Memo: *Heads will Troll: An analysis of links shared by IRA Twitter accounts*
Author: Joshua A. Tucker[1]

Note: This memo draws from a number of ongoing projects within the NYU Social Media and Political Participation (SMaPP) lab, including work that is co-authored with Leon Yin, Franziska Keller, Yevgeniy Golovchenko, Cody Buntain, Gregory Eady, Megan Brown, Sergey Sanovich, Denis Stukal, Richard Bonneau, and Jonathan Nagler.[2] When sections draw text directly from other publications or works in progress, it is noted at the start of the section; if you want to quote from these sections, please cite the original work.

Executive Summary: The purpose of this memo is to examine the strategy of the Russian Internet Research Agency (IRA) "troll" accounts in the 2016 US elections. I situates this activity within a larger framework for conceptualizing the ways in which (competitive) authoritarian regimes can respond to online opposition domestically, and the ways in which some of these same tools can be used to conduct foreign operations. After summarizing extant research from other research teams, I report on recent analyses from the SMaPP lab of the "links" shared in the tweets of IRA trolls. We examine both the types of links shared (e.g., links to "junk" news vs. local news sources vs. national news sources) as well as the political orientation of the media sources to which the IRA accounts linked. We aim to assess whether the empirical data better supports the (somewhat) prevalent media narrative that the IRA was simply trying to exacerbate existing political divisions by playing both sides of the game as opposed to trying to help Donald Trump do better in the election. Preliminary analyses suggest a somewhat sophisticated strategy of relying on local news sources, sharing links to both liberal and conservative media sources for most of the campaign, but with an increase in links to conservative sources close to the election and an overwhelmingly of links to conservative (as opposed to liberal or moderate) YouTube videos.

Regime Options for Responding to Online Opposition

In Sanovich et al. (2018), we lay out a tripartite classification strategy for the ways in which authoritarian and competitive authoritarian regimes can respond to online opposition: offline response; online restriction of access to content; and online engagement.[3] We start with this information because it is important to realize that Russian external operations in the online domain were not simply crafted *de novo*; they take place in a larger context of domestic online strategy that predated the 2016 US elections.[4]

---

[1] Professor of Politics, Director Jordan Center for the Advanced Study of Russia, Co-Director, Social Media and Political Participation (SMaPP) Lab: (https://smappnyu.org/), New York University.
[2] https://smappnyu.org/wp-content/uploads/2018/11/SMaPP_Data_Report_2018_01_IRA_Links_1.pdf; Sanovich et al. 2018; Golovchenko et al. (nd).
[3] Our classification system, while not identical to, overlaps substantially with Roberts (2018) classification systems of *fear, frictions,* and *flooding*. For a concise summary of her approach, see Tucker et al. (2018).
[4] For more details on Russian domestic strategy in this realm and its development throughout the 21st century, see especially the appendix to Sanovich et al. (2018) as well as Sanovich (2018). Of course, Russia's strategy in responding to domestic online opposition itself takes place against a backdrop of a long line of Soviet tactics for addressing domestic opposition; for interesting discussion in this regard, see the video of an event on *kompromat* at

Offline responses refer to actions that can be taken offline in an attempt to counter potential opposition online. One category of response would be action in the legal sphere, such as requests for platforms to remove information, changes to legal codes to address questions of liability for online content and government powers to order the removal of content, or attempts at changing the ownership of online platforms to install directors/owners that are more loyal to the state. Offline responses can also involve tactics more commonly associated with repression, such as threatening online opposition figures, arresting them, or perpetrating acts of violence against them.

Online restriction of access to content refers to what we would more typically think of as censorship. Here the goal is to prevent people from seeing potentially objectionable content by blocking access to it. Tactics to do so can be more blunt -- including shutting off access to the internet, blocking access to particular platforms (e.g. Facebook) or websites (e.g. Al Jazeera), or, if the state wants to avoid direct attribution, denial of service (DDoS) attacks to accomplish similar goals for short periods of time – or targeted, such as removing/blocking individual posts. As Roberts (2018) notes, this type of censorship does not even have to be dichotomous (e.g., the post/site/platform either is available or not), it can also involve inserting "friction" into the process of finding information by simply slowing down the accessibility of information.

Online engagement is perhaps a more novel tactic of the digital age, and involves trying to shape the online conversation in a more pro-regime direction. The variety of actors that can be employed to do so are presented in the following section, but run the gamut from forms of political engagement on the part of elites that would in many ways be considered normatively desirable in democracies to the much more nefarious techniques relying on deception that formed the basis of the IRA efforts in the 2016 US election.

Types of Online Actors for Advancing State Interests

Regimes can employ a variety of actors in an attempt to shape domestic online conversations, some of which can also be harnessed for attempting to shape foreign online conversations as well:

Legitimate government actors truthfully identifying themselves. Perhaps the most obvious – and most forthright – way for state actors to attempt to shape the online conversation is to simply join in that conversation as themselves. Indeed, part of the original promised of the "e-government" era was that it would reduce barriers between citizens and elites and create a two-way conversation that was previously impossible.[5]

NYU Jordan Center for the Advanced study of Russia featuring Keith Darden, Katy Pierce and Miriam ???: ADD LINK.

[5] See for example the Twitter feed of Cory Booker, who is renowned for not only having a large number of Twitter follows – like numerous politicians – but for repeatedly engaging in conversations with them. I myself experienced this once when I was able to secure an interview with a Booker staff member by simply tweeting at Booker one evening while he was traveling back from Washington DC to New Jersey.

Legitimate actors outside the state truthfully identifying themselves. The state can of course benefit from actors outside the state who share pro-regime views online. To the extent that this is truly organic it can help shape the online conversation in a way in which the state appreciates, although it is a little more difficult to call this a regime "strategy" if the regime truly exerts no control over the process. As the regime begins to exert more control over the process – say through sending instructions to youth groups or "patriotic" citizens – the extent to which such actors can be part of the regime's tool kit increases, but of course the transparency of such arrangement can become an issue. Such concerns are further exacerbated when actors are paid on behalf of the state to advocate for the state in public online fora.

State (or state funded) actors who do not legitimately identify themselves or falsely represent their identities.

Finally, we come to the unambiguously deceptive form of online actor: someone who portrays themselves as someone they are not. This can take the form of simply masking one's affiliation as an agent of the state, or can take on more pernicious forms of adopting entirely fake persona in an effort to feign membership in a particular community. As we learned in 2016, this can also involve an agent from one country attempting to portray themselves as a resident of another country.

Bots vs Cyborgs vs. Humans/Trolls

It is also important to note that deceptive accounts can be controlled in a variety of ways. If the account produces content entirely based on automated algorithms, the convention is to refer to the account as a "bot". If a human being controls a deceptive account being put to a political purpose, we generally call that account a "troll"; certainly, this is exactly to what people are referring when the discuss the behavior of the "IRA troll accounts": accounts controlled by Russian actors attempting to participate in political activity in the context of the 2016 election by pretending to be something other than a Russian state actor.[6] Cyborgs are accounts that produce both automated and human generated content. There is no further discussion of cyborgs in this memo, but it is likely that closer examination will reveal some of the Russia IRA troll accounts to actually be cyborgs.

How do We Locate Bots and Trolls?

Finding political bots online is a whole research area unto itself, but for the purpose of this memo – to contrast with the source of our information about Russian IRA trolls – we provide a very brief conceptual overview here. Basically, bot detection techniques revolve around having some "ground truth" of identified bot accounts, using machine learning models to learn the features that distinguish bots from non-bot accounts, and then using those models to identify additional

[6] Note that the term "troll" is also applied more generically to online – almost always anonymous – actors that harass others online. Sometimes this can be done purely for the enjoyment of the troll; other times it is done for expressly partisan purposes. Moreover, some trolls will claim to be acting as part larger ideological framework, e.g., fighting against a hypocritical conformist online culture, even when such activity is not partisan in the classical sense. For the remainder of this memo, we use this term to as described in the text of the memo.

accounts that are predicted to be bots.[7] Finding "ground truth" examples of bots, however, is a challenging task. One option is to rely on leaked data and/or collaborate with someone who has actually built political bots; our approach has been to train human beings to identify accounts that look as if they are bots.

The reason these types of methods can work well is because we believe bots provide a distinctive cyber-footprint due to their automated nature. While in theory it should be possible to use similar methods to detect troll accounts, extant work on the Russian IRA trolls has involved a very different approach to identifying these accounts: a list of 3,841 Twitter accounts identified by Twitter as being Russian IRA accounts (which entered the public domain when they became part of the congressional record following testimony by Twitter executives), and then a subsequent collection of all the Tweets from these accounts released by Twitter.[8]

The advantage of relying on the list released by Twitter in our (and others) analyses of Russian IRA troll behavior is that we are able to skip the step in the bot detection methods of having to train machine learning models; nor do we have to figure out a way to come up with our own "ground truth" of which accounts really are Russian troll accounts. The downside, however, is that we are completely reliant on Twitter's methods for identifying troll accounts (which we do not really know), and – if one uses the data released by Twitter – then you are also trusting Twitter to have released all of the tweets. This is worth keeping in mind while reading the rest of the report.

What Have We Learned From Other Analyses of Russian Troll Data?[9]

According to the official assessment published by the Office of the Director of National Intelligence in 2017, the Kremlin's strategy during the 2016 Election was not only "to undermine public faith in the US democratic process," but also to damage Hillary Clinton's campaign and to support Donald Trump's campaign. The methodology of the investigation by intelligence services remains opaque because it is unclear how the information on the Internet Research Agency was collected (Department of Justice, 2018a). However, the same indictment simultaneously points towards an opposing direction. The Department of Justice accuses Russian entities of organizing protests against Trump, such as the "Charlotte Against Trump" protest on November 19, 2016 in North Carolina (Department of Justice, 2018a, 23). In line with this, the Department of Justice accuses the Russian agency of "spread[ing] distrust towards the candidates and the political systemin general" (Department of Justice, 2018a, 6). This has become the central debate over activity of Russian trolls: were they trying to "sow discord", elect Trump (or at least reduce the size of Clinton's victory if electing Trump was unlikely), or some combination of both?

[7] See in particular Botometer (https://botometer.iuni.iu.edu/#!/) and Stukal et al. (2017). These are known as "supervised learning" methods. It is of course possible to work in the opposite direction, by identifying features one would think would likely identity a bot, using those features to find potential bots, and then deciding whether the model seems to be correctly finding bots.

[8] https://blog.twitter.com/official/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html

[9] This section is a condensed version of materials found in the current draft of Golovchenko et al. (nd).

Like the Department of Justice, academic research has also presented a mixed picture. Zannettou et al. (2018) analyzed 27K tweets from one-thousand Russian trolls on Twitter's list. The researchers report that at least 10.3% of the trolls portray themselves as Trump supporters by writing "trump" and "maga" in their profile names and self-descriptions (Zannettou et al., 2018, 4). Badawy, Ferrara and Lerman (2018) analyzed tweets from 221 Russian trolls disclosed by Congress together with users who engaged with the trolls. Using an automated approach to estimating account ideology, they argue that 107 of the trolls are "liberal" while 108 are "conservative"[5]. This supports the hypothesis that the Internet Research Agency promoted both the both sides of the ideological spectrum. However, they also show that the conservative trolls are more active because they posted 844 original tweets while the liberal trolls posted only 44 original tweets in the researchers' dataset. Furthermore, they argue that the trolls had a "mostly conservative, pro-Trump agenda" based on the text analysis of the most frequent, stemmed words in the tweet corpus posted by the trolls. The same paper argues that conservative users are 31 times more likely to retweet the trolls than liberal users.

Linvill and Warren (2018)'s study also suggests that right-wing IRA trolls on Twitter produced more content than left-wing trolls; although, the difference is not as large as suggested by Badawy, Ferrara and Lerman (2018)'s work. Of the 1,311 troll accounts analyzed by Linvill andWarren (2018), 617 were manually coded by the authors as right-wing and 230 as leftwing, while the remaining 464 were classified using other categories such as as "Fearmonger", "Hashtag Gamer" and "News Feed". Although left-wing trolls were more active on average, the Russian agency generated approximately 64 percent more tweets using right-wing accounts than the left-wing accounts on an aggregate level (716 and 437 tweets per day respectively).

Where Zannettou et al. (2018) and Linvill and Warren (2018) focus on the individual characteristics of troll accounts, Steward, Arif and Starbird (2018) examine the Internet Research Agency trolls through a network perspective. The researchers in this study identified 96 trolls (from the aforementioned list released by the Congress) who posted tweets related to both Black Lives Matters and shootings. They show that the troll accounts successfully infiltrated both the left-leaning and right-leaning parts of the retweet network, thus adding to the pre-existing polarization between the two groups. However, they also argue that the fake accounts are slightly more prevalent in the left-leaning cluster of retweets (Steward, Arif and Starbird, 2018, 2-5).

Most recently, Howard et al. (2018) analyzed a range of data sets across platforms of content posted by IRA and argued that the IRA did indeed target both liberals and conservatives on Facebook and Instagram and noted that that "the IRA sought to energize conservatives around Trump's campaign and encourage the cynicism of other voters in an attempt to neutralize their vote" (Howard et al., 2018, 32). In other words, the goal behind engaging with liberals and conservatives was to support the conservative side of the presidential campaign. In their analysis of the IRA's presence on Twitter, Howard et al. (2018) find that the agency targeted conservatives more than liberals in early 2015; however, the gap closed later that year. Liberal and conservative levels of activity were similar throughout the presidential campaign until early 2017, where-after the authors observe a surge in activity targeted towards the conservatives (Howard et al., 2018, 26). According to an analysis for the United States Senate Intelligence Committee, carried out by New Knowledge, 96 percent of the content on 1,107 videos uploaded on the 17 fake YouTube channels by the IRA were thematically related to "Black LivesMatter

police brutality" (DiResta, 2018, 16). This data is not yet publicly available, nor does it entail the IRA's use of YouTube content generated by ordinary users outside of the Russian agency.

SMaPP Lab Research on Links Shared by Russia Trolls: Sources of Links[10]

In our research at the SMaPP lab, we have taken a different approach by focusing not on the text shared by the IRA trolls nor the networks in which they were enmeshed, but rather on the *links* shared by the troll accounts.

Our first study drew on the dataset shared online by Twitter's Elections Integrity Initiative of more than 9 million tweets sent by approximately 3,600 IRA-linked accounts in 2016. As we were primarily interested in online activity in the US and tweets containing links, we discarded tweets without links and accounts that posted in Russian, leaving us with 556 accounts that tweeted approximately 209,000 links between January 2016 and November, 2016.[11]

We also collected data for two relevant comparison groups -- politically engaged users and random users -- over the same period of activity (January through November 2016). The sample of random Twitter users contains 1,344 accounts that tweeted approximately 106,000 links; the sample of politically engaged users encompasses 1,952 accounts that shared roughly 437,000 URLs.[12] We found that:

- IRA-operated accounts shared **1.5 times** as many links to **junk news** websites, such as *Breitbart*, *Truthfeed* or *Raw Story,* as the most active comparison groups, but the overall share of junk news among all links shared was low (6 percent).[13]

- The percentage of junk news websites the IRA shared **spiked sharply** in September and October 2016, that is, in the weeks immediately leading up to the November 8, 2016 presidential elections.

- The troll accounts heavily banked on **local news** outlets when sharing articles online: 30 percent of all links led to local media content, potentially exploiting the added trust that local news sources enjoy in the US as well as highlighting true events that might be thought to increase political polarization. This was 15 times the proportion of local news

[10] The text in this section is a lightly edited version of text found in Yin et al. 2018 (https://smappnyu.org/wp-content/uploads/2018/11/SMaPP_Data_Report_2018_01_IRA_Links_1.pdf).
[11] Many URLs in tweets are shortened by link shorteners (e.g bit.ly, ow.ly). As a pre-processing step, we unshorten all such links using a new Python library created for this purpose and now available as an Open Source package, *urlExpander*.
[12] Politically engaged users are those that tweeted about Donald Trump or Hillary Clinton at least twice during the 2016 presidential election race. The sample contains 1,952 accounts that shared 437,091 URLs. The group of random users is determined by a random number generator that validates if a user with that ID exists on Twitter. The sample contains 1,344 accounts that tweeted 106,416 links. Because the IRA sample and the comparison groups are not of the same size (there are more politically interested and random users than IRA accounts), we only compare *relative* activity, not total activity, across groups.
[13] For our "junk news" category, we use Merrick College's OpenSources dataset to identify websites that are known to produce content containing entirely false information, extreme bias, conspiracy theories, hate-based discrimination, clickbait, rumors, state-sponsored news, or junk science.

shared by our political interested comparison group, and **90** times as much as our random sample of US Twitter users.

- At least 27 IRA accounts **posed as local media** outlets and were responsible for 80 percent of the local news content shared. These accounts heavily relied on social media managers to automate their activity.

- Whether or not a state was a **"swing state"** in the presidential election does not appear to explain whether it was heavily featured in the distribution of local news by IRA accounts.

- National news sites most frequently shared by IRA-linked accounts include the Congress-focused website *The Hill,* the *Washington Post* and the *Chicago Tribune.* The conservative news outlet *Fox News* (6th most popular) was shared more often than the *New York Times* (8th most popular). The IRA's favorite websites differ from those shared by the comparison groups: Both political users and random users most frequently shared *CNN* content, followed by the *New York Times*. *Fox News* only comes in as the 8th most popular national news outlet for politically engaged users, and ranks 9th for random users.

Preliminary SMaPP Lab Research on Ideology of Links[14]

In a second (in progress and preliminary) study, we draw on a large collection of politically-oriented tweets collected in the lead-up to and after the 2016 US Presidential Election. This data set was collected through Twitter's streaming API from November 5, 2015 to December 31, 2017. From this collection, we extracted all tweets and retweets authored by the IRA accounts identified by Twitter, yielding 108,781 tweets from 1,052 unique IRA accounts. From these tweets, we extracted 30,662 unique URLs sent by IRA members across 2,002 unique domains, 10,450 of which link to news stories by national media organizations, and 855 of which link to YouTube videos across 315 YouTube channels.

We then use an automated method (described in Golovchenko et al. (nd) and Eady et al. (nd)) to estimate the ideology of different news websites based on sharing behaviors of members of Congress and the general public.[15] This method allows us to produce a continuous measure of ideology, but for the purpose of the table below, we then collapse media sites into bins for "Liberal", "Moderate", and "Conservative", such that any media site with a score less than the Washington Post is liberal, any media site with a score greater than the Wall Street Journal as conservative, and anything in between is moderate. We are currently working on an automated version of estimating the ideology of YouTube channels as well, but as this method is still in development, for the current study we relied on human coding (three trained undergraduates). The intercoder reliability of this process is not at the level we would like (while there is a .9

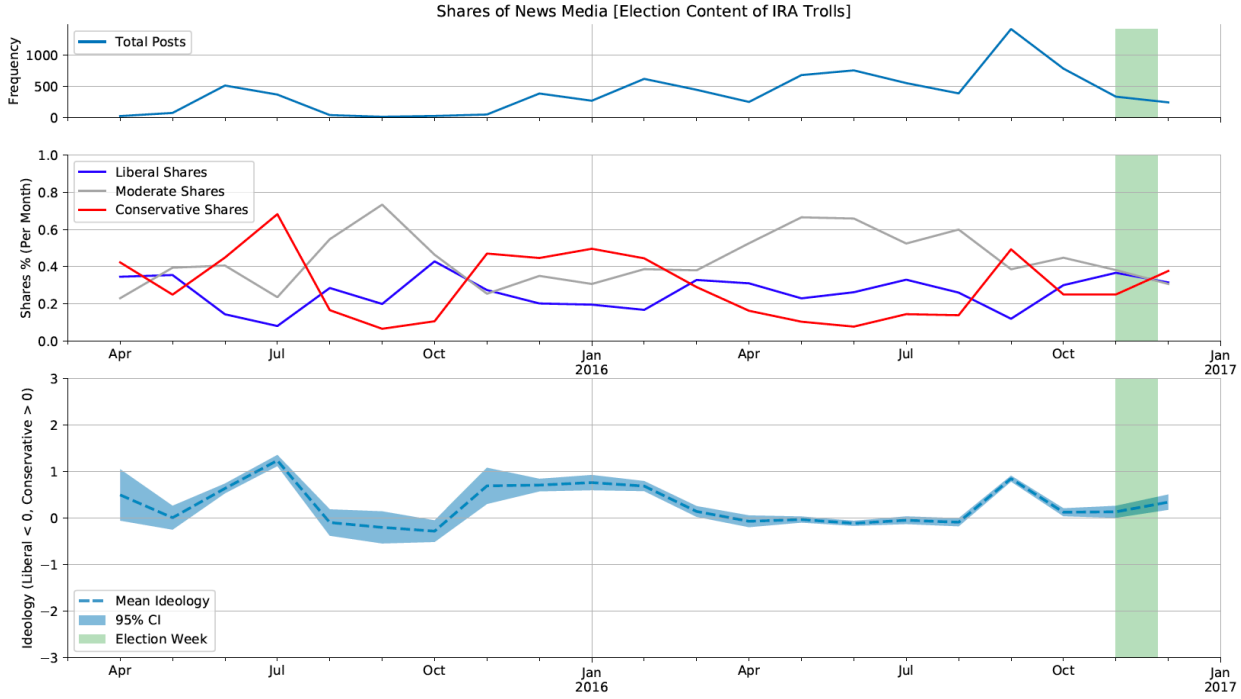
[14] The table, the figure, and a portion of the text in the first two paragraphs are taken from the current draft of Golovchenko et al. (nd).

[15] For those familiar with NOMINATE to estimate the ideologies of members of a legislature based on roll call voting, the method works in a similar way: here, the members of Congress and the general public are "voting" whether to share a link to a given domain or not. This allows us to simultaneously estimate the position of legislators (which serves as a useful validation check) and the media domains on the same scale.

intercoder reliability score for whether the channel contains political content, the coding of that ideological content as liberal, conservative, moderate or unclear only has an intercoder reliability score of .6, reflecting what is probably an inherently difficult task), the findings appear to be stark enough that they are unlikely to change:

	News Media URLs	YouTube URLs
Liberal	24%	17%
Moderate	42%	2%
Conservative	34%	81%

We see from the table two interesting patterns. First, matching the previous literature suggesting that Russian troll accounts were playing on both the left and right side of the political spectrum, the links to news media sites are distributed across the ideological spectrum, although there are significantly more shares to sites to the right of the *Wall Street Journal* than there are to sites to the right of the *New York Times*. However, the patterns from YouTube do not replicate this distribution: instead, our student coders labeled the vast majority of videos to which the IRA trolls were linking as conservative.



In addition to observing aggregate level counts of links across the entire time period, we can also track this activity dynamically. The figure above shows the total sharing of media links (but not YouTube channels) across the entire time period (top panel), the proportion of shares to conservative vs. moderate vs. liberal sites (middle panel) and the average ideology across all shares (bottom panel) by month. Perhaps most interestingly, we see that for the spring and summer of 2016, IRA trolls were particularly to share links to *moderate* news sources. This

seems to change in the months immediately prior to the election, when the share of conservative and liberal links begins to increase relative to the moderate links. There is a particularly sharp increase in more conservative content in September of 2016 (which we can see in the increase in number of posts, the increase in proportion of conservative posts, and the increase in the average ideology across all posts, although see the caveat about this spike below). Nevertheless, it is worth noting how in general, the average ideology of links shared hovers close to moderate.[16]

There are a number of important caveats while interpreting these results. First and foremost, just because you are sharing a link to moderate or liberal news source does not mean that the link is not intended to help elect Donald Trump; it is possible to share a *Washington Post* story about Clinton's emails or a *Boston Herald* story about illegal immigrants being arrested on rape charges. Further research is therefore necessary to disentangle the content of the articles shared by IRA trolls from the sources shared. But at the very least, the findings do seem consistent with an attempt by the IRA not to simply share links to far right content. Second, new data released by Twitter suggests that it is possible that the September spike in Conservative link shares was caused by miscoding a set of Venezuelan trolls as Russian trolls in the original Twitter release of accounts; we will update our findings accordingly once we can confirm whether this is the case. But the general trend of more sources from liberal and conservative accounts as the election approaches does not depend on the September conservative spike. Finally, we want to reiterate that coding the ideology of YouTube channels is a difficult and very much in progress process, so these findings should be treated as especially preliminary.

Overall, though, our findings regarding link sharing largely reflect the conclusions for the original studies of troll types, tweets, and network behavior: the trolls accounts appear to be playing a sophisticated strategy to appear "reliable" to different portions of the electorate (e.g., sharing news from liberal and conservative sources; sharing a lot of links to local news sources), while on balance sharing more information from conservative sources than liberal sources, and especially so when it came to YouTube videos.

[16] This is not a function of how we trained the model, as we deliberately *excluded* the troll accounts when training our ideology model.

Works Cited:

Golovchenko, Yevgeniy, Cody Buntain, Gregory Eady, Megan Brown, and Joshua A. Tucker. Nd. "Ideology of IRA Troll Links", *working paper.*

DiResta, Renee, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, Ben Johnson. 2018. "The Tactics Tropes of the Internet Research Agency."

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.".

Linvill, DL and PL Warren. 2018. "Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building." pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf

Roberts, Margaret. 2018. *Censored: Distraction and Diversion inside China's Great Firewall*. Princeton, NJ: Princeton University Press.

Sanovich, Sergey, Denis Stukal, and Joshua A. Tucker. 2018. "Turning the Virtual Tables: . Government Strategies for Addressing Online Opposition with an Application to Russia". *Comparative Politics*. 50(3): 435-54.

Steward, L., Ahmer Arif and Kate Starbird. 2018. "Examining Trolls and Polarization with a Retweet Network."MIS2 Proceedings, Marina Del Rey, CA.

Stukal, Denis, Sergey Sanovich, Richard Bonneau, and Joshua A. Tucker. 2017. "Detecting Bots on Russian Political Twitter", with. *Big Data*. 5(4): 310-324.

Tucker, Joshua A., Pablo Barberá, Margaret Roberts, and Yannis Theocharis 2017. "From Liberation to Turmoil: Social Media and Democracy",. *The Journal of Democracy* 28(4): 46-59.

Yin, Leon, Franziska Roscher, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2018. "Your Friendly Neighborhood Troll: The Internet Research Agency's Use of Local and Fake News in the 2016 US Presidential Campaign". *SMaPP Data Report 2018:3*.

Zannettou, Savvas, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini and Jeremy Blackburn. 2018. "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web." arXiv, January 28, 2018. https://arxiv.org/abs/1801.09288.