Analyzing Political Communities on Reddit

Kristy Duong
Computer Science
Stanford University
kristy5@stanford.edu

Henry Lin
Computer Science
Stanford University
henryln1@stanford.edu

Sharman Tan
Computer Science
Stanford University
sharmant@stanford.edu

Abstract

In recent years, the political atmosphere in the United States has become more strained and divisive, particularly since the campaign runs for the 2016 election that included President Donald Trump. Social networking sites like Reddit have led to easier and more rapid information dissemination, and given the copious amounts of data available on these websites, serve as an excellent source to better understand polarization of users and communities across time. We performed analysis on several politically motivated subreddits, including r/The_Donald and r/politics, and we look at the communities formed across these subreddits as users engage with one another in political discourse. Ultimately, we found that both users and subreddits could both be clustered into distinct communities based on interaction and user overlap along each community. The temporal aspect of social networks also played a big factor, and we used this to further examine long-term users and their level of engagement on the site.

1. Introduction

The Internet's booming popularity over the past several decades has led to the creation of popular social networking sites such as Reddit, allowing for a convenient forum for discourse and interaction. One of the most polarizing topics is politics, and the recent political atmosphere in the United States exemplifies this, particularly with the 2016 and 2018 election cycles. Many issues in the government have devolved into votes along party lines rather than a bipartisan solution that leaves both major parties satisfied, and social websites like Reddit may provide some insight as to why across the aisle conversations have become so warped.

In this paper, we explore political communities and their users on Reddit, and through graph analysis, we try to elucidate some of the interesting aspects of these communities in isolation, and in relation to one another. In order to do this, we analyze Reddit data dating all the way back to 2014, but we concentrate on efforts on the years of 2016, 2017, and 2018. We select several political subreddits to focus our analysis on, and we further zoom in by looking at the users that frequent these communities. By delving deeper into these communities, we can look at how user engagement develops over time and how closely related communities can become as users come and go. To help with our research, we build several graphs to highlight subreddit relationships and user interactions, and we employ techniques like spectral community detection and natural language processing to better understand what binds and separates these groups.

2. Related Work

2.1. Time-Varying Graphs and Social Network Analysis

Social networks are dynamic structures that are constantly changing overtime. Prior work done by Santoro *et al*. [8] introduces and summarizes several atemporal and temporal metrics for analyzing time-varying graphs. Atemporal parameters, including density, clustering coefficient, and modularity, can be used on static graphics and its evolution can be seen by examining the metrics from a sequence of static graphs. Meanwhile, temporal indicators examine a sequence of time-varying graphs restricted to a *lifetime* and include distance, diameter, and centrality. These metrics allow us to detect community structures and closeness, as well as user impact and information dissemination.

2.2. Community Identity and User Engagement in a Multi-Community Landscape

Reddit has been a prominent player in the world of online communities for over a decade now, and there has been a significant amount of analysis already done on its communities. In [9], the authors introduce several new metrics, *distinctiveness* and *dynamicity (DYN)*, to help better understand the discussion with a community and its effects on user retention and engagement. The former is a look into how specialized the topic is within the community; in other words, it attempts to quantify the level of jargon a community uses. *DYN*, on the other hand, quantifies how quickly a community changes its discussion topics, measuring how stable topics are over time. We employ both of these metrics to help better understand the communities we are interested in.

2.3. Language use as a reflection of socialization in online communities

Although the graph properties of a network representing data from subreddits may provide us with important insight about graph structure, the language usage in subreddit comments can reveal more specific properties of subreddits and users over time in the context of politics. In [4], Nguyen and Rose introduce various metrics to measure language usage over time and between communities. These metrics include *Kullback-Leibler (KL) divergence* and *Spearman's Rank Correlation Coefficient (SRCC)*. These metrics use word frequencies and rankings to evaluate the how language usage changes and how it might converge, possibly due to socialization in online communities. They also use these metrics to predict user retention rates. All of these metrics are relevant in the context of political subeddits, and we compute both *KL divergence* and *SRCC* to evaluate language usage over time and between subreddits.

3. Dataset

Reddit is an American social news aggregation platform that allows users to discuss various topics ranging from politics to gaming and to react to content using an up- and down-vote system. r/The_Donald was created June 27, 2015 and currently has approximately 667,000 subscribers. When r/The_Donald was initially conceived, its community description was as follows: "Following the news related to Donald Trump during his presidential run. Media hit pieces from the left and the right will be vetted. Interesting topics include polling, campaign-related comments, reactions and push backs." However, since then, their community description has drastically changed and is as follows: "The_Donald is a never-ending rally dedicated to the 45th President of the United States, Donald J. Trump." This community is our main subreddit of interest given the polarized atmosphere the American political system is currently in and the rapid growth of the community over the past several years. For comparison and to track change, we build our political subreddit community from the following subreddits: r/The_Donald, r/PoliticalDiscussion, r/politics, r/socialism, r/Libertarian, r/NeutralPolitics, r/Ask_Politics,

r/AskTrumpSupporters, r/moderatepolitics, r/democrats, r/Conservative, r/Republican, and r/Liberal. Our data comes from a publicly available repository of Reddit stored as compressed json files from across the years. The data contains, but is not limited to, all users, comments, score for each comment (based on up- and down-votes), controversiality score for each comment, and timestamp for each comment.

4. Methods and Evaluation

4.1. Data Preprocessing

Because the dataset we are relying upon simply gives compressed json files divided into monthly chunks, we built a data pipeline to help reduce the file size and remove anything unnecessary. For each month, the compressed file was anywhere from 5-10 gigabytes (GB), and for each of these files, we extracted the comments associated with the subreddits we are interested in (r/politics, r/Republican, r/The_Donald, etc.) along with some important metadata including but not limited to score, author, and timestamp. This information was then written to a csv file that we could later access instead of the original file. Doing this, we managed to shrink the files from several GB to at most several hundred megabytes, a large improvement to help speed up computation later on.

We ultimately ended up processing over a billion Reddit comments over the 34 month period from January 2016 through October 2018 for our data. At the time of analysis, November 2018 comments had not been scraped yet, so we excluded this month. We decided on this period of time starting in January 2016 because it was when the election cycle began in earnest within the United States, and a subreddit like r/The_Donald really exploded in popularity and visibility.

4.2. Graph Construction

The first graph we construct is a weighted, undirected bipartite graph between the subreddit communities and individual users. We constructed this graph on a monthly basis, meaning that for each month, there is a separate graph for the users that were active in that month. For an edge to exist between an user and a community, the user must have commented at least once in that community that month. The edge weight is the total number of comments by that user. This graph allows us to gain an initial understanding as to what the clustering of nodes and communities are, which we measure using metrics such as density and clustering coefficient.

The second graph we construct is a community relation graph that elucidates more clearly how closely two communities are related by the number of common users. Again, we do this on a monthly basis. The nodes are individual communities and the weights are the number of users that have commented at least once in each community that month.

Lastly, we created a user interactions graph to highlight how often different users commented on the same topics, an indication of shared interest. The nodes of these graphs are different users and the edge weights for any pair of users is the number of times they have commented in the same submission/thread, regardless of community. This means that it is possible for two users to comment on the same submissions over multiple subreddits (e.g. r/politics and r/democrats). Due to the scale of the dataset, even after limiting to solely political subreddits, we further scaled down the users by limiting it to consistent users, accounts that had commented at least once for 12 consecutive months. Doing this removed any users that only participated for a brief period of time and limited our analysis to accounts that had consistently engaged within their communities over an extended period of time. This brought the number of users to about 20,000, a much more manageable size to calculate our metrics on.

In order to evaluate the user interactions graph's metrics, we create a null model for a particular month of the graph (Feb. 2016) by edge rewiring. We rewire edges while calculating the

clustering coefficient and stop rewiring when the clustering coefficient converges (in our case, 0.29 (Table 3)). The resulting graph is a null model that has the same degree distribution as our original user interactions graph but is otherwise random.

4.3. Language Content Processing and Metrics

In order to analyze the language in subreddit comments, we first clean the comments by tokenizing the, removing stop words, removing punctuation and case, and stemming the words using the Porter-Stemmer algorithm [6]. Then, we computed a word distribution of the 100 most common words in November of 2014-2016 in each of the subreddits of interest. We normalize the word counts over the total number of words. Once we have the word distributions, we are ready to compute our metrics: Kullback-Leibler (KL) divergence, Spearman's Rank Correlation Coefficient (SRCC), and dynamicity (DYN).

Dynamicity (DYN) is a look into how stable or volatile a community is when looking at the common topic trends over time [9]. For a community with high DYN, their topics of discussion will vary as new interests pop up and fade away. Conversely, a stable community will possess a very low DYN value. The value itself is calculated using a volatility metric for each word spoken within a certain time frame compared to the entire history of that words frequency (computed as PMI). If a word is occurring more often than usual at this time step, then it's volatility score will increase. In the equation below, w is a word and t is the time period of interest while T represents the entire frequency history of the word in question in community c. The DYN is then the average of all word volatility scores for a time period.

$$V_{c_t}(w) = log(\frac{P_{ct}(w)}{P_{cT}(w)})$$

KL divergence, represented by the formula below, measures the difference between two given distributions. Larger values indicate bigger differences in distribution, and P represents the true distribution. A KL divergence score of 0 indicates identical distributions.

$$KL(P,Q) = \sum_i P(i)log(\frac{P(i)}{Q(i)})$$

Unlike KL divergence, SRCC does not involve the difference between word counts of different time periods and instead measures the similarity of words relative to each other. In the formula below, d_i is the difference between the ranks of word i in the two rankings and n is the total number of words. A SRCC score of 1 indicates identical rankings.

$$SRCC = 1 - 6\sum_i \frac{d_i^2}{n(n^2-1)}$$

4.4. Evaluation

After computing graph metrics (clustering coefficients, densities, average degrees, the number of connected components, and the number of edges) over time, we compute the same for a null model and compare the values of the metrics for each of the models to gauge their significance.

The Temporal PageRank algorithm tells us which nodes are most important at various time periods, and we make sense of these results by looking into that node's activity on Reddit in general and hypothesizing why it was chosen.

For our community detection algorithms, we use manual evaluation to in order to verify our results. Since each node contained the username of the Reddit account associated, we are able to look up the Reddit comment histories of these accounts and determine their areas of activity, and by doing this, we are able to better understand the highlighted communities detected by the Louvain algorithm [2].

We evaluate our results from processing the language usage of subreddit posts by observing

Bipartite Graph Metrics		
Year	Number Nodes	Number Edges
2014	48,281	53,620
2015	68,617	76,916
2016	266,635	325,969
2017	425,531	519,846
2018	459,233	550,520

Table 1: Nodes and edges in the bipartite graph corresponding to users and connections to subreddits.

Subreddit Size Metrics		
Year	The_Donald	politics
2014	0	39,288
2015	386	56,571
2016	104,967	173,968
2017	113,809	295,154
2018	99,819	340,632

Table 2: Number of active users

Clustering Coefficient		
Month	Original	Configuration
Feb 2016	0.65	0.29

Table 3: Comparison to Configuration Model

overall word distributions that appear and verifying that these word distributions represent the major themes and attitudes of the subreddits and time periods.

5. Results

5.1. Graph Metrics Analysis

Looking at the data from 2016, we visualized the average number of comments per month and retention rate for recurring users across several subreddits of interest. We classified a user as active if they had commented at least once in any given month. This is important to note as many

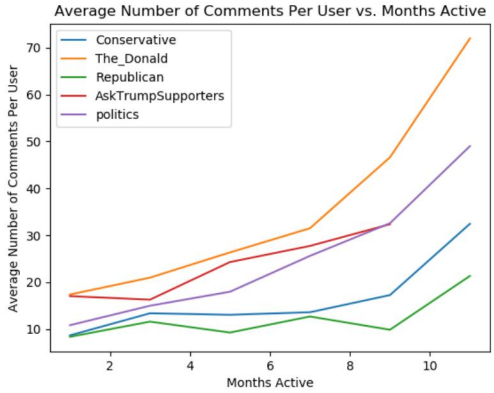


Figure 1: Comments/User vs. Time in Community

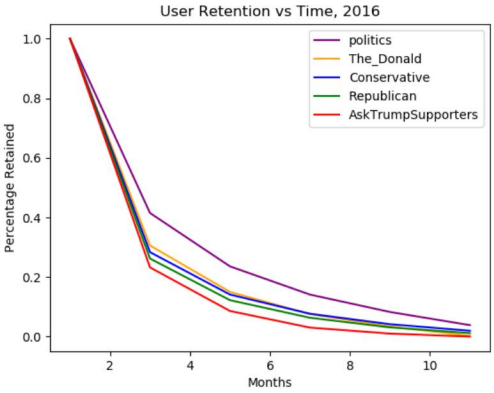


Figure 2: User Retention over Time

Reddit users tend to simply browse and refrain from actively participating. We then tracked when a user first joined a community and their subsequent comment counts in the following months. In Figure 1, we see this trend, with the x-axis representing the number of months they had been part of the community, and the y-axis being the average number of comments for users who had been active for x months. This shows a clear trend of users becoming more and more involved across time, and the most significant increase occurs with r/The_Donald, even amongst conservative communities. We contrast this with the user retention rate displayed in Figure 2. We calculated user retention as the number of users active

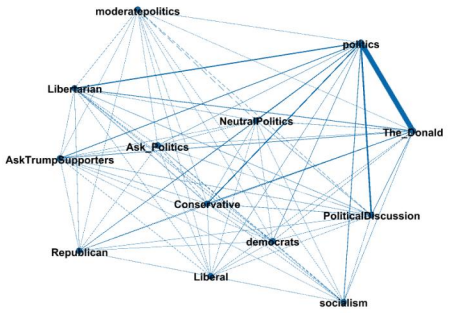


Figure 3: Number of Common Users between Communities, November 2017

for i-th months over the users active for at least 1 month, shown below.

$$Ret(t_i) = \frac{Users(t_i)}{User(t_1)}$$

Aside from r/politics, there is an almost equivalent drop in retention for the conservative subreddits, suggesting that these subreddits do equally well in maintaining user interest across time, and less than 10% of users actually participate for a full year in a community. We suspect that the higher retention rate of r/politics comes from the fact that until very recently, Reddit automatically subscribed new accounts to r/politics, and thus, it is naturally a more visible community compared to the others on this graph.

In Figure 3, we provide a visualization of the number of overlapping users between the various political subreddits. For any two communities, we considered an user active in both if they had, in a specific month, commented at least once in both communities. We can see that there is a significant amount of overlap between r/politics and r/The_Donald in the number of sheer users, but it is important to remember that those two subreddits also boast the greatest number of subscribed users.

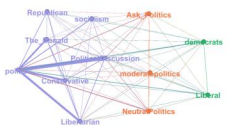
To analyze how the properties of the user interaction graph change over time, we computed the clustering coefficient (Figure 11), average degree (Figure 14), density (Figure 16), number of connected components (Figure 15), number of nodes (Figure 17), and number of edges (Figure 18) for every other month of 2016. We consider the same users for each month of 2016, but in each month, some number of users are isolated (have degree 0). This may mean those users never commented directly on a thread (instead commented in response to other comments), or the people who those users would have been connected to were not people who were active throughout 2016. Figure 17 shows that we account for the same 17291 users in each month of 2016, and among those users some number of them are isolated.

The clustering coefficient decreased over time, To evaluate the clustering coefficient of the user interaction graph, we compared its value in February 2016 (0.65) to that of the null model we produced by rewiring edges (0.29) (Table 3). We only make this comparison for February 2016 because of computing clustering coefficients and other metrics for our networks are extremely computationally expensive. The fact that the clustering coefficient is significantly higher in the user interaction graph compared to the null model indicated that there are significant clusters in the graph, spurring our work involving community detection algorithms to identify and analyze these clusters.

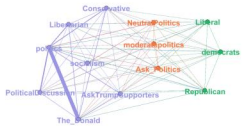
The average degree increased over time. This is expected of networks as they evolve, because networks generally get denser over time as the number of edges grows faster than the number of nodes. From Figures 17 and 18, we see that the number of nodes increases by a factor of 16 over 2016, while the number of edges increases by a factor of 7. Figure 16 confirms that the density increases over time, increasing steadily from 0.02 to 0.05 over 2016.

The number of connected components decreased over time, going from 4 to 1 and staying at 1 starting in June 2016. This makes sense –

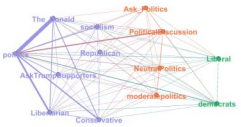
in every month, we are considering the same total number of nodes (17291 nodes), but over time the number of isolated nodes decreases. This means that over the year, more and more of the users become involved by commenting on more posts. Therefore, although the graph may have started with 4 connected components, as more users comment and become interconnected, by June 2016 the graph becomes connected (disregarding completely isolated nodes with degree 0).



(a) January 2016



(b) January 2017



(c) January 2018

Figure 4: Community Detection on Subreddits, 2016-2018

5.2. Community Detection

We applied community detection via spectral clustering on the community relation graphs like the one displayed in Figure 3 to try and partition the subreddits into distinct communities based on the subreddits' political stances and how users come and go in communities. We tested the optimal number of communities by running the algorithm detecting from 2 to 9 communities, and we found that 3 resulted in the highest modularity score (Figure 12). In Figure 4, we provide the partition for January of 2016, 2017, and 2018. One thing to note is that in 2016, r/AskTrumpSupporters did not yet exist which is

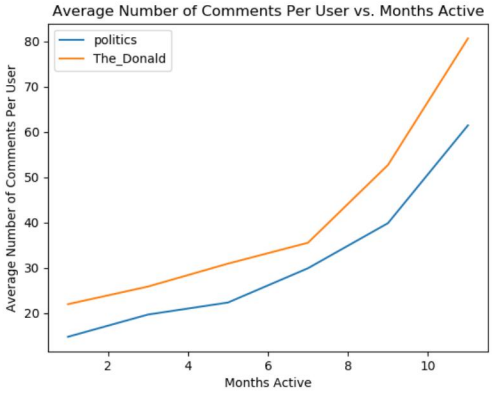


Figure 5: Common Users of r/politics and r/The_Donald

why there is one less node there. While upon initial inspection there seems to be a clear divide into liberal, moderate, and conservative, this does not perfectly describe the clusterings. Rather, it simply highlights the fact that a large number of users tend to frequent both r/politics and conservative communities while on the democratic end, they tend to avoid r/politics despite it being a community open to anyone interested in politics.

While Figure 1 shows the amount of participation versus time in community, one of the primary aspects of online political communities is how they impact one another. To better understand this, Figure 5 expands on Figure 1 and specifically looks at common users between r/politics and r/The_Donald using the same metric. Again, we see the same increase in comments as a user stays in a community, but the slope from the first month active to the final month of activity is steeper, with approximately ten more comments in each subreddit for users by the eleventh month. This provides a strong baseline moving forward, as we now know there are active users that frequent both communities, and by tracking these users, we can understand how their preferences change over time.

For our user interactions graph, we used both spectral clustering and the Louvain technique to detect community structure amongst the users

[2]. The optimal number of clusters for February 2016 was 15 (Figure 13). We perform spectral clustering just for February 2016 and June 2016 (results for both months are near identical) because spectral clustering on our large networks is extremely computationally expensive. Figure 8 shows that spectral clustering simply clusters the vast majority of users into the same cluster, leaving the other clusters with very few points. However, the Louvain technique (Figure 7) resulted in very distinct structures amongst the users, particularly the divide between liberal users and Trump supporters. We also provided labels for the largest communities detected. The algorithm did split up the two ends of the political spectrum into multiple groups however, as noted by the multiple labels for *Liberal-leaning* and *Trump Supporters*. We suspect this is because there are a significant number of users of both groups that break out of their communities and attempt to engage in communities that exist between the two groups like r/Political_Discussion and r/AskTrumpSupporters.

However, Figure 7 indicates one particularly active month of political discussion, since in most other months in 2017, the graph structure looked more like Figure 6, in which there are two primary communities, representing liberals and conservatives. Evidently, the Louvain technique resulted in much more meaningful communities than spectral clustering, and this may be because our data does not meet the assumptions of spectral clustering (relations are not transitive or dataset is noisy).

5.3. Temporal PageRank Analysis

We also computed temporal PageRank [7] on our user graphs from 2017. One interesting point to note is that within the top ten 'most important' users, 90% of users were within the blue and black communities shown in Figure 7, suggesting that users that engage more with both of the liberal and conservative groups are more important. Contextually, this makes sense because users in

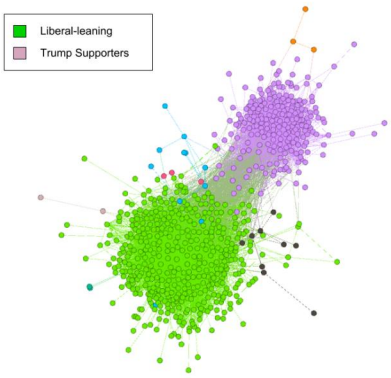


Figure 6: Community Detection via Louvain on Users, October 2017

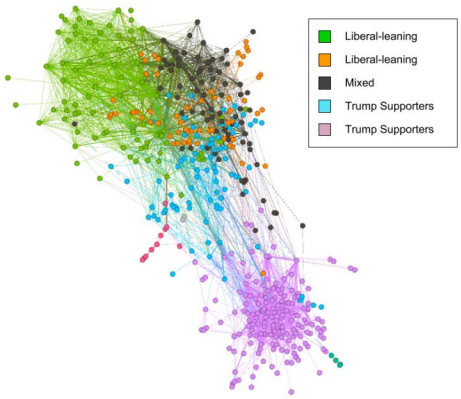


Figure 7: Community Detection via Louvain on Users, December 2017

the mixed group may include swing voters, and users that choose to engage with both sides interact with or have access to more users, increasing their importance in the network.

5.4. Language Content Analysis

To compute KL, SRCC, and DYN, we found word distributions representing the 100 most common words and the number of occurrences. We graphed the word distributions of each subreddit for November of 2014, 2015, and 2016. Figures 9 and 10 represent three of the subreddits' word distributions that are intuitively the most significant.

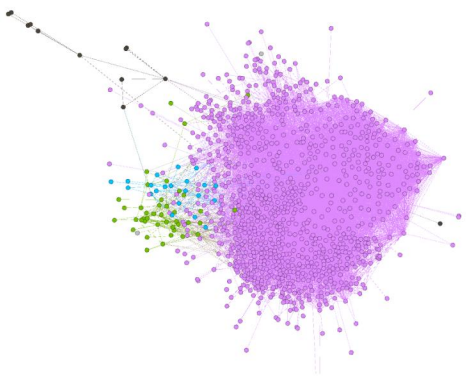


Figure 8: Community Detection via Spectral Clustering on Users, February 2016

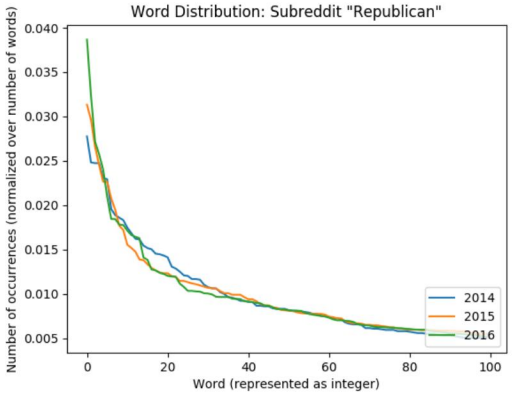


Figure 9: r/Republican Word Distributions

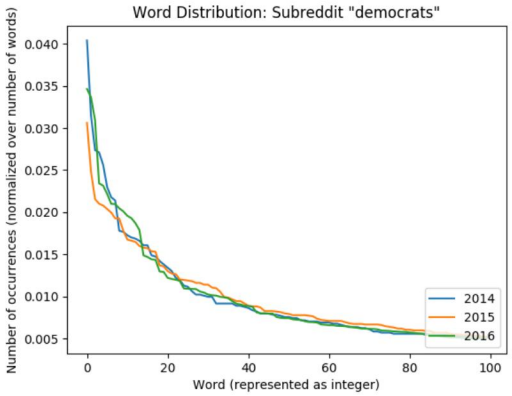


Figure 10: r/democrats Subreddit Word Distributions

Notice that in each of the three figures, there is a year whose most common word's normalized count is the highest. In Figure 9, the word "trump" topped all other most common words of November 2014 and 2015 to be the highest occurring word represented by the first point in the green curve. This indicates how significant Donald Trump became in the transition between November 2015 and 2016. In Figure 10, the highest point in the blue curve is not particularly significant, because it simply represents the stem word "peopl," which is generally one of the most common words. However, upon inspecting the words and their frequencies in r/democrats, we notice that the word "vote" ranks first, third, and second in the three time periods, unlike the other subreddits. This may indicate a stronger emphasis on encouraging people to vote (possibly for specific candidates) among Democrats.

From inspecting the rankings of words r/Libertarian, we observed that "fuck" was the most common word in November 2015, an intriguing finding considering "fuck" is never in the top three most common words in any other subreddit. Also, we noticed that in November 2016, the top three ranked words of r/The_Donald were "news," "fake," and "cnn," words that were not in the 100 most common words of any other subreddit. This indicates the presence of notable events in November 2016 involving Donald Trump. Therefore, from the word distribution graphs and the rankings of words, we can find intuitive implications about events in specific time frames and the different language usages of different subreddits.

Using the word distributions, we computed KL divergence, SRCC, and DYN for each of the subreddits over time. Each of these metrics are displayed in Table 4. All the KL divergence scores are greater than 0 and all the SRCC scores are less than (but very close to) 1, so the distributions do change over time. The subreddits r/moderatepolitics and r/The_Donald have the highest KL divergence scores. This may be true

of r/moderatepolitics because it has the fewest number of comments (Table 5) compared to the other subreddits, and so small changes in word distributions can result in high KL divergence scores. As for r/The_Donald, we previously noticed that r/The_Donald has a significantly different word distribution in November 2016 compared to November 2015. This may be attributed to the 96,502% increase in the number of comments in r/The_Donald (Table 5) between November 2015 and 2016 and the influence of significant events on the words in comments.

The SRCC scores generally indicate a relatively small but significant level of difference between the word occurrence rankings, mostly between 0.6 and 0.9. The lowest SRCC score 0.42382 (Table 4) comes from r/moderatepolitics, but, similar to the KL divergence score, this may be because r/moderatepolitics has the least comments. Most of the dynamicity scores we computed were small negative numbers, meaning that in general for each time period, the time period's word occurrences were less frequent compared to all of our history (November 2014-2016).

6. Conclusion and Future Work

Overall, our results do show some clear and perhaps expected results of the Reddit political community that in many ways reflect that of the real world. We found that amongst the political communities chosen, there is a distinct clustering into several different factions as shown earlier in Figure 4, and this clustering often times mirrors the ideologies of the communities themselves. Our analysis of user interactions through comments also highlights the polarized atmosphere in online discourse at the moment. We see noted that many of the communities detected amongst users through the Louvain algorithm looked like Figure 6, where each end of the political spectrum is abundantly clear. We found at least one example of a more fragmented month though, where we can also see users that clearly engage with both types of users (Figure 7). This result is further corroborated by our evaluation via Temporal PagerRank with nodes existing between liberals and conservatives receiving a higher score.

There are many directions to explore in the future as Reddit continues to generate massive amounts of data perfect for analysis. One area we would like to expand on in future work is the natural language processing, as while subreddits may talk about the same thing, the tone and manner in which they talk about it may differ drastically. This would be an interesting area to compare against the amount of user retention and also detected communities. Additionally, r/The_Donald remains a relatively new community in the Reddit sphere, and an analysis over a longer period of time would be interesting, especially as the next presidential election approaches. Expanding on that, topic discourse and user engagement are both aspects of a community highly impacted by the real world, and research into whether or not these things can predict future events would be a worthwhile avenue to explore.

7. Code Repository

All code from data preprocessing to evaluation metric calculation is located at https://github.com/henryln1/CS224W. We did not upload our cleaned data into this repository due to the size, but the original data can be found at https://files.pushshift.io/reddit/comments/.

8. Acknowledgements

All graph visualization were generated using the Gephi software [1]. We performed our graph construction and analysis using the Networkx Python library [3]. The spectral clustering community detection was done via scikit-learn [5], a machine learning library in Python.

We would also like to thank Professor Jure Leskovec and the TAs of CS224W for the rewarding class and providing useful feedback along the way.

References

[1] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[3] A. Hagberg, P. Swart, and D. S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[4] D. Nguyen and C. Rose. Language use as a reflection of socialization in online communities, 2011.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[7] P. Rozenshtein and A. Gionis. Temporal pagerank. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 674–689. Springer, 2016.

[8] N. Santoro, W. Quattrociocchi, P. Flocchini, A. Casteigts, and F. Amblard. Time-varying graphs and social network analysis: Temporal indicators and metrics, 2011.

[9] J. Zhang, W. Hamilton, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec. Community identity and user engagement in a multi-community landscape. 2017.

9. Appendix

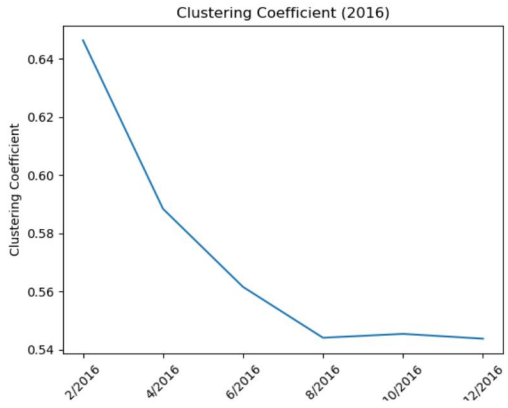


Figure 11: User Graph: Clustering Coefficient (2016)

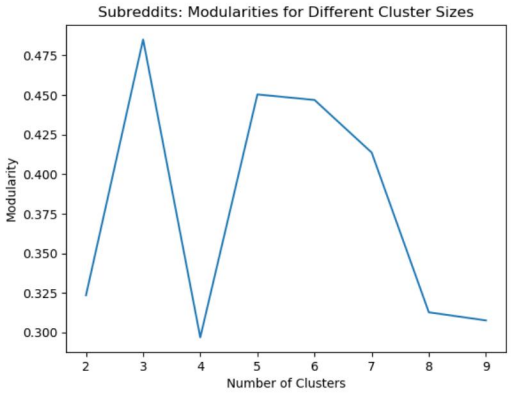


Figure 12: Subreddit Modularity Scores by Number of Communities

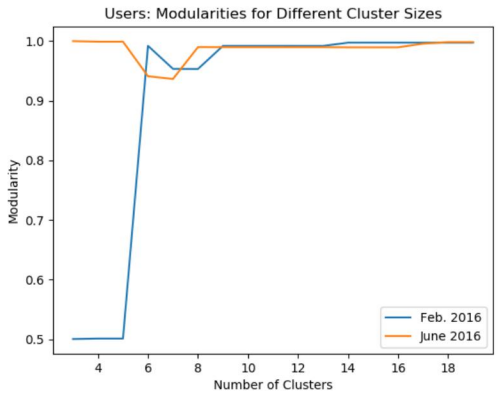


Figure 13: User Modularity Scores By Number of Communities

Subreddit Language Usage Metrics Over Time (Top 100 common words)								
	11/2014, 11/2015	**11/2015, 11/2016**	**11/2014, 11/2015**	**11/2015, 11/2016**	**11/2014**	**11/2015**	**11/2016**	**11/2014, 11/2015, 11/2016**
Subreddit \ Metric	**KL**	**KL**	**SRCC**	**SRCC**	**DYN**	**DYN**	**DYN**	**Mean DYN**
Ask Politics	0.32918	0.309493	0.746223	0.83478	-0.11729	-0.014363	-0.048576	-0.060077
AskTrumpSupporters	x	x	x	x	x	x	x	x
Conservative	0.25953	0.26863	0.82023	0.82824	-0.14287	0.008349	-0.072705	-0.069077
democrats	0.59617	0.50170	0.67150	0.66494	-0.18985	-0.020254	-0.085411	-0.098504
Liberal	0.31588	0.36909	0.76298	0.75586	-0.14698	-0.028670	-0.102111	-0.092586
Libertarian	0.28420	0.37745	0.86436	0.85108	-0.084849	-0.122062	-0.027214	-0.078042
moderatepolitics	0.96412	0.85152	0.67531	0.42382	-0.22811	-0.158857	-0.191224	-0.192731
NeutralPolitics	0.73541	0.67384	0.61701	0.73919	-0.10640	-0.170250	-0.093066	-0.123240
PoliticalDiscussion	0.23407	0.39716	0.89917	0.81288	-0.12552	-0.057104	-0.034505	-0.072377
politics	0.17678	0.37504	0.87572	0.70397	-0.266252	-0.134684	-0.028002	-0.142979
Republican	0.36791	0.29238	0.72002	0.78767	-0.220867	-0.025024	0.121848	-0.122580
socialism	0.156337	0.21072	0.93504	0.89174	-0.087446	-0.037187	-0.009977	-0.044870
The Donald	x	0.82529	x	0.667339	x	-0.247955	-0.001636	-0.12480

Table 4: Metrics for Subreddit Language Usage

Percent Increase in Number of Comments Per Subreddit (11/2014,11/2015,11/2016)			
Subreddit	**# Comments (11/14)**	**# Comments (% Increase 11/14-11/15)**	**# Comments (% Increase 11/15-11/16)**
Ask Politics	2802	3036 **(+8%)**	6982 **(+130%)**
AskTrumpSupporters	x	x	49206
Conservative	14731	27973 **(+89%)**	50731 **(+81%)**
democrats	1225	2074 **(+69%)**	9023 **(+335%)**
Liberal	2425	2424 **(-0.0004%)**	3591 **(+48%)**
Libertarian	30529	35292 **(+16%)**	52678 **(+49%)**
moderatepolitics	622	200 **(-32%)**	1177 **(+489%)**
NeutralPolitics	2020	4354 **(+116%)**	12765 **(+193%)**
PoliticalDiscussion	23335	63115 **(+170%)**	178275 **(+182%)**
politics	256783	505031 **(+97%)**	2654644 **(+426%)**
Republican	2109	3721 **(+76%)**	5682 **(+53%)**
socialism	16498	17628 **(+7%)**	29142 **(+65%)**
The Donald	x	2304	2225716 **(+96,502%)**

Table 5: Percent Increase in Number of Comments in Subreddits

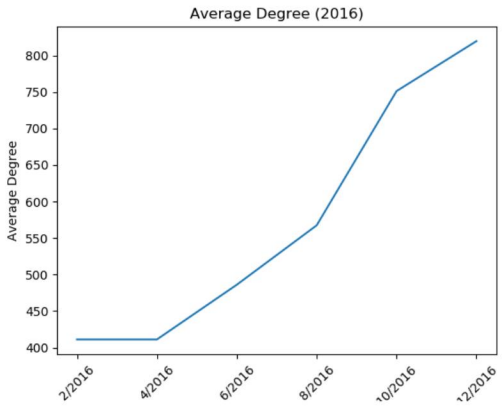


Figure 14: User Graph: Average Degree (2016)

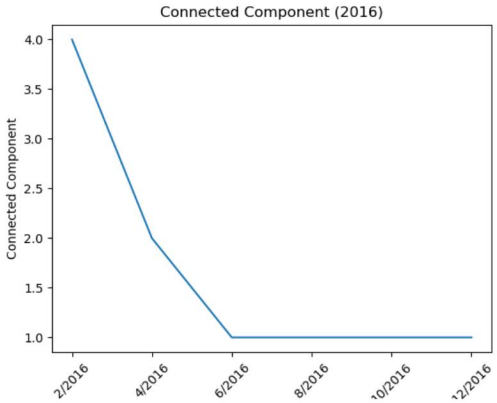


Figure 15: User Graph: Number of Connected Components (2016)

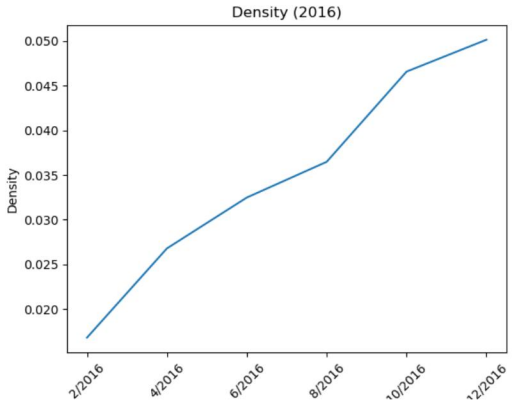


Figure 16: User Graph: Density (2016)

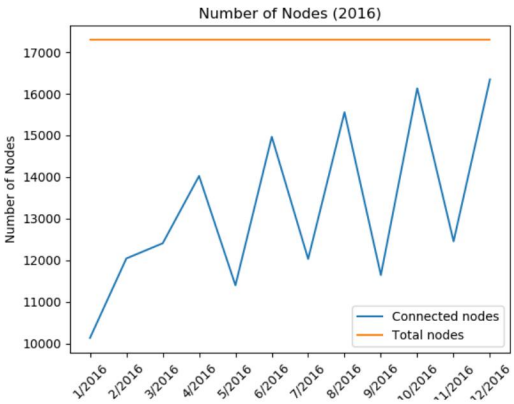


Figure 17: User Graph: Number of Nodes (2016)

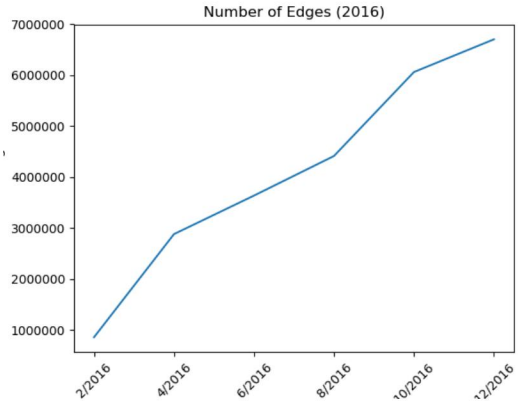


Figure 18: User Graph: Number of Edges (2016)