# I. Project Objective and Motivation

Social media is gaining popularity in recent years as it enables users to share ideas and interact with each other freely without any physical and spatial constraints. A user on one side of the world can easily comment or repost the article posted by another user on the other side of the world with a simple click of button. Users can also collaborate with each other without even knowing the other party. Such convenience allows social media to develop rapidly and it is taking a more and more prevalent role in our daily life. As a result, social network has drawn great attention from researchers. They study the evolvement of trendy topics and users' behaviors on the social network to generate meaningful insights.

Among all the trendy topics discussed on the social media platforms, politics is always an evergreen hot topic, especially in the US. Through social platforms such as Facebook, Twitter and forums, citizens have ample opportunities to participate in the discussion about political news and political candidates, especially during election season. In fact, social media platforms have become an important battleground during elections. In the 2016 US presidential election, investment in digital ads exceeds $1 billion, accounts for 9.5% of the overall political ad spending[4]. Previous researches have studied using digital platforms as a testbed to analyze voters' behavior before or after election. Among all the research areas related to political behaviors on social media, political party affiliation identification is gaining more attention in recent years. The results of such analysis have significant impacts as it can be used to launch targeted digital marketing campaigns.

In our research, we first examine the overall network structure of Reddit network. Since there are numerous topics discussed over the platform, community detection is performed to group subreddits that share similar topics. In this project, we zoom in to analyze the community focused on the 2016 US presidential election to study Reddit's role during the election. To support our analysis, we draw inspiration from various political science studies on the presidential candidates to understand each candidate's position and voters' profile. The table below briefly summarizes the above information. [2][3][5]

| Party | Democratic Party | | Republican Party |
|---|---|---|---|
| Nominee | Hillary Clinton | Bernie Sanders | Donald Trump |
| Formally launched as a major candidate | April 12, 2015 | April 30, 2015 | June 16, 2015 |
| Proposal | Issues of expanding racial, LGBT, and women's right, raising wages and ensuring equal pay for women, and improving healthcare | Issues of income and wealth inequality | Issues of immigration and border security, e.g. banning of foreign Muslims entering the US. |
| Voters' profile | Middle aged and older female voters, and voters from minority races | White voters under 40 years old | Male, white, blue-collar voters and those without college degrees |

*Table 1 Summary of Political Parties and Candidates*

# II. Description of Datasets

## Reddit Hyperlink Network Dataset[8]

Reddit is an online forum which allow users to share content or interact with other users by commenting on their posts. Communities covering different topics are known as "subreddits". The Reddit Hyperlink Network dataset, extracted from SNAP Stanford, represents the connection between subreddits. In this network, nodes represent subreddits. The directed edges represent hyperlinks appeared in a post in the source subreddit citing posts in the target subreddit. The edge also has a sentimental attribute of "-1" if the source subreddit post is negative towards the target subreddit post, and "1" if it is neutral or positive. Other information provided by the dataset includes the timestamp when the source posts are created and text properties of the source post. In total, there are 55,863 number of nodes together with 858,490 number of edges and the timespan ranges from

January 2014 to April 2017. To supplement the analysis, we also scraped comments and submissions from Reddit through Pushshift API.

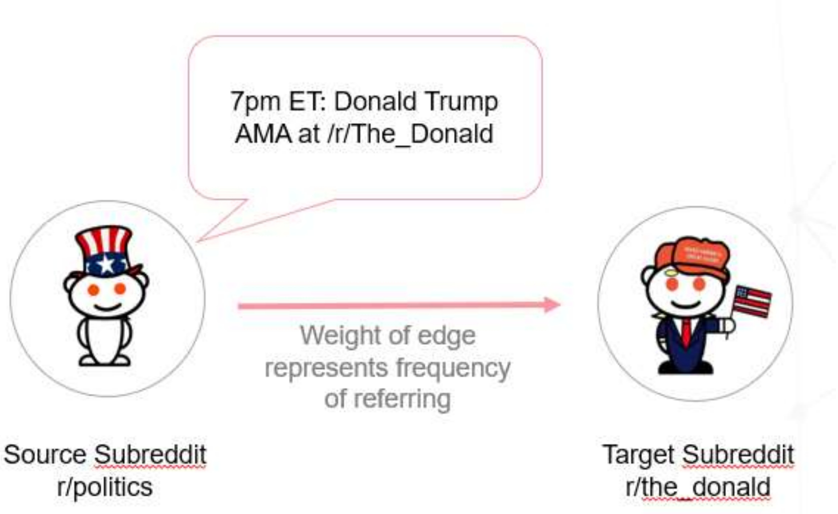
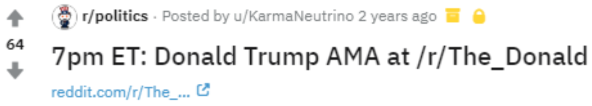

7pm ET: Donald Trump AMA at /r/The_Donald
Weight of edge represents frequency of referring
Source Subreddit
r/politics
Target Subreddit
r/the_donald

*Figure 1 Diagram Representing the Connection Between Nodes*


r/politics · Posted by u/KarmaNeutrino 2 years ago

64

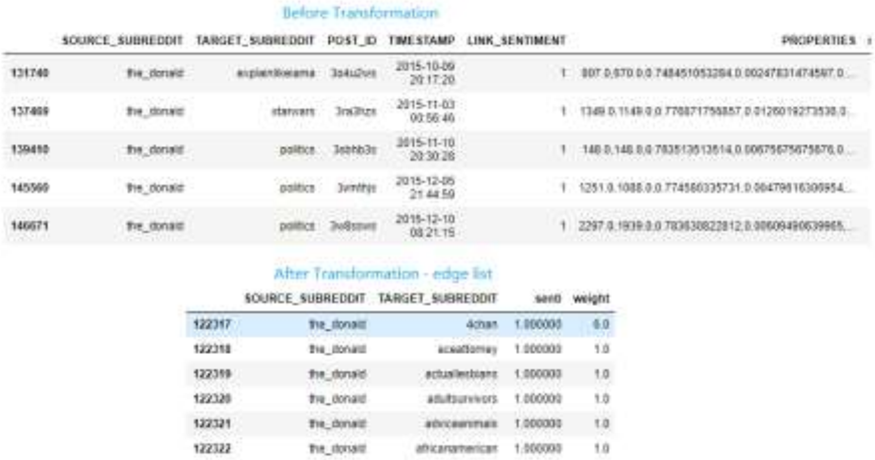
7pm ET: Donald Trump AMA at /r/The_Donald

reddit.com/r/The_...

*Figure 2 A Sample Submission Containing Hyperlink*

## Data transformation

Two separate CSV files are provided, one is the network of hyperlinks extracted in the body of posts while the other is the network of hyperlinks extracted in the title of posts. In our analysis, the file with hyperlinks extracted in the body is used. The dataset contains 35,776 subreddits and 137,821 connections. To further reduce the dimension of the dataset, we aggregate them by computing the frequency of source subreddit referring to target subreddit. The frequency of reference is defined as the weight of hyperlinks. In addition, we aggregate sentimental attribute between the target and source subreddit by taking the mean value. The transformed edge list only contains four columns named source subreddit, target subreddit, weight and mean of sentiment.


Before Transformation

| | SOURCE_SUBREDDIT | TARGET_SUBREDDIT | POST_ID | TIMESTAMP | LINK_SENTIMENT | PROPERTIES |
|---|---|---|---|---|---|---|
| 131740 | the_donald | explainlikeama | 3o4u2vs | 2015-10-09 20:17:20 | 1 | 307.0,270.0,0.748451053284,0.00247831474597,0... |
| 137469 | the_donald | starwars | 3ra3hze | 2015-11-03 00:56:46 | 1 | 1349.0,1149.0,0.770871758857,0.0126019273538,0... |
| 139490 | the_donald | politics | 3sbhb3c | 2015-11-10 20:30:28 | 1 | 148.0,148.0,0.783513513514,0.00675675675676,0... |
| 145569 | the_donald | politics | 3vmthjs | 2015-12-05 21:44:59 | 1 | 1251.0,1088.0,0.774580335731,0.00479616306954... |
| 146671 | the_donald | politics | 3w8sowt | 2015-12-10 08:21:15 | 1 | 2297.0,1939.0,0.783630822812,0.00609490639965,... |

After Transformation - edge list

| | SOURCE_SUBREDDIT | TARGET_SUBREDDIT | senti | weight |
|---|---|---|---|---|
| 122317 | the_donald | 4chan | 1.000000 | 6.0 |
| 122318 | the_donald | aceattorney | 1.000000 | 1.0 |
| 122319 | the_donald | actuallesbians | 1.000000 | 1.0 |
| 122320 | the_donald | adultsurvivors | 1.000000 | 1.0 |
| 122321 | the_donald | africaanimals | 1.000000 | 1.0 |
| 122322 | the_donald | africanamerican | 1.000000 | 1.0 |

*Figure 3 Transformation of Raw Data*

# III. Approach

The following analysis is performed.

- Network influence:
  a. Study the distribution of centrality in the network. Is centrality evenly distributed among nodes or are there a small number of nodes with extremely high centrality? Summarize the network features.
  b. Compare the various centrality measures including degree centrality, closeness centrality~~,~~ betweenness centrality, eigenvector centrality and PageRank centrality. Centrality score of various methods are computed using NetworkX package. For centrality measures other than degree centrality, weight of edges is used as parameter “weight”.
  c. Identify “important” subreddits – nodes that have frequent interactions with other nodes. Explore the characteristics of such subreddits.
- Community detection:
  a. Identify groups of subreddits where there are active interactions among them and explore if they share similar topics. We use the label_propagation_communities function of NetworkX to perform community detection. Weight of edges is included in the parameter.
  b. Examine the US politics community in details. Identify important nodes within the community (with Pagerank). Study how the interactions within the community evolves with time.

# IV. Results

## Influential Analysis

### Centrality

We calculate the **degree (in degree and out degree) centrality, PageRank, betweenness centrality, closeness centrality, eigenvector centrality** of all subreddits within the dataset. The top 5 subreddits with highest centrality score from each method are shown in the table below.

| | In-degree Centrality | Out-degree Centrality | PageRank | Betweenness Centrality | Closeness Centrality | Eigenvector Centrality |
|---|---|---|---|---|---|---|
| **1** | askreddit | subredditdrama | askreddit | askreddit | askreddit | askreddit |
| **2** | iama | copypasta | iama | subredditdrama | iama | iama |
| **3** | pics | drama | videos | iama | videos | videos |
| **4** | videos | subredditoftheday | pics | outoftheloop | pics | pics |
| **5** | todayilearned | outoftheloop | leagueoflegends | writingprompts | todayilearned | todayilearned |

*Table 2 Subreddits with High Centrality*

As we can see from Table 2, the results are similar regardless of the method used. We observe the following characteristics:

1. Some subreddits like SubredditDrama, CopyPasta, Subredditoftheday are the hubs of the Reddit network. They are not related to any particular topics. Instead, they act more like a content aggregation site where users share and discuss submissions from other subreddits.
2. Subreddits with high importance often have high user activity and high number of subscribers.
3. Subreddits with high centrality score may also indicate the trendy topic during the period of study. For example, subreddit LeagueOfLegends has the 5th highest Pagerank score, suggesting that the game League of Legend is popular between 2014 to 2017.

Moreover, the distribution of centrality shows high skewness. A small number of subreddits have extremely high centrality score while majority of subreddits have low centrality score. The distribution may suggest that the Reddit network is a scale-free network.

## Scale-free Network

Scale-free Network is a network whose distribution of degree follows power law. In the real world, most of the networks are the scale-free network, e.g. airline network, social network, interbank payment network etc. Scale-free network have the following characteristics[9].

1. The distribution of degree of the nodes in scale-free network follows the power law distribution, indicating that nodes with high degree are rare and most of the nodes are of low degree.
2. Scale-free network is robust against random attacks/failures but is fragile against target attracts. Due to high skewness of the degree distribution, random attacks/failures are more likely to happen on nodes with relatively low importance. However, targeted attacks targeting the central nodes can easily break down the entire network, because the central nodes are well connected to other central nodes and majority of the small nodes.
3. One of the possible generative mechanism of scale-free network is through preferential attachment. Preferential attachment refers to the mechanism that new nodes in a network are more likely to form connections with existing popular nodes in the network. In the context of Reddit network, it is intuitive that a newly created subreddit is more likely to refer to posts from other popular subreddits.

## Power Law Distribution

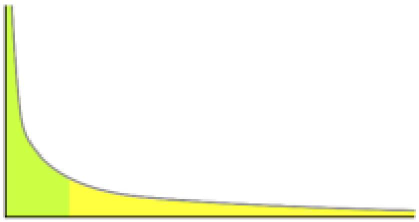


*Figure 4 Power Law Distribution*

In empirical contexts, Power Law Distribution can be represented by function: $f(x) = \alpha x^{-k}$. After performing logarithm transformation on both sides, a linear relationship is obtained: $\log(y) = -k\log(\alpha x)$.

## Distribution of Degree

To verify the power law distribution, we compute the in degree and out degree cumulative frequency respectively. The following two diagrams are obtained after performing the logarithm transformation on the cumulative frequency of indegree and outdegree. Both the log-transformed cumulative frequency of indegree and outdegree shows approximately a linear relationship. Therefore, we conclude that the Reddit network is a scale-free network.

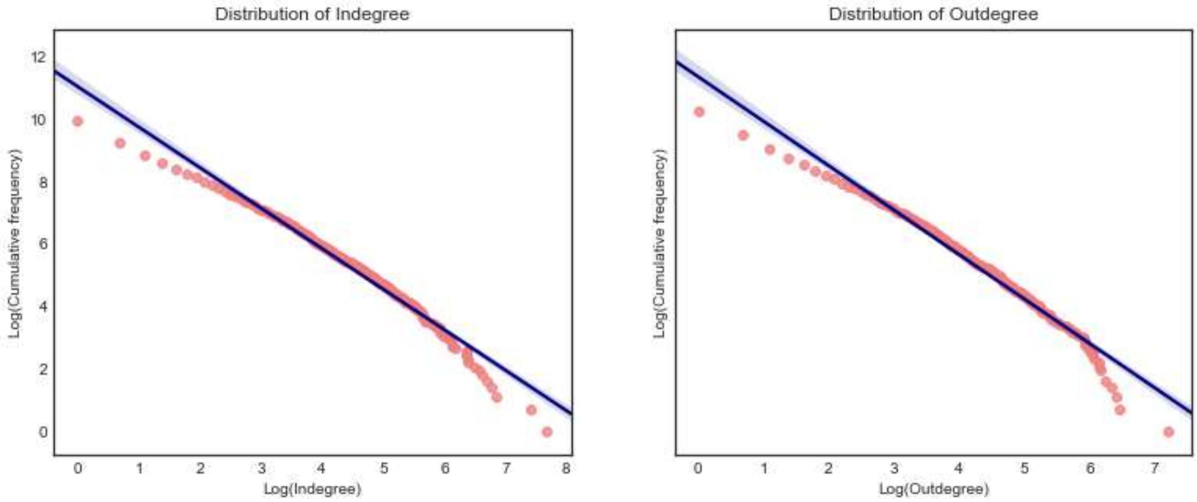


*Figure 5 Distribution of Indegree and Outdegree (Log Transformed)*

We perform goodness of fit to statistically prove that the distribution of indegree and outdegree follows power law. To achieve this, Python packages powerlaw and distribution_compare are used. The result shows that the p value is much smaller than 0.05, indicating that with a significant level of 5%, our conclusion is 5statistically sound.

## Community Detection

Performing community detection on such a large scale network is resource demanding. Moreover, as we proved in previous section that the centrality distribution of reddit network is highly skew and most of the subreddits are of low importance. Therefore, we decrease the size of the network with the following steps.

1. Remove edges with weight less than 5
2. Remove isolated nodes
3. Select the largest Strongly Connected Component.

In the 3$^{rd}$ step, we use the API called strongly_connected_components in NetworkX. Basically, this API uses Tarjan's algorithm with Nuutila's modifications to return the nodes with the maximum score.

Initially, we have 35776 Subreddits with 137821 connections. After the pre-processing method above, only 3092 subreddits with 6381 connections are left.

Next, we perform community detection for the cleaned subreddit network with undirected modularity-based agglomerative algorithms. 375 communities are formed and each community revolves around one particular topic. For example, There is Soccer community whose subreddit are all about discussion on soccer, like soccer matches and the players performance.

The graph below shows the distribution of community size. We notice that around 90% of the communities have less than 5 community members. Also, some communities can be grouped into a super topic. For example, there are soccer, basketball, baseball and other sports communities. They share the same common topic, sport.

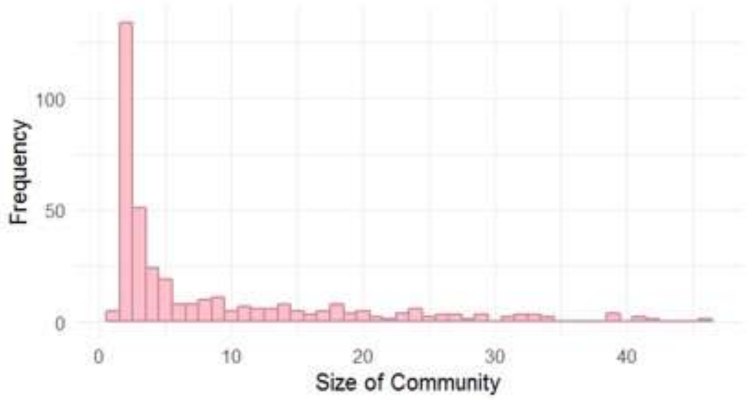


*Figure 6 Distribution of Community Size*

Therefore, we group the existing communities by their common topics. We select the top 100 communities with largest community size and perform topic tagging to each of the communities.  This process is done by manually identifying the super topic based on the subreddits' name and content. Using the example above, communities about soccer, basketball and baseball are all tagged as sports.

The distribution of topics of communities shows that over one quarter of the communities belong to the topic gaming. Other popular common topics are hobby, technology and so on.

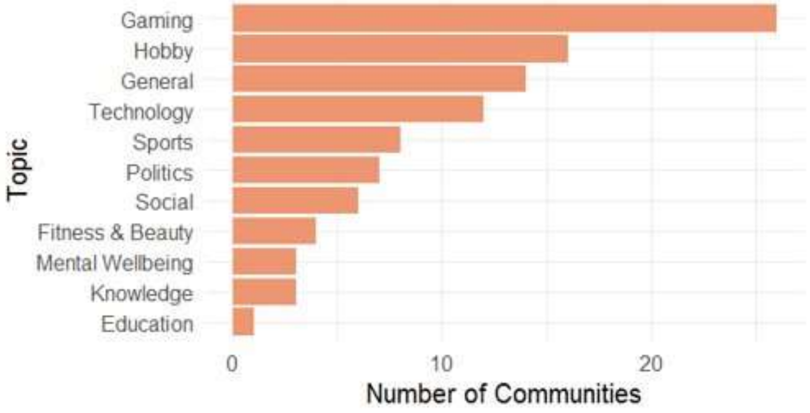


*Figure 7 Distribution of Topics of Communities*

## The US Politics Community

We select the US politics community for further analysis. The graph below shows the visual representation of the US politics community. The size of node represents the page rank of the subreddit. The width of edge represents the weight of edge while the colour of edge represents the average link sentiment.

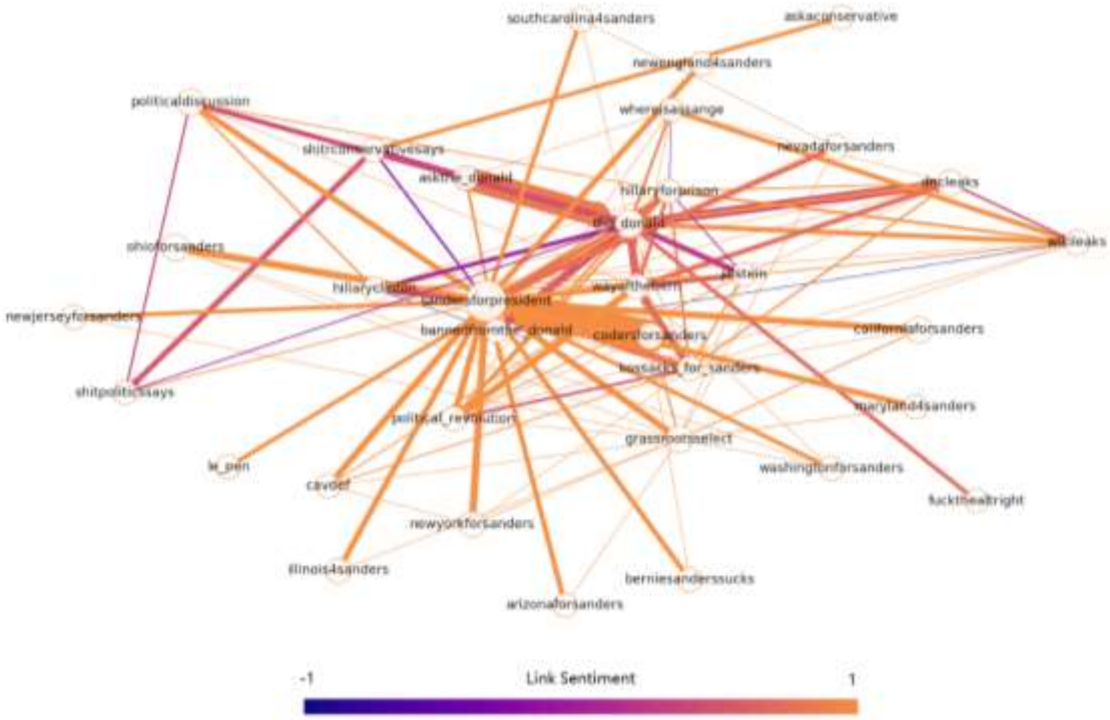


*Figure 8 Visual Representation of US Politics Community*

Looking at the overall network structure of the community, we have the following observations.

- Out of the 34 nodes in the community, 16 subreddits are the supportive subreddits of Democratic candidate Bernie Sanders. Moreover, 3 out of the 5 subreddits with the highest Pagerank score are dedicated to support Bernie Sanders. This reveals that Bernie Sanders turns out to be the most popular presidential candidate on Reddit, though Hillary Clinton and Donald Trump is more well-known by the public. The Sanders phenomenon can possibly be explained by the demographic profile of Reddit's users. Young, white males form a large part of the user base. Pew Research Center reported that 71% of Reddit news users are male, 59% belong to the age group between 18 and 29 years old and 47% identify as liberal. In contrast, the US population comprises 49% of males. 22% are between 18 and 29 years old and only 24% identify as liberal[6]. In fact, Sanders' supporters on Reddit provides powerful support for him. In 2016, Reddit users raised over $1 million of fund to support his political campaign[1].
- The link sentiment between The_Donald, a subreddit devoted to Republican Party's nominee Donald Trump and HillaryClinton, a subreddit devoted to Democratic Party's nominee Hillary Clinton is negative. Similarly, negative link is observed between The_Donald and JillStein, a subreddit devoted to Green Party's nominee Jill Stein. This is hardly surprising as Donald Trump, Hillary Clinton and Jill Stein are political opponents.
- Other than subreddits that are devoted to a particular political party or political candidate, another group of subreddits that are included in the community are related to discussion of election fraud and WikiLeaks. The 2016 Democratic National Communitee email leak could explain the presence of such subreddits in the community. It shows that the 2016 DNC email leak is a frequently discussed topic relating to political candidates.

Next, we examine how the interactions within the community change over time. By grouping the edges by date of posting, we plot network of the community by quarters and compare the difference.

In the first quarter of 2016, the discussion mainly revolve around subreddit SandersForPresident, which is devoted to Bernie Sanders. As the voting for Democratic Party primary starts from Feb 2016. Reddit users are actively involved in discussion about the two Democratic candidates Bernie Sanders and Hillary Clinton. Many Sanders' supporters use Reddit as digital platform to support Sanders' progressive candidacy. For example, we observe strong connection between subreddit SandersForPresident and CodersForSanders. CodersForSanders is a grassroot network form by technology enthusiasts who build websites and mobile applications to promote information about Sanders and voting of primaries.

Starting from the second quarter of 2016 onwards, we notice more discussion on Donald Trump and Hillary Clinton. In the third quarter of 2016, Hillary Clinton was announced as the Democratic Nominee during the Democratic National Convention. SandersForPresident was shut down after DNC due to the flooding of trolling posts and negativity in the subreddit. Sanders' supporters migrate to subreddit WayOfTheBern and Political_Revolution to continue to support Sanders. In the last quarter of 2016, the focus of political discussion on Reddit is largely shifted towards Donald Trump and Hillary Clinton. Subreddit The_Donald evolves as the centre of the network.

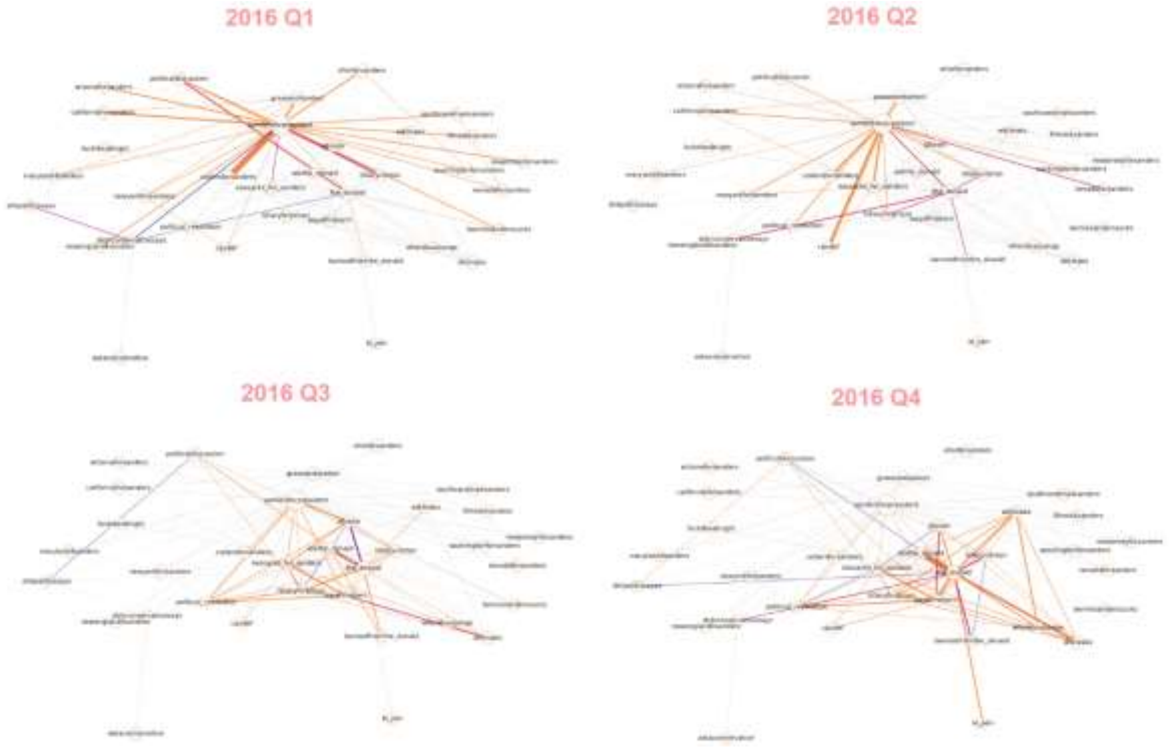


*Figure 9 The Evolvement of US Politics Community in 2016*

The change in Pagerank reveals similar patterns. The importance of subreddit SanderForPresident in the network dropped significantly after June 2016 as Sanders was defeated by Hillary Clinton. The Pagerank score of HillaryClinton and The_Donald increases after Hillary Clinton and Donald Trump were confirmed as the nominee of their respective Party. Considering Reddit is dominated by while male users, it is not surprised to see that the importance of subreddit HillaryClinton is constantly much lower than that of Sanders' and Trump's subreddits.

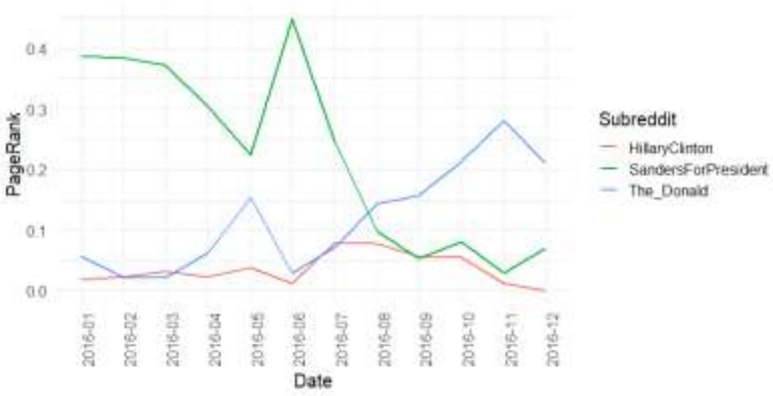


*Figure 10 Trend of Pagerank Score of Subreddits*

To further validate our analysis, we look at the popularity trend of the three political candidates in the entire Reddit network. The trend is alike to what we observe from the US politics community. In the first half of 2016, the number of comments mentioning Sanders exceeds the number comments mentioning Hillary and Trump. Starting from June 2016, Sanders' popularity drops and attention starts to shift to Trump and Hillary. The peak observed in July 2016 is due to realising of results of nomination. The discussion about Trump and Hillary spikes again in November because of the actual Election Day.

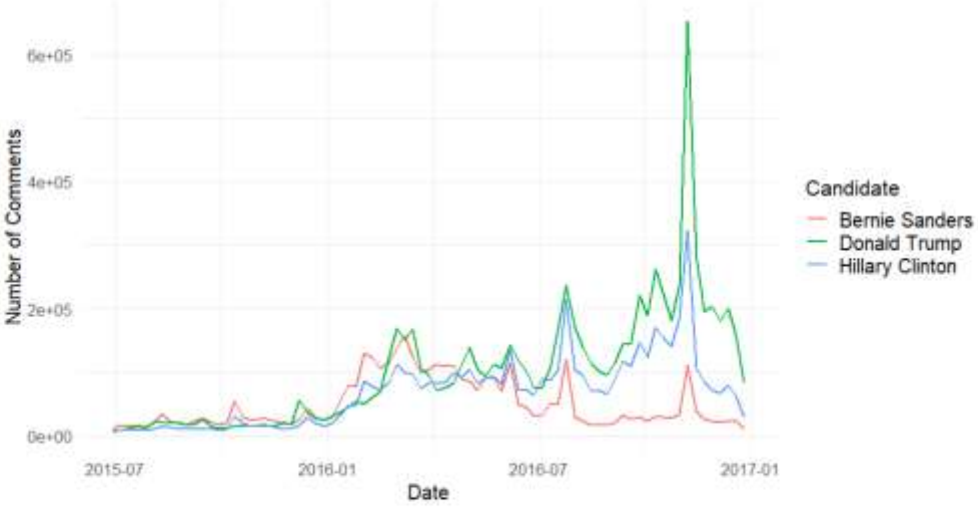


*Figure 11 Popularity of Political Candidates on Reddit*

## V. Discussion of Findings

In summary, we discover and prove that the Reddit network is a scale-free network with a small number of hubs present in the network. The high skewness of degree and centrality indicates that information diffuses at a much faster speed if nodes with high centrality are the sources of information. The conclusion provides us with insights in launching effective marketing campaign or curbing the spreading of malicious information.

The results of community detection show that subreddits discussing same topics form communities as they are often closely connected. A detailed examination of the US politics community reveals the political behaviour of Reddit users and Reddit's significance during the election season. The change of interactions within the community reflects the progression of events in the political landscape. In the digital era, social media will overtakes traditional media to become the new platform for political discussion. User behaviours on social media may even shape the development of certain critical political events. Therefore, study of patterns and behaviours of political discussion on social media has great significant.

In addition, our project has the following potential areas for future work.

- The community detection is performed using the entire dataset which means all edges - hyperlink references spanning from 2014 to 2017 are included. However, as social network changes at unprecedented speed, a trendy topic of yesterday may not be a trendy topic of today. Performing community detection after splitting the dataset by a time interval (e.g. quarter) may provide a more precise result.
- The analysis can be more in-depth if the sentiment and the subject of posts in the US politics community can be analysed. For example, a post can be identified as "negative towards Donald Trump".

# VI. Contributions:

Lin Bei: Proving of Scale-free Network, data cleaning for text analysis and generation of word cloud for selected subreddits

Li Jing: Topic tagging of communities, analysis and visualize several communities, e.g. Sports and technology.

Sun Qiansheng: Computation of centrality, Data Cleaning, TF-IDF research

Wu Jing: Explored topic modelling on scraped submission, analysis and visualisation on different communities.

Zhu Wanyi: Overseeing and planning, Reddit scraping, community detection, topic tagging of communities, analysis and visualisation of US political community, sentiment analysis and topic modelling of posts in selected subreddits (results not used)

# Reference

[1] Aloe, J. (2016, January 27). Bernie Sanders' Reddit supporters raise over $1 million. Retrieved April 27, 2019, from https://www.burlingtonfreepress.com/story/news/politics/2016/01/26/bernie-sanders-reddit-efforts/79302264/

[2] Bernie Sanders 2016 presidential campaign. Retrieved 23 April 2019, from https://en.wikipedia.org/wiki/Bernie_Sanders_2016_presidential_campaign

[3] Donald Trump 2016 presidential campaign. Retrieved 23 April 2019, from https://en.wikipedia.org/wiki/Donald_Trump_2016_presidential_campaign

[4] Greiff, F. (2015, August 18). 2016 Election Digital Ad Spending Will Break $1 Billion. Retrieved April 27, 2019, from https://adage.com/article/digital/2016-election-digital-spend-break-1-billion/299992

[5] Hillary Clinton 2016 presidential campaign. Retrieved 23 April 2019, from https://en.wikipedia.org/wiki/Hillary_Clinton_2016_presidential_campaign

[6] Mitchell, A., Holcomb, J., Barthel, M., & Stocking, G. (2016, May 26). Reddit news users more likely to be male, young and digital in their news preferences. Retrieved April 27, 2019, from https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

[7] Pushshift/api. (2018, September 19). Retrieved April 27, 2019, from https://github.com/pushshift/api

[8] Social Network: Reddit Hyperlink Network. (n.d.). Retrieved April 27, 2019, from https://snap.stanford.edu/data/soc-RedditHyperlinks.html

[9] Stephen, A. T., & Toubia, O. (2009). Explaining the power-law degree distribution in a social commerce network. *Social Networks*, *31*(4), 262-270.