Characterizing Banned Subreddits by Network Attributes

Bryce Cai
Stanford University
Stanford, CA 94305
bcai@stanford.edu

Sean Decker
Stanford University
Stanford, CA 94305
skdecker@stanford.edu

Crystal Zheng
Stanford University
Stanford, CA 94305
czheng11@stanford.edu

Abstract

As "the front page of the internet," Reddit in recent years has started to crack down on "toxic" behaviors, including hate speech, bullying, etc. The primary method by which Reddit is combating this is by the banning of subreddits whose users exhibit these behaviors. However, differentiating subreddits that should be banned from others is a hard problem, and with new subreddits often immediately replacing banned subreddits, there is a need to make this process of identifying subreddits that should be banned more automated. In this paper, we propose using not only subreddit features for classifying if it is to be banned, but also the network structure of Reddit. We compare a simple logistic regression model of classification with a local classifier and discuss the useful network characteristics of the Reddit cross-link network for this classification problem.

1 Introduction

In 2012, then-Reddit CEO Yishan Wong famously said, “We stand for free speech. This means we are not going to ban distasteful subreddits. We will not ban legal content even if we find it odious or if we personally condemn it." Since then, however, Reddit.com has grown into the “front page of the internet.” As of 2018, Reddit had approximately 330 million monthly active users and more than 138,000 active subreddits, making it one of the most popular online social platforms in the world (1). Previous work has largely focused on characterizing inter-community interactions, individual user behaviors, and conflicts of Reddit as a whole or individual subreddits. However, investigating banned vs. non-banned subreddits in terms of network characteristics has not been previously researched. With the expansion of Reddit’s reach and power has come renewed controversy around how Reddit moderates content. In 2019, to fight against hate speech, bullying, and many other behaviors that are seen as “toxic”, Reddit’s main tool is now banning subreddits. Given the nature of Reddit, oftentimes it is difficult to identify which subreddits may or may not be banned currently or in the future.

Thus, we were motivated to investigate the network characteristics of banned subreddits compared to non-banned subreddits and whether we can utilize network characteristics to help classify banned vs. non-banned subreddits. Overall, by using the ability to identify subreddits similar to previously banned ones and then further being able to categorize and analyze these banned subreddits, Reddit could take a more proactive approach to moderating the content on its site. Although there were other papers that analyze Reddit bans and their effect on the website (2), we took a novel approach by treating this problem as a graph problem.

2 Related Work

Past research has been conducted to try to categorize and analyze Reddit networks from different perspectives, utilizing various techniques.

Preprint. Under review.

2.1 Community Interaction and Conflict on the Web

Kumar et al. created a model that accurately identified and predicted mobilizations by examining common motifs in networks of comments and using a novel adaptation of PageRank to categorize successful defensive responses to mobilizations (3). However, Kumar et al.'s techniques heavily relied on cross-links to identify intra-community conflict, which neglected communities that did not form cross-links. Kumar's work informed us that certain characteristics of intra-community interactions could help in identifying conflict and other characteristics that could be useful in identifying banned subreddits.

2.2 Temporal Analysis of Reddit Networks via Role Embeddings

Grayson and Green investigated the change in user roles in Reddit subreddits over time using graph embedding methods but were limited in their evaluation, finding difficulty in generalizing their findings on the differences between loyal and vagrant users to the entirety of Reddit (4). Furthermore, they used visual examination of PCA for the nature of community and user roles, which could also be evaluated in more depth and on additional subreddits such as AskReddits, Debate Reddits, Questions Reddits (where roles were potentially more distinguished to allow for further comparisons). Grayson and Green's experience in having limited generalizability highlighted the fact that higher-level analysis on multiple subreddits or larger graph network was necessary to identify certain differences between community and user behavior/roles.

2.3 Characterizing Conversation Patterns on Reddit

Choi et al. analyzed the structures of conversations on Reddit by exploring the tree-like structures of comment chains. Their findings illustrate the utility of a few metrics on general comment structure on Reddit on a wide array of analysis (5). However, Choi et al. seemed to extrapolate findings on vast amounts of topics from a mere three-metric analysis that was inherently flawed because of confounding causes leading to similar comment tree structures. Choi et al.'s research highlighted methods to extract structure from Reddit conversations.

2.4 Cross-Community Influence in Discussion Fora

Belak et al. sought to identify groups that were vulnerable to influence by analyzing cross-community level influence of a very large system, investigating whether the levels of impact between communities differ and how these levels evolve over time (6). They developed a novel framework for cross-community impact analysis based on structural features derived from a dynamic repleto graph, which was useful on data that has little external info.

3 Methodology

3.1 Dataset

For this project, we used raw Reddit data obtained from files.pushshift.io, which contained a publicly available archive of Reddit submissions with the entirety of comments, moderators, subreddits and accounts that was updated monthly (8). In order to constrain the scope and utilize a reasonable amount of accessible compute resources, we limited the time frame to the range of May 23, 2018, 12:19:21 UTC to June 26, 2018, 19:57:21 UTC. Out of 118,886 subreddits contained in this dataset (i.e., subreddits with at least one submission within the timeframe present in this dataset), 143 were banned or eventually banned.

To interact with this dataset, we used Google Cloud Platform's BigQuery, which allowed us to apply SQL queries on the dataset without needs to load it locally (10). We queried the following tables: pushshift.rt_reddit.comments, pushshift.rt_reddit.submissions. While providing ease of data access, BigQuery's window into the PushShift dataset only contained data from May to August 2018; hence, the aforementioned limit on the time range was applied both to constrain the dataset's size and to normalize the two queried tables' sampled ranges.

Taking the Reddit comment data as an example, this data was stored as an SQL table that included the following fields:

Field	Datatype
author	STRING
body	STRING
created_utc	TIMESTAMP
subreddit	STRING
subreddit_id	INTEGER

Figure 1: Notable fields of the BigQuery's "pushshift.rt_reddit.comments" table.

In order to narrow our scope and conduct deeper graph network analysis, we utilized SNAP's Social Network: Reddit Hyperlink Network, which was comprised of publicly extracted available Reddit data between January 2014 and April 2017 (9). The subreddit hyperlink network was directed, signed, temporal, and attributed. The dataset contained 55,863 subreddits (nodes), 858,490 edges (hyperlinks between subreddits), edge weights of either -1 or +1, and text property vectors. We primarily used data found in soc-redditHyperlinks-title.tsv, which was a network of subreddit-to-subreddit hyperlinks extracted from hyperlinks in the title of the post.

We also compiled a dataset containing the names of banned subreddits from a dedicated subreddit post (11), which listed the names of 1932 banned subreddits. We also added to the banned subreddits dataset by manually checking likely banned subreddits determined from filtering subreddit names of the SNAP dataset on whether they contained a swear word (12) and inputting the associated url to reddit to see whether they were banned or not. We searched through 600 subreddits in this manner and found an additional 25. Ultimately, we aggregated a list of 1957 banned subreddits.

3.2 Classification

For our baseline, we used a supervised learning model, Logistic Regression, in order to classify whether or not a subreddit is banned trained on features extracted from the network. Features were extracted from the PushShift BigQuery dataset. Using Standard SQL in BigQuery, for each subreddit, we extracted the following features:

- total number of subscribers
- total number of submissions
- total number of unique submission authors
- average count of submissions per unique author
- fraction of posts marked not safe for work
- total number of comments
- total number of unique comment authors
- average count of comments per unique author
- average count of comments per submission
- average sentiment score (unweighted)
- average sentiment magnitude
- average sentiment score, weighted by magnitude

Sentiment analysis was conducted using Google Cloud's Natural Language API, which calculated a sentiment score (ranging from -1 to 1) and magnitude (a non-negative value) for a body of input text, indicating positivity/negativity and emotional weight of the input; only one significant figure was provided for both scores. To ease computation time while maintaining a large sample size, at most 250 comments were sampled for sentiment analysis for each subreddit.

If no comments were found for a subreddit, all comment-related feature values were set to 0.

Upon incorporation of graph analysis, the following features were appended to the feature vector for the subreddits with data available in both the PushShift and SNAP datasets:

- clustering coefficient of subreddit node

- in-degree of subreddit node
- out-degree of subreddit node

3.3 Cosine Similarity

In order to calculate cosine similarity to measure how similar 2 nodes were according to their feature vectors x and y, we used the following equation:

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\|_2 \cdot \|y\|_2} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$$

We chose 3 basic local features: the degree of target node, the number of edges in the egonet of the target node, and the number of edges that connect the egonet of the target node and the rest of the graph. We then recursively generated more features using mean and sum. For 2 iterations, at each iteration, we concatenated the mean and sum of all of the node's neighbors' feature vectors with the feature vector of the target node.

$$\bar{V}_u^{(1)} = \left[\bar{V}_u; \frac{1}{|N(u)|} \sum_{v \in N(u)} \tilde{V}_v; \sum_{v \in N(u)} \tilde{V}_v \right] \in \mathbb{R}^9$$

where $N(u)$ is the set of u's neighbors in the graph and $\bar{V}_u^{(1)}$ is the updated feature vector for node u.

4 Experiments Results/ Discussion

4.1 Logistic Regression

As a baseline model for predicting if a subreddit would be banned in the near future, we used a simple logistic regression model which featurized each subreddit on the basis of the features listed in the Features subsection above.

Feature vectors for each of the 143 banned subreddits in the PushShift databases were collected as well as for 500 normal (non-banned) subreddits sampled randomly from the 118,743 remaining subreddits in the databases. To aid regression, features were normalized (to zero mean and unit variance) for the entire dataset along each feature.

Logistic regression was conducted using the defaults specified in the scikit-learn machine learning Python package (ℓ_2 norm penalization, stopping criteria tolerance of $1 \cdot 10^{-4}$, 1.0 regularization strength, L-BFGS solver, maximum 250 iterations) and trained on 70% of the dataset, randomly selected while maintaining the proportion between banned and normal subreddits in the dataset. The remaining 30% was reserved for testing the logistic regression.

On average, logistic regression produced a model that was 76% accurate. As indicated by the confusion matrix in Figure 4.1, the model had significant (albeit far from perfect) accuracy, suggesting that features in comment text and frequency within a subreddit could indicate whether the subreddit would be banned, even within the small sample of comments available. The likelihood that a subreddit labeled as a banned subreddit by logistic regression was actually banned was far higher than the proportion of banned subreddits within the sample, suggesting that simple analysis of comment content and frequency (via logistic regression) could identify candidate subreddits likely to be banned.

However, far more false negatives were found than false positives; in fact, the number of false negatives tended to outnumber the number of false positives. This was likely because many subreddits in the dataset were small and/or new subreddits that did not contain many subscribers and, in many cases, did not contain any comments, leading to a feature vector containing many zeros. While many non-banned subreddits fell under this category, a significant portion of non-banned subreddits did as well. As seen by the examples of the "rstatistics" and "lifeafterhate" subreddits in Figure 4.1, such subreddits had similar sparse feature vectors and were subsequently likely to be categorized under the same label by logistic regression. The vast majority of such subreddits were not banned, so the significant fraction of banned subreddits with this small, comment-less structure were subsequently labeled incorrectly. This suggested that in order to correctly identify all subreddits likely to be banned, further analysis would be needed, such as an analysis of the users active in that subreddit (and whether they frequented banned subreddits).

Type	Predicted normal	Predicted banned
Normal	144	6
Banned	29	13

Figure 2: Confusion matrix generated by the results of simple logistic regression.

	Subreddit	
Feature	rstatistics (non-banned)	lifeafterhate (banned)
Subscriber count	1	11
Submission count	1	1
Average submissions per author	1	1
Fraction of submissions marked NSFW	0	0
Unique submission author count	1	1
Comment count	0	0
Unique comment author count	0	0
Average comments per author	0	0
Average comments per submission	0	0
Average sentiment score	0	0
Average sentiment magnitude	0	0
Average sentiment, weighted by magnitude	0	0

Figure 3: Feature vectors (non-normalized) for examples of small non-banned and banned subreddits.

4.2 Adding Graph Features

To improve upon this simple logistic regression classifier, we proposed abstracting Reddit as a graph, in order to exploit network attributes for the featurization of subreddits.

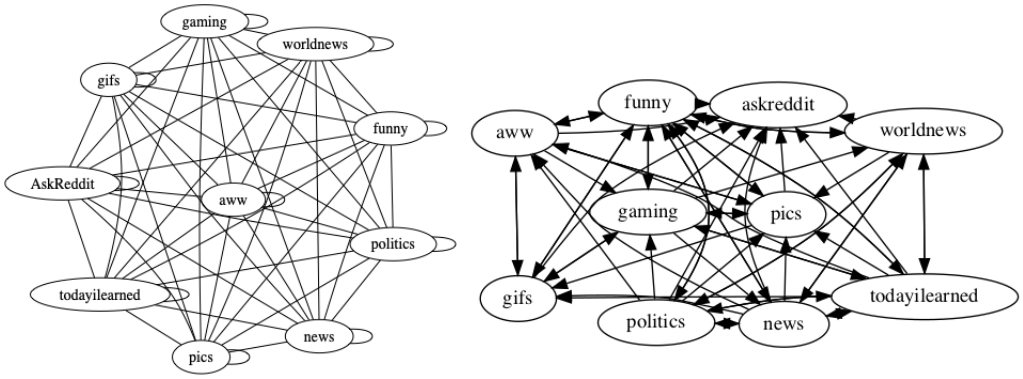

Figure 4: Left: Graph of 10 largest subreddits, with relations defined as the number of shared commenters between the subreddits. This graph was built using Google Big Query queries and visualized using SNAP.py. Right: Graph of 10 largest subreddits with relations defined by crossposts. As shown by using crossposts as relations, this subgraph was not fully connected.

Reddit can be abstracted as a network in many ways; in this project, we first tried to abstract Reddit with nodes as subreddits and undirected edges between nodes as weighted by the number of users that commented in both subreddits. To build these graphs, we used SQL queries to Bigquery (10). In working with this graph, however, we found it to be unwieldy and not terribly descriptive because it was too well connected. As seen from Figure 4, abstracting Reddit in this way led to many fully connected components, which were not as descriptive for relationship finding as maybe a more sparse graph.

Instead of defining relationships as the number of shared commenters, we then defined relations by crossposts. Specifically, we defined directed edges from a source subreddit to a destination subreddit as indicating that a post in the source subreddit links to the destination subreddit. We built these graphs using the SNAP Social Network: Reddit Hyperlink Network dataset (9). Defining relationships like this decreased the number of fully connected components Figure 4 and gave us a richer graph structure to work off of.

4.3 Initial Graph Exploration

Before using this graph structure to enhance our classifier, we explored it in order to gain some insights into how to best take advantage of it. Some simple statistics on the Reddit Hyperlink Network:

- The number of nodes in the network: 55,863
- The number of directed edges in the network: 858,490
- Clustering Coefficient: 0.0174
- The number of triads: 1329278

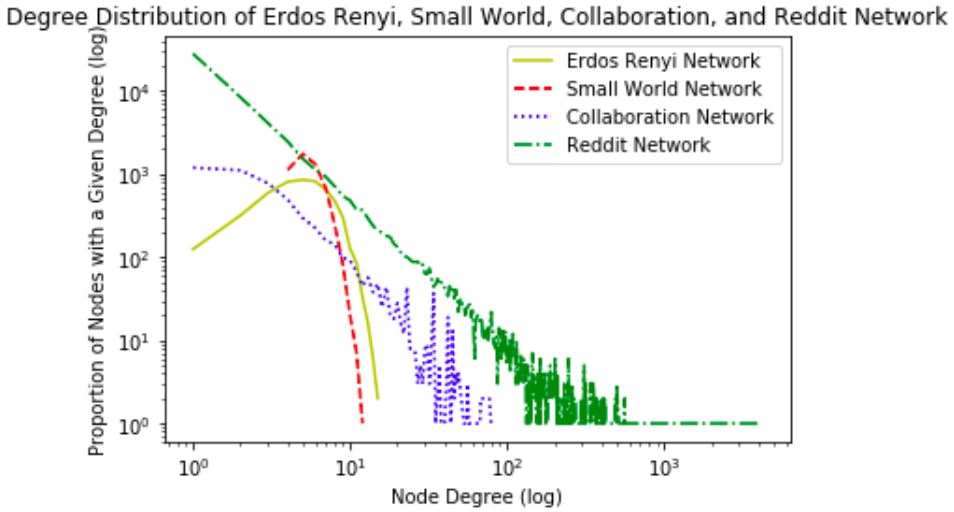

Figure 5: Graph of the degree distribution of the Reddit Network against an Erdos Renyi, Small World, and Collaboration network. It can be seen that the Reddit network most closely resembles the Collaboration network.

As a first step, we plotted the degree distribution of the Reddit network against an Erdos Renyi, a Small World, and a research paper Collaboration Network (9) to compare the distribution, as show in Figure 5. From these graph comparisons, we can see that the Reddit network is most similar to the Collaboration Network, and thus this indicates that it is a scale-free network, which follows a power law distribution. This seems to indicate that the Reddit network is distributed similar to other social networks, however, note that it also has a relatively small clustering coefficient of around 0.0174. Thus is much less than the clustering coefficient of the Collaboration Network: 0.529636.

4.4 Cosine Similarity

After analyzing this Reddit network at a high level, we moved on to analyzing the nodes themselves. Specifically we moved on to investigate the cosine similarity of various banned and not-banned subreddits. In order to further elucidate cosine similarity in the context of this network structure, we narrowed our focus onto the banned subreddit r/incel. From Figure 6, we see that using only basic features, we do not find subreddits that seem related to a banned subreddit, e.g. r/love, r/dance, but using recursive feature generation we find subreddits that tend to seem more closely related to a banned subreddit, e.g. r/fffffffuuuuuuuuuuu, r/darknetmarkets. Note even that r/darknetmarkets is a banned subreddit itself.

With Basic Features			
Rank	Subreddit	Cosine Similarity	Banned (Yes or No)
1	ghostbc	0.9999995255441468	No
2	love	0.9999995057275693	No
3	keepournetfree	0.9999992914319293	No
4	dance	0.9999988896950565	No
5	clubesteban	0.9999982373294889	No
6	lacrosse	0.999998070311967	No
7	insurance	0.9999980524679625	No
8	credibledefense	0.9999977672462803	No
9	wichita	0.9999959089855327	No
10	southbend	0.9999953872218518	No

With Recursive Features			
Rank	Subreddit	Cosine Similarity	Banned (Yes or No)
1	darksouls	0.9997389975728118	No
2	scotch	0.9997125608448031	No
3	globaloffensivetrade	0.9986806305859485	No
4	chivalrygame	0.9980109252969892	No
5	assistance	0.9971550521050944	No
6	skyrimmods	0.9967314748530173	No
7	darknetmarkets	0.9964563197962923	Yes
8	ffffffffuuuuuuuuuuu	0.9959668387371923	No
9	moviesinthemaking	0.9957072787214674	No
10	malehairadvice	0.9939993847591253	No

Figure 6: Top: The subreddits most similar to r/incel using only the basic feature, degree of node, number of edges in egonet, number of edges that connect egonet to the rest of the graph. Bottom: The subreddits most similar to r/incel using 3 iterations of recursive feature generation, starting eache feature vector as being just a subreddit's basic features. Note that the most cosine similar subreddits found using recursive features seem to be more similar to what would be in a banned subreddit, and we even find another banned subreddit: darknetmarkets.

This leads us to believe that recursive feature generation could indeed be useful to help classify subreddits. On the other hand, we see from Figure 7 that it seems that a majority of subreddits were cosine similar to r/incels, even after recursive feature generation. Thus, banned subreddits may still be difficult to distinguish from other subreddits even with the incorporation of calculated cosine similarity from recursive feature generation.

4.5 Local Classifier

After investigating the structure of the Reddit Network, we moved on to adding graph features into our logistic regression in order to make it into a Local Classifier. For each node present in the graph with a generated feature vector (i.e., from the list of banned subreddits and sampled non-banned subreddits present in the PushShift database), the clustering coefficient, in-degree, and out-degree of the node were added to its feature vector. As this analysis could only be done for subreddits with data available in both the SNAP and PushShift datasets, filters were applied to limit the scope only to these subreddits; this limited the dataset to encompass only 44 banned and 109 sampled non-banned (normal) subreddits.

Re-applying logistic regression to these new results, however, revealed no significant changes. As before, the new classifier obtained 76% accuracy, and as shown by the new confusion matrix in Figure 8, the ratio of correct predictions, false negatives, and false positives remained the same as before. In fact, limiting the feature vectors to just the new features resulted in a classifier produced by logistic

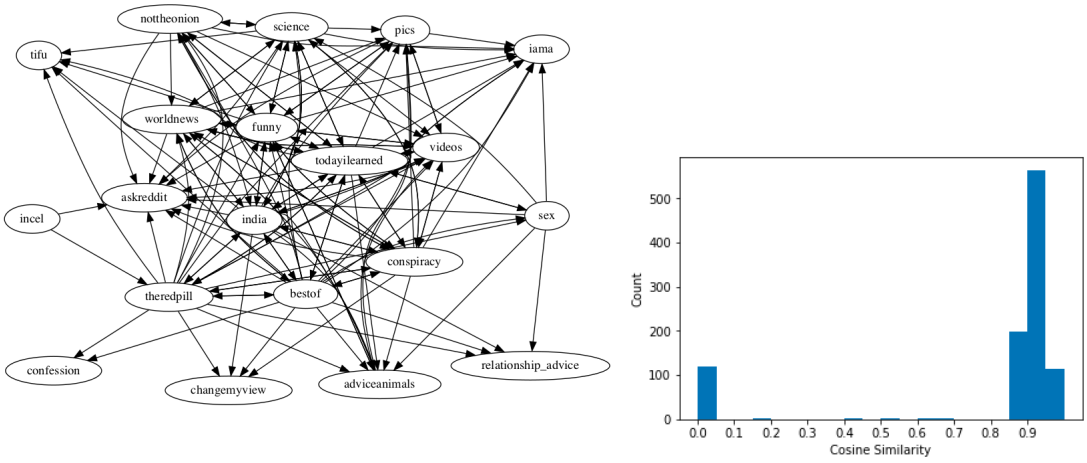

Figure 7: Left: 20 node neighborhood of r/incel, a banned subreddit. Right: Histogram of cosine similarity between a random selection of 1,000 nodes from the Reddit Network.

regression that classified all vectors as non-banned, indicating that graph structure likely had little to no bearing on whether a subreddit was banned or not.

However, as the working sample of subreddits was relatively small in size as a result of being constrained to subreddits only available in both the PushShift and SNAP datasets (both of which were fundamentally limited in scope themselves), this result may have been the result of a problematic working dataset.

Type	Predicted normal	Predicted banned
Normal	31	2
Banned	9	4

Figure 8: Confusion matrix generated by the results of logistic regression with graph features.

5 Conclusion

Preliminary results from higher-level network analysis resulting from cosine similarity revealed that cosine similarity, particularly with recursive features, or other network characteristics could potentially better capture the nature of banned subreddits and their relations with other subreddits. However, it was interesting to note that even with recursive feature generation, a majority of subreddits remained cosine similar to the representative banned subreddit, r/incel. We also successfully implemented logistic regression and a local classifier incorporating additional features drawn from network characteristics. We did not observe much of a difference between the accuracy of the two models, which could indicate that the network structure information may not help with the classification of banned vs. non-banned subreddits.

To build on top of this project, it would be best to start from the beginning and build a more well suited dataset for this task. In this project, we utilized 3 distinct datasets: PushShift dataset (8) for featurization of subreddits, the SNAP dataset (9) for creation of the Reddit network, and a Reddit dataset of various banned subreddits / a list of subreddits that we compiled on our own for labelling subreddits as banned or not banned. This reliance on various datasets was problematic because we could only use data for subreddits that were contained in all three of our datasets. To improve on this, it would be best to maybe continually cache PushShift datasets and labels for banned subreddits over some time period. In this way, one can be sure to have enough training and testing data for building their models.

We hypothesize that because of our lack of a large dataset to train and build off of, our Local Classifier was not able to outperform classification done by our Logistic Regression. From our Graph Exploration section, using network attributes for this classification task did seem promising. Specifically in our exploration of cosine similarity, we showed that the use of recursive feature generation did seem to lead to the discovery of more similar subreddits. Therefore, more research is needed in order to come to a conclusion on the usefulness of network attributes in this classification task, and further research would entail gathering a better dataset. A different selection of features for nodes beyond clustering coefficient and node degree is perhaps needed as well.

With a more complete dataset, it would then also be possible to build more complex models for this problem, including Relational Classifiers or a GCN, and do more intensive network analysis to understand better the topography of Reddit.

References

[1] App, S. (2019) Reddit Is More Influential in Higher Ed Than You Think. Retrieved November 7, 2019, from https://www.insidehighered.com/blogs/call-action-marketing-and-communications-higher-education/reddit-more-influential-higher-ed.

[2] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstien, E. Gilbert. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech". Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 31 (November 2017), 22 pages. https://doi.org/10.1145/3134666. http://delivery.acm.org/10.1145/3140000/3134666/a031-chandrasekharan.pdf.

[3] S. Kumar, W. Hamilton, J. Leskovec, D. Jurafsky. Community Interaction and Conflict on the Web. Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18 (2018): n. pag. Crossref. Web.

[4] S. Grayson, D. Green. Temporal Analysis of Reddit Networks via Role Embeddings. (2019). Web. arXiv:1908.05192

[5] D. Choi, J. Han, T. Chung, Y. Ahn, B. Chun, T. Kwon. Characterizing Conversation Patterns in Reddit: From the Perspectives of Content Properties and User Participation Behaviors. (2015) Web. https://dl.acm.org/citation.cfm?id=2817959

[6] V. Belak, S. Lam, C. Hayes. Cross-Community Influence in Discussion Fora. Digital Enterprise Research Institute, NUI Galway IDA Business Park, Lower Dangan Galway, Ireland. 34 Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media. (2012) Web.

[7] A. Robertson. "Reddit has broadened its anti-harassment rules and banned a major incel forum." The Verge. Sept. 30, 2019. https://www.theverge.com/2019/9/30/20891920/reddit-harassment-bullying-threats-new-policy-change-rules-subreddits.

[8] J. Baumgartner. Reddit data dump. (2019) http://files.pushshift.io/reddit/

[9] Leskovec, Jure and Krevl, Andrej. "Social Network: Reddit Hyperlink Network." SNAP: A General-Purpose Network Analysis and Graph-Mining Library, *Stanford Large Network Dataset Collection*, http://snap.stanford.edu/data/soc-RedditHyperlinks.html.

[10] PushShift Reddit BigQuery dataset. https://github.com/pushshift/google_bigquery .

[11] List of all known banned subreddits sorted alphabetically and by reason. (2019, July 22). Retrieved from https://www.reddit.com/r/reclassified.

[12] Banned Word List. (2009). Retrieved from http://www.bannedwordlist.com/.

[13] H. Habib, M. Bin Musa, F. Zaffar, R. Nithyanand. "To Act or React?" https://arxiv.org/pdf/1906.11932.pdf.