Touching Trolls – How Real People Respond to a Coordinated Information Operation and Why They’re So Nice

Darren L. Linvill
Associate Professor
Clemson University
Department of Communication

Patrick L. Warren
Associate Professor
Clemson University
John E. Walker Department of Economics

Amanda E. Moore
Sr. Manager of Integrated Marketing and Partnerships
Loggerhead Marinelife Center
Abstract
From 2015 to 2017, the Russian-affiliated Internet Research Agency operated a large, clandestine online influence operation that targeted political conversation in the United States. The goal of this research is to better understand how real social media users interacted with IRA persona so that we may gain insight into their agenda-building efforts. We combine hand coding with natural-language processing to measure the ways in which real users talked with and about the accounts making up the IRA operation, in the month before the 2016 U.S. Presidential Election. We find that these interactions were overwhelming positive, with most mentions being supportive of the persona, belying the standard characterization of these persona as “trolls”. When we measure the success of these accounts, in terms of engagement measures like follower growth or retweet rates, we find that on days that the persona's mentions were very (but not unanimously) supportive, they received more engagement than when their mentions were more combative. This pattern suggests that a strategy of building homophilic connections with like-minded people was central to the IRA campaign.
Introduction

State affiliated disinformation has been and continues to be an issue across many social media platforms. The presence of such disinformation online became public 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). Central to the IRA’s attempt to affect the election was a sophisticated campaign using social-media platforms to encourage division, discontent, and disconnection with reality among potential U.S. voters. Since 2016, however, state affiliated disinformation has expanded. Operations targeting U.S. social media users have continued to be identified which are believed to stem from Russian sources (Popken, Engel, Benyon-Tinker, & Ghosh, 2019), but also Iran (Newcomer, Wagner, & Sebenius, 2019) and China (Cohen, Wells, & McGinty, 2019) among others.

The work of the IRA and similar state-affiliated actors online has moved traditional propaganda and disinformation efforts into a new space. Social media has the potential for a broader reach and lower cost of entry for potential bad actors relative to other media. As it is more interactive than past modes of social influence, it is important to look beyond the message form and source of social media messages to the processes at work. For this reason, past work exploring state-affiliated social-media disinformation has examined it through the lens of agenda-building (Linvill et al., 2019).

Cobb and Elder (1971) defined agenda-building as the process by which actors endeavor to move issues from their own agenda onto the agendas of policy makers. Since its original
conceptualization, agenda-building has been applied in a variety of ways. Denham (2010) defined three distinct but overlapping areas of work done using differing conceptualizations of agenda building: policy agenda building, media agenda building, and public agenda building. Policy agenda building addresses how issues are created, expanded upon, and consequently enter the public policy agenda. Media agenda building explores how media agendas are built, “reflecting institutional imperatives and an ongoing negotiation between media personnel and their sources of information” (p. 311). Lastly, public agenda building applies to “behavioral responses to mass and interpersonal communication. Examples of such responses might include voting for a particular policy action, attending an event, or offering financial support to a social movement” (p. 316). It is this final category, public agenda building, that relates most directly to social media disinformation.

While IRA social-media activity prior to the 2016 U.S. Presidential campaign was directed at both journalists (Lukito & Wells, 2018) and politicians (Gallagher, 2018), overall, it seems probable the IRA’s chief efforts were directed at public agenda building. The IRA focused on differing, often seemingly contrary agenda in their disinformation campaigns. Influencing votes, attendance at events, and support for particular social movements were central, however, in all of their efforts (Shane & Mazzetti, 2018). Parmelee (2014) found that political posts on Twitter have the potential to influence mass publics. Similarly, Kahne and Bowyer (2018) found that online activity fosters political participation. It is through such mechanisms that the IRA likely hoped to influence the United States, its people, and its government.

The goal of this research is to better understand how real social-media users interacted with IRA trolls so that we may gain insight into their agenda-building efforts. Due to Twitter’s credible transparency regarding disinformation on its platform (Gadde & Roth, 2018), more is known to
date about Russian 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 purported themselves 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 the 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 which the IRA seemed to focus on in their campaign (Broniatowski et al., 2018; Yan et al., 2019; Strudwicke & Grant, 2020).

While some scholars have argued the IRA’s activity played a significant, perhaps even pivotal, role in the outcome of the 2016 election (Jamieson, 2018), little work has explored quantifiable effects of their efforts. One exception is the work of Bail et al. (2020); this research found no evidence interaction with IRA Twitter accounts from late 2017 had any impact on user’s political attitudes or actions. This research 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 goal of IRA activity. Other possible strategic goals the IRA may have maintained include ideological entrenchment, the sowing of distrust for alternate views, agenda building around particular issues, or, perhaps most likely, some combination of various goals. Analysis of actual user engagement with IRA accounts may help us better understand actual goals of the IRA’s online activity and how they worked to achieve them.

To date, 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), scholars have yet to examine the nature of user engagement with fake social media accounts such as those
created by the IRA. There is a lack of understanding regarding how users respond when unknowingly confronted with fake, professionally built social media accounts sowing disinformation. Deeper knowledge of the behavior real users express in response to social media disinformation may help us to infer the tactics and strategy that organizations such as the Russian IRA employ to reach and influence their targets. This knowledge may also help us infer what that desired effect may be. Freelon et al. (2020) argued it would be valuable to explore user responses to IRA trolls, specifically suggesting examination of user responses on Twitter as many such responses may be a valuable tool still accessible to researchers. With this in mind, this study aimed to explore the following research question:

RQ1: How do real users engage with Internet Research Agency 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 manufactured twitter accounts. 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. As Linvill and Warren (2019) argued, IRA trolls “don’t go to social media looking for a fight; they go looking for new best friends.” Hence:

H1: The majority of user engagement with IRA trolls will be supportive in nature.
Our study employed Linvill and Warren’s (2020a) typology of 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. 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. A large number 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. Our use of these categories enabled our final two research questions.

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.
Even within account types, there is also some evidence that the IRA accounts’ behavior varies over time (Linvill & Warren, 2020b). This variation over time includes both volume of output and mix of networked behavior. The timing also seems to vary by account type. Thus, it might also be important to investigate how engagement varies by both account type and time, therefore:

RQ 2: How, if at all, does the volume and character of engagement with Internet Research Agency English language troll accounts vary in the month prior to the 2016 U.S. presidential election

(A) by the type of the troll account being mentioned, and

(B) over time.

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. Do trolls, as the saying goes, catch more flies with honey than with vinegar? With this in mind we ask:

RQ 3: How, if at all, does the volume and character of engagement with Internet Research Agency 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?
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 (Gentzkow, Wong, & Zhang, 2018), or motivates the targeted group to engage in some desired action, such as voting (Dellavigna & Kaplan, 2007; Gils, Muller, and Prufer, 2020). It is also consistent with related theories of persuasion. Social judgement 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. Cognitive dissonance theory (Festinger, 1957), meanwhile, teaches us that situations in which individuals are faced with possibly changing their attitudes or beliefs are uncomfortable and the theory suggests we therefore seek experience to reduce discomfort and reinforce existing beliefs. Together, these theories and models suggest two further hypotheses, related to RQ2 and RQ3:

H2: Tweets from ideologically specialized oriented troll accounts (left trolls and right trolls) will be targeted in such a way to 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.

Method
Data collection

Salesforce’s Social Studio platform enabled keyword searches of tweets mentioning known IRA Twitter account handles. A list of these account handles was released on June 18, 2018 by the U.S. 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 ran up to and overlapped with the 2016 U.S. Presidential election – a period when the IRA was highly active. The combined searches resulted in 117,626 replies and 96,539 quote tweets which included an IRA troll account name, together with another 38,333 tweets that simply mentioned a troll. All these mentions were downloaded for analysis.

Figure 1 presents the trends in mentions of IRA accounts on Twitter by account type over time. The conversation is dominated by mentions of right troll accounts, which are, in turn, dominated by mentions of @TEN_GOP. But the other accounts and account types also received hundreds of mentions on many days. Figure A1 provides additional detail. Although mention counts randomly vary from day to day, there are substantial spikes (1) on election day, for right trolls, (2) around November 1st, for left trolls, and (3) on October 9th and 19th (the final two Presidential debates), for the hashtag gamers.

Qualitative Hand Labelling

To start our qualitative analysis, we examined random selections of data as recommended by Corbin and Strauss (2015). First, we read tweets to get a sense of the data. From there, we
conducted unrestricted open coding, working together to examine, break down, compare, and conceptualize data. From this process we identified meaningful patterns.

Second, we conducted axial coding by comparing and reducing these patterns. Four broad categories were identified through axial coding (see Results). As this process continued, we created definitions for each category and clarified their meaning. 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 includes an active link to the original tweet on the Twitter platform). In many cases, reading through the Twitter thread which a tweet was a part of (i.e. the conversation the tweet engaged in) was necessary for proper interpretation. It should be noted, however, that this task was made difficult given that the IRA accounts and their tweets are suspended and 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 engaged in peer debriefing, the use of a code book, and intercoder reliability. Peer debriefing (Creswell & Miller, 2000) involves recruiting an individual familiar with the phenomenon being explored but external to the research team to play the role of devil’s advocate. A code book was developed using the definitions and example tweets developed and identified during axial coding. The use of a code book served as a stable representation of the coding analysis to serve as a reference throughout the coding process (Creswell & Poth, 2018). Employing this code book, the three members of the research team coded randomly selected sets of 50 tweets. After each set was coded we compared results and refined our analysis. This process was continued until a Krippendorff’s alpha reliability of .76 was met (Krippendorff, 2004). Following the successful completion of reliability analysis a random sample
of 5000 tweets was selected and distributed among the research team for analysis. Of these tweets, 4316 mentioned right troll accounts, 483 mentioned left troll accounts, 58 mentioned hashtag gamer accounts, and 143 mentioned news feed accounts. The research team placed each of the 5000 sample tweets into one of the four categories or labeled it as unknown.

**Predictive Binomial Labelling**

Addressing RQ2 and RQ3 effectively requires a more extensive data set of labelled tweets than can feasibly be implemented by hand. Fortunately, machine-learning techniques exist to predictively extend a hand-labelled sample to a larger population of similar documents. As the overwhelmingly dominant label in our sample was “Support,” we reduced this predictive labelling exercise to a simply binomial problem of predicting whether the tweet would have been labelled as “support” or not.

Formally, we estimate a L2-normed (penalized) logistic regression where the probability that a given tweet would have been labelled 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 choose 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 use the entire 236,744 tweet corpus to build a word embedding space, where each word that is used at least 5 times in the overall corpus is projected into a 75-dimensional vector. 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 collapse from the word level to the tweet level by taking the average across words in the tweet along each of the 75 dimensions, in order to capture an overall 75-dimensional summary of what the tweet is about.

We estimate this model on a training sub-sample of 4000 of the labelled tweets and apply it to the 1000 held back tweets in order 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. Figure 1 presents a locally linear nonparametric regression relating our predicted probability of a tweet being supportive to the share of tweets that actually are (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 under-prediction of supportiveness for relatively high levels of predicted supportiveness. As our predictions seem quite accurate, on average, and we will average many tweets in our metrics of supportiveness, this estimated proxy could serve our needs.

We apply this model to the entire 236,744 tweets in our unlabeled dataset in order to calculate a predicted probability of being supportive for each mention. We will treat these predicted probabilities as data when analyzing RQ2.\textsuperscript{1} Specifically, we present the average predicted probability of support by account time over time.

Relating Mention Mix to Engagement

\footnote{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.}
To evaluate the relationship between supportive mentions and engagement, we shift 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 three ways: follower-count growth rates, the number of retweets received by tweets by the account on that day, and number of total mentions the account receives.

For each IRA troll account that appears in the Twitter database which is active in this period and which appears in the mention dataset, we can 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. Formally, we conduct regressions of the form

\[ y_{it} = a_i + b_t + B_l 1(p_{it} < 0.25) + B_h 1(p_{it} > 0.5 \& p_{it} < 0.75) + B_{vh} 1(p_{it} > 0.75) + e_{it}, \]

where \( y_{it} \) represents one of our metrics of engagement, \( p_{it} \) represents our metric of supportiveness, \( a_i \) are 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 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 there will certainly be autocorrelation in unmodelled determinants of engagement (\( e_{it} \)), we will cluster standard errors by account for inference.
There are several sample-selection issues that arise in this setting. First, to calculate the mention positivity estimate requires that the account receives some mentions. Second, to calculate follower count growth rates, requires observing follower counts over multiple days. Given how our data are derived, where we observe follower counts only when the account tweets, that essentially requires the account to be sufficiently active. 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 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. We can only measure the patterns in these jointly-determined correlations.

**Results**

**Qualitative Results (RQ1)**

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. 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) or contained links that were no longer active and therefore lacked full context for interpretation. Qualitative analysis placed the remaining tweets (n = 4489) into one of four categories: attack, support, troll whistling, and unrelated comment. 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 4489 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 !” Another example of an indirect attack included a October 31, 2016 tweet “hahahaha oh wow. The social justice is strong in this one...” in response to another user on a thread started by the newsfeed troll account @ChicagoDailyNew.
Support \((n = 3459, 76.1\%)\). 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, 2018).

This category included 216 tweets which were indirect support \((6.2\% \text{ of the category})\). 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, but one in which the troll account was no longer involved. Some of these tweets defended the troll account from users who criticized it. 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. Many of the tweets in this category, however, simply further engaged in discussion initiated by an IRA troll account. This included an October 25, 2016 tweet, “police hiring really needs to be reformed. Who in the hell is in charge of hiring and training these losers?” This tweet was in response to another user who quote tweeted left troll @BleepThePolice’s tweet about the New York Police Department’s involvement in the death of a child.
Troll whistling ($n = 126, 2.8\%$). Tweets in this category included no significant message, but mention a list of account handles, including the troll account. This is a common tactic on Twitter done to call the mentioned accounts’ attention to a particular 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 tweet “Hamburger Tom sounds like a mafia name.” This was in reply to a tweet from the right troll @Jenn_Abrams and 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 in a message to another user on a right troll thread regarding how to best use Twitter: “Psst, you're supposed to reply after the name of who you're addressing.”

H1 is confirmed 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.
In the hand-coded data, the distribution of these categories does vary slightly across account type. Figure 3 presents the mention category shares across account types. Right and left trolls have similar distributions, although they do differ statistically, taken as a whole. Looking specifically at supportiveness, left trolls and right trolls have statistically and substantively similar shares of supportiveness, while hashtag gamers and newsfeed have lower shares of supportiveness, both substantively and statistically. This provides some evidence in support of \textbf{H2}.

\textbf{Quantitative Analysis}

\textbf{RQ2}

Figure 4 illustrates how our estimate of the supportive mention share over time and 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. For hashtag gamers, the estimated share of supportive tweets is both lower and less consistent, with a spike in the share of supportive mentions in the day or two right 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-section gap between newsfeeds and the ideological trolls, however, does not seem 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 \textbf{qualified support for H2}.

\textbf{RQ3}
The relationship between supportive mentions and metrics of engagement depends on the category of the troll account. Figure 5 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 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, in any case. For hashtag gamers, accounts-days with maximal growth are those with a more even mix of estimated supportiveness.

Table 1 presents the regression version of these results, as specified in equation (1). It shows that, for left and right trolls, account-days in which mention supportiveness is unusually very high are also those account-days in which those accounts grow unusually fast. Relative to days in which the level of support is moderately low (25 to 50 percent supportive), accounts on very high support days between .5 and .8 percentage points higher daily growth rates. For these account types, no other contrasts are statistically significant. For newsfeeds, there is no significant relationship between growth and mentions supportiveness. For hashtag gamers, high mention-
supportiveness days correspond to high follower growth rates, about a third a percentage point more than in low mention-supportiveness days.

Figure 6 and Table 2 present the parallel analysis for a different metric of engagement, retweet rates (transformed as \( \log(1+x) \)). 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 percent supportive. In regressions, there is no statistically significant pattern for left trolls, but right trolls receive about 40 percent more retweets on very positive mention days. In contrast to the first metric, there is a substantial relationship for newsfeeds between supportive mentions and retweets, with about 70 percent more retweets on positive mention days. Hashtag gamers get about 20 percent more, although very high mention days include smaller and not statistically significant larger retweet rates for both of these last two account types.

Finally, Figure 7 and Table 3 present the parallel analysis for our final measure of engagement, mention counts (transformed as \( \log(1+x) \)). The patterns for this metric of engagement are nearly identical as the retweet metric, with very similar locally-linear graphs, and a similar pattern of regression coefficients. The two marked differences is that very low supportive-mention days are particularly bad for mention engagement, for all types but left trolls, and, for hashtag gamers, very high supportive-mentions days also have relatively low mention levels, but both of these results should be interpreted with care, as extreme results in support estimates may be directly driven by low number of mentions.

We interpret these results as **qualified support for H3**, where ideological types seem to benefit from high levels of supportive mentions, as long as those mentions are exclusively supportive. For hashtag gamers, by contrast, a more mixed response of moderate supportiveness seems to be more associated with engagement. For newsfeeds, 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 or mention engagement, however, the pattern seems to fall between newsfeeds and the left/right troll types, with moderately high levels of support being mostly highly related to high levels of engagement.

Finally, the predictions that supportiveness and engagement would be most highly correlated for the ideological types is confirmed.

Discussion

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). Our findings suggest that, by such a definition, the name “Russian troll” may be inappropriate for the work of the IRA. In the month before the 2016 election, a period doubtless of critical importance to the organization, IRA posts solicited what were overwhelmingly positive responses from users on Twitter. It can be inferred that these responses are a result of some combination of the nature of IRA messaging and who they target with said messages.

This pattern of positivity should be understood in the context of identity group infiltration (Arif, Steward, & Starbird, 2018; Freelon & Lekot, 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 passionate and motivated groups, with strong ethnic or political identities. While this study could not examine the relative effect of IRA messaging on different audiences, our findings suggest their posts were carefully tailored to their specific, chosen audience. It is unlikely messages from left or right trolls would have had the same
outcomes had they been distributed to a more general public. 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 were beginning to transition 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 network. So high levels of supportiveness we find, are—in some sense—supporting not only the Trolls themselves but also the content originators who 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 worth note is the relationship between the percentage of IRA tweets receiving supportive response and the volume of IRA tweets seen in Figure 7. 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. Their tactics have changed in the ensuing years (Alba,
2020). 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. Future research will need to explore these concerns. None-the-less, this study clearly demonstrates the proverb that one catches more flies with honey than with vinegar has a universal truth that also applies to social media disinformation.
References


doi:10.1080/15358593.2010.502593


doi:10.1080/10584609.2020.1718257

https://10.37016/mr-2020-011


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


Figures and Tables

Figure 1. Daily Mentions of Troll Accounts of Indicated Type over Time.

Figure 2. Smoothed Actual Fraction Supportive and Function of Predicted Probability of Being Supportive (Left) and Distribution of Predicted Probabilities (Right) in n=1000 Hold-Out Sample
Figure 3. Mention Type Shares by Troll Account Type, Hand-labelled Sample.
Figure 4. Mean Probability Supportive of Daily Mentions of Troll Accounts of Indicated Type over Time.
Figure 5. Locally Linearly Smoothed Percent Change in Followers (Left) and Count of Account Days (Right) by Daily Mean Probability Supportive.
Figure 6. Locally Linearly Smoothed Retweets of Tweets Produced by Troll Accounts (Left) and Count of Account Days (Right) by Daily Mean Probability Supportive.
Figure 7. Locally Linearly Smoothed Mentions of Troll Accounts (Left) and Count of Account Days (Right) by Daily Mean Probability Supportive.
Table 1. Average Probability Supportive and Follower Count Growth Rate

<table>
<thead>
<tr>
<th>Pct. Support</th>
<th>Left Troll</th>
<th>Right Troll</th>
<th>Newsfeed</th>
<th>Hashtag Gamer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0--25</td>
<td>0.29</td>
<td>1.25</td>
<td>-0.0</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>(0.39)</td>
<td>(1.07)</td>
<td>(0.02)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>50--75</td>
<td>0.16</td>
<td>0.11</td>
<td>0.01</td>
<td>0.37**</td>
</tr>
<tr>
<td></td>
<td>(0.28)</td>
<td>(0.18)</td>
<td>(0.03)</td>
<td>(0.16)</td>
</tr>
<tr>
<td>75--100</td>
<td>0.83*</td>
<td>0.53**</td>
<td>0.01</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>(0.48)</td>
<td>(0.21)</td>
<td>(0.01)</td>
<td>(0.36)</td>
</tr>
</tbody>
</table>

Account FE Yes Yes Yes Yes
Day FE Yes Yes Yes Yes
Obs. 744 670 669 556

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 25 to 50 percent 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 (see sec. x). *:p<0.10, **:p<0.05, ***:p<0.01.
### Table 2. Average Probability Supportive and Retweets

<table>
<thead>
<tr>
<th>Pct. Support</th>
<th>Left Troll</th>
<th>Right Troll</th>
<th>Newsfeed</th>
<th>Hashtag Gamer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0--25</td>
<td>0.0</td>
<td>-0.07</td>
<td>-0.55</td>
<td>-0.15</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.24)</td>
<td>(0.57)</td>
<td>(0.18)</td>
</tr>
<tr>
<td>50--75</td>
<td>0.04</td>
<td>0.2</td>
<td>0.7**</td>
<td>0.22**</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.22)</td>
<td>(0.36)</td>
<td>(0.13)</td>
</tr>
<tr>
<td>75--100</td>
<td>-0.03</td>
<td>0.44*</td>
<td>0.39</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.23)</td>
<td>(0.37)</td>
<td>(0.15)</td>
</tr>
</tbody>
</table>

Account FE | Yes | Yes | Yes | Yes
Day FE     | Yes | Yes | Yes | Yes
Obs.       | 767 | 689 | 693 | 561

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 25 to 50 percent supportive. Standard errors are clustered at the level of the account. Accounts are omitted on days that they are not mentioned did not tweet. *:p<0.1, **:*p<0.05, ***:*p<0.01.
### Table 3. Average Probability Supportive and Mentions

<table>
<thead>
<tr>
<th>Pct. Support</th>
<th>Left Troll</th>
<th>Right Troll</th>
<th>Newsfeed</th>
<th>Hashtag Gamer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0--25</td>
<td>0.09</td>
<td>-0.36**</td>
<td>-1.22**</td>
<td>-0.33***</td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.18)</td>
<td>(0.6)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>50--75</td>
<td>0.28**</td>
<td>0.13</td>
<td>0.61**</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.08)</td>
<td>(0.29)</td>
<td>(0.1)</td>
</tr>
<tr>
<td>75--100</td>
<td>0.13</td>
<td>0.13</td>
<td>-0.05</td>
<td>-0.31**</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.33)</td>
<td>(0.15)</td>
</tr>
</tbody>
</table>

Account FE: Yes, Yes, Yes, Yes
Day FE: Yes, Yes, Yes, Yes
Obs.: 744, 670, 669, 556

Note: Panel regression relating aggregate mention count for the account each 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 25 to 50 percent supportive. Standard errors are clustered at the level of the account. Accounts are omitted on days that they are not mentioned. *:p<0.10, **:p<0.05, ***:p<0.01.
Figure A1. Timeline of Mentions of IRA Accounts of the Indicated Type/Account.
Figure A2. Fraction Supportive of Daily Mentions in Labelled Sample of Troll Accounts of Indicated Type over Time.