FULL-LENGTH RESEARCH ARTICLE

Talking to Trolls—How Users Respond to a Coordinated Information Operation and Why They're So Supportive

Darren L. Linvill[1], Patrick L. Warren[2], & Amanda E. Moore[1]

[1]Department of Communication, Clemson University, Clemson, SC, USA
[2]John E. Walker Department of Economics, Clemson University, Clemson, SC, USA

This research explored how users interacted with inauthentic social media accounts with the goal of gaining insight into tactics employed by state-backed disinformation efforts. We combine hand coding with natural-language processing to measure the ways in which users talked with and about the accounts employed by the Russian-affiliated Internet Research Agency in the month before the 2016 U.S. Election. We find that user mentions were overwhelming supportive of the IRA accounts, belying the standard characterization of these personas as "trolls." This pattern is particularly strong for the more ideological troll types, suggesting that a strategy of building homophilic connections with like-minded people was central to the IRA campaign. This strategy seems to work—on days that the personas' mentions were more supportive, they received more engagement.

Lay Summary

Through this research we gained insight into tactics employed by state-backed social media disinformation. With that goal, we explored user interactions with inauthentic Twitter accounts. We used multiple procedures to measure the ways in which users talked with and about the accounts employed by the Russian-affiliated Internet Research Agency in the month before the 2016 U.S. Election. We found that users were overwhelming supportive of the IRA accounts, a fact that calls into question the standard representation of these accounts as "trolls." Users were particularly supportive of the accounts that pretended to be part of a particular ideological group (on both the left and right), supporting arguments that a strategy of building connections with like-minded people was central to the IRA campaign. This strategy seems to work—on days that the Russian accounts received more support they also received more engagement.

Keywords: SIDE model, Disinformation, Internet Research Agency, User Engagement, Twitter

https://doi.org/10.1093/jcmc/zmab022

Corresponding author: Darren L. Linvill; e-mail: darrenl@clemson.edu
Editorial Record: First manuscript received on 12 April 2021; Revisions received on 31 August 2021; Accepted by 20 October 2021

Introduction

The issue of state-affiliated social media disinformation became prominent following the 2016 U.S. Presidential campaign and the indictment of 13 Russian nationals for interfering in the election (*United States of America v. Internet Research Agency LLC*). These individuals worked in varying capacities for the St Petersburg-based Internet Research Agency (IRA), an organization owned by Russian oligarch Yevgeny Prigozhin and widely held as a tool of the Russian state (Shane & Mazzetti, 2018). In part through social media, the IRA encouraged division, discontent, and disconnection with reality among potential U.S. voters. While past IRA social-media activity has been directed at both journalists (Lukito & Wells, 2018) and politicians (Gallagher, 2018), overall, their primary targets on social media seemed to be social identity groups. The IRA focused on differing, often seemingly contrary agenda in their disinformation campaigns, varying depending on the particular type of persona they employed (Linvill & Warren, 2020a). Influencing votes, attendance at events, and support for particular social movements were central, however, in all of their efforts (Shane & Mazzetti, 2018).

Since 2016, state-affiliated disinformation has expanded. Operations targeting U.S. social media users have continued to be identified which are believed to stem not only from Russian sources (Popken et al., 2019), but also Iran (NewComber et al., 2019) and China (Cohen et al., 2019) among others. Social media gives bad actors the potential for a broader reach and lower cost of entry relative to other media, facts which have moved traditional propaganda and disinformation efforts into this space. As it is more interactive than past modes of social influence, it is important to look beyond the message form and source to the processes at work.

IRA social media accounts and the operatives running them are often referred to derisively as "trolls" (Shane & Mazzetti, 2018). Internet trolling is commonly defined by activity such as online name calling or harassment, perhaps including language strewn with profanity, racism, and sexism (Cheng, et al., 2017), generally with a goal of causing conflict. Below we will detail why this definition may not give us a full understanding of the behavior of some state affiliated, professional social media disinformation campaigns. In this study, we examined how social-media users engaged with inauthentic Twitter accounts created by the Russian IRA. Below, we will show, through an initial qualitative phase of research, that when users engaged with the ideologically oriented IRA persona, their messaging was largely supportive in nature. Building from this qualitative stage of research we conducted quantitative analysis on a larger corpus of data. Through this analysis, we found a positive relationship between the level of supportiveness that IRA accounts received for their messaging and both the level of engagement they received as measured by retweets and the growth of the IRA accounts as measured by followers. These findings illustrate the important role positive interactions with users can play in coordinated information operation tactics on social media.

Literature Review

One useful lens we can employ to understand IRA engagement with online communities is SIDE, the social identity model of deindividuation effects (Reicher et al., 1995). Building from social identity theory (Israel & Tajfel, 1972), SIDE suggests that the anonymity found in computer-mediated communication shifts the relative salience of personal versus social identity to focus on the social. Klein et al. (2007) extended SIDE and considered how individuals could strategically present their identities in ways intended to influence groups. They defined both *identity consolidation* as well as *identity mobilization* functions of identity performance. Identity consolidation involves pursuing in-group

acceptance through behaviors such as marking oneself as a group member and expressing ideas that are emblematic of the group's shared identity. Identity mobilization, on the other hand, is the use of identity performance to strategically mobilize in-group members into supporting specific political goals. They stress that not only is identity performance to in-group members central to mobilization "but also that one of the key ways in which large social categories of people can be mobilized to create social change is through the strategic performance of social identities" (p. 36).

Due to Twitter's credible transparency, more is public about disinformation on Twitter as compared to other platforms. Twitter has released data sets of content from a range of state-affiliated actors, including approximately 2.8 million English-language tweets produced by the IRA from accounts that claimed to be operated by U.S. nationals or organizations ("trolls") between the start of 2015 and the end of 2017. Most work exploring this content has been descriptive, focusing on the identities troll accounts took on, the content they produced, or the tactics they employed (DiResta et al., 2018; Howard et al., 2018; Linvill et al, 2019; Linvill & Warren, 2020a). Previous research has also examined specific issues on which the IRA focused (Broniatowski et al., 2018; Strudwicke & Grant, 2020). While some research has explored the nature of the accounts who engaged with the IRA (Freelon & Lokot, 2020; Zhang et al., 2021) no work has yet closely examined the individual messaging of user responses to IRA accounts.

Despite extensive research exploring IRA efforts, little work has explored quantifiable effects of their campaigns. One exception is the work of Bail et al. (2020); this research found no evidence that interaction with IRA Twitter accounts from late 2017 had any impact on user's political attitudes or actions. Bail and colleagues did find, however, that IRA accounts interacted most often with users possessing strong ideological homophily within their Twitter network. It is possible ideological change was not a primary IRA goal; other goals may have included ideological entrenchment, the sowing of distrust, agenda building around particular issues, or, perhaps most likely, some combination of various goals.

While research has explored the spread of fake news and the role of user engagement in this process on social media (Grinberg et al., 2019; Guess et al., 2019), there is a lack of academic work regarding how users respond when unknowingly confronted with inauthentic accounts. Deeper knowledge of users' behaviors may help us to infer the goals, tactics, and strategy that state actors employ to reach and influence their targets. Freelon et al. (2020) argued that it would be valuable to explore user responses to IRA trolls, specifically suggesting examination of user responses on Twitter as many such responses are still accessible. With this in mind, we aimed to explore the following research question:

> *RQ1: How do real users engage with IRA English language troll accounts in the month prior to the 2016 U.S. presidential election?*

As previously stated, Bail et al. (2020) found strong ideological homophily within the Twitter networks of individuals engaging with IRA twitter accounts. Zhang et al. (2021) found that this homophily extended also to the media networks these users engaged with. IRA accounts engage in networks of like-minded individuals, and it seems likely they may do so as "fellow travelers" purporting to be of a similar ideology to those with whom they engage. But that does not necessarily imply that most interactions with trolls are positive. For example, Bail et al. note that "observing a trusted social-media contact argue with an IRA account that impersonates a member of the opposing political party, for example, could thus contribute to stereotypes about the other side" (p. 244).

But Linvill and Warren (2019), in contrast, argued that IRA trolls, "don't go to social media looking for a fight; they go looking for new best friends." This would be consistent with engaging in the

identity consolidation function of identity performance (Klein et al., 2007). If they were successful in doing so, SIDE suggests that their messaging (even if negative in tone) would be received by the accounts they choose to engage with in a positive manner. Previous research employing SIDE has shown that social media's lack of individuating cues, "not only enhanced group cohesiveness but also prompted individuals to polarize their position in the direction of group norms" (Lee, 2007, p. 399). Hence:

H1: Most engagement with IRA trolls will be supportive in nature.

We employed Linvill and Warren's (2020a) typology of these IRA trolls to facilitate data analysis. Linvill and Warren identified four broad types of English-language accounts active in the month prior to the 2016 election: *right trolls, left trolls, news feeds*, and *hashtag gamers*. This typology was defined at the account level and captured the dominant persona that the account expressed throughout its existence. This categorization depends on the output of the troll account, alone, and does not consider how the troll is received or mentioned by other accounts. Right trolls often expressed nativist and right-leaning populist messages, typically employing hashtags engaged with by similar but genuine Twitter users, including #tcot, #RedNationRising, and #MAGA. Left trolls expressed liberal views and typically focused on cultural and racial identity. Many of these accounts engaged in conversation with the Black Lives Matter community. News feed troll accounts aggregated real, local news. They had names specific to a city, such as @TodayPittsburgh and @OnlineMemphis, and tweeted news specific to the city whose name they adopted. Finally, hashtag gamer troll accounts were dedicated to playing word games. Hashtag games, popular among many genuine Twitter users, involve posting a hashtag and then answering the question implied by the post, e.g., "#ThingsILearnedFromCartoons You can get your head blown off by a cannon and completely recover in five minutes." IRA trolls both organized and took part in hashtag games.

Answers to our qualitative research question facilitate further inquiry best explored quantitatively. Following on RQ1, it is important to understand how the character of real users' responses to disinformation may change depending on who is viewing the messaging. Given we know IRA trolls demonstrate a great deal of homophily in their networks (i.e., IRA accounts with left-leaning persona tend to engage with liberal Twitter users and IRA accounts with right-leaning persona tend to engage with conservative users), we can begin to explore this question by examining how engagement varies by troll type, therefore:

RQ 2: How, if at all, does the volume and character of engagement with IRA English language troll accounts vary in the month prior to the 2016 U.S. presidential election by the type of the troll account being mentioned.

Finally, we will use what we learn from RQs 1 and 2 to better understand what relationship, if any, the character of responses has with the level of engagement and prominence the trolls receive. User engagement and increased following is important for many coordinated disinformation campaigns as it broadens that campaign's reach and influence. It is therefore important to understand if trolls, as the saying goes, catch more flies with honey than with vinegar (honey, in this case, being a matter of perspective)? With this in mind, we ask:

RQ 3: How, if at all, does the volume and character of engagement with IRA English language troll accounts in the month prior to the 2016 U.S. presidential election relate to the number of responses and followers the troll accounts receive?

The potential effectiveness of targeted positioning is consistent with many complementary models of media competition, in which outlets position themselves by sending messages that cater to the prior tastes/beliefs of their target group, either because it increases their willingness to pay for that content (Mullanthian & Shleifer, 2005), makes them trust the content more, or motivates the targeted group to engage in some desired action, such as voting (Dellavigna & Kaplan, 2007). Targeted positioning is also consistent with related theories of persuasion. Social judgment theory (Sherif et al., 1965), for instance, suggests that individuals' ego involvement in particular issues, how important particular issues are to their identity, play an important role in how likely they are to accept a message. Finally, as previously stated, SIDE suggests the deindividuation which occurs on social media enhances group cohesion and polarization (Lee, 2007). Together, these suggest two further hypotheses, related to RQ2 and RQ3:

H2: Tweets from ideologically oriented troll accounts (left trolls and right trolls) will receive a greater share of supportive responses than will more generalized troll accounts (hashtag gamers and newsfeeds).

H3: Tweets from troll accounts which receive more positive engagement from real users will result in greater engagement and troll account growth, and that will be especially true for ideologically specialized troll accounts.

Data Collection

Saleforce's Social Studio platform enabled keyword searches of tweets mentioning known IRA Twitter account handles.[1] A list of accounts identified by Twitter as associated with the IRA was released on June 18, 2018 by the United States. House Intelligence Committee (Permanent Select Committee on Intelligence, 2018). These handles were searched using Social Studio and all replies to and quote tweets of these accounts were collected through keyword searches using the account names as the keyword. The search was conducted for all tweets between October 8, 2016 and November 10, 2016. This period overlapped with the 2016 U.S. Presidential election—a period when the IRA was highly active. The method of data collection constrained our ability to collect large periods of data and so the month prior to and overlapping with the election was chosen for its likely importance to the IRA campaign. The combined searches resulted in 117,626 replies and 80,276 quote tweets which included an IRA troll account name, together with another 13,708 tweets that simply included a troll name in the content of a tweet. Together, these three categories will be referred to as "mentions." All these mentions were downloaded for analysis and we removed any mentions that were produced by accounts on the IRA list.

To analyze our data, we applied the account types for these accounts as defined and applied by Linvill and Warren (2020a). We found that the data is dominated by mentions of right troll accounts ($n = 186{,}476$), which are, in turn, dominated by mentions of @TEN_GOP ($n = 130{,}476$). But the other accounts and account types also received hundreds of mentions on many days (16,190 mentions of left trolls, 4,933 of news feeds, and 2,273 of hashtag gamers). Although mention counts vary from day to day, there are substantial spikes (a) on election day, for right trolls, (b) around November 1st, for left trolls, and (c) on October 9th and 19th (the final two Presidential debates), for the hashtag gamers. These spikes in mentions are consistent with spikes in IRA output shown by Linvill and Warren (2020a).

Qualitative Methods

Our first research question asked about the nature of how genuine users engaged with IRA accounts. To explore this question, we conducted qualitative analysis of the mention data, first examining random selections as recommended by Corbin and Strauss (2015). We read tweets, working together to examine, break down, and conceptualize data and get a sense of meaningful patterns. Four broad categories were identified (see "Results" section). As this process continued, we created definitions for each category and identified exemplar posts. Tweets were read in context with an understanding of the account type being replied to as well as any content to which a tweet had an active link (Social Studio data provides an active link to the original tweet on the Twitter platform). In many cases, reading through the Twitter conversation which a tweet was a part of was necessary for proper interpretation. It should be noted, however, that this task was made difficult given that the IRA accounts were suspended and their tweets no longer appear on the platform (though threads of conversations replying to these tweets do still appear). Tweets were coded using the information available and categorized based on what the coder felt was the most likely intent of the tweet.

To help assure the reliability of our analysis, we developed a code book and worked to establish an acceptable intercoder reliability. The use of a code book detailing clear definitions and sample tweets for each category served as a stable representation of the coding analysis throughout the process (Creswell & Poth, 2018). We coded randomly selected sets of 50 tweets, placing each tweet into one of the four categories or labeling it as unknown. After each set was coded, we compared results and refined our analysis. This process was continued until a Krippendorff's alpha reliability of .76 (Krippendorff, 2004) was met between three coders over five categories: *attack, support, troll whistling, unrelated comment*, and *unknown*. A random sample of 5,000 tweets was then selected and distributed for analysis. Of these tweets, 4,316 mentioned right troll accounts, 483 mentioned left troll accounts, 58 mentioned hashtag gamer accounts, and 143 mentioned news feed accounts.

Qualitative Results (RQ1)

A total of 511 (10.2%) of the sample tweets from the month prior to the 2016 U.S. Presidential election could not be placed into a category and were labeled *unknown*. Almost all of these tweets were part of Twitter conversations that could no longer be viewed in full (as the IRA accounts and their tweets do not appear on the platform the same as any other accounts in a thread that have since been suspended or deleted) or contained links that were no longer active and therefore lacked full context for interpretation. Qualitative analysis placed the remaining tweets ($n = 4{,}489$) into one of four categories described below. Tweets were placed into the category with which they best matched. Note, all example tweets are presented verbatim, including errors. Percentages given below are as a percentage of the 4,489 coded tweets.

Attack ($n = 614$, 13.5%)

These tweets generally or directly attacked the ideas, ideology, or worldview espoused by the mentioned troll account. An example of a tweet in this category is the October 26, 2016 tweet "@TEN_GOP You are being a disgrace to the great state of TN." This tweet was directed at the IRA troll account @TEN_GOP, a right troll which purported to be the unofficial Twitter account of the Tennessee Republican Party. Another example, also directed at @TEN_GOP, is a November 8, 2016 tweet "Y'all actin so helpless. Like you can't just go tell somebody working there that your machine

don't work." This tweet was in response to a video of a malfunctioning voting machine posted by the IRA account. A final example of a direct attack tweet is the October 25, 2016 tweet directed at a left troll account: "@BlackNewsOutlet When Black folks stop treating each other like crap & empower each other racism can loose its grip. BUT selfishness rules."

This category included 85 tweets which were indirect attacks (13.8% of the category). These tweets mentioned the troll account, but were directed at a different user. This is typically because the tweet was part of a conversation thread started by the troll account, but one in which the troll account was no longer involved. An example of such a tweet is the October 22, 2016 tweet "In make believe land". This tweet took place on a thread started by the right troll @TEN_GOP, but was in response to another account who tweeted "This is not a Presidential Election, this is Coup, by the Corrupt Democrat & Republican Government, against Trump & American's!"

Support ($n = 3,459$, 76.1%)

These tweets generally or directly supported the ideas, ideology, or worldview espoused by the mentioned troll account. An example of a tweet in this category was the November 1, 2016 tweet "In NY theirs more illegals than citizens on some job sites. But they'd probably scramble a the sight of a Trump rally". This tweet was a reply to @TheFoundingSon, a right troll which purported to be an anti-immigration, Trump supporter. Another example included the November 2, 2016 response to @BlackMattersUS, "Another nightmare that will only be remembered by the family. So many instances going down, it's hard to keep up." @BlackMattersUS was a left troll account which purported to be the twitter account of an organization which was part of the broader Black Lives Matter movement. That organization, Black Matters US, was itself a creation of the IRA (Albanesius, 2019).

This category included 216 tweets which were indirect support (6.2% of the category). These tweets mentioned the troll account, but were directed at a different user. As with indirect attacks, this is typically because the tweet was part of a conversation thread started by the troll account. These tweets typically defended the troll account, the troll accounts professed ideological views, or members of the ideological group the troll account professed to be a part of. This included the October 10, 2016 tweet "Who has DJT assaulted? Did you see evidence no one else has? Your argument is flawed. Try harder." This tweet was in response to another user suggesting then-candidate Trump had committed sexual assault.

Troll Whistling ($n = 126$, 2.8%)

Tweets in this category included a list of account handles but no significant message. This is common practice to call the mentioned accounts' attention to a post. It can be done to call others to either attack or support another user's comment. Examples of this include an October 10, 2016 tweet, "RT@jturnershow @blacktalkradio @tariqnasheed @NateParker @TheBlackChannel @Allblackmedia @BlackNews @AndyBLACKnews @BlackNewsOutlet @newsone". This tweet included the left troll account @BlackNewsOutlet and was intended to call attention to a linked article about Texas prisons banning books by Malcolm X. Another example includes the October 15, 2016 tweet, "Righton @LouDobbs #TrumpAllTheWay @RTDNEWS @CatNamedLily @JudgeJeanine @avanconia @gjathanas @JacquelinIsBest @Nvr4Get91101 @lilacbananas23". This tweet included the right troll account @JacquelinIsBest and was intended to call attention to a tweet from Fox News' Lou Dobbs critical of candidate Hillary Clinton.

Unrelated Comment ($n = 290$, 6.5%)

These tweets were not clearly supportive or critical of the mentioned troll account's ideas, ideology, or worldview. An example of this was the November 1, 2016 reply to right troll @Jenn_Abrams, "Hamburger Tom sounds like a mafia name," which appeared simply as an amusing observation. Tweets in this category often contained messages with no relationship to the discussion involving the troll account. These tweets included advice given to another user on a right troll's thread regarding how to best use Twitter: "Psst, you're supposed to reply after the name of who you're addressing."

We find support for H1 by this hand-coded sample of tweets. A large majority of mentions are supportive of the troll account, more than five-times as many as attack the troll. The distribution of these categories does vary slightly across account type in this data (Reject null of jointly identical distributions at $p < .02$). Figure 1 presents the mention category shares across account types. Right and Left trolls have similar distributions, although they do differ statistically ($p < .01$). Looking specifically at supportiveness, however, left and right trolls have statistically ($p > .10$) and substantively similar shares of supportiveness, while hashtag gamers and news feeds have lower shares of supportiveness, both substantively and statistically ($p < .05$).

Quantitative Methods

Predictive Binomial Labeling

Addressing RQ2 and RQ3 requires a more extensive data set of labeled tweets than can feasibly be implemented by hand. We applied machine-learning techniques to address this challenge. As the overwhelmingly dominant label in our sample was "Support," we reduced this predictive labeling exercise to a binomial problem of predicting the probability that a given tweet would have been labeled as "support" if investigated by hand. Our goal is purely predictive, to build a model that relates the features of the tweets to the "support" label to apply that model to the broader corpus. The partial relationships between the independent variables in the model and the outcomes are not causal, or even particularly interesting. Rather our interest is in the predicted probabilities. Trends in these probabilities can then be used to investigate how supportiveness varies over time and type in finer detail.

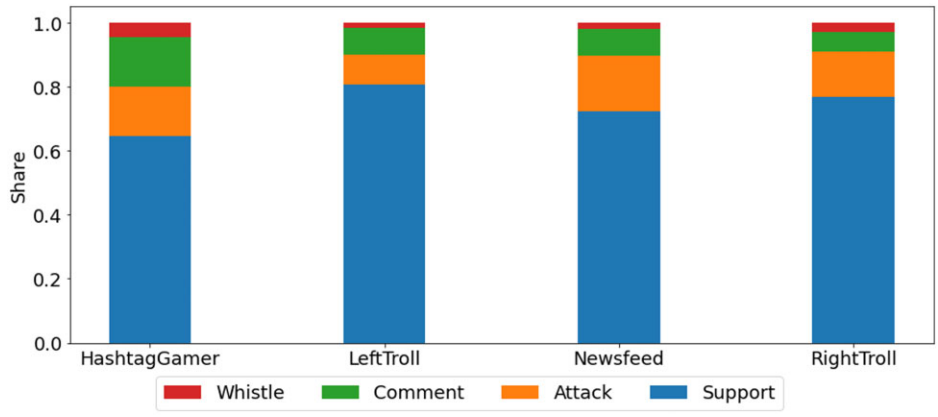

Figure 1 Mention Type Shares by Troll Account Type, Hand-labelled Sample.

Formally, we estimated a L2-normed (penalized) logistic regression where the probability that a given tweet would have been labeled as "support" is a function of a constant and 80 variables: 75 variables representing the mix of words it contains (details below), the number of mentions included in the tweet, the tweet's length in words, and three dummy variables indicating whether the tweets has exactly 0 non-mention words, exactly 1 non-mention word, and exactly 2 non-mention words. We chose the penalization parameter by 5-fold cross-validation, where we choose the parameter that minimizes logged loss.

To transform the words into numerical objects that represent meaning, we used the entire 210,492 tweet corpus to build a word embedding space, where each word that is used at least five times in the overall corpus is projected into a 75-dimensional vector.[2] For details on this method, see Joulin et al. (2016). Intuitively, words that are often used in semantically similar ways (modelled by where they sit in relation to other words, in the whole corpus of tweets) are given vectors that are geometrically close. We then collapsed from the word level to the tweet level by taking the average across words in the tweet along each of the 75 dimensions, to capture an overall 75-dimensional summary of what the tweet is about.

We estimated this model on a training sub-sample of 4,000 labeled tweets and apply it to the 1,000 held back tweets to evaluate the quality of the predictions. In this hold-out sample, the predicted probability correlates well with the actual share of tweets that are coded by hand as supportive.[3] Figure 2 presents a locally linear nonparametric regression relating our predicted probability of a tweet being supportive to the share of tweets that are labeled as supportive (left axis), together with a histogram for the distribution of these predicted probabilities (right axis). In the hold-back sample, the relationship between predicted and actual probabilities is nearly linear, with a slight underprediction of supportiveness for relatively high levels of predicted supportiveness. If we impose linearity, the coefficient on the predicted probability in a linear probability model at the level of individual tweets is 1.005, with a standard error of .019.

In our application, we aggregate many tweets together to build a proxy for the general supportiveness of the real users' interactions with trolls. The key feature we need these predicted probabilities to satisfy is that tweets with higher (lower) predicted probabilities really are, on average, more (less) likely to be judged supportive. This goal contrasts, for example, with a researcher who needs a highly discriminatory

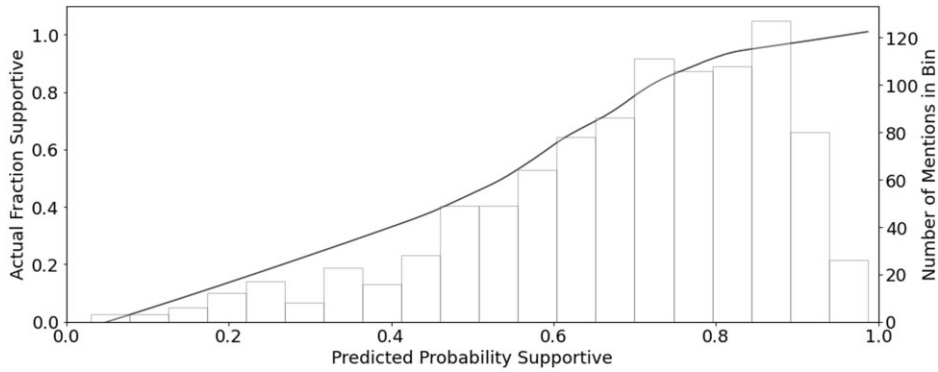

Figure 2 Smoothed Actual Fraction Supportive and Function of Predicted Probability of Being Supportive (Left) and Distribution of Predicted Probabilities (Right) in $n = 1{,}000$ Hold-Out Sample.

classifier that ends up with very extreme posterior probabilities for any given tweet. As Figure 2 makes clear, our predictions are quite accurate, on average, and so this proxy can serve our needs.

We applied this model to the entire 210,492 tweets in our unlabeled dataset to calculate a predicted probability of being labeled supportive for each mention. We will treat these predicted probabilities as data when analyzing RQ2.[4] Specifically, we present the average predicted probability of support by account type over time.

Relating Mention Mix to Engagement

To evaluate the relationship between supportive mentions and engagement, we shifted our unit of observation from the tweet level to the account by day level to analyze the relationship between the mix of responses a troll account receives on a day and the level of engagement it receives, where engagement is measured in two ways: follower-count growth rates and the number of retweets received by tweets by the account on that day.

For each IRA account that is in the Twitter database, is active in this period, and appears in the mention dataset, we calculate an estimate of the growth rate in followers throughout this period (for details, see Linvill & Warren, 2020b). To investigate the relationship between positive interactions and engagement, we conduct a panel regression to correlate positivity with engagement. We estimate this relationship separately for each troll type, as they engage with very different audiences and use very different strategies throughout their life-cycles (Freelon & Lokot, 2020; Linvill & Warren, 2020a, 2020b). Formally, we conduct regressions of the form

$$y_{it} = a_i + b_t + B_{l1}(0.4 <= p_{it} < 0.6) + B_{h1}(0.6 <= p_{it} < 0.8) + B_{vh1}(p_{it} >= 0.8) + e_{it}, \quad (1)$$

where y_it represents one of our metrics of engagement, p_it represents our metric of supportiveness, a_i are troll account fixed effects, b_t are time fixed effects, and 1() is an indicator function for various ranges of estimated supportive share. Our interest is in the coefficients B_l, for low levels of supportiveness, B_h, for high levels of supportiveness, and B_vh, for very high levels of supportiveness, which we interpret as the (partial) relationship between supportiveness and engagement. As we do not expect monotonic relationships between supportiveness and engagement (and the non-parametric cross-sectional results suggest important non-monotonicity), we bin the supportiveness range to allow for a flexible relationship. As there will be autocorrelation in unmodelled determinants of engagement (e_it), we cluster standard errors by account for inference (Cameron & Miller, 2015).

These fixed-effect estimates deliver a particular form of correlation that might be particularly useful for characterizing the patterns in the data in a focused way. The inclusion of account-specific fixed effects guarantees that the correlations we observe are not driven, for example, by simple account-level heterogeneity. We are not simply measuring whether accounts that get a greater share of supportive mentions grow faster (for example). Rather, we are measuring whether on days that an account gets more supportive mentions than it usually does, it also grows faster than it usually does. Similarly, day fixed effects guarantee that the correlations we document are not driven by simple time-series variation. So, on days when an account gets more supportive mentions than other accounts do, on the same days, does that account grow faster than its peers? The two-way fixed effects model does both adjustments simultaneously.

There are several sample-selection issues that arise in this setting. First, calculating mention positivity estimates requires that the account receives some mentions. Second, calculating follower-count growth rates requires observing follower counts over multiple days. However, we observe follower counts only when the account tweets. This requires the account to be sufficiently active to observe follower counts on subsequent days (we could extrapolate growth across multiple days, which gives substantially similar results). To the extent that this selection is unrelated to supportiveness, it will not bias our estimates. But if, for instance, relatively inactive accounts (which may not show up in our data) were getting both less supportive mentions and presumably low follow growth rates, we would underestimate the true the relationship between supportiveness and engagement.

Furthermore, there is no sense in which these partial relationships are causal. Rather, we interpret the supportiveness estimate as a proxy for the underlying tenor of the interactions, which are jointly determined by many decisions made by both the IRA account and its interlocutors. Whatever these underlying decisions are, they are unobservable to us, and they drive both supportiveness and engagement. Even lagging the supportiveness metric or calculating impulse/response functions, as in Zhang et al. (2021), cannot avoid this essential weakness of our setting and approach. We can only measure the patterns in these jointly-determined correlations. We do not observe the underlying complex of behavior that drives it.

Quantitative Results

RQ2

Our second research question asked how the character of user engagement with IRA accounts varied depending on the troll account type. Figure 3 illustrates how our estimate of the supportive mention share varies in our data across account types in the full set of labeled and unlabeled tweets. The estimated rate of support is very stable around 70% for right and left trolls. Except for the week leading up to the election, it is also similar for the newsfeeds, but in that week it dips (statistically below the right trolls, $p < .05$, in the first 5 days of November but otherwise statistically indistinguishable). For hashtag gamers, the estimated share of supportive tweets is both lower (statistically below the right trolls, $p < .05$ on nearly every day in the sample) and less consistent, but the share of supportive

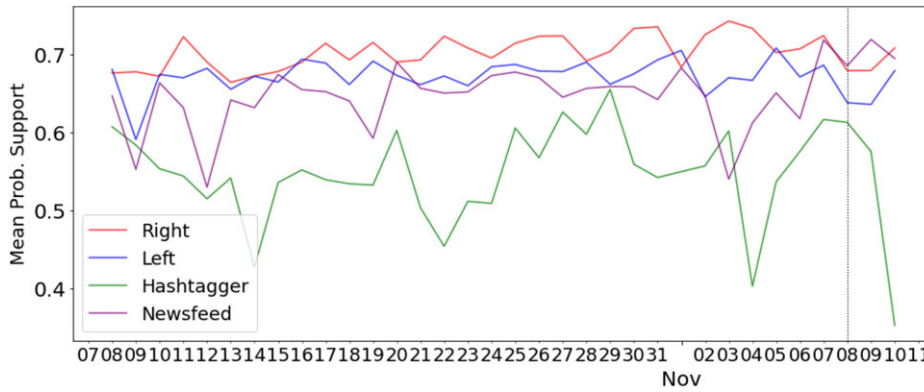

Figure 3 Mean Probability Supportive of Daily Mentions of Troll Accounts of Indicated Type over Time.

mentions also drops in the week immediately before the election. With respect to H2, there are consistent differences between the ideological types (left trolls and right trolls) and the hashtag gamers, in the predicted direction, in terms of estimated supportiveness of mentions. The cross-sectional gap between news feeds and the ideological trolls, however, is not consistent over time. Rather, it is driven by a small period of time immediately before the election, some evidence against H2. We interpret this confluence as *qualified support for H2*.

RQ3

Our third research question asks how the character of engagement by users with IRA accounts may influence the number of responses and followers that IRA accounts receive. The relationship between supportive mentions and metrics of engagement depends on the category of the troll account. Figure 4 presents locally linear regressions relating the estimated share of supportive mentions an account receives and the daily growth rate of its follower counts (left axes), with one regression for each troll account type, separated into two separate panels, the top for left and right trolls and the bottom for newsfeeds and hashtag gamers. It also includes histograms of the estimated share supportive by account-day (right axes), for accounts of the indicated type.

For both right and left trolls, high, but not unanimous, rates of support predict high follower growth rates. The highest (mean) growth rates occur on days in which we estimate mentions around

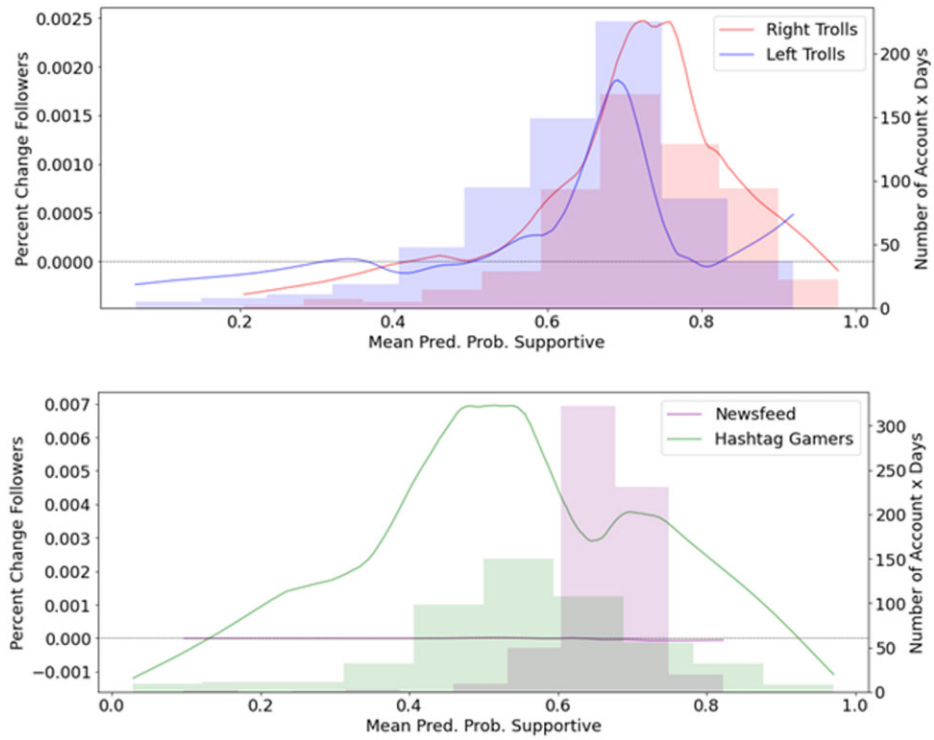

Figure 4 Locally Linearly Smoothed Percent Change in Followers (Left) and Count of Account Days (Right) by Daily Mean Probability Supportive.

to be about 70% supportive. The observed distributions of supportive shares are both centered about that follower-growth maximizing rate of support. For newsfeeds, in contrast, there seems to be no relationship between supportive mentions and follower growth, but, in this period, newsfeeds were growing very little, overall. For hashtag gamers, account-days with maximal growth are those with a more even mix of estimated supportiveness.

Table 1 presents the two-way fixed effects regression version of these results, as specified in equation (1).[5] It shows that, at least for right trolls, account-days in which mention supportiveness is unusually high for that account are also those account-days in which those accounts grow unusually fast. Relative to days in which the level of support is low (0–40% supportive), accounts on high support days have between .5 and .8 percentage points higher daily growth rates. For the other three account types, there is no significant relationship between growth and mentions supportiveness, when controlling for date and account fixed-effects. For left trolls, the point estimates are suggestively positive, while for hashtag gamers they are suggestively negative, but the standard errors are large.

Figure 5 and Table 2 present the parallel analysis for a different metric of engagement, retweets (transformed as log(1+x)). These regressions include more account-day observations, because the outcome variable does not require that we observe follower counts on subsequent days. On this metric, the locally linear smoothed patterns for left and right trolls are similar, both to each other and to the first metric of engagement, with peaks near 70% supportive. In regressions, however, there is no statistically significant relationship between supportiveness and retweets across days within an account, for left trolls, but, again, right trolls receive about 40% more retweets on very positive mention days. In contrast to the first metric, there is a substantial relationship for newsfeeds between supportive mentions and

Table 1 Average Probability Supportive and Follower Count Growth Rate

Pct. Support	Left Troll	Right Troll	Newsfeed	Hashtag Gamer
40–60	0.27	0.76*	−0.02	−0.00
	(0.24)	(0.33)	(0.03)	(0.41)
60–80	0.01	0.51†	−0.02	−0.24
	(0.20)	(0.29)	(0.02)	(0.37)
80–100	0.75	0.54†	−0.04	−0.58
	(0.77)	(0.29)	(0.03)	(0.35)
Account FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
Obs.	502	434	434	147

Note: Panel regression relating daily follower growth rate and mean estimated supportiveness of mentions the account receives, for IRA accounts of the indicated type. All regressions include day and account fixed effects, and dummies representing an estimated probability of support between the indicated range. The omitted range is 0–40% supportive. Standard errors are clustered at the level of the account. Accounts are omitted on days that they are not mentioned or where follower growth rates are not available.
†$p < .10$,
*$p < .05$,
**$p < .01$.

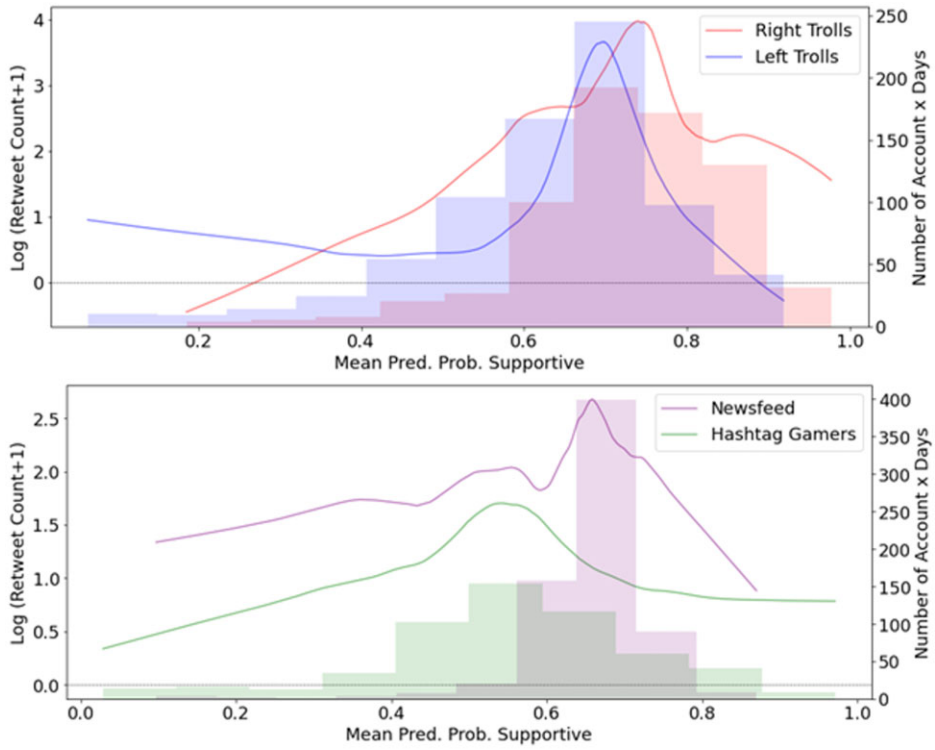

Figure 5 Locally Linearly Smoothed Retweets of Tweets Produced by Troll Accounts (Left) and Count of Account Days (Right) by Daily Mean Probability Supportive.

retweets, with about 40% more retweets on positive mention days. Hashtag gamers have no statistically significant pattern in the relationship between high-support days and retweets.

We interpret these results as *qualified support for H3*. The pattern is clear in the cross-section, where ideological types seem to benefit from high levels of supportive mentions. For hashtag gamers, in contrast, a more mixed response of moderate supportiveness seems to be more associated with engagement. For news feeds, the pattern seems to depend on how we measure engagement. For follower growth, there appears to be no consistent relationship between supportiveness and engagement, with the caveat that the overall levels of growth are quite low for this troll type. For retweet engagement, however, the pattern looks more like the left/right troll types, with moderately high levels of support being mostly highly related to high levels of engagement.

When we adjust for account and day fixed effects, this basic pattern holds up for right trolls (which had, by far, the most mentions), but is much murkier for the other types.

Discussion

This research has shown that "Russian trolls" may be a somewhat misleading label for the IRA. Rather than anger and confrontation, our analysis shows that, consistent with SIDE (Reicher, et al., 1995), Twitter users most often responded to IRA messaging in a way that supported group ideological cohesion and was supportive of the identity of the IRA persona being responded to. These findings

Table 2 Average Probability Supportive and Retweets

Pct. Support	Left Troll	Right Troll	Newsfeed	Hashtag Gamer
40–60	−0.01	0.22	0.36†	0.24
	(0.12)	(0.33)	(0.21)	(0.15)
60–90	0.07	0.31*	0.41†	0.21
	(0.10)	(0.14)	(0.22)	(0.22)
80–100	−0.14	0.43**	0.31	−0.13
	(0.19)	(0.14)	(0.22)	(0.19)
Account FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
Obs.	673	658	522	428

Note: Panel regression relating aggregate retweet count for tweets created by the account each given day and mean estimated supportiveness of mentions the account receives, for IRA accounts of the indicated type. All regressions include day and account fixed effects, and dummies representing an estimated probability of support between the indicated range. The omitted range is 0–40% supportive. Standard errors are clustered at the level of the account. Accounts are omitted on days that they are not mentioned or did not tweet.
†$p < .10$,
*$p < .05$,
**$p < .01$.

suggest that acts of *identity consolidation* as well as *identity mobilization* among specific online communities may be central to the IRA's social media efforts. Our findings are important to future disinformation research because, as Klein et al.'s research employing SIDE points out, "a long tradition of research in social identity suggests that social action emanates from social identity" (p. 41). The ideological IRA trolls received the most supportive mentions of our troll types (.70 mean-predicted probability supportive for Right Trolls and .67 for Left Trolls). If we take a random sample of 100,000 political tweets that include the word "Trump" or "Clinton" from early October, 2016, and apply the predictive model we developed for this article to the subset that contain mentions, we find a mean-predicted probability of supportiveness of about .68. So extreme ideological IRA trolls get about the same level of support as typical politics-related mentions on Twitter in the same era. This is despite the relatively extremist persona IRA trolls adopted and counter to common expectations regarding the goals of online trolls. It is, however, consistent with SIDE and the IRA's use of strategic performance of identity (Klein et al., 2007). It seems that the IRA trolls were successful in appearing to be "fellow travelers" and used this to reach a wider audience.

In the month before the 2016 election, a likely critical period to the organization, IRA Twitter posts received overwhelmingly positive responses. It can be inferred that these responses are a result of some combination of the nature of IRA tweets and who they targeted. This is not, it should be pointed out, meant to suggest that the trolls are themselves using positive messaging, the content of the troll messaging was not a part of this research. We know, in fact, that a great deal of troll messaging was negative in tone, engaged in name calling, and focused on divisive issues (Linvill & Warren,

2020a). An important aspect of identity consolidation is distinguishing one's identity from other groups, and this can reasonably be accomplished through negative messaging. Regardless of the nature of the messages, however, the trolls seem to have reached receptive and supportive audiences. Their negative messaging seems to have been appreciated.

This pattern of supportive responses should be understood in the context of identity group infiltration (Arif et al., 2018; Freelon & Lokot, 2020; Freelon et al, 2020). The account types that both had the most supportive mentions, and those that benefited the most from the most supportive mentions, were those that involved infiltrating ideologically motivated online communities with strong ethnic or political identities. While this study could not examine the effect of IRA messaging on audiences, our findings suggest that their posts found support amongst specific, chosen audiences. It is unlikely messages from left or right trolls would have had the same outcomes had they been distributed to a more general public or across the ideological divide. The less specialized accounts, which tried to appeal to broader audiences, both had less supportive mentions *and* demonstrated a weaker correlation between strongly supportive mentions and engagement.

These patterns should also be understood in the context of the life-cycle of IRA troll accounts (Linvill & Warren, 2020b). By the month before the election, the left and right troll accounts transitioned from the "growth" phase, where the primary goal seems to be picking up followers, to the "amplification" phase, where the primary goal seems to be sharing content (through retweets) from like-minded accounts outside the IRA. So high levels of supportiveness we find are—in some sense—supporting not only the Trolls themselves but also the content originators whom the trolls are amplifying. To the extent that those originators are, themselves, experiencing increased engagement, we do not measure it in this study.

One final finding of this research is the relationship between the percentage of IRA tweets receiving supportive response and the volume of IRA tweets seen in Figures 4 and 5. This relationship suggests that the IRA may target the level of supportiveness that is most beneficial in receiving more followers and greater engagement. Given that the IRA is known to have performed assessments of their own work and to have teams dedicated to work such as search engine optimization (*United States of America v. Internet Research Agency LLC*) it seems possible that this relationship we demonstrate is not by accident but rather by design.

In extending findings from this research, it is important to keep in mind the relatively short sample period undertaken and the limitations this suggests. We examined data from only the month before and the few days overlapping with the 2016 U.S. Presidential election. While this was surely an important, even culminating, moment of the IRA operation, it is still simply a snapshot in time of one ongoing disinformation campaign. Even in this campaign, Linvill and Warren (2020b) have shown that IRA Twitter account behavior varied greatly through different periods. Further, the IRA is only one of many groups operating to spread disinformation on social media and what we learn from studying their tactics may tell us little about the work of others. Nonetheless, this study clearly demonstrates the proverb that one catches more flies with honey than with vinegar has a kernel of truth as applied to social-media disinformation.

Data Availability

Anonymized original data as well derived data can be found here: https://github.com/patrick-lee-warren/talking_trolls

Notes

1. At the time this search was conducted Social Studio maintained a real-time archive of Twitter output by all non-protected accounts, even if the tweets were subsequently deleted and or accounts subsequently suspended. The tweets are gathered near the time they are produced (often within minutes, but sometimes at a slightly later period).
2. As we are training this word embedding on our corpus, alone, rather than a larger super-set of millions of Tweets, we prefer the relatively parsimonious 75-dimensional approach, rather than the 150–300 dimensional embeddings used when training the embedding on larger (but more heterogeneous) datasets (Yang et al., 2018). Fifty-dimensional and 100-dimensional in-corpus embeddings give very similar patterns of supportiveness across account types and over time, and similar relationships between supportiveness and engagements, but the predictions are not as accurate in the hold-back sample.
3. In a prediction-normed confusion matrix for this classifier, the 0.5 predicted probability threshold has some predictive power. Seventy-seven percent of the messages predicted as supportive are hand-coded as supportive and 66% of those predicted not to be supportive are hand-coded something other than supportive. The 50-dimension version of the embedding is 75% or 64% accurate, and the 100-dimension version is 75% or 66% accurate. But rather than use the binomial prediction, we use the raw predicted probabilities when pooling to the account-day level, below, to make best use of the intensive margin of our predictions.
4. From one point of view, that's correct, as they are pre-specified deterministic transformation of data. They are, from that perspective, simply a deterministic proxy for support and should be treated as data. From another point of view, they are estimates of some unobservable underlying variables (true probability of a human coder defining this tweet as supportive), and as estimates we should recognize the uncertainty around those estimates when estimating the uncertainty about further statistics. When we use these proxies as independent variables, below, their imperfection will bias our estimates toward zero, as measurement error in this proxy will attenuate the relationships relative to what we would find if we had hand coded the full set of 210k tweets.
5. Regressions without account and day fixed effects mirror the figures, exactly, so are not presented, here.

References

Albanesius, C. (2019, December 17). *Russian disinformation targeted African-Americans to sow division*. PCMAG. Retrieved from https://www.pcmag.com/news/365517/russian-disinformation-targeted-african-americans-to-sow-div

Arif, A, Stewart, L., & Starbird, K., (2018) Acting the part: Examining information operations within #BlackLivesMatter discourse. *PACMHCI.* 2, Computer-Supported Cooperative Work. Article 20.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., . . . Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *PNAS*, 117, 243–250. doi:10.1073/pnas.1906420116

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., . . . Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108, 1378–1384. doi:10.2105/AJPH.2018.304567.

Cameron, C. & Miller, D. (2015) A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50, 317–372. doi:10.3368/jhr.50.2.317

Cheng, J., Danescu-Niculescu-Mizil, C., Leskovec, J., & Bernstein, M. (2017). Anyone can become a troll. *American Scientist*, 105, 152. doi:10.1511/2017.105.3.152

Cohen, B., Wells, G., & McGinty, T. (2019, October 16). How one tweet turned pro-China trolls against the NBA. Retrieved from https://www.wsj.com/articles/how-one-tweet-turned-pro-china-trolls-against-the-nba-11571238943

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research*. Thousand Oaks, CA: Sage.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry & research design: Choosing among five approaches*. Thousand Oaks, CA: Sage.

Dellavigna, S. & Kaplan, E. (2007). The Fox news effect: Media bias and voting. *Quarterly Journal of Economics*, 122, 1187–1234. doi:10.1162/qjec.122.3.1187

DiResta, R., Shaffer, D., Ruppel, B., Sullivan, D., Matney, R., Fox, R., . . . Johnson, B. (2018). *The tactics & tropes of the Internet Research Agency. Technical report*, New Knowledge Foundation.

Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., & Adams, K. (2020). Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review*. doi: 10.1177/0894439320914853

Freelon, D. & Lokot, T. (2020). Russian Twitter disinformation campaigns reach across the American political spectrum. *The Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-003

Gallagher, N. K. (2018, August 2). *Russian-backed troll Twitter accounts targeted Maine politicians*. >Portland Press Herald. Retrieved from https://www.pressherald.com/2018/08/02/russian-backed-troll-twitter-bots-targeted-maine-politicians/

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. Presidential election. *Science*, 363, 374–378. doi:10.1126/science.aau2706

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5, 1–8, doi:10.1126/sciadv.aau4586

Howard, P. N., Ganesh, B., and Liotsiou, D. (2018). *The IRA, social media and political polarization in the United States, 2012-2018*. Technical report, Oxford Internet Institute, Computational Propaganda Project.

Israel, J., & Tajfel H.. (1972). *The context of social psychology*. London: Academic Press.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016) Bag of tricks for efficient text classification. *arXiv preprint*: 1607.01759

Klein, O., Spears, R., & Reicher, S. (2007). Social identity performance: Extending the strategic side of SIDE. *Personality and Social Psychology Review*, 11, 28–45. doi:10.1177/1088868306294588

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30, 411–433. doi:10.1111/j.1468-2958.2004.Tb00738.x

Lee, E. (2007). Deindividuation effects on group polarization in computer-mediated communication: The role of group identification, public-self-awareness, and perceived argument quality. *Journal of Communication*, 57, 385–403. doi:10.1111/j.1460-2466.2007.00348.x

Linvill, D. L., Boatwright, B. C., Grant, W. J., & Warren, P. L. (2019). "THE RUSSIANS ARE HACKING MY BRAIN!" investigating Russia's Internet Research Agency Twitter tactics during the 2016 United States presidential campaign. *Computers in Human Behavior*, 99, 292–300. doi: 10.1016/j.chb.2019.05.027

Linvill, D. L., & Warren, P. L. (2019, November 25). *That uplifting tweet you just shared? A Russian troll sent it. Rolling Stone.* Retrieved from https://www.rollingstone.com/politics/politics-features/russia-troll-2020-election-interference-twitter-916482/

Linvill, D. L., & Warren, P. L. (2020a). Troll factories: Manufacturing specialized disinformation on Twitter. *Political Communication.* doi:10.1080/10584609.2020.1718257

Linvill, D. L., & Warren, P. L. (2020b). *Engaging with others: How the IRA coordinated information operation made friends. HKS Misinformation Review.* https://10.37016/mr-2020-011

Lukito, J., & Wells, C. (2018). Most major outlets have used Russian tweets as sources for partisan opinion: Study. *Columbia Journalism Review*. Retrieved from https://www.cjr.org/analysis/tweets-russia-news.php.

Mullanthian, S. & Shleifer, A. (2005) The market for news. *American Economic Review*, 95, 1031–1053. doi:10.1257/0002828054825619

NewComber, E., Wagner, K., & Sebenius, A. (2019, October 21). Facebook identifies Iranian, Russian influence campaigns. Retrieved from https://www.bloomberg.com/news/articles/2019-10-21/facebook-identifies-iranian-russian-foreign-influence-campaigns

Permanent Select Committee on Intelligence. (2018, June 18). Schiff statement on release of Twitter ads, accounts and data. Retrieved from: https://democrats-intelligence.house.gov/news/documentsingle.aspx?DocumentID=396

Popken, B., Engel, R., Benyon-Tinker, K., & Ghosh, M. (2019, August 8). Russia-linked Twitter accounts promoted 'doxxing' over racial tension videos. Retrieved from https://www.nbcnews.com/tech/tech-news/russia-linked-twitter-accounts-promoted-doxxing-over-racial-tension-videos-n1040596

Reicher, S. D., Spears, R. & Postmes, T. (1995) A social identity model of deindividuation phenomena. *European Review of Social Psychology*, 6, 161–198, doi:10.1080/14792779443000049

Shane, S. & Mazzetti, M. (2018, September 20). The plot to subvert an election. *The New York Times*, Center insert, 1–11. Retrieved from: https://www.nytimes.com/interactive/2018/09/20/us/politics/russia-interference-election-trump-clinton.html

Sherif, M., Sherif, C., & Nebergall, R. (1965). *Attitude and attitude change: The social judgment-involvement approach.* Philadelphia, PA: W. B. Saunders.

Strudwicke, I. J., & Grant, W. J. (2020). #JunkScience: Investigating pseudoscience disinformation in the Russian Internet Research Agency tweets. *Public Understanding of Science*, 29, 459–472 doi: 10.1177/0963662520935071

United States of America v. Internet Research Agency LLC. District of Columbia (2018). Retrieved from https://www.justice.gov/file/1035477/download

Yang, X., Macdonald, C., & Ounis, I. (2018). Using word embeddings in Twitter election classification. *Information Retrieval Journal*, 21, 183–207 doi:10.1007/s10791-017-9319-5

Zhang, Y., Lukito, J., Su, M., Suk, J., Xia, J., Kim, S.J., . . . Wells, C. (2021) Assembling the networks and audiences of disinformation: How successful Russian IRA Twitter accounts built their followings, 2015–2017, *Journal of Communication*, 71, 305–331, https://doi.org/10.1093/joc/jqaa042