Abstract

This paper researches a selection of Tweets obtained in by NBC in the wake of the 2016 election interference campaign by Russian assets. Using social network analysis, topic modeling, and other machine learning techniques we aim to understand the messages and factors which differentiated effective trolls from non-effective trolls. The implication for our work was determining what and how trolls were able to build influence, serving as a blueprint from which to formulate countermeasures. While we were unsuccessful in our effort to identify successful traits, suggestions for future research are reviewed.

1. Introduction

In the lead up to the 2016 presidential election Russia's "Internet Research Agency" (IRA) worked to gain influence on various social media platforms. Since as early as 2009 Russia has used social media to influence domestic politics [4]. The U.S. election of 2016 represents the dramatic realization that these efforts were also occurring abroad. While their role and impact has been debated by those who had the most to gain by the troll activities, the US intelligence agencies stand united in their indictment of the IRA. The trolls aim was to polarize the US electorate and sow disinformation, capitalizing on various conspiracy theories and extremist rhetoric [2]. Much of their efforts were in support of Republican candidates and then candidate for President Donald Trump. However, trolls also targeted non-Republican voters as part of a suppression campaign to lower voter turnout. Former Director of National Intelligence Dan Coats testified that Russia considers its efforts to be successful [2]. The trolls used multiple platforms in their efforts and touted actions to reduce trolling as evidence of bias. In one case, when a troll "black lives matter" Facebook account was suspended, the trolls took to twitter to allege racism and bias at Facebook [4]. Facebooks limited API has challenged efforts in understanding troll actives on the platform. Likewise, Google had provided limited and non-machine readable data into their activities which included buying advertisements. Twitter, by contrast, provided the names of almost 3,000 Twitter accounts which were believed to be connected to the influence effort. These files were released as part of the House Intelligence Committee investigation into Russian influence activities. Once Twitter identified the offending accounts they were suspended, which deleted account data from public view. Additionally, as part of their terms of service, Twitter demands that any other sources of suspended account data be deleted as well. This has limited the ability of researchers to understand the full scale of the effort on Twitter. The methodology used to rebuild the dataset is poorly understood. Developers have been hesitant to publically discuss their efforts for fear of being found in violation of terms and losing access.

Despite the revelations from the 2016 election, little public effort has been under taken to counter it. Responsibility for limiting these attempts rests largely on the social media companies themselves. The current administration refuses to acknowledge the well-established facts surrounding Russian interference. More recent investigations and congressional testimony are clear that this effort continues even today. As recently as 2018 US Cyber Command took Russian trolls offline to limit interference in the 2018 midterm election [5]. Still, this action does not represent a concerted effort or policy to prevent future influence. Rather, it's merely an offensive thrust from an otherwise defensive posture.

Our paper seeks to understand and quantify the efforts of the Russian influence campaign during the 2016 election. In doing so, we aim to provide a blueprint of what makes a troll effective through analysis of their social network graph and messaging. Using machine learning models, we attempt to predict features which may indicate troll success and hopefully lay the groundwork for developing methods to counteract troll behavior. This paper is broken into five additional parts 2: review of current literature, 3: materials and methods, 4: results, 5: discussion, and finally 6: conclusion.

2. Review of Current Literature

Insert lit review here.

Reserved for lit review.

3. Materials and Methods

The dataset used in our analysis was obtained from Kaggle [1]. As part of the congressional investigation into Russian meddling in the 2016 election, social media giant Twitter released the screen names of 3,000 suspected Russian troll accounts it had deleted. NBC news worked to rebuild a database of the deleted tweets and collected information on 200,000 tweets and just under 400 users [2]. Loading the dataset into R studio we plotted the user creation and tweet counts (Figure 1). In an attempt to understand variation in troll activity, large spikes in the number of tweets were manually related to major 2016 election news events [3]. We chose the three largest spikes in tweets to further focus our analysis: An unknown event occurring on 9-17-2016, the release of the DNC WikiLeaks emails on 10-7-2016, and Election Day 11-8-2016. The analysis for each date can be broken into three main components: 1. Network analysis, 2. Feature Modeling, and 3. Natural Language processing.

3.1 Network Analysis

Tweets for each date were put into separate subsets and parsed to identify retweets. The user sending the tweet was parsed as the sending user, the user mentioned in the retweet was parsed as the receiving user. Missing values and those who did not retweet and mention another user were removed. The resulting list of users was then compared against the second provided dataset of known trolls. This aimed to differentiate troll and non-troll users from within the retweet mentions. The sender to receiver relationship comprised the edge list (who retweeted who and how many times) which was then translated into a directed retweet network with a node attribute of troll / non-troll. Self-ties were removed and centrality metrics were calculated.

To better understand troll effectiveness the following centrality metrics were used:

1. Degree centrality: A simple reflect of the number of connections a node has used as an overall engagement measure of popularity. In our directed network, indegree (popularity) and outdegree (socialization) are also used.
2. Betweenness centrality: The total number of shortest paths that run through the node, used to determine the "information brokers" of a network.

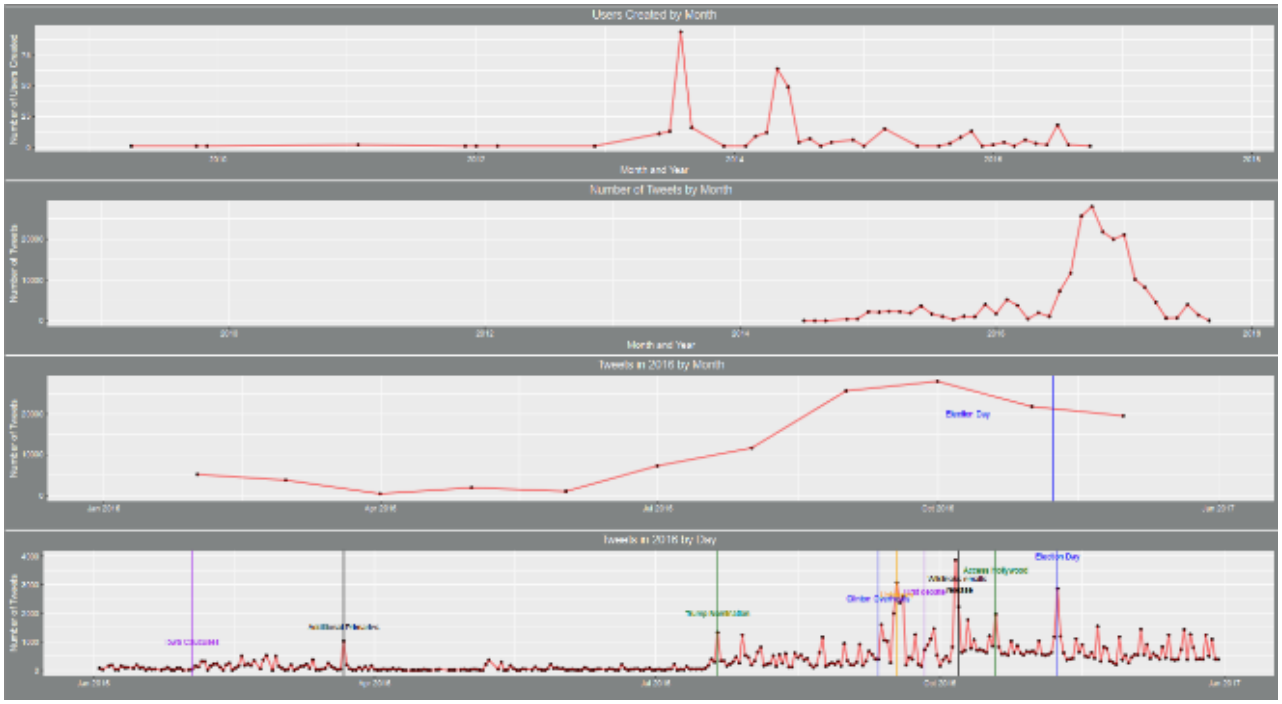

Figure 1: plotting the dataset overview. Top: Users created by month, with many being created in late 2013 and early 2014. Second from top: number of tweets sent by troll accounts by month. Third from top: number of troll tweets sent in 2016 in relation to November 8th 2016, Election Day. Bottom: number of troll tweets by day overlaid by various "major" media events.

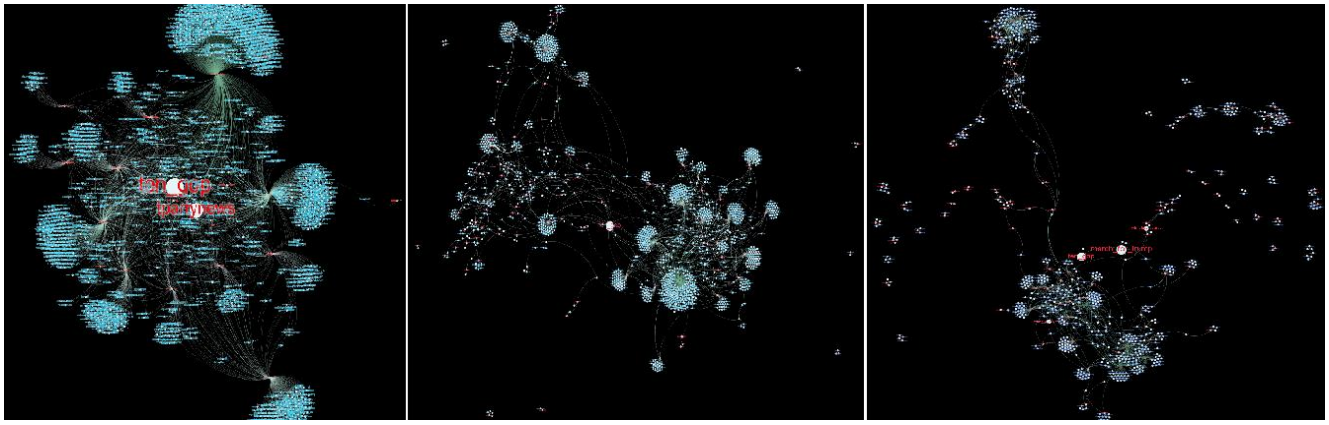

Figure 2: Network graph images from Geph using Force Atlas layout algorithm with node size according to betweenness. Red nodes are known trolls, blue nodes are non-trolls. Left: Network from unknown event 9-17-2016. Middle: The release of the DNC WikiLeaks emails on 10-7-2016. Right: Election Day 11-8-2016. Note: Larger images attached at the end of the document.

3. Eigenvalue: A combination metric of connections, minimum number of hops, and shortest connections

In attempting to determine effective messages and behaviors of trolls, our analysis focused on degree and betweenness measures. A fourth metric, closeness centrality, was not used given the highly disconnected nature of our network. These networks were then exported from R as graphml files and loaded into Gephi for visualization (figure 2). Networks were visualized using the Force Atlas layout algorithm. Troll nodes have user names colored in red, with non-troll users in blue. Node size reflects betweenness centrality.

3.2 Feature Modeling

An assumption of our initial analysis was that influence and popularity of the most successful trolls would trend positively as we approached the election. To make this determination, we evaluated all tweets in advance of our three chosen dates. Centrality metrics were recalculated and troll users were evaluated across dates to create a comparison of the early and late stages of the troll's campaign. For example, degree centrality was calculated on the Twitter network for all tweets sent on or before 9-17-2016, which was subtracted from the degree centrality for all tweets on or before 10-7-2016. This resulted in the percent difference shown in period delta 1, Figure 3. Likewise, degree for all tweets sent on or before 10-7-2016 was subtracted from the degree for all tweets sent on or before 11-8-2016 (period delta 2, Figure 3). Changes in centrality indegree and outdegree were also calculated (Figure 4).

Additional features were calculated in the dataset:

1. percentRT: The percent of total tweets which were retweeted and not original content.

Figure 3: Average centrality metric comparison for trolls in the early delta1 (green) and later delta2 (red) phases of the campaign. Left: Troll percent change in degree. Center: Troll percent change in betweenness. Right: Troll percent change in eigenvalue.

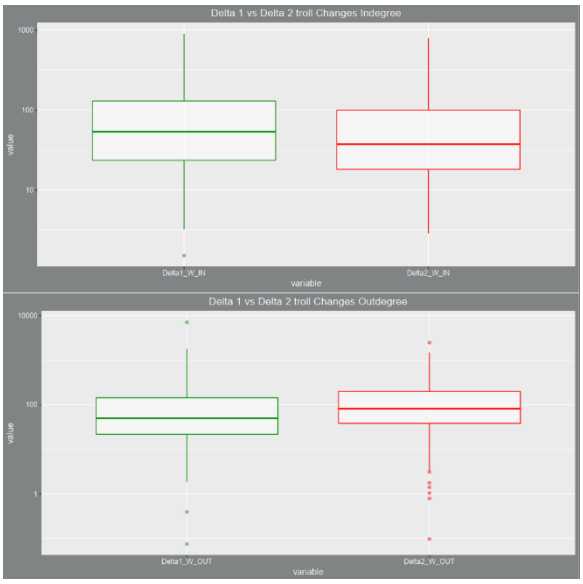

Figure 4: Top changes in troll indegree from early to later phases of the campaign. Bottom: changes in troll outdegree.

2. Retweet_count_mean: For each troll the average number of times their tweets were retweeted / shared.
3. Favorite_count_mean: For each troll the average number of "favorites" their tweets received.

In an effort to predict troll behavior and identify significant features indicating effectiveness, network metrics together with the additional features were modeled against percentRT. Features were combined into a correlation table, assessed for near zero variance, and preprocessed by centering and scaling values for increased regression performance using the caret package. Data was randomly sampled into a 60/40 training/testing subset and evaluated via General Linear Model (GLM), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest (RF) models for best performance.

3.3 Natural Language Processing

For each of the three day specific troll tweet subsets, natural language processing was performed using Latent Dirichlet Allocation (LDA). Tweets were preprocessed using the tm_map function of the R 'tm' text mining framework package to lower case values and to remove punctuation, numbers, and white space. Text was further processed for stemming and stop word removal as defined by the package. Additional text artifacts from web links such as "rt" "amp" and "http" were removed manually before creation of the Document Term Matrix. Given the size of the resulting matrices (a vector of 370Gb) only non-zero documents were retained in the text corpus. With preprocessing complete, the corpus for each date was analyzed to find two topics (k=2) and the top ten terms plotted (Figure 5).

To understand the messages of the most influential trolls, cluster edge betweenness was employed for community detection. Each subset was evaluated as an undirected graph to identify the number of communities in a given subset. Trolls were assigned a corresponding community number and the average betweenness centrality metric calculated for each community. The troll community with the highest average betweenness centrality metric was thought of as a "community of information brokers." To understand the message coming from these influential trolls, text analysis was employed as outlined above. As a benefit of the smaller text corpus for this limited community, LDA tuning was used to

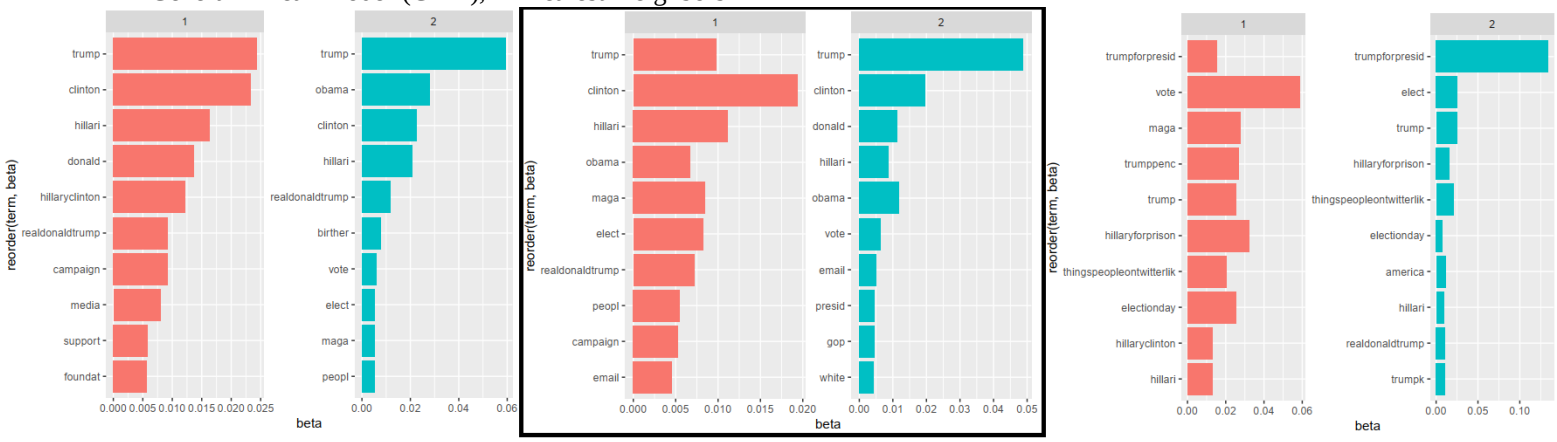

Figure 5: Top 10 terms by beta for each of the two topics analyzed per subset. Left: Terms from unknown event 9-17-2016. Middle: Top terms from the release of the DNC WikiLeaks emails on 10-7-2016. Right: Top terms from Election Day 11-8-2016.

determine the appropriate number of topics from k=2 to k=15 using associated performance metrics "CaoJuan2009," "Arun2010," "Griffiths2004," and "Deveaud2014".

4. Results

In evaluating the network results the outsized influence of a handful of trolls becomes clear (Figure 2). On 9-17-2016 trolls "ten_gop" and "tpartynews" dominate the graph on account of high betweenness measures. Centrally located in the graph, other troll accounts form a ring like structure around them and tweet outwards to relatively isolated non-troll groups of nodes. Topic analysis from that day doesn't reveal the main new story and fails to suggest why such a large spike in tweets was observed outside of election type messaging. In the high beta terms for Topic 1, we see equal references to Trump and Clinton with other terms focusing on campaigns, media, support, and "foundat." Sample tweets from this time shows a focus on the allegation of media bias against Trump, but also furthering suspicions and conspiracies about the Clinton Foundation. In Topic 2, we see a reference to the Obama "birther" conspiracy. Both topics demonstrate a known tactic of the trolls to stoke discord across established divisions with conspiracy and misinformation (Figure 5, Table 1).

On the network graph from 10-7-2019 we again see "ten_gop" as central to the network, with other lesser trolls tweeting out to other unique communities of non-trolls (Figure 2). Topic 1 terms center around references to "Clinton," "campaign," and "email" which aligns with the timing of the WikiLeaks DNC and Podesta emails release. Topic 2 top terms focus on "Trump," "email," and "vote." This suggests an effort to push the email leak story to leverage support for voting Trump (Figure 5, Table 1).

On Election Day 11-8-2016 "ten_gop" is joined by "march_for_trump" with the similar information flow out to disconnected groups of non-trolls (Figure 2). The top term in topic 1 is no surprise, simply "vote" followed by popular hashtags such as "#HillaryForPrison," "#ElectionDay," and even non-election related trends such as "#ThingsPeople-OnTwitterLike." Topic 2 is dominated by a hashtag term "#TrumpForPresident." Other terms show significantly less, but level beta scores for a patchwork of Election Day of hashtags (Figure 5, Table 1).

While the average percent change in degree and eigenvalue increased in the later part of the campaign, betweenness slightly decreased (figure 3). In evaluating degree more closely, we see outdegree as the primary driver of its increase (figure 4).

In evaluation of centrality metrics against percentRT, a measure of how frequently original content was shared, Random Forest was chosen for its best performance (mean RMSE = 0.44, Figure 7). RF feature importance suggested "retweet_count_mean" was of greatest importance follow by degree, "favorite_count_mean" and betweenness (figure 8). However, the final model reflected its composition of insignificant variables and showed poor performance (RMSE = 0.486, R2 = 0.5598, Figure 6).

Focusing in on the community of information brokers for each day, we were not able to identify terms that were materially different from the topics of the wider population. Additionally, the increased number of topics suggested through LDA tuning obscured understanding of the more dominate theme.

5. Discussion

Results support what has been established by other researchers and media investigations into the 2016 election. Russian trolls targeted groups of users

	Topic 1	Topic 2
9/17/2016	@realDonaldTrump "Hillary Clinton has zero record to run on - unless you call corruption positive.." - @IngrahamAngle	RT @jturnershow: Black Birther Proves Hillary & CNN Shows Pure Racism With Handling Of Obama, Cruz Birther (Vetting) Issue #MAGA RT https:/…
10/7/2016	RT @JohnFromCranber: Clintons Used Clinton Foundation as Personal 'Piggy Bank' https://t.co/wwleitqyyG …Corruptocrats #NeverHillary #tcot h…	RT @movement_trump: Retweet if you can't wait for Donald Trump to replace Barrack Obama as President! https://t.co/GgPnaVf2Fk
11/8/2016	@realDonaldTrump #MAGA #TrumpPence16 #HillaryForPrison2016 #BuildTheWall to #TrumpForPresident	#ThingsPeopleOnTwitterLike Puppies Covered with bacon Drizzled with Nutella

Table 1: A sample of tweets from each day which reflect the primary topic terms.

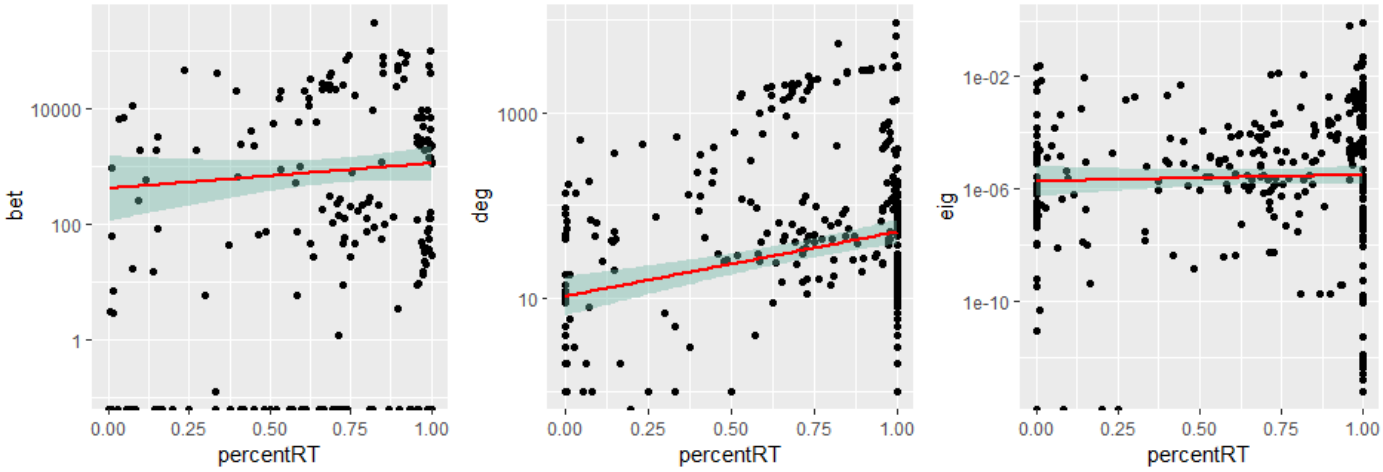

Figure 6: Scatter plot of centrality metrics against percentRT with linear model trending and 95% confidence interval. Left: Betweenness (p = 0.221), middle: Degree (p = 0.110), right: Eigenvalue (p = 0.855)

with polarizing messages and conspiracy theories. It should be noted, that the average troll retweeted 70% of their content. While we did not find a relationship between percent of retweeted content and degree, betweenness or eigenvalue, it suggests that the trolls were not tweeting a purpose built agenda. Rather, they were re-sharing and perpetuating content from other sources. We were surprised in failing to identify a troll or group of trolls whose popularity and influence grew consistently across the campaign. No single troll was conserved in the top 5% for positive degree change across the delta1 and delta2 time periods. It is interesting to note that outdegree specifically contributes to degree centrality increase later in the campaign (Figure 4). While indegree is thought of as a form of popularity as a count of ties coming in, outdegree is the number of ties directed to others. This would corroborate with the overall increase in tweets towards the election (Figure 1). Further researcher into the significance should seek to control for the number of tweets sent to determine the effect of volume verses an increasing number of outbound connections.

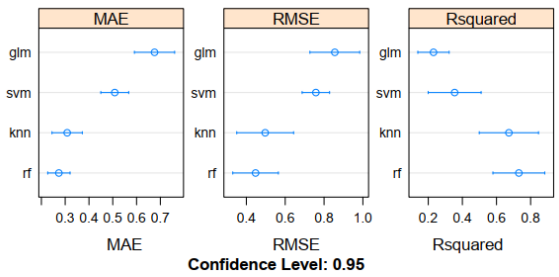

Figure 7: Performance of tested models in predicting percentRT. Random forest chosen for lowest mean RMSE = 0.44, mean R2 = 0.73.

Further study would be needed to review trolls growth in network metrics across the timeframes of the campaign. Doing so may allow a greater understanding of the factors which contributed to gaining popularity and influence. This may also inform on any hierarchy which may have existing among the trolls or the cascading of topics and terms. Upon review of the overall network structure across campaign, it's apparent that high betweenness trolls connected other lesser trolls who then targeted unique subsets of non-trolls. This structure would support the ability to target different messages for different groups of non-trolls, targeting divisions to increase polarization. Our use of community detection methods did not reveal this level of nuance. However, further study would support understanding which topics were distributed to various subgroups of non-trolls.

Missing information from the recreated dataset impacted the effectiveness to create additional feature measures to understand factors of troll success. For example, many of the most successful trolls as measured by betweenness centrality lacked follower counts. It would otherwise be interesting to quantify any relationship between follower counts and network

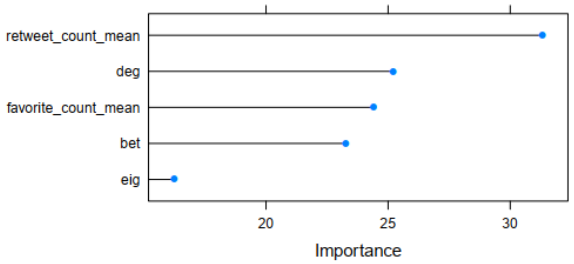

Figure 8: Variable importance as determine by Random Forest

metrics. Additionally we were not able to evaluate the list of followers for known trolls. This may have shed light on the relationships between trolls and more clearly quantify the roll of non-trolls in gaining influence. This type of insights may help contribute to a more significant model of troll success.

6. Conclusion

While our research failed to draw definitive conclusions into characteristics of successful trolls, we have successfully quantified their characteristics of their network graph. Additionally we've offered insight into the primary messages with which they sought to influence the 2016 election. This messaging is in line with expectations from literature and news reports. Finally we propose a number of methods, including more carefully following single trolls across the dataset, and calculation of additional features, which may contribute to the success of future studies seeking to model troll characteristics and behavior.

References

[1] NBC, "Russian Troll Tweets," Kaggle, 15 Feb 2018. [Online]. Available: https://www.kaggle.com/vikasg/russian-troll-tweets. [Accessed 9 Dec 2019].

[2] B. Popken, "Twitter deleted 200,000 Russian troll tweets. Read them here.," NBC, 14 Feb 2018. [Online]. Available: https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731. [Accessed 9 Dec 2019].

[3] P. Bump, "Timeline: What we know about Trump's campaign, Russia and the investigation of the two," The Washington Post, 8 Dec 2017. [Online]. Available: https://www.washingtonpost.com/news/politics/wp/2017/05/30/timeline-what-we-know-about-trumps-campaign-russia-and-the-investigation-of-the-two/. [Accessed 9 Dec 2019].

[4] S. Gallagher, "Massive scale of Russian election trolling revealed in draft Senate report," Arstechnica, 17 Dec 2018. [Online]. Available: https://arstechnica.com/information-technology/2018/12/senate-committee-report-details-how-russians-boosted-trump-across-all-social-media/. [Accessed 12 Dec 2019].

[5] S. Gallagher, "Report: US Cyber Command took Russian trolls offline during midterms," Arstechnica, 27 Feb 2019. [Online]. Available: https://arstechnica.com/information-technology/2019/02/report-us-cyber-command-took-russian-trolls-offline-during-midterms/. [Accessed 12 Dec 2019].

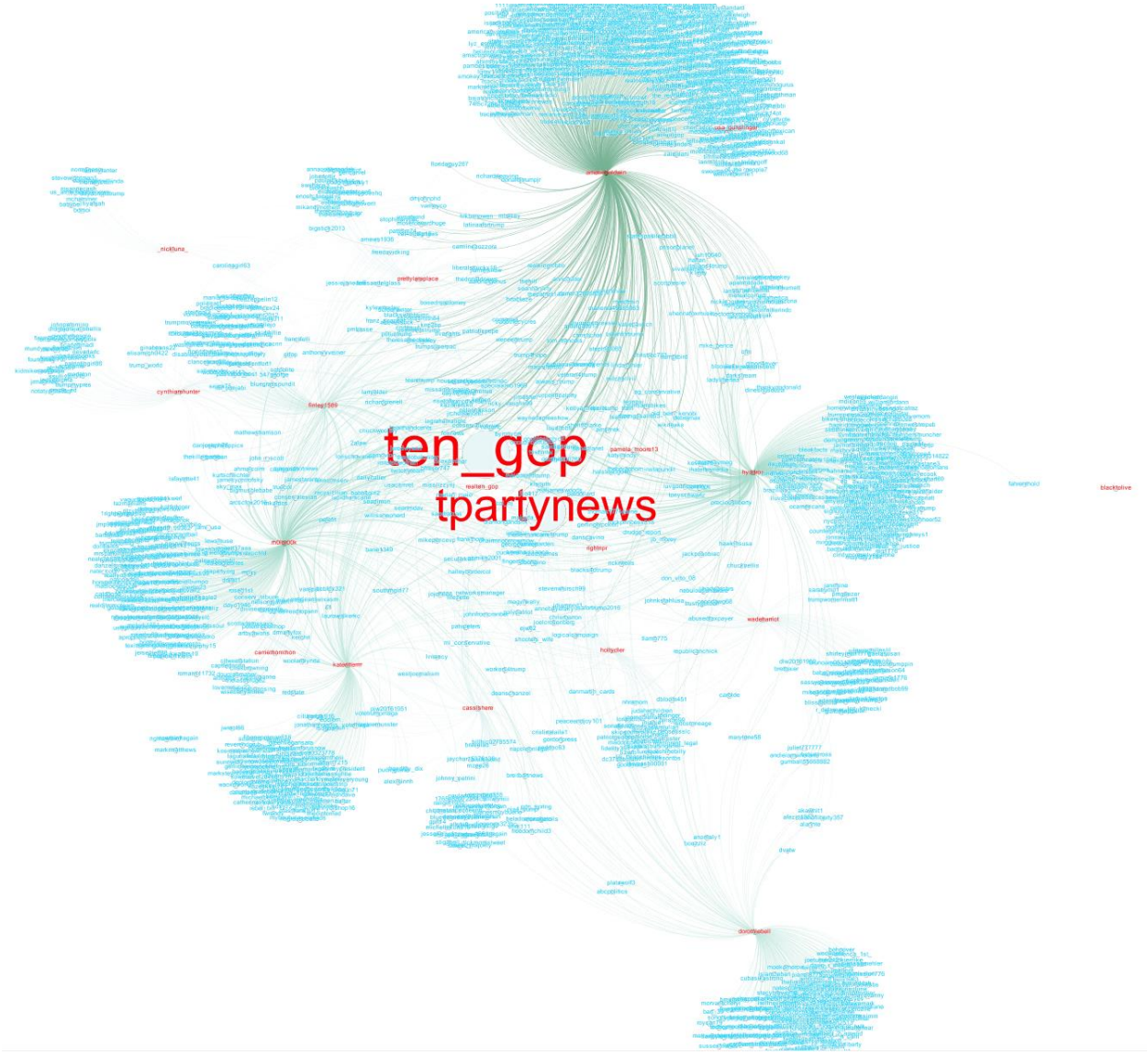

Figure 2 (Left, enlarged): Network graph images from Gephi using Force Atlas layout algorithm with node size according to betweenness. Red nodes are known trolls, blue nodes are non-trolls. Network from unknown event 9-17-2016.

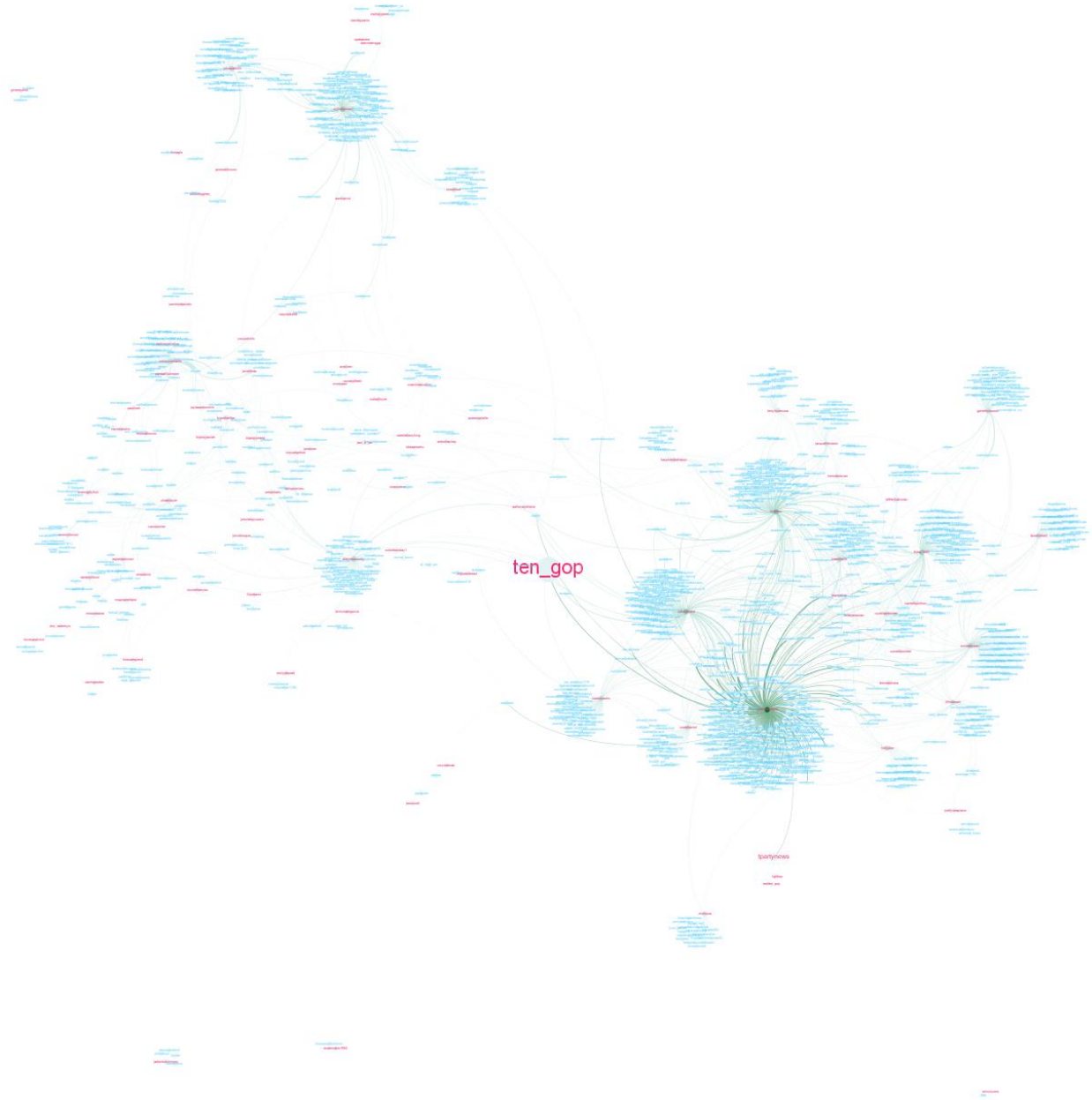

Figure 2 (Middle, enlarged): Network graph images from Gephi using Force Atlas layout algorithm with node size according to betweenness. Red nodes are known trolls, blue nodes are non-trolls. The release of the DNC WikiLeaks emails on 10-7-2016.

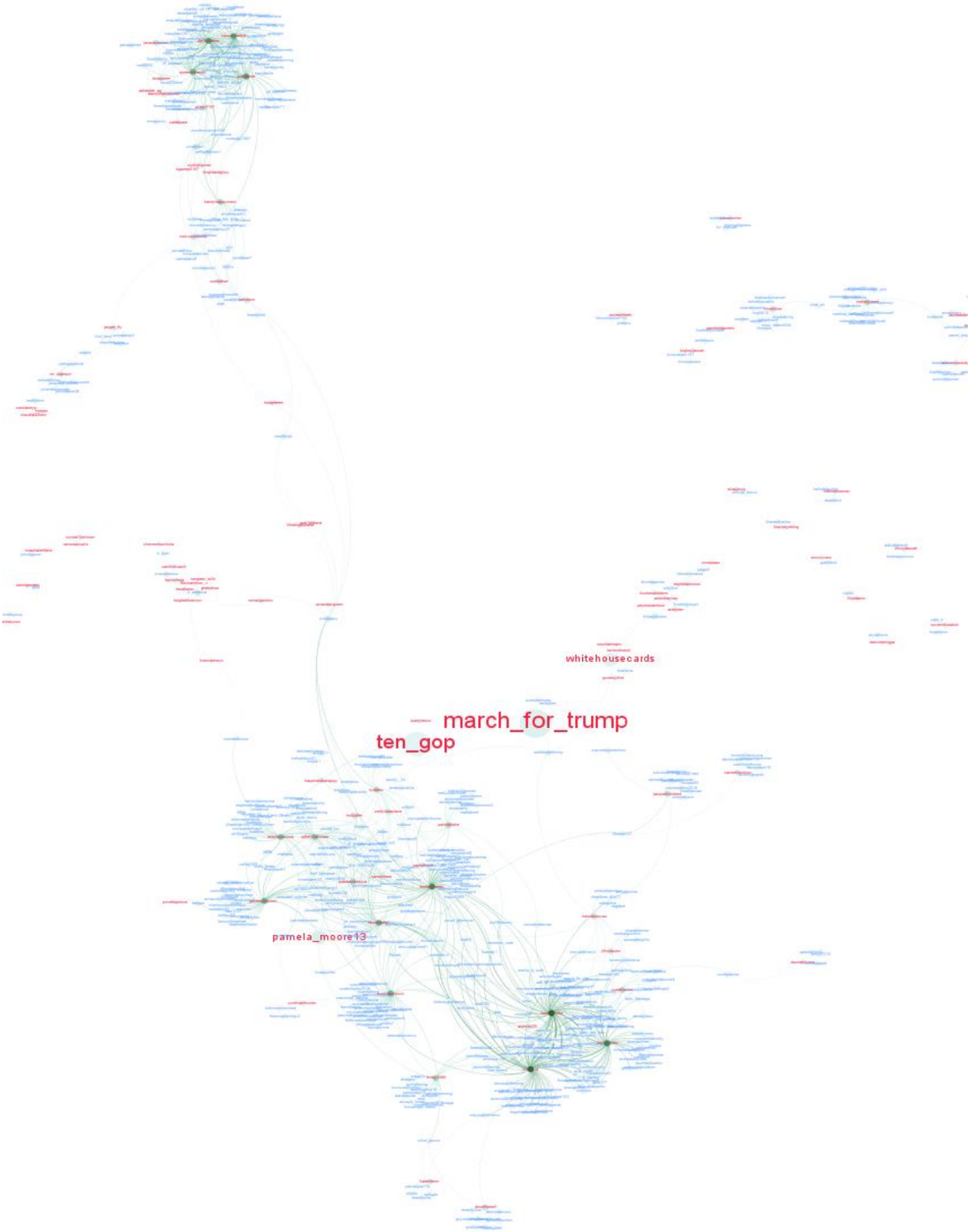

Figure 2 (Right, enlarged): Network graph images from Gephi using Force Atlas layout algorithm with node size according to betweenness. Red nodes are known trolls, blue nodes are non-trolls. Election Day 11-8-2016.