reduced the audience of these subreddits while still allowing access to the subreddit for those forming the community responsible for the content. We propose that the nature, characteristics, and interactions of a community strongly contribute to the eventual quarantine of an entire subreddit. We expect that the investigation of these communities within the Reddit social network will enable us to glean insights regarding how user interactions within their community are able to provoke, sustain, or even exacerbate offensive and toxic behavior.

In the remaining sections of this proposal, we review three research papers addressing various concepts salient to our area of research. We discuss how they relate to our topic, and use them as a starting point to develop the specific research question we hope to explore. Finally, we aim to address and answer this question, through our analyses of various quarantined and non-quarantined subreddits.

2. Related Work

A significant amount of research has been previously conducted on different online communities. Fast and Horvitz[3] discovered that controversial Reddit communities with diverse opinions have a greater likelihood of hosting negative dogmatic language. Their research allowed them to determine not only which conversation topics are most likely to give birth to dogmatic comments, but also how dogmatic users were able to shape the nature of a conversation. Building upon their work, we aim to additionally examine the relationships between comments and varying levels of dogmatism, or communities and the number of dogmatic comments they contain.

Ganley and Lampe[4] investigate the effect of network configuration on social capital by examining the social news website, Slashdot. Similar analyses can be applied to other sites, such as Reddit. To build upon their research, we aim to look into whether the core group of high Karma users is concentrated in a handful of large subcommunities or in many smaller subcommunities.

Hamilton et. al.[5] formalize a measure of loyalty as a Reddit "user-community" relation and find edge density and activity assortativity differences in loyal and unloyal networks. Ultimately, they analyze the behaviors that loyal and unloyal users display to predict future user loyalty, finding several features that are strongly predictive of loyalty, such as comment language, post language, and post score that users interact with. A key criticism is that the research does not account for per-community differences that may lead to differing "loyalty" scores. When we extend this work, we might experiment with different normalization techniques that quantify post activity as a standardized per-user metric, without biasing for frequency of posting.

3. Methods

3.1 Dataset

In this project, we utilized Reddit comments data available from pushshift. Specifically, we used psaw, which is a python library that wraps pushshift.io, an aggregation for reddit comments and submissions data. We selected 53 non-quarantined and 13 quarantined subreddits from the top 100 subreddits. We considered the past 100,000 comments for each subreddit in our analysis, as opposed to performing a time-frame based analysis, which might be biased by community size. With this dataset, we simplify our computation, allowing us to focus on analysis instead of sharding data or setting up distributed system. It should be noted that there is a very small number of quarantined subreddits, and that we have chosen to analyze data from all quarantined subreddits that have substantial activity.

We use the following comment fields from the data returned by pushshift:: author, body, created_utc, id, link_id, parent_id, replies, subreddit.

3.2 Basic Interaction and Negative Sentiment Graphs

Before running any experiments or analysis, we sought to characterize the interactions in our subsets of subreddits. Using the dataset described above, we constructed two distinct networks for each quarantined and non-quarantined subreddit.

The first network represented basic interactions, and was built as follows: for each snapshot (past 100,000 comments) under each subreddit, we constructed an interaction graph where the nodes are users that have commented in this timeframe, and edges exist between two nodes A and B if user A and B "interacted" with each other in the timeframe. We define this "interaction" to be that one of the users commented on either a submission or a comment of the other user.

The second network utilized TextBlob, a Python library that uses NLP to process textual data, in order to model negative interactions in the network based on sentiment analysis. In this network, nodes represent users who have participated in a negative "interaction," with the same definition of interaction as above. However, unlike the basic interaction graph, this network only places an edge between users if the comments they have exchanged were negative in sentiment.

In total, we constructed 132 graphs.

3.3 Network Analyses

In order to determine which features were characteristic of quarantined and non-quarantined subreddits, we attempted to examine several network characteristics for the basic interaction and negative interaction graphs. Specifically, we looked at number of nodes, number of edges, average clustering coefficient, average degree, standard deviation of degree, average neighbor degree, average pagerank score, and number of connected components. In order to normalize our values, we divided the average degree, neighbor degree and standard deviation of degree by the total number of nodes in the graph. We also examined the proportion of nodes and edges in the negative interaction graph to the nodes and edges in the basic interaction graph to determine how much negativity exists in a particular subreddit.

After computing these statistics for all the graphs we had constructed, we sought to analyze and classify statistically significant differences between the quarantined and non-quarantined subreddits. To do so, we compared the average value for each network characteristic, for each type of subreddit.

Our hypothesis was that the network structure of quarantined graphs would prove to be significantly different from that of non-quarantined graphs. Specifically, we predict greater interaction with negative sentiment, as well as the presence of fewer, larger communities as opposed to many, smaller communities. This hypothesis stems from social research that motivated this project.

3.4 Classification Model

To understand whether network characteristics and structures would be able to accurately determine and/or predict quarantined subreddits, we chose to develop a machine learning classification model based on logistic regression. To correctly assess which features to use in our model, we conducted feature analysis by modeling the relationship between a particular network feature and the network's quarantine status.

We understand there are several limitations to constructing a logistic regression model with only 66 data points. Unfortunately, the nature of our data and experiment restricts us from expanding this sample size, since the set of all quarantined subreddits with activity is extremely limited.

4. Results and Findings

4.1 Statistical Analysis

Our aim in this project was to determine what network properties are characteristic of subreddit communities that have been quarantined. In order to achieve this, we sought to first represent our 66 subreddits as basic interaction and negative interaction networks. We then conducted network analyses on the subreddits, attempting to characterize them by their properties.

Following this analysis, we calculated the average value over all network characteristics, over both types of interaction graphs -- basic and negative sentiment -- for quarantined and non-quarantined subreddits. We chose to normalize the network characteristics that were dependent on network size by dividing their values by $2m$. In this way, we hoped to account for different population sizes. The characteristics that are normalized in this way have a star next to their name.

The data we derived is captured in the tables below.

Basic Interaction Graphs

Subreddit Type	Non-quarantined	Quarantined
Average Number of Nodes	43526.83018867926	8917.538461538461
Average Number of Edges	71816.83018867923	38046.53846153846
Average Clustering Coefficient*	3.4805202655718355E-7	3.91510432620651E-5
Average Pagerank	2.4018867924528298E-5	6.784615384615385E-4
Number of Connected Components*	0.0018674656500381392	8.020806346106047E-4
Average Degree Centrality*	5.98940015757457E-10	1.873602731099544E-6
Average Neighbor Degree*	0.004341491606055393	0.01676750179550442
Average Degree*	2.3944624207179045E-5	6.783962018730194E-4
Standard Deviation of Degree*	2.101970036645883E-4	0.002243713580331534

We analyzed a few more statistics when analyzing characteristics of the negative interaction graphs. Namely, we calculated the proportion of nodes and edges in the negative interaction graph to the number of nodes and edges in the basic interaction graph. As before, we normalize certain characteristics that are starred, to account for differences in network size.

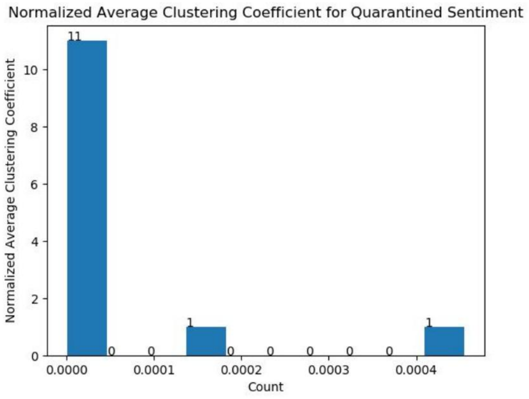
Negative Interaction Graphs

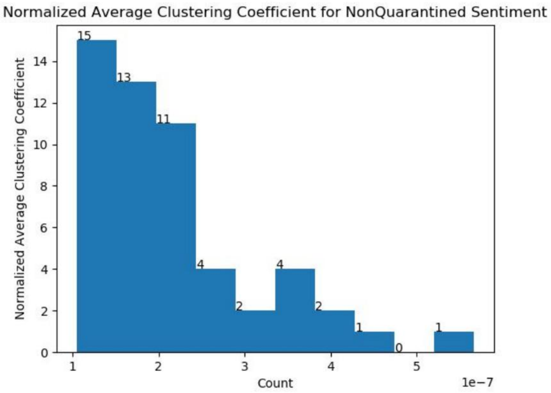
Subreddit	Non-quarantined	Quarantined
Proportion of Negative Nodes	0.37901340878597745	0.5627616907553795
Proportion of Negative Edges	0.2333284621907044	0.31094156861324357
Average Number of Nodes	16214.735849056602	5045.307692307692
Average Number of Edges	16729.716981132085	11441.307692307693
Average Clustering Coefficient*	2.1890131146126312E-7	5.0661806354324466E-5
Average Pagerank Score	6.452830188679245E-5	0.0013486923076923077
Number of Connected Components*	0.05419080060490571	0.009645730207768436

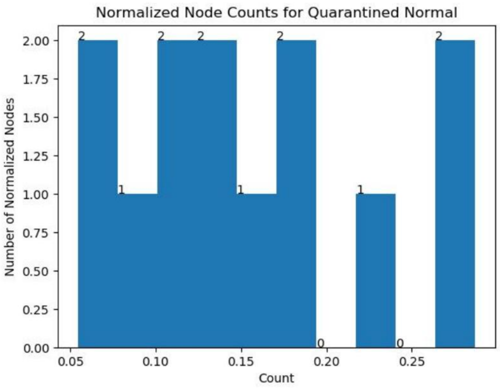
Average Degree Centrality*	4.427358103646631E-9	8.485641420766565E-6
Average Neighbor Degree*	0.0029716913518318335	0.019474406336036552
Average Degree*	6.459082390609511E-5	0.0013487204540186097
Standard Deviation of Degree*	2.8810267035251784E-4	0.0035420507208297008

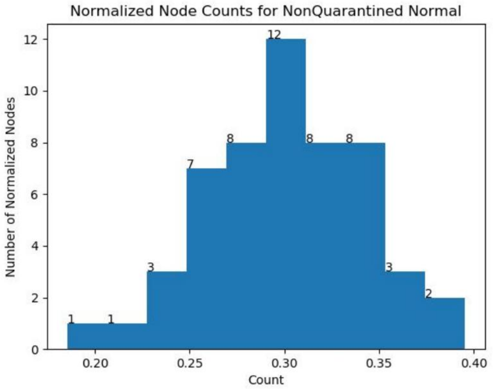
Comparing the network statistics for each type of graph to one another, we found that the negative interaction graphs yielded significantly more differences than the basic interaction graphs. Upon further analysis, we concluded that there exists some relationship between quarantined graphs and negative interactions.

We also examined differences in variability for a given network characteristic. We found particularly salient differences for normalized clustering coefficients of the negative interaction networks and normalized population counts for both basic and negative interaction networks. The histograms capturing these differences are below:


Normalized Average Clustering Coefficient for Quarantined Sentiment


Normalized Average Clustering Coefficient for NonQuarantined Sentiment


Normalized Node Counts for Quarantined Normal


Normalized Node Counts for NonQuarantined Normal

Our results support the hypothesis that differences exist between the communities of quarantined and non-quarantined subreddits. Namely, our analysis allows us to glean insights regarding the presence of:

1. **Greater Negative Interaction in Quarantined Communities:** The percentage of users that engaged in *some* negative interaction was significantly higher in quarantined communities than in non-quarantined communities: 56% and 38%, respectively.

2. **Groups Clustering Together:** On average, nodes in the quarantined subreddits had 100x as high of a clustering coefficient than nodes in the non-quarantined subreddit, for the basic interaction networks. For the negative interaction network, the difference in clustering coefficient jumped to 200x as high. This difference indicates a greater likelihood of groups clustering together in the quarantined subreddit, especially when negative sentiment comments are involved.

3. **Interactions With Other Users:** The average degree in quarantined subreddits was 28x greater than that of that for non-quarantined subreddits for the basic interaction network, and about 20x greater for the negative interaction network. This means that community members in quarantined subreddits were more likely to interact with one another by commenting on each other's posts, even when only looking at negative interactions.

4. **Number of Connected Components:** The negative interaction and basic interaction networks for quarantined subreddits contained 5x fewer connected components than that of the non-quarantined subreddits.

This characteristic analysis seems to suggest that subreddits in danger of being quarantined consist of fewer communities than subreddits that are not quarantined, and yet have greater interaction within those communities. The prevalence of fewer and more tightly knit communities may contribute to the toxicity that eventually propels Reddit to quarantine a particular community. In general, there is a greater likelihood that any two nodes have interacted with one another in the quarantined subreddits than there is in the non-quarantined subreddits.

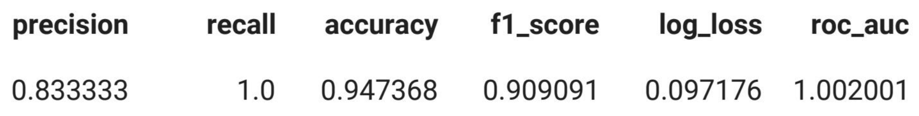
4.2 Logistic Regression

By nature, our focus on quarantined and non-quarantined subreddits dramatically reduces our number of data points, as there are only a handful of quarantined subreddits that are active and can be represented as interaction networks. That being said, we wanted to investigate what would happen if we created a classifier using logistic regression to characterize subreddits, so we used our limited subset to do just that.

In order to determine which features to train and evaluate our model with, we analyzed the characteristics from 4.1. We found the most significant features to be a mixture of individual network characteristics, as well as ratios that combined information about both the basic interaction networks and the negative interaction networks. Ultimately, we chose to focus on the ratio of nodes in the negative interaction network to nodes in the normal interaction network, the ratio of edges in the negative interaction network to edges in the normal interaction network, the normalized average degree and normalized average neighbor degree of the normal interaction network, and the ratio of normalized average clustering coefficient of the negative interaction network to the normalized average clustering coefficient of the basic interaction network.

We discovered that our accuracy and precision were trivially high, due to class imbalance. This is because our majority class was non-quarantined subreddits, and a classifier that simply assigns the majority class is bound to be highly accurate.

We have provided our evaluation statistics below:



precision	recall	accuracy	f1_score	log_loss	roc_auc
0.833333	1.0	0.947368	0.909091	0.097176	1.002001

As can be seen, our model performs extremely well with only a few data points, but this reveals less about our model than we'd like, thanks to class imbalance. If we were to repeat this work, we would need to ensure we have a sufficiently large sample size, as well as have an equivalent number of quarantined and non-quarantined subreddits.

5. Conclusions

Overall, we have found that our results support the finding that there exists an inherent difference in the network structure of quarantined and non-quarantined subreddits. We have found that a combination of both normal and negative interaction activity are able to characterize these differences. Specifically, quarantined subreddits are more heavily skewed towards containing negative interactions, and generally show greater clustering in their communities than non-quarantined subreddits.

6. Limitations and Further Work

A significant limitation in our research is the imbalance between the number of quarantined subreddits and the number of non-quarantined subreddits we looked at. Quarantined subreddits are limited in number, and active quarantined subreddits with enough user activity to create a significant interaction graph are even more limited. As a result, we were only able to sample 13 quarantined subreddits whereas we sampled 53 non-quarantined subreddits. To address this issue, we would ideally be able to find additional active quarantined subreddits we may have missed.

This uneven proportion of quarantined and non-quarantined subreddits also affects our logistic regression findings, as mentioned previously. A classifier that simply assigns the majority class will end up being highly accurate. To have meaningful logistic regression results, we would want to increase our sample size and ensure that we have an equal number of quarantined and non-quarantined subreddits.

In future work, we could consider including additional network and non-network features in our analyses and logistic regression. Such features could include the subreddit name, comment polarity, and proportion of negative comments made by the average subreddit user. Additional analyses could include identifying pairs who engage in retaliation with each other.

References

1. Lagorio-Chafkin, Christine. “How Charlottesville Forced Reddit to Clean up Its Act.” *The Guardian*, Guardian News and Media, 23 Sept. 2018

2. Auerbach, David. “How Reddit Can Solve Its Hate Speech Problem-Without Banning Hate Speech.” *Slate Magazine*, 14 July 2015.

3. Fast, Ethan, and Eric Horvitz. "Identifying Dogmatism in Social Media: Signals and Models." arXiv preprint arXiv:1609.00425 (2016).

4. Ganley and Lampe, 2009. The ties that bind: Social network principles in online communities.

5. W. L. Hamilton, J. Zhang, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec. 2017. Loyalty in Online Communities. ArXiv e-prints (March 2017). arXiv:1703.03386

6. Boe, Bryce. "PRAW: The Python Reddit API Wrapper." PRAW, 2018. Web. 8.