Social Network Analysis of r/The_Donald

Husam Abdul-Kafi (hsa2136), Hari Devaraj (sd2920), Ethan Liu (yl3156)

May 6, 2017

1 Introduction

Social news websites like Reddit can reflect public opinions that are often not reported by the mainstream media. The_Donald subreddit has been an online hub for the supporters of Donald Trump since its creation in June 2015. Political pundits and analysts have painted the picture of the mainstream Trump supporters based on polls and interviews, but The_Donald subreddit may reflect a different set of identities and characteristics of Trump supporters because users tend to be much less reserved on anonymous platforms such as Reddit[1]. The main objective for this project is to see who supports Donald Trump on Reddit by looking at the other subreddits his supporters post in and the nature by which these different subreddits are linked linked together by visits from users of The_Donald.

For the propose of this project, we focus mostly on power users of The_Donald and how their activities connect other related subreddits. We explore changes in the subreddit network before and after two major events: election day and inauguration day. By tracking the power users of The_Donald, we aim to identify the changes of their reddit habits with respect to major events. From this, we hope to make inference and draw conclusions on the identities of the most powerful Trump supporters online.

2 Data

Reddit is very open about data sharing. A Reddit user named Stuck_In_The_Matrix[2] posted a large data dump of Reddit's entire history to /r/datasets. Another user

---

[1]J. Suler, "The Online Disinhibition Effect," CyberPsychology & Behavior, vol. 7, no. 3, pp. 321326, 2004.

[2]https://www.reddit.com/r/datasets/comments/3mg812/full_reddit_submission_corpus_now_available_2006/

named fhoffa dumped the data onto Google BigQuery[3]. BigQuery was an ideal location for such a data set since it was designed to quickly run SQL queries over billions of rows and it can return data in both a csv and REST-like format.

The Reddit data dump available via Google BigQuery covers the time period from December 2015 - March 2017. Our queries initially looked for power users of The_Donald with more than 2000 points in The_Donald subreddit and at least 1500 points in another subreddit. We define power users this way so we can be sure that they have influential posts in both The_Donald and another subreddit. We included the SQL query that we used in BigQuery in the source code attached to this report. This data dump was formatted in the following way:

[user of t_d] [top subreddit s] [total_score in s] [# posts in s] [rank]

Later, with the intent of trying to observe the change power users of The_Donald made over a couple of time periods, we ran SQL queries to get user behaviors for the two weeks before and after an event. Because our time window became shorter, we redefined our power users to be those who have more than 500 points in The_Donald and at least 100 points in another subreddit within the time window. This gave us data sets where The_Donald is connected to other subreddits via these power users. Then, for every user, we draw an edge between the subreddits that they contributed to. This means that, if the user had s1, s2, s3 as their top three subreddits, we would create the following undirected edges: (s1, s2), (s1, s3), (s2, s3) by making a graph data set with the following headers:

[subreddit s1] [subreddit s2] [# times power users posted in both subreddits]

This created a nice format to be read into igraph for our later attempts at graph modeling and visualizations. The number of times power users posted in both subreddits is interpreted as the weight on the subreddit-subreddit edges. The_Donald subreddit, in this data set, is by definition connected to every other subreddit. We would filter out The_Donald subreddit in our modeling step.

3   Method

We identified two major events – the election and the inauguration – to study their effect on the activities of power users in The_Donald subreddit. We set the time window of investigation to be two weeks before and after (one month in total) the

---

[3]https://www.reddit.com/r/bigquery/comments/3mv82i/dataset_reddits_full_post_history_shared_on/

event. After getting the data from BigQuery and creating the graph representation, we proceed with exploratory analysis and graph modeling.

For the exploratory analysis, we fist filter out all the default subreddits. These subreddits are not appropriately indicative of The_Donald users' habits as the default subreddits are subreddits that are visible to all users when they fist make their account. Users are by 'default' subscribed to these subreddits, so posts in these groups automatically get more views. Our data became significantly less noisy when we take out these subreddits, and generally the prevalent default subreddits were unrelated to politics or Donald Trump (they included /r/aww and /r/AdviceAnimals).Then we looked at the top 20 subreddit that power users of The_Donald are most actively involved. We visualized this in terms of histograms where the height of bins corresponds to the number of users. By running this exploratory analysis for the two time windows before and after our events of interest, we hope to get a rough sense of what those users are like before we delve into the graph modeling analysis of subreddits.

For the graph modeling part, we examine the connectivity of the subreddits that power users of The_Donald post in. We filter out nodes with edge weight less than 3 as well as nodes corresponding to the default subreddits. This filtering gives us confidence that those subreddits are most strongly connected to The_Donald since multiple power users are posting in them. Then the graphs are visualized so we can see the specific connections among the subreddits.

We ran both exploratory analysis and graph modeling for each event so that we could get both a high level view of power users activities and how their activities are connected to each other. Results from these analysis are included in the next section.

4 Results

4.1 Election Day

Our preliminary analysis on the time period two weeks before and after election day are shown in Figure 1 and Figure 2. Before election day, Trump supporters seemed to concentrate their attention on the American election - the most popular cross-subreddit for The_Donald users is HillaryForPrison and conspiracy (where a lot of the PizzaGate "scandal" was developing). After Donald Trump's victory, however, we see that their attention shifted to both nationalism in Europe, and trolling and attacking Hillary Clinton supporters - the most popular cross-subreddits include Le_Pen and HillaryMeltdown.
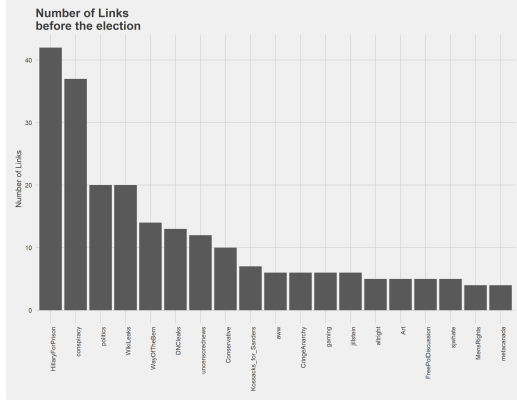
Figure 1: Before Election Day [link]   Figure 2: After Election Day [link]
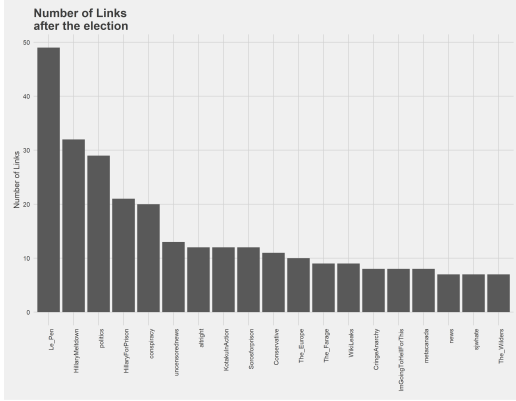
With our weighted graph models, which are presented in Figure 3 and Figure 4, we can see how the networks are connected. Before the election, we can see clearly that subreddits concerning European and American politics are two distinct connected components. This indicates that there are two distinct userbases of The_Donald – Americans involved in American politics and Europeans involved in European Politics. The only thing that unites them is The_Donald. On the other hand, European and American subreddits evolved towards a single connected component after election. This seems to be because the European side got involved in subreddits such as HillaryMeltdown, bringing both components together.
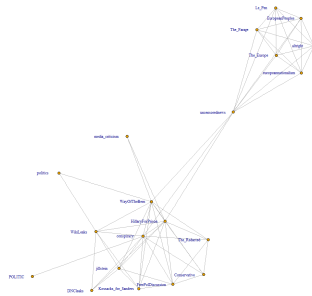

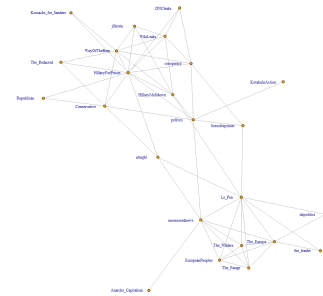Figure 3: Before Election Day [link]   Figure 4: After Election Day [link]

Another possible explanation could be that Donald Trump supporters, galvanized by his big win in America, feel the desire to encourage populism in Europe as well. While this is a compelling theory, it doesn't seem to really be corroborated by the change in the graph structure between Figure 3 and Figure 4. The European contingent is still mainly together, but the new subreddits it reaches in Figure 4 are

SorosForPrison and HillaryMeltDown. This suggests to us that mainstream American supporters aren't branching into European subreddits because much of the large subreddits of the American component remain isolated from the European contingent.

4.2 Inauguration Day

Our preliminary analysis on the time period two weeks before and after election day are shown in Figure 5 and Figure 6. The two most popular cross-subreddits for these two time periods are r/conspiracy and r/politics. This is pretty standard for The_Donald, and there doesn't seem to be any strong conclusion that can be drawn from these charts.
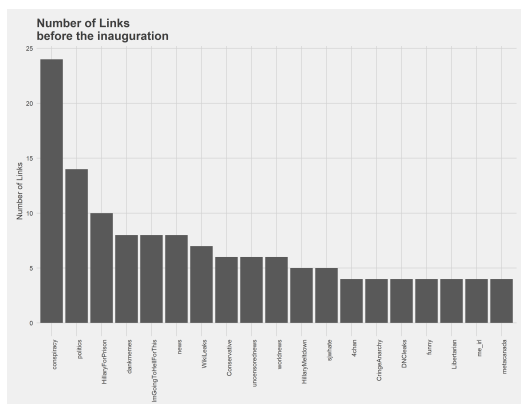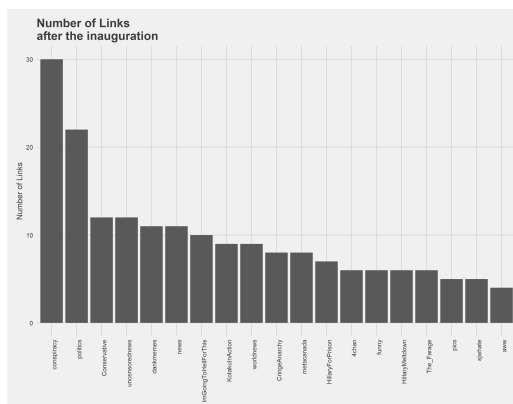

Figure 5: Before Inauguration [link]   Figure 6: After Inauguration [link]

Our weighted graph models before and after inauguration are presented in Figure 7 and Figure 8. These network graphs are much, much sparser than during the election. Very few subreddits appear connected to The_Donald. This is probably explained by the fact that the posts in general on The_Donald did not maintain the same momentum before and after the inauguration as compared to before and after the election. Further study is needed on the thresholds used to create these graphs.
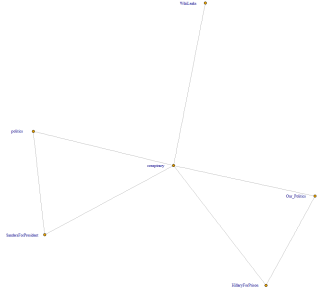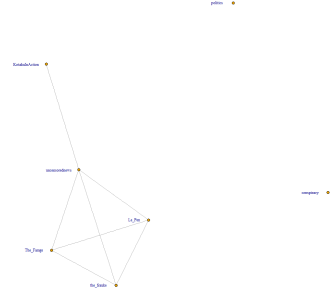
Figure 7: Before Inauguration [link]   Figure 8: After Inauguration [link]

4.3 Algorithm Changes

In Figure 9, the total daily score on The_Donald posts is plotted. The way Reddit works[4] is that there is a "Front Page" that consists of the top posts from the default subreddits. This page appears to users that are logged out, and to users that log in but don't change their default settings. There is also a page called r/all that consists of the top post form all subreddits, not just the defaults. In any particular subreddit, the mods can "sticky" a post so it appears at the top of the subreddit's page (and artificially gain much higher visibility). The_Donald routinely "stickies" posts on their front page so they get lots of upvotes and consequently reach r/all. They did this so often that people grew sick of seeing The_Donald on r/all. So, on or around November 30th[5], Reddit changed their r/all algorithm to specifically eliminate The_Donald's abuse of sticky posts.

On or around February 6th[6] Reddit introduced a new change their algorithm which was basically designed to get rid of The_Donald from r/all altogether. They introduced a rule where if a subreddit has been filtered out of r/all by a certain number of users, it will be permabanned from r/all. So now, if you go to r/all, you won't find any The_Donald posts at all.

We attempted to see how these rule changes affected the total score garnered per day on The_Donald. The data clearly shows that this algorithm change did not decrease The_Donald's total scores[7]. Surprisingly, it seemed the second rule change

---

[4]This description of Reddit does not explain the use of r/popular, which was introduced after our analysis window

[5]https://www.reddit.com/r/announcements/comments/5frg1n/tifu_by_editing_some_comments_and_creating_an/

[6]https://www.reddit.com/r/modnews/comments/5sghb1/introducing_popular/

[7]The massive spike at the end of November was caused by a witch-hunt, evidence-gathering event on The_Donald against the CEO of Reddit. link

actually *increased* total scores. Looking back, the purpose of the rule changes was to stop people *uninterested* in The_Donald from seeing those posts in their feeds. These uninterested people are more likely to *downvote* The_Donald posts they see in r/all[8]. In fact, the moderators turned off the ability to downvote posts as they appear on The_Donald's front page[9]. So, after the second rule change (which occurred just after the high point caused by inauguration day), the chances someone downvotes a The_Donald post basically went to zero[10]. That is the reason caused the spike of score since the rule change.
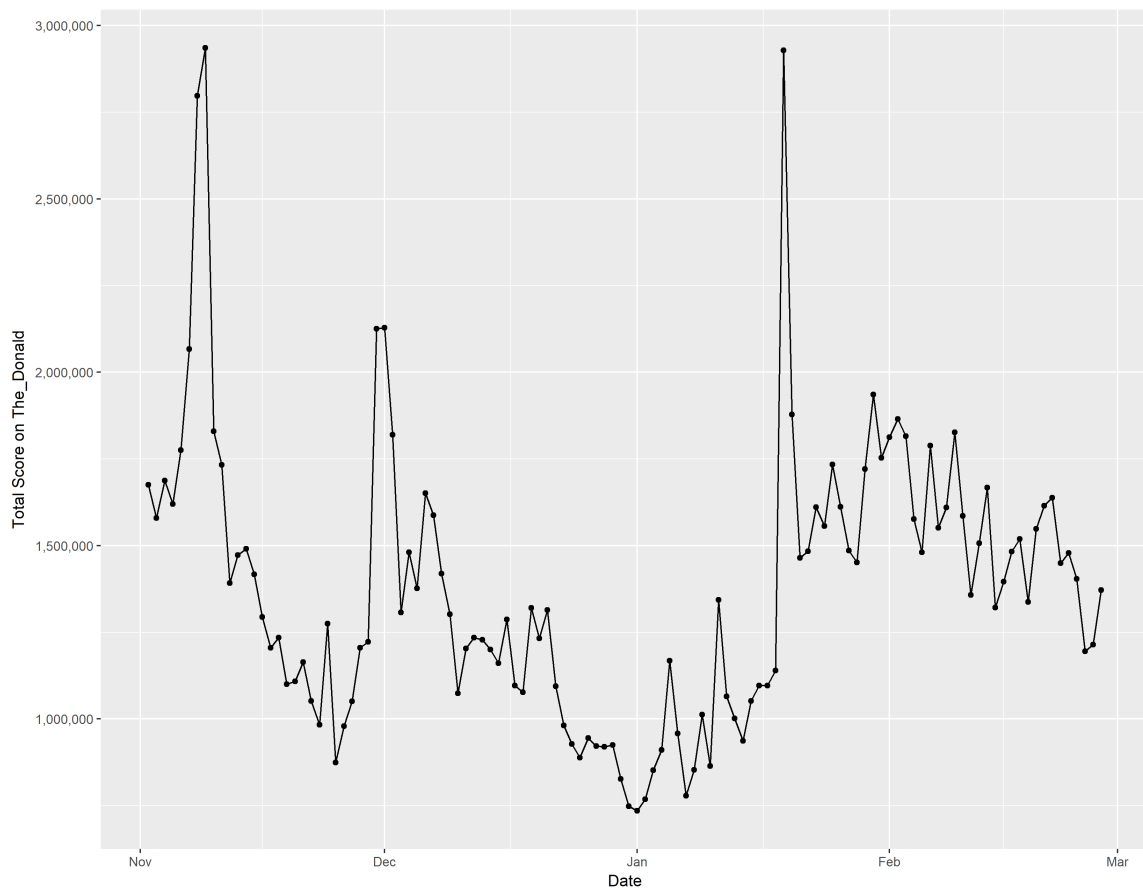

Figure 9: Total Score on The Donald Per Day [link]

---

[8]https://www.reddit.com/r/EnoughTrumpSpam/comments/5fsl1v/number_of_people_filtering_the_donald_the_fuck/

[9]This was done with a custom css style sheet - e.g. the downvote button is gone only when a user is viewing The_Donald's front page

[10]Technically, people still could if they browsed on mobile, but no non-The_Donald user would go out of their way to just downvote

5   Conclusion

We set out to analyze Donald Trump's biggest internet fan club and find out whether we can draw conclusions about the habits of the users there and outside forces affected its popularity. We saw that The_Donald users were very active in the HillaryForPrison immediately prior to the election, and moved straight to HillaryMeltdown after the election, where they congregated to make fun of sad Hillary fans (both of those subreddits are shadows of their former selves now). We also observed that there were distinct subgroups visiting The_Donald before and after the election, mainly European Nationalists and American Conservatives. Both groups joined in the game of making fun of sad Hillary fans after the election.

We also analyzed how Reddit's algorithmic changes affected the scores garnered by posts on The_Donald. We saw that Reddit's attempts to de-emphasize The_Donald posts led to the creation of a sort of echo chamber that led to an overall increase in the scores of The_Donald posts.

Link to Code