Tracking Kremlin-linked Propaganda Efforts on Social Media During the 2016 U.S. Presidential Election

A Preprint

Shouyang Wang
The Information School
University of Washington
Seattle, WA 98195
wang20@uw.edu

June 7, 2022

Abstract

Following the rise of social media platforms, many information challenges have emerged — one class of challenges centers around political manipulation and information operations that aim to erode social cohesion. Research topics related to disinformation remain a challenge due to the contextual nature of human languages. Qualitative analysis techniques are often impractical to perform repetitively due to the sheer volume of data. This paper describes a graph-based framework that dissects the content of a copious amount of propaganda tweets disseminated by the Internet Research Agency in the one-year timeline leading up to the 2016 presidential election date. We apply community detection to discover Russian trolls that distribute similar hashtags and descriptive analysis to measure the troll network. We found that right-wing content tends to be more diverse in topics. In contrast, left-wing content follows a specific pattern that exploits America's racial wounds. This study provides a more detailed understanding of the trolling content disseminated by the IRA that amplifies political polarization.

***K*eywords** Twitter Hashtags · Russian IRA · Text Network Analysis · Community Detection

1 Introduction and Related Work

The Internet Research Agency, IRA, a Kremlin-linked organization, was accused of conducting strategic social media operations that sabotaged the 2016 Presidential Election. The US Intelligence Committee found the IRA's activities were part of a "*broader, sophisticated, and ongoing information warfare campaign designed to sow discord in American politics and society* (USA v. Internet Research Agency LLC. et al. 2018)."

Despite the classified nature of the state-sponsored online information campaign, the intentions of the IRA are not entirely obscure to researchers outside of the intelligence community. Stewart and colleagues investigated how the Russian trolling messages were disseminated and amplified in the context of the Black Lives Matter movement. Previous analysis results reveal that the IRA aims to destabilize democracy by exacerbating political polarization and amplifying extreme political views that divide the country (Stewart, Arif, and Starbird 2018).

Other researchers show that the trolls targeted many different kinds of US communities. Notably, communities that are sensitive to racial and immigration issues and extreme conservatives (Howard et al. 2018). A copious amount of trolling content was infiltrated into both liberal and conservative discussions. As Pomerantsev, a Soviet-born British journalist would put it:

> *"Unlike in the Cold War, when Soviets largely supported leftist groups, a fluid approach to ideology now allows the Kremlin to simultaneously back far-left and far-right movements, greens, anti-globalists and financial elites. The aim is to exacerbate divides and create an echo chamber of Kremlin support. (Pomerantsev and Weiss 2014)"*

Furthermore, a mix of qualitative and quantitative studies have been conducted on a large scale to study the IRA's political agenda leveraging Twitter data. Researchers identified different trolling categories using unrestricted open coding and axial coding (Linvill and Warren 2020).

Leveraging curated historical Twitter data published by Linvill and Warren, we combine Social Network Analysis and Natural Language Processing (Bail 2016) to link IRA Twitter handles based on hashtag sharing. We apply descriptive social network analysis techniques to measure the network. Further, we partition users into groups using unsupervised network analysis techniques (Blondel et al. 2008). Top TF-IDF hashtags are scrutinized to understand the latent themes across communities.

The goal of this study is to explore the latent structures of the topics disseminated by the IRA. We present a more detailed and granular description of the left and right-wing trolling content. Manipulating a presidential election is detrimental to a democratic society. We offer insights that better inform U.S. policymakers, voters, and social media users to identify political trolling.

2 Data & Methods

In 2018, the U.S. House Intelligence Committee released a list of Twitter handles that are tied to the Russian IRA. Based on this list, 3 million IRA troll tweets and the corresponding metadata were collected, curated, and published (Linvill and Warren 2020). The structured CSV file we use contains 5 columns:

Table 1: Data Description (Linvill and Warren 2020)

Column Name	Description
author	IRA Twitter handle
content	Tweet content
publish_date	Timestamp when the Tweet is published
followers	Number of followers
account_type	Trolling category (i.e. Right Wing)

Leveraging the dataset, the analysis is conducted based on a year's worth of troll tweets leading up to the 2016 U.S. presential election. All hashtags within tweets are filtered out. 10 out of 657 accounts did not use any hashtags in the span of the one-year timeline leading to the 2016 election date, and we eliminate those users in the analysis. Further, 558 out of 657 accounts can be categorized into either left wing or right wing accounts. We label the non-political-leaning accounts as "other". Here we show an example of what the data looks like after wrangling. The author column represents the usernames of Russian troll accounts. The hashtag column contains all of the hashtags tweeted by the account from November 8, 2015, to November 8, 2016.

	author	hashtags
0	4EVER_SUSAN	#Raiders #Carr #RaiderNation #Raiders #MikeAnd...
1	4MYSQUAD	#blacklivesmatter #blm #equality #equalrights ...
2	AANTIRACIST	#targets #iceisis #opiceisis #DemDebate #MLKDa...
3	ABOUTPOLIT	#myfirstTweet #BlackLivesMatter #ConservativeB...
4	ACAB_ZONE	#Arizona #ACAB #Police #Tarantino #CopWatch #C...

Figure 1: Filtered Hashtags

We then construct a network of IRA Twitter handles based on hashtag sharing. The nodes represent IRA troll accounts, whereas the edges represent hashtag sharing.

A hashtag carries a unique meaning on Twitter – it is heavily utilized to denote a specific topic that makes it easier for viewers to categorize. We view a hashtag as the component of a tweet that captures the theme behind the whole message.

We connect two Twitter handles if and only if they have shared hashtags. The edge weight of a connection is mathematically defined as the sum of the TF-IDF (Bail 2016) of all overlapping hashtags. The TF-IDF equation assigns high scores to rare words with high discriminating power and low scores to terms that commonly occur within a corpus. In this model, sharing many rare hashtags that only a group of people tweets significantly boosts tie strength among group members. However, sharing common hashtags only have a minor impact on tie strengths.

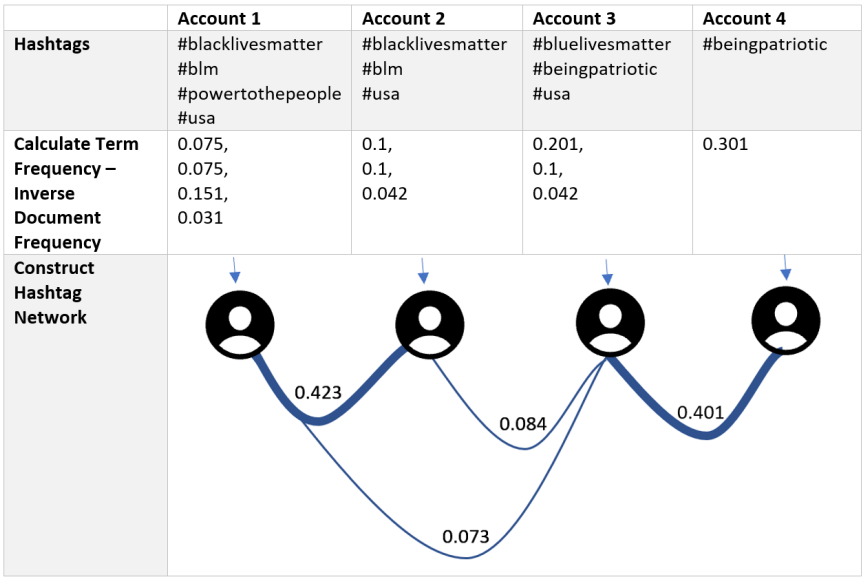

Figure 2: Example: Constructing a Hashtag Network

Classical network analysis techniques are employed to measure the network. A graph partition algorithm for large networks, the Louvain method (Blondel et al. 2008), is employed to detect the number of communities and the members within each community. The Louvain method works by optimizing modularity - a statistic that measures the strength of division in a network (Newman 2010). A community is defined as a subset of nodes among whom there are relatively frequent and intense ties (Wasserman and Faust 1994).

Twitter handles within the same community share many overlapping hashtags, indicating they have been tweeting about similar topics leading to the election. The number of nodes within each community represents the number of accounts the IRA assigns to each broad topic group.

3 Results

3.1 Descriptive Network Analysis

The constructed hashtag network is an undirected and weighted graph, and it has the following properties:

Metric	Magnitude
# nodes	647
# edges	145839
density	0.698
diameter	4
average degree	209
average path length	1.3

In contrast to most observed social networks in the real world, the constructed hashtag network has a much higher density, indicating that hashtag sharing happens exceptionally frequently. The average degree, or the average number of connections per vertex, is also high. The network also exhibits a low average path length, suggesting that the number of edges between two nodes tends to be small.

3.2 Unsupervised Network Clustering

Twitter handles are further grouped using the Louvain clustering algorithm. Seven distinct communities were identified using the unsupervised learning algorithm.

Community Id	# Nodes	# edges	Description
2	194 (29.98%)	17114 (11.73%)	Right Wing
3	171 (26.43%)	12676 (8.69%)	Left Wing
4	92 (14.21%)	3504 (2.4%)	Other
1	83 (12.83%)	3293 (2.26%)	Right Wing
7	63 (9.89%)	1835 (1.26%)	Right Wing
5	34 (5.56%)	560 (0.38%)	Right Wing
6	10 (1.55%)	27 (0.02%)	Right Wing

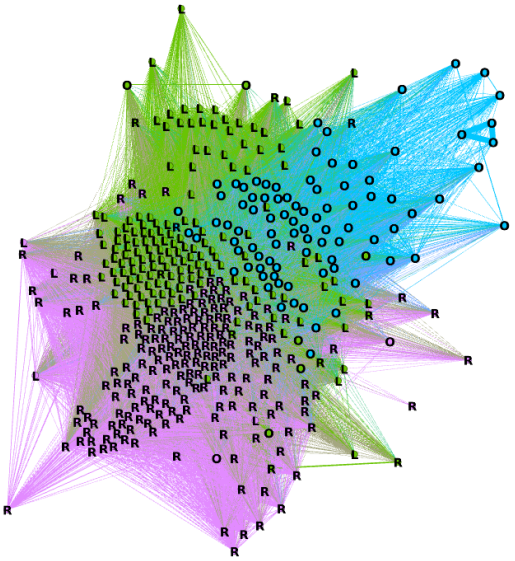

Figure 3: Results Based on the Louvain Method (Community 2,3,4 Shown, Colored by Community Id)

Recall the dataset contains a label column that indicates the political leaning of an account. The labels are generated by performing qualitative analysis (unrestricted open coding and axial coding) on the entire tweet (Linvill and Warren 2020).

Method	# Right Wing	# Left Wing	# Other
Qualitative Analysis (Linvill & Warren)	362	188	97
Hashtag Network Analysis	384	171	92
Percentage Accuracy	93.9%	90.0%	94.8%

We use this column as the labels of the nodes. Notice that the results from the Louvain algorithm accurately align with the labels: purple nodes tend to be right-wing accounts (R), whereas green nodes tend to be left-wing accounts (L). Blue nodes are other types of accounts (O). We observed one community dominated

by left-wing accounts (Community 3), 5 communities dominated by right-wing accounts (Community 2, 1, 5, 6, 7), and 1 community dominated by other types of accounts (Community 4).

Table 5: Top TF-IDF hashtags from the top 3 communites

Community 2	Community 3	Community 4
#veteransday	#blacklivesmatter	#makeamovieedible
#bluelivesmatter	#freekevincooper	#signsyouareamerican
#veteransus	#nopolicebrutality	#myamazonwishlist
#beingpatriotic	#pray4charleston	#obamanextjob
#texassecede	#justice4aiyana	#damnitjimim

Table 6: Communities 1, 5, 6, 7 (Top 5 TF-IDF hashtags)

Community 6	Community 5	Community 1	Comomunity 7
#demndebate	#gopstop	#abcnews	#tcot
#kochfarm	#gopdebatequestion	#abcgopdebate	#pjnet
#turkey	#jebbush	#refugee	#ccot
#turkeyaggressor	#foxbusinessdebate	#news	#conservative
#usda	#rednationrising	#abc	#maga

Tables 4 and 5 show the hashtags with the highest TF-IDF scores from all communities. In community 2, we can observe that many of the hashtags are pro-veteran and pro-police. Moreover, we observe themes that incite strong patriotic feelings and separatism.

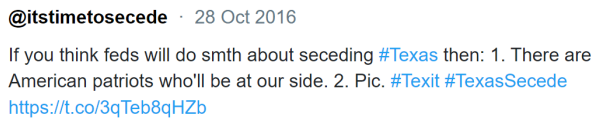

Figure 4: Example Tweet that contains the hashtags from Community 2

In community 3, we see entirely different hashtags that align with left-wing ideologies, such as social and racial justice, inequality, and support for the black lives matter movements. Hashtags such as

- #nopolicebrutality
- #justice4aiyana (Aiyana, an African American girl, was killed by a Detroit police officer)
- #freekevincooper (a presumably innocent African American prisoner)
- #pray4charleston (white supremacist terrorist mass shooting, nine African Americans were killed)

are all related to the black lives matter movement.

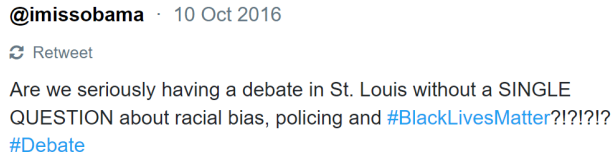

Figure 5: Example Tweet that contains the hashtags from Community 3

In community 7, we see many hashtags related to conservatism and right-wing ideologies.

- #tcot Top Conservatives on Twitter.

- #pjnet The Patriot Journalist Network
- #ccot Conservative Christians on Twitter
- #maga Make America Great Again

Figure 6: Example Tweet that contains the hashtags from Community 7

Community 5 and 1 are about the Republican Party presidential debates.

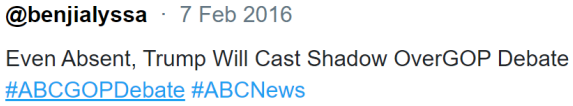

Figure 7: Example Tweet that contains the hashtags from Community 5 and 1

Community 6 is about foodborne illness outbreaks in the US. This community is relatively small with only 10 members.

Figure 8: Example Tweet that contains the hashtags from Community 6

All of the URLs from the tweets within community 6 are linked to one of the following pages:

Figure 9: Community 6: URL 1

Figure 10: Community 6: URL 2

4 Discussion

We construct an undirected network of IRA Twitter handles based on hashtag sharing. The edge weight of a connection is mathematically defined as the sum of the TF-IDF score of all overlapping hashtags two accounts share (Bail 2016). In this model, sharing hashtags would result in 1) a connection between two vertices and 2) marginally increased tie strength. However, not all shared hashtags are assigned equal boosts in edge weight. This approach paves the way for the Louvain community detection algorithm (Blondel et al. 2008) that takes edge weights into account. Several R and Python packages are used to perform data wrangling and analysis. The network is then visualized using Gephi (Bastian, Heymann, and Jacomy 2009).

The findings from this study offer several original perspectives on the latent structure of the content posted by the IRA.

- Analyzing one year's worth of trolling tweets leading up to the 2016 presidential election reveals seven different communities varying in size. Six out of the seven communities have political leaning.
- One out of the seven communities is left-wing leaning, and five communities are right-wing leaning. This suggests right-wing trolls disseminate a much broader range of topics and are significantly larger in terms of community members.
- Analyzing high TF-IDF hashtags within each community reveals left-wing trolls tend to discuss the shooting and police-related incidents that trigger community-wide racial conflict and disturbances explicitly targeting the African American communities. We see a typical pattern in the top TF-IDF hashtags within this community: the victims are all African Americans, and the death of the victims are the results of either police brutality or extreme racism and domestic terrorism.
- Right-wing trolls do not have a specific topic pattern like the left-wing trolls. The topics right-wing trolls spread include advocacy of Conservative and Christian beliefs on Twitter, MAGA, GOP Republican Party presidential debate, police support, US military veterans, patriotism, foodborne illness outbreaks and Texas exit polls.
- By studying top TF-IDF hashtags alone, we can observe a contradiction (i.e., Community 2 vs. Community 3: pro-police vs. anti-police) of hashtag topics disseminated by the identified troll accounts. Preliminary results from hashtag network analysis reinforce previous findings that the IRA tries to spread polarized information and views (Stewart et al. 2018, Howard et al. 2018, Linvill and Warren 2020, Pomerantsev and Weiss 2014).
- The network visualization with previously labeled data suggests that the methodology can be used to differentiate right-wing accounts from left-wing accounts with relatively high accuracy. Given a new dataset, we can predict the number of Twitter handles the IRA assigned to left-wing versus right-wing trolling by examining the hashtags within each community.
- The highly scalable quantitative approach has an advantage when qualitative coding is impractical due to the sheer volume of data that needs to be analyzed manually. Researchers will not always have the opportunity to obtain curated and well-labeled information. It is only sensible to assume that the Kremlin's propaganda machine will spread trolling content at an even larger scale and faster rate in the foreseeable future. We provide an extremely fast way of using computers to analyze, summarize and extract insights from millions of tweets related to propaganda and disinformation without performing qualitative analysis.
- By identifying members within different communities, we break down a large text corpus into many segments, making qualitative analysis more manageable. Qualitative analysis and topic modeling techniques can be further applied to each individual community.

5 Limitations and Future Work

To our knowledge, this study has several key limitations. The limitations we describe in this section also demonstrate improvements that can be made in the future.

First, the dataset we used for analysis is a subset of the original data. It only contains a year's worth of tweets leading up to the 2016 election. The interpretations of this research should only be applied to the topics discussed during the 2016 election by the (now banned) Russian troll accounts. It is expected that the IRA will adopt new strategies to infiltrate the US social media network via other covert means.

Second, we assume that hashtags do an excellent job capturing the content of the tweets. The construction of the hashtag sharing network has one main disadvantage: not all tweets contain a hashtag, and not all troll accounts heavily utilize hashtags. 10 out of 657 accounts (1.5%) did not use any hashtags in the span

of the one-year timeline leading to the 2016 election date. 52 out of 657 accounts (7.9%) used less than 10 hashtags. It is sensible to assume that a tweet can contain a topic of interest without including hashtag information. A text network can be constructed based on raw textual data rather than filtered-out hashtags. One challenge with using raw tweet content is that social media textual data are known for their messiness. The data cleaning method directly impacts the final results. Constructing a hashtag network does not require significant data cleaning (i.e., removing stop words, URLs, emojis, filtering based on part-of-speech tagging, etc.). Hence, a hashtag network is a simplified approach and should be regarded as a very high-level summary of the actual content being disseminated.

Lastly, it is highly appealing for researchers to apply network analysis and NLP techniques to study (dis)information propagation and monitor social movements on social media. Ultimately, a significant amount of qualitative inspection is mandatory to get valuable and practical insights that are useful to policymakers because much disinformation is contextual. TF-IDF and other text preprocessing methods are quantitative statistical techniques that summarize large-scale human language data. When grouping users, the Louvain community detection method does not consider semantics, word order, or linguistic nuances. Not all propaganda is disinformation, and there is so much truth mixed with lies, which makes it very challenging to use one model or approach to accurately depict the whole agenda of foreign propaganda machines. Meaningful research mandates multidisciplinary collaboration and rich literature reviews from both social and data sciences perspectives.

References

Bail, Christopher Andrew. 2016. "Combining Natural Language Processing and Network Analysis to Examine How Advocacy Organizations Stimulate Conversation on Social Media." *Proceedings of the National Academy of Sciences* 113 (42): 11823–28. https://doi.org/10.1073/pnas.1607151113.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008.

Howard, Philip N., Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. 2018. *The IRA, Social Media and Political Polarization in the United States, 2012-2018.* Project on Computational Propaganda.

Linvill, Darren L., and Patrick L. Warren. 2020. "Troll Factories: Manufacturing Specialized Disinformation on Twitter." *Political Communication* 37 (4): 447–67. https://doi.org/10.1080/10584609.2020.1718257.

Newman, Mark. 2010. *Networks: An Introduction.* USA: Oxford University Press, Inc. https://dl.acm.org/doi/10.5555/1809753.

Pomerantsev, Peter, and Michael Weiss. 2014. Institute of Modern Russia (New York, N.Y.). https://lccn.loc.gov/2015433465.

Stewart, Leo Graiden, Ahmer Arif, and Kate Starbird. 2018. "Examining Trolls and Polarization with a Retweet Network." In *Proc. ACM WSDM.* https://faculty.washington.edu/kstarbi/examining-trolls-polarization.pdf.

USA v. Internet Research Agency LLC. et al. 2018. United States district court for the district of Columbia. https://www.justice.gov/file/1035477/download.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications.* Structural Analysis in the Social Sciences. Cambridge University Press. https://doi.org/10.1017/CBO9780511815478.