Touched by the Trolls: How and Why a Coordinated Information Operation Interacts with Outsiders

Darren L. Linvill  Patrick L. Warren*

October 6, 2019

Abstract

From 2015 through 2017, the Russian-government affiliated Internet Research Agency produced nearly 2.8 million English-language tweets from accounts that purported to be operated by U.S. nationals or organizations ("trolls"). Almost half of the trolls’ output were retweets of or replies to other accounts, overwhelmingly from outside the network. We analyze the characteristics of outside accounts that were targeted by the trolls in this way, and how this behavior changed over the life of the operation, in order to infer what role contacts with outsiders played in the trolls’ propaganda strategy. We document the three stage life-cycle of these externally-oriented trolls: introduction, growth, and amplification. In a quasi-experiment of the amplification stage, we estimate that in the month leading up to the 2016 U.S. Presidential election, the trolls induced about 4 million additional tweets from and 3 million additional followers for the 25,000 unique accounts they amplified, an impact that rivals the direct output of the trolls themselves over the entire three-year campaign.

*Linvill: Department of Communication, Clemson University. Email: darrenl@g.clemson.edu
Warren: John E. Walker Department of Economics, Clemson University. Email: patrick.lee.warren@gmail.com. Address: 222 Sirrine Hall, Clemson, SC, 29634. This research was supported by a grant from the Charles Koch Foundation.
1 Introduction

Attempts by the Russian-government affiliated Internet Research Agency (IRA) to influence Western political dialogue on social media are now well known. The organization’s work was comprised of millions of social media posts across every major platform (Timberg and Romm, 2018). Because of the laudable transparency of Twitter (Roth and Gadde, 2018), more is publicly known to date about the content of Russian disinformation on their platform than most others. From the beginning of 2015 through the end of 2017 Twitter has released approximately 2.8 million English-language tweets produced by the IRA from accounts that purported to be operated by U.S. nationals or organizations ("trolls").

Most work exploring this content has focused on the identities the troll accounts took on and the content they produced (DiResta et al., 2018; Howard et al., 2018; Linvill and Warren, 2018). Some research has examined specific issues which the IRA seemed to focus on in their campaign (Broniatowski et al., 2018; Yan et al., 2019). This emphasis on content is common among analysis of disinformation, more generally (King et al., 2017).

In this paper, instead, we focus attention on “networked output” and on the other Twitter users with whom the trolls interacted. We document who the IRA accounts engaged with, how that engagement changed over time, and estimate some effects this engagement had on these accounts. Specifically, we analyze the timing and type of networked contacts and the characteristics of outside accounts in order to infer the role these activities played in the trolls’ overall propaganda strategy. This focus is distinct form a stream of research that asks who shared the content the trolls produced (Badaway et al., 2018b,a) and what online communities they were networked with Stewart et al. (2018).

The IRA accounts were thematic, and the accounts that interacted most with outsiders fell into one of three themes: Right Troll, Left Troll, and Hash-tag Gamers. Figure 1 plots the daily tweet output from accounts of these three types (data from Linvill and Warren (2018)). Left Trolls represented themselves as individuals and organizations from the far left of the U.S. political distribution, mostly members of the Black Lives Matters community.
Right Trolls similarly represented themselves as from the far right—extremely pro-gun, anti-immigrant, and pro-Trump. Finally, Hashtag Gamers operated and/or participated in a social-media improvisation game in which players created and promoted jokes and commentary on some pre-specified topic of conversation (e.g. #RejectedDebateTopics or #ObamasNextJob). These accounts were active intermittently, and seemed to be particularly active around salient pre-scheduled political events, such as the Presidential debates. The first panel of Figure 2 presents the most prominent words in the account descriptions that the IRA accounts of these three types. For details on these themes (and the two others, News Feeds and Fearmongers), see Linvill and Warren (2018).

Over 135,000 unique accounts were retweeted by the trolls throughout the campaign, and 87 percent of those accounts still existed as of January 22, 2019 (when we collected the account details, see below). The surviving accounts had median follower count of 2,241, median following count of 1,057, and median tweet count of 22,072. 12.9 percent of the surviving accounts were verified. We compare these statistics to those from accounts referenced in a random sample of retweets/replies of political tweets that were produced between Oct 1 and Oct 7, 2016. Compared to those retweeted by the trolls, these random political retweeted accounts were larger (median follower count of 5,398), more prominent (21.3 percent verified), and more active (median tweet count of 34,206). This pattern holds for other ways that the trolls interacted with external accounts, such as replies and mentions. These consistent differences suggest that the Trolls chose to interact primarily with relatively small accounts.

But exactly how the trolls interacted with outsiders varied considerably over both calendar time and the life-cycle of the troll account. We identify three periods of networked activity: introductory, growth, and amplification. During the introductory period, the trolls established the identify of their accounts and tried to place them in specific sub-networks. This period is characterized by high levels of replying, both among troll accounts and between those troll accounts and outside accounts that are thematically aligned with the trolls. During the growth period, the troll accounts gained followers from the organic network through a mixture of retweets and “original” (or plagiarized) content.
Finally, during the amplification period, the troll network used its, by now well established, accounts in order to amplify the content and prominence of outside accounts of their choosing. Amplification took place both directly, by increasing the potential audience for a retweeted message, and indirectly, by increasing the activity and extending the network of the retweeted account. The amplification period is characterized by both a jump in total output and a shift into nearly exclusively retweeting of outside accounts. We see this period as the culmination of the life-cycle of these troll accounts.

To document the consequences of this amplification stage, we investigate follower, following, and output counts for external accounts that were retweeted during the first day of the amplification period. This transition occurred on October 6th, 2016, one month before the 2016 U.S. Presidential election and one day before the Wikileaks release of John Podesta’s hacked emails. Using a difference-in-differences strategy with a matched set of control accounts that tweeted about politics but were not retweeted by the trolls, we estimate that the retweeting of 4000 external accounts on October 6th led to 500,000 additional followers and 600,000 additional tweets over the following 4 days for those external accounts. Extending these estimates to the 25,000 unique accounts retweeted in the month before the election increases the overall impact to 3 million additional tweets and 4 million additional followers. If these estimates are accurate, the amplification activity in the final month of the 2016 election could have had a more significant contribution to the political conversation on Twitter than all the direct activity the IRA produced over the 3 years documented here, combined. This raises the possibility of a new type of agenda setting, where a substantial impact of an informational operation is how it changes the voices that are prominent in a political conversations and, thereby, indirectly changes the tenor of the conversation.

2 Data

The data for this analysis come from four sources. First, the output of the IRA-affiliated accounts and the identities of the accounts they retweet and reply to come from Twitter’s January 2019, update to their October 2018 release of
the full tweet output of 3613 accounts that they linked to the IRA (Gadde, 2019). Second, we draw on our prior work to provide the categorization of these accounts into thematic types (Linvill and Warren, 2018).

Third, we derive follower, following, and update counts of the IRA accounts over time from Social Studio, a social media analytics platform that collects a nearly real-time record of Twitter output. These data are available from the moment each tweet is gathered, usually immediately after the tweet is posted, but must the interpolated for days in which no tweets are produced/scraped. We also use Social Studio to identify a random sample of non-IRA accounts that tweet about Trump or Clinton in the first week of October 2016 and track the same statistics about those accounts throughout the first 10 days of October 2016.

Finally, we extract account characteristics and descriptions from the Twitter User Look-Up API, as of January 2019. These characteristics include follower, following, and update counts; account description and location; and whether the account is verified. These features are available for any account that still exists and is not suspended as of January 22, 2019. We gather this information for all external accounts linked to by the IRA and for the “control” accounts identified in the Social Studio data.
3 Networked Output

Although the focus of this study is external networked activity, and this external activity represents the majority of networked activity, it is useful context to understand how the IRA networked their accounts internally. Figure 3 illustrates the number of times troll accounts retweeted and replied to other troll accounts of the indicated type, in each week, throughout 2015-2018. The top 4 timelines in each panel illustrate how retweets and replies were used internally, where the titles above the timelines indicate the type of troll that originally produced the content that was retweeted/replied to, and the colors of the lines indicate the type of troll doing the retweeting/replying, with blue indicating tweets by Left Trolls, red for Right Trolls, and green for Hashtag Gamers. Note that the scales of these graphs vary widely.

The internal mentions of Right Trolls, Left Trolls, and Hashtag Gamers originated overwhelmingly from other accounts of the same type. This homophilic networked activity likely served a dual purpose of elevating the profile and authenticity of the originating accounts and of reducing the costs of producing novel content on the retweeting accounts. Internal replies to other trolls of the same type, in particular, spiked in mid-2015, for the Left and Right Trolls, creating a false sense of these accounts being active participants in a conversation.

A fourth account type, News Feeds, which purported to be legitimate providers of local news for over 30 large U.S. cities, originated no retweets or replies but were retweeted, themselves, by the Right and Left Trolls in very large numbers in mid-2015. Again, the goal of these retweets was likely to establish the authenticity of these putative local-news sources. These News Feeds were also replied to by the Left Troll accounts in substantial numbers in late 2015.

The wider timelines at the bottom of each panel indicate the patterns of networked activity directed at accounts outside the IRA network. The type of troll account producing the content is, again, indicated by the color of the timelines. These external replies and retweets are the primary subject of this inquiry.
3.1 Targeted Account Characteristics

Before analyzing the strategy underlying the timelines of networked output, we will first summarize several qualitative and quantitative characteristics of the accounts “touched by the trolls.” We begin with the qualitative, establishing the thematic character of these accounts, before turning to a quantitative reckoning of their size and levels of activity.

Although it is not appropriate to apply the typology we introduced for the IRA-affiliated accounts to the outside accounts, we can nevertheless see strong evidence of homophilic networked activity with these external retweets and replies. Figure 2 presents word clouds of the account descriptions associated with several sets of accounts. The first panel presents an output-weighted word cloud of the most common words included in the final account description of the actual IRA trolls of the indicated types, where the size of the word indicates its prominence. The second panel presents similar word clouds for the accounts that were retweeted by troll accounts of the indicated types, where the account descriptions were gathered in January 2019. There is significant overlap in these sets. For example, “Black” is a very prominent term in both the descriptions of the Left Trolls and the accounts retweeted and replied to by Left Trolls, as are “MAGA” and “Conservative” for Right Trolls and those retweeted and replied to by the Right Trolls, and “Game”, “Hashtag”, and “Host” for the Hashtag Gamers.

But there are some consistently prominent words in the descriptions of retweeted accounts that are not prominent in the original troll accounts. These include several terms that indicate that the account is a member of the media, including words like “News,” “Author,” “Host,” and “Radio.” These words are even more prominent in the word clouds representing the descriptions of the accounts that the trolls replied to. This combination of similarities and differences suggests that the trolls targeted ideological similar accounts, to some extent, but perhaps also tried to target media sources, especially when replying.

Quantitatively, it is useful to compare the accounts retweeted and replied to by the trolls to a set of representative politically engaged accounts, in order
to recognize if their behavior differs substantially from “average.” To do so, we collected a random sample of 2 million tweets that mentioned “Trump” or “Clinton” in the first week in October, 2016. From these tweets we selected the (1.3 million) retweets and (10.2 thousand) replies and gathered the account information for the accounts that were retweeted or replied to. This set of 17,215 accounts (2,830 for replies) forms the basis of comparison for our quantitative analysis.

The first quantitative feature we display is the distribution of January 2019 follower counts among accounts retweeted and replied to by the trolls. These, together with the follower counts of random political accounts (in black) are displayed in Figure 4 and summarized in Table 1, where the histograms for the troll retweets are colored consistent will all the earlier graphs (blue for Left, red for Right, green for Hashtag Gamers). The follower counts for both the trolls and the controls are single peaked and fairly close to log normally distributed, with fat right tails. The accounts retweeted and replied to by trolls are much smaller, in terms of followers, than those retweeted by random political accounts. The difference is larger for replies, but that difference is mostly driven by the fact that random accounts reply to bigger accounts than they retweet, on average, while the trolls show only small and inconsistent differences.

Consistent with these differences in follower counts, we also see large differences between the fraction of targeted accounts that are verified, with all troll types retweeting and replying to a smaller share of verified accounts than random political accounts do. Moreover, both trolls and random political accounts reply to a greater share of verified accounts than they retweet.

Finally, we can measure the activity levels of the networked accounts by looking at the number of status updates (tweets) those accounts have produced. For retweets, these results are broadly consistent with the other factors, the trolls retweet and reply to less active accounts than random political accounts do.

To sum up, the trolls focused their networked activity on accounts that were ideologically/thematically similar to themselves, which is consistent with their in-network activity, but they also included some official or media outlets. The
Table 1: Networked Account Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Left Troll</th>
<th>Hashtag Gamer</th>
<th>Right Troll</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Retweets</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median Follower Count (000s)</td>
<td>2.4</td>
<td>1.0</td>
<td>3.4</td>
<td>5.4</td>
</tr>
<tr>
<td>Mean Follower Count (000s)</td>
<td>130</td>
<td>94</td>
<td>131</td>
<td>149</td>
</tr>
<tr>
<td>Pct. Verified</td>
<td>16.6</td>
<td>5.4</td>
<td>14.2</td>
<td>21.3</td>
</tr>
<tr>
<td>Median Status Count (000s)</td>
<td>26</td>
<td>12.7</td>
<td>24</td>
<td>34</td>
</tr>
<tr>
<td>Mean Status Count (000s)</td>
<td>58</td>
<td>36</td>
<td>55</td>
<td>74</td>
</tr>
<tr>
<td><strong>Replies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median Follower Count (000s)</td>
<td>2.6</td>
<td>1.6</td>
<td>2.8</td>
<td>7.9</td>
</tr>
<tr>
<td>Mean Follower Count (000s)</td>
<td>954</td>
<td>318</td>
<td>537</td>
<td>667</td>
</tr>
<tr>
<td>Pct. Verified</td>
<td>23.5</td>
<td>8.7</td>
<td>19.7</td>
<td>32.7</td>
</tr>
<tr>
<td>Median Status Count (000s)</td>
<td>24</td>
<td>20</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td>Mean Status Count (000s)</td>
<td>61</td>
<td>47</td>
<td>51</td>
<td>69</td>
</tr>
</tbody>
</table>

Note: Statistics as of January 22, 2019 for accounts retweeted and replied to by IRA trolls of the indicated type or which had tweets they created mentioning ”Trump” or ”Clinton” retweeted or replied to in Oct 1–Oct 9, 2016. Accounts that were suspended or deleted before that date are not included.
thematic similarity was stronger for retweets than replies, which seemed more focused toward (formal or informal) media accounts. Despite the inclusion of these public-facing accounts, the set of accounts retweeted and replied to be the trolls were relatively obscure. Compared to random political accounts, the trolls produced networked output that connects to smaller, less prominent, and less active accounts.

This pattern raises further questions. If the goal was to ”get out the word”, why focus external contact on relatively unimportant accounts? An analysis of the timeline of the trolls’ external activity, to which we now turn, suggests an answer– that the goal of some of this external activity, especially the retweets, may have been to raise the prominence and activity levels of the external accounts.

3.2 Three regimes- Introduction, Growth, and Amplification

Figure 6 presents the trolls’ externally networked activity as shares of overall activity within each week of the sample. Each panel represents the networked output of trolls of the indicated type, with external retweet shares in blue and reply shares in orange.

There are several apparent regimes in the timeline of external networked activity. A spell of significant retweeting occurs in the summer of 2015, and this behavior aligns with the period during which the Left and Right Trolls were retweeting the News Feed accounts and, to a lesser extent, themselves. This retweeting spell is coincident with a wave of reply activity by the Right Trolls, which continues into the fall of 2015, when they are joined by the Left Trolls, which also began replying to News Feeds. This period’s external retweet activity peaks in June, and external reply activity peaks and ends in late August or early September 2015. We will designate June through September of 2015 as the “introductory” period, and hypothesize that this activity had the goal of introducing the troll accounts to a wider variety of potential followers and to establish their reputations as a type.

The Right Trolls engage in a second period of similar activity in the sum-
mer of 2016, which may seem at odds with the hypothesis of introduction. But as presented in the first panel of Figure 7, which represents the distribution of the date of first tweet for accounts of each troll type, this second introductory spell for the Right Trolls is consistent with a new set of Right Troll accounts who first appear in the beginning of 2016. The other troll types had a single account creation spike at the beginning 2015 and, therefore, a single introductory period.

From the end of the introductory period through August, 2016, the trolls engaged in what we will term the “growth period.” In this period, external retweet activity is moderate, while external reply activity slowly declines. Left Trolls and Hashtag Gamers engaged in significant internal retweeting, but no types did significant internal replying.

Most importantly, as the name indicates, the average estimated follower counts of troll accounts of each type increased throughout this period. There are two ways in which these follower counts are estimates. First, we observe follower counts at the moments the tweets were harvested by Social Studio, often but not always soon after the tweet was produced. For the dates between the harvest dates, we do not observe follower counts, so we simply project forward the last observed count. As follower counts tend to rise over time, these estimates are, on average, biased downward. The second way they are estimates is in our measure of whether the account still exists. All of the IRA-affiliated accounts are suspended by Twitter by the end of our period, but we do not know the exact date each account was suspended. Instead, we remove the account from our sample for any date after its last tweet (displayed in the bottom frame of Figure 7), although accounts could have persisted beyond their last tweets. Figure 8 illustrates the interpolated mean follower counts for active trolls of each troll type from the middle of 2015 through October 2016.

Finally, the Right and Left Trolls transition into the third regime, what we will call the “amplification period”, in mid-September (for the Right Trolls) and early October (for the Left Trolls), of 2016. In this final period, activity spikes dramatically, accompanied by an upward shift in external contact, almost entirely from retweets. Beginning on October 6, 2016, both Right and Left Trolls changed their behavior along multiple dimensions, including
tweeting at much higher rates, more consistent tweet production, and almost exclusively retweeting from outside the network. This change continued until May, 2017, when a number of accounts suddenly stopped tweeting. We will argue in detail, below, that this period’s strategic change amplified the messages, increased the activity level, and increased the prominence of accounts from outside the network.

The Hashtag Gamers, in contrast, never make a transition from growth to amplification. Rather, they slowly decrease their output and external share throughout 2017, up to their last set of tweets in late July 2017.

The next two sections will explore, in detail, the networked activity of the trolls during these three periods, provide full details on the tactics involved, and estimate the impact of this activity.

### 3.3 The Life-Cycle of a Troll

The timeline of data in the prior section provided some suggestive evidence about the way that IRA trolls used networked behavior to introduce and build their accounts over the months of the campaign. But simply looking over time potentially obscures important elements of the strategy, as we know the mix of active accounts can also change substantially, as Figure 7 indicates. In this subsection, instead, we will focus in on how an account’s activity varies over the life of the account, during the introductory and growth phases. As the behavior changes so dramatically across all Left- and Right-Troll accounts in the Amplification Period, we will delay our analysis of that period to section 4.

There are two ways to measure the maturity of an account, as a function of the account’s output and as a function of calendar time. From the output perspective, “young” accounts are those that have not yet tweeted very much, while from the calendar perspective, “young” accounts are those that produced their first tweet a short time ago.\(^1\) An advantage of the output perspective is that accounts that lie dormant do not progress on this measure. A disadvan-

---

\(^1\)With these accounts, using the account creation date as the “birth” of the account is often very misleading, as many of the accounts had been created years before they produced their first tweet.
tage is that it misses other activity that the accounts could engage in, such as liking, following, or direct messaging. It turns out that the behavior of the IRA accounts appears to be much more consistent when viewed from the output perspective, so we will rely on this perspective in this section. Results from the calendar perspective (available by request) share broadly similar patterns but are much less precise, suggesting that output is a better proxy for maturity.

Figures 9 and 10 illustrate how the IRA’s internal and external networked activity varied over the life-cycle of their accounts during the introductory and growth phases. In both figures, we limit our investigation to the first 1500 tweets of accounts that produced at least 1500 tweets. The histograms, with scales on the right, track the number of observations per 15-tweet interval. The sample size declines near the top of the range, as some accounts do not reach 1500 tweets before September 1, 2016, and the tweets after that date are excluded as part of the amplification period. The results are nearly identical if we limit to accounts that reach 1500 before that date. The dots represent the share of tweets in that bin of the indicated activity type (external retweets, for example). The dot near 0.15 for 500 updates in the Left-Troll internal reply panel of Figure 9, for example, means that for Left Trolls about 15 percent of the tweets numbered between 500 and 515 are internal replies. The shares in the four panels for each troll type need not sum to one, as there is an omitted category of non-networked (basic tweet) content.

There are a few patterns that are immediately apparent from these graphs. First, networked activity is quite rare in the first 250 or so tweets. Second, beyond that threshold, there is rapid growth in both internal and external retweet activity for both partisan troll types, while the retweet activity for Hashtag Gamers grows more slowly and smoothly. Third, the internal retweet activity for both partisan troll types peaks at around 500 tweets before declining back to relatively low shares. This pattern is in contrast to the pattern for external retweet activity which stays high throughout the span, flattening out at about 50% for Right Troll while rising to above 60% for Left Trolls. Fourth, the pattern for internal replies is dramatically different for Left and Right Trolls, with Left Trolls following a similar pattern as they do for internal retweets, rising and falling, but Right Trolls hardly engaging in any at
all. Most of this internal replying by Left Trolls is directed at the News Feed accounts. Finally, external replying seems to be low and consistent across the output levels, but with Right Trolls consistently engaging in a slightly greater share of this activity than Left Trolls.

Beginning the introductory phase with 250 original tweets mixed with a handful of external replies makes the newly created trolls look more realistic as it begins establishing its character and place within a community. Often, these early tweets are simply quotes or generic canned phrases copied from other accounts. Once the account looks modestly realistic, it can establish its theme by inserting itself in internal and external networks through retweeting ideologically-consistent content and replying/commenting on the conversations in those networks. Having established that reputation, replying and retweeting internally has limited value in growing the account. By the time the account has accrued 1000 tweets, internal activity has limited value for the partisan types, so they reduce their level of that activity to very low levels.

Hashtag Gamers, in contrast, retweet at monotonically increasing rates both internally and externally as their accounts mature. This is more evidence that they use external behavior in a qualitatively different way than the ideological types do, mostly to directly amplify content, rather than accounts. For them, it may be as simple as the benefits of retweeting increasing in the number of followers, which, in turn, increases with output.

In addition to looking at how the mix of activities changes with the account’s life cycle, we can also investigate how the mix of accounts targeted with those external activities change. Figure 11 has a similar structure to the preceding ones, but it displays how the present-day follower count of the externally networked accounts changes over the life-cycle of IRA accounts of the indicated type. We are still looking at the first 1500 tweets, restricting to prior to September 1, 2016, but sample size changes over the life cycle, because the share of tweets dedicated to the indicated networked activity changes (see Figure 10). Since replies are relatively infrequent, the bin size has been increased to 75-tweet bands, to estimate the means more precisely.

In each of these graphs, the dots represent the average prominence of the accounts targeted, represented by the mean (log) follower count. An alterna-
tive representation of prominence, the fraction verified, gives nearly identical results.

For external replies (panel (a)), there appears to be a negative relationship between the update count of the Left Troll accounts and the prominence of the accounts to which they reply. For other account types, there appears to be no relationship.

For external retweets (panel (b)), there seems to be a strongly negative relationship between the update count of Right Troll accounts and the prominence of the external accounts they retweet. For Hashtag Gamers, there also appears to be an initial negative relationship, which seems to flatten out above 1000 tweets. The relationship for Left Trolls is more complicated. Up to about 800 tweets, there appears to be no relationship between tweet count and prominence, but between 800 and 1000 tweets, there is a jump in the prominence of retweeted accounts by about 1 log point. From that point on, this new higher prominence persists throughout the output range. This pattern is for Left Trolls tracks with the share of activity devoted to external retweets, presented in panel (b) of Figure 10.

Overall, the general downward relationship suggests that more mature troll accounts target smaller external accounts. The difference in the exact pattern of results for the three troll types remains a question.

4 Amplification

In this final analytical section, we describe the behavior and impact of the Left and Right Trolls during the Amplification period. When a troll account retweets another account, either internal or external to the troll network, they are directly amplifying the account being retweeted; that account’s message has the potential to be read not only by its own followers but also by the troll’s followers. But retweeting not only has this direct effect, it also has indirect and potentially lasting amplification effects. The retweeted accounts may become more engaged on the platform, increasing their rate of activity. As the retweeted account reaches a broader audience, they may also pick up additional followers, and, potentially, become connected with different, more
central, or broader networks with which they may engage. These direct and indirect amplifications may combine to give retweeted accounts incrementally greater social influence on Twitter.

Although the Right Trolls have a spike in retweet activity in mid September, we will concentrate our analysis around the October 6th spike, at which point the behavior of both ideological types changes so markedly. That date was the largest single day of output of English-language content throughout the entire analysis period, and was particularly marked for Left Trolls. It is, therefore, worth pausing to consider why that date may have been selected for such a marked change in strategy.

Thursday, October 6, 2016 was not a particularly busy day in the U.S. national or world affairs. Although the Presidential election loomed a month out, the biggest story of the day was the pending arrival of Hurricane Matthew, which was expected to batter the southeastern U.S. over the coming weekend. But it turns out that it was also the calm before the biggest political storm to hit the U.S. in a long time.

On October 7, a series of hugely important political events occurred. First, at 3pm EDT, the White House released a joint statement from the Department of Homeland Security and Office of the Director of National Intelligence, making the first public claim that the Russian government was behind the election hacking that had targeted the Clinton campaign and the DNC, specifically claiming that “only Russia’s senior-most officials could have authorized” the operation. At 4pm, the Washington Post released the story and audio from an Access Hollywood tape that included then-candidate Trump making lewd comments about assaulting women. Many within the GOP felt the tape spelled the end to the campaign. At 4:32pm, Wikileaks released the first trove of emails that the Russian military intelligence had illegally obtained by hacking the email account of John Podesta, Hillary Clinton’s campaign chairman. This trove included several emails that put then-candidate Clinton in a poor light, especially to Bernie Sanders supporters, include quotations from her paid speeches to Wall Street indicating that she had a “public position and a private

\[https://www.huffpost.com/entry/yahoo-64-hours-october-american-politics\]
position”\(^3\) and one suggesting that she may have had early access to a primary debate question for the March 23 Democratic Town Hall.\(^4\) October 7th is also the Russian President Vladimir Putin’s birthday. October 8th was exactly one month before the election. October 9th was the second Presidential debate.

Given this coincidence of events, it is impossible to determine which, if any, were related to the sudden change in the ideological trolls’ external activity strategy on October 6th. But an argument can be made that the email release is the most likely candidate. First, we know from the Special Counsel’s investigation that the emails originated with the GRU, and they had expressed preferences over the timing of their release.\(^5\) Second, the flood of retweets on October 6th, continuing up to and through the election were dominated by Left Trolls. The accounts retweeted by the Left Trolls would be likely targets for the information contained in the hacked Podesta emails, wavering Sanders-Clinton voters who may have already been weakly supportive of her candidacy. Linvill et al. (2019) shows that the messages contained in Left Troll tweets in the month leading up to the election were quite ambivalent about Clinton’s candidacy. These sorts of messages would be complementary to the information included in the hacked emails.

The other alternative triggers seem less plausible. There is no evidence that the IRA had any advance knowledge of the Access Hollywood tape. A preemptive amplification strategy to shape the response to the release of the intelligence assessment seems hard to understand. Such a strategy would likely have focused on the Right Trolls, as it is the strong pro-Trump partisans who would have been most likely to react with suspicion to such a claim from the Obama White House. Activating accounts associated with the left of the Democratic party, and with Black Lives Matter, makes little sense in that context. Finally, it is hard to understand why this shift would have been made in anticipation of the second debate. We see no similar strategic shifts around either of the other two Presidential debates, and much of the activity


\(^4\)https://money.cnn.com/2016/10/11/media/donna-brazile-wikileaks-question/

immediately surrounding pre-scheduled political events such as the debates is
carried out by the Hashtag Gamers, rather than the ideological types.

Stated cleanly, we hypothesize that the amplification strategy that began
on October 6th was pursued in order to raise the activity level and prominence
of organic accounts that would naturally act (and react) in ways that aligned
with the trolls’ purposes. Specifically, in the weeks leading up to the election,
they sought to amplify accounts that would react poorly to the information
contained in the Podesta emails. In the rest of this section we will show that
the amplification succeeded.

4.1 The October 6 Amplification “Experiment”

On October 6, 2016, the trolls retweeted 4,249 unique accounts from outside
the network. Of these, 2,790 satisfied the following additional inclusion criteria
and therefore comprise our treatment group. First, they were active on Octo-
ber 1st or 2nd, on October 4th or 5th, and on October 8th or 9th, so follower
counts are available in those periods. Second, they had non-zero followers on
all those days and were not retweeted by the trolls on October 1st through
5th. Finally, they were not verified. From our sample of political accounts in
the first week of October, we identified an additional 2906 accounts that were
not retweeted by the trolls and satisfied those same criteria. For each of these
accounts we calculated a growth rate in follower, following, and output from
the earliest tweet in the Oct 1-2 period to the latest tweet in the Oct 4-5 period
and another growth rate from the latest tweet in the Oct 4-5 period to the
latest tweet in the Oct 8-9 period, where the growth rates was normalized by
the number of days between the tweets, in order to calculate a percent growth
per day.

Once these treated accounts and potential control accounts are identified
and collected, we use two forms of matching to estimate the effective of being
amplified on account features. In the simplest version, each treated unit is
matched with the closest control unit in terms of pre-period (Oct 1-5) growth
rate in followers, where ties are broken at random. In this nearest-neighbor
approach, each control account can potentially serve as control for multiple
treated units. In fact, only 1268 unique control accounts are selected to be matched with the 2790 treated units and the control accounts that match with more treatment units carry a proportionately heavier weight in the analysis.

This approach is straightforward to implement, but it has three important weaknesses. First, it discards many control observations, because they are not matched with a treated unit. This exclusion can include control units that are identical to included control units along the matching variable, or nearly so, and their exclusion reduces precision. Second, since we require every treated unit to be matched with a control unit, some of those treated units might end up being matched with a control that is quite different along the matching variable, since there are no similar controls available. Finally, if you want to match in multiple dimensions, you have to specify a weight to apply to each dimensions in order to have a well-defined “nearest” neighbor.

To address these shortcomings, we also include a Coarsened Exact Matching method, where we match on both pre-experiment follower growth rates and pre-experiment follower counts. Under this method, the matching variables are partitioned into intervals and the average of all control observations within the same set of intervals serves as the control outcome for all treated units in those intervals (Iacus et al., 2012). The width of the intervals is chosen according to Scott’s Rule, which makes it a function of the density of observations. With this approach, there can be isolated treated units that have no controls and are, therefore, excluded from the analysis. Under this method, 2839 of the 2988 potential control units end up contributing the estimates, as do 2764 of the 2790 treated units.

Table 2 presents the results of this “experiment”. The first panel reports the mean levels of the three account characteristics that we investigate: followers, following, and number of account updates, broken out for the treatment and matched control groups for each approach to matching. When we nearest-neighbor match on follower growth rates, alone, we also end up with fairly good balance between treatment and control in pre-treatment follower counts, but significant imbalance in following and update counts. The control units follow fewer accounts but have tweeted more than the treatment accounts. Note, also, that it is not appropriate to compare these account statistics to those
presented in Table 1, as those statistics represent the January 2019 values for accounts that still existed at that time. The statistics presented in this table are the levels in October 2016 and require that the accounts existed between Oct 1 and Oct 9, 2016, only.

The second panel of Table 2 presents the growth rates of the three account characteristics for the treatment and matched control accounts in the pre-treatment period (Oct 1-5) and the post-treatment period (Oct 5-9). The growth rates have all been calculated as percent growth per day. For followers, 0.29 percent growth per day is equivalent to about 10 percent growth per month. The pre-period growth rates for following are a bit smaller than that, while growth rates in update counts are a bit larger. The pre-period growth rates in followers for the treatment and control are very similar, by construction, but the differences in the other pre-period growth rates are also close, with control units growing slightly faster.

For the nearest-neighbor matches, in the post-treatment period, the follower growth rate for the treatment and control units diverge, with treated unit experiencing a 0.07 percentage point increase in daily growth rates and control units experiencing a 0.09 percentage point decline. On net, we estimate the average effect of treatment to be a 0.16 percentage point increase in the daily growth rate in followers, which is presented in the first row of the third panel, together with the standard error on this estimate. The pattern for the CEM analysis is similar, but with a smaller (but still very significant) diff-in-diff estimate (0.10).

These effects are estimated over a 4-day period (Oct 5-9). Over this time period, treated accounts actually increase their follower count by about 1.48%. We estimate that their counterfactual growth in the absence of amplification would have been of about 0.8%. For the average account size, this a difference of about 140 followers.

The following count growth rates, in contrast, moved approximately together in the post-treatment period, for both matching methodologies. For Nearest Neighbor matching, the treated accounts increased their growth rates 0.05 percentage points, while the control units increased theirs 0.07, for no statistically significant difference.
Table 2: Amplification Experiment

<table>
<thead>
<tr>
<th></th>
<th>Followers</th>
<th>Following</th>
<th>Updates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Levels (000s), mean (std. dev.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Treat</td>
<td>20.8</td>
<td>17.0</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>(116)</td>
<td>(97.5)</td>
<td>(33.3)</td>
</tr>
<tr>
<td>Pre-Control</td>
<td>22.6</td>
<td>16.0</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>(100)</td>
<td>(97.6)</td>
<td>(14.9)</td>
</tr>
<tr>
<td>Percent Change/day, mean (std. dev.)</td>
<td>0.29</td>
<td>0.24</td>
<td>0.18</td>
</tr>
<tr>
<td>Pre-Treat</td>
<td>0.24</td>
<td>0.21</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>(1.43)</td>
<td>(1.46)</td>
<td>(0.86)</td>
</tr>
<tr>
<td>Pre-Control</td>
<td>0.27</td>
<td>0.24</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>(1.10)</td>
<td>(0.94)</td>
<td>(0.94)</td>
</tr>
<tr>
<td>Post-Treat</td>
<td>0.36</td>
<td>0.32</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>(1.29)</td>
<td>(1.05)</td>
<td>(1.33)</td>
</tr>
<tr>
<td>Post-Control</td>
<td>0.18</td>
<td>0.22</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>(0.59)</td>
<td>(0.77)</td>
<td>(2.06)</td>
</tr>
<tr>
<td>Estimated Effect (std. error)</td>
<td>0.16**</td>
<td>0.10**</td>
<td>−0.02</td>
</tr>
<tr>
<td>Diff-in-Diff</td>
<td>0.02</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>Match Method</td>
<td>NN</td>
<td>CEM</td>
<td>NN</td>
</tr>
<tr>
<td>N Control</td>
<td>1268</td>
<td>2839</td>
<td>1268</td>
</tr>
<tr>
<td>N Treat</td>
<td>2790</td>
<td>2764</td>
<td>2790</td>
</tr>
</tbody>
</table>

Levels and percent changes per day for accounts retweeted by Trolls and controls matched on Oct 1-5. Pre-Treat and Pre-Control means represent the Oct 1-5 period and Post-Treat and Post-Control means represent the Oct 5-9 period. Standard deviations in parentheses, in the first two panels, and standard error in parentheses for the difference-in-difference estimates in the bottom panel. NN represents Nearest Neighbor matching and CEM represents Coarsened Exact Matching, both matched on Oct 1-5 follower growth rate and CEM also matched on initial follower count. Statistical significance represented by **: 0.01, *:0.05.
Finally, the update count growth rates also diverged for both matching methodologies. For Nearest Neighbors matching, the treated accounts increased their update growth rate about 0.08 percentage points, while the control accounts increased theirs approximately 0.02 percentage points. On net, we estimate the average effect of treatment to be a 0.06 percentage point increase in the daily growth rate in updates. Over the four day period, this is about 150 additional tweets per treated account.

Figure 12 explores whether these differences-in-differences estimates vary by the size of the IRA accounts that were doing the Oct 6 retweeting. We use Nearest Neighbor matching for this analysis. Because some outside accounts were retweeted by multiple IRA accounts, we sum the follower count for all retweeting accounts as a metric of the “size” of the retweet. The left graphs present the percent-change results, while the right graphs show the parallel results in level changes, rather than percent changes. The line is a locally-linear regression for the relationship between retweeting follower counts and the diff-in-diff estimate, with 95% confidence bands shaded. In the range of around 1000 followers (7 on log scale of the graph) there seems to be a positive relationship between retweeting account size and the targeted account picking up followers/engaging in more output, but further followers seems to have little additional impact. The one exception is that if we look at updates from the levels perspective, where bigger accounts do seem to encourage consistently more activity.

Our identification assumption is that the accounts that were retweeted by the trolls would have continued to trend with their control units on average if it had not been for the retweets on October 6. There are multiple threats to this identification assumption. First, although we chose control units from among accounts tweeting on political topics, not all political accounts are the same, so the treated units might be subject to common shocks to which the control accounts are not (or vice versa). Second, even if the treatment accounts were pulled from the same population as the control units, they may still have been deferentially subject to shocks other than the troll retweets in either the pre- or post-period. The most obvious source of these sorts of differential shocks are viral tweets. If, for instance, these accounts came to the trolls’ attention
because the tweet had quickly spread on October 6, then that spread, itself, may have had a direct impact on the accounts outcomes that is independent of the trolls’ intervention.

To investigate this alternative explanation, we looked at all the tweets that were retweeted by trolls on October 6th. We then gathered all other "genuine" retweets of these tweets occurring between the beginning of October 4th and the end of October 8th. On average, those tweets were retweeted about 59 times in that 5-day period, with a median of 4. On average, they received 23.5 (32.4 percent) of these retweets prior to the troll retweet, with a median pre-troll retweet count of 1. We take this as evidence that the trolls are engaging relatively early in the “life-cycle” of these tweets, rather than simply retweeting those that have already gone “viral.”

With these caveats, the difference-in-difference results are consistent with the IRA troll intervention increasing both the level of activity of amplified accounts and the reach of those account as measured by the number of followers. On October 6th alone, the trolls retweeted over 4000 accounts. If these results extend to all accounts retweeted by the trolls that day, that single day of amplification would lead to over 500,000 additional followers for those retweeted accounts over the next 4 days, and induce them to produce over 600,000 tweets they would otherwise have not produced. In the month leading up to the election, nearly 25,000 unique external accounts were retweeted by the trolls. If these result extend to all of them, those retweets would result in more than 3 million additional followers and nearly 4 million addition tweets. In fact, the amplification period went on well beyond the election, through the beginning of May, 2017, with nearly 100,000 external accounts retweeted. The impact of the whole amplification process could be enormous. Furthermore, our analysis has ignored multiple interactions with the same account and any portion of the effect that persists beyond 4 days.

5 Conclusions

This paper has documented the external behavior of the Russian-government-affiliated Internet Research Agency on Twitter from 2015 through 2017. Work
to date has analyzed the direct output of these accounts, and has been interested in accounts external to the troll network only to the extent that they shared the content produced by the IRA. This research supports Starbird et al.’s (2019) observations regarding the importance of considering "the role of online crowds (unwitting and otherwise) in spreading disinformation and political propaganda" (p. 4). Far from being passive social media users useful only as an audience for troll content, external accounts were, in fact, an additional tool to be employed and amplified. In total, our findings suggest amplification of external accounts was an important, possibly even central, goal of troll activity.

We have shown that the external outreach behavior of IRA Twitter accounts was very homophilic. The IRA-operated accounts linked to external accounts that seemed to share a theme with the account doing the linking. Right Trolls, for example, retweeted and replied to other accounts from the extreme right of the U.S. political distribution, accounts which shared the anti-immigrant and pro-Trump ideology that the troll account was presenting. Left Trolls, similarly, linked to non-IRA accounts that were politically left and, more specifically, identified as black activist.

Relative to randomly selected retweeted political accounts, the IRA linked to slightly smaller accounts, which were less likely to be verified. This suggests that the IRA targeted less prominent accounts than the average politically active Twitter user does. This pattern seems consistent with the amplification behavior we observe in the final period of the accounts’ life-cycle.

The external behavior varied over the life-cycle of the troll account, and over calendar time, and can be broadly split into three periods: introductory, growth, and (for Right and Left Trolls) amplification. During the introductory period, trolls produced “original” content and replied to other accounts, developing their character and placing themselves in the network. As the accounts matured, they transition into engaging in a mix of “original” content and external retweets, with internal networked activity and external replies declining to low levels. Throughout this period the follower counts of the trolls grew quite consistently. The Right Trolls and Hashtag Gamers also shift to retweeting smaller accounts as they mature. Finally, on October 6th, both Right and
Left Trolls shift to the “amplification period”, shifting almost entirely into retweeting outside accounts, an activity in which they persist until, at least, the beginning of May, 2017.

From a quasi-experiment around the sudden change in Oct 6, we can evaluate the potential consequences of this amplification effort. Relative to matched control accounts, amplified accounts significantly increase their output and expand their set of followers. In the month between the shift to amplification and the election, amplified accounts could have been induced to produce nearly 4 million additional tweets and gained over 3 million additional followers.

If these estimates are accurate, the amplification activity in the final month of the 2016 election could have had a greater impact on the political conversation on Twitter than all the direct activity the IRA produced over the 3 years documented here, combined. Furthermore, unlike the direct activity, which Twitter has removed, about 90% of the accounts retweeted by the trolls still remain active on Twitter today. If those accounts are still messaging in ways the IRA finds useful, that amplification investment might be yield dividends into the indefinite future.

References


Gadde, V. Roth, Y. (2019). Enabling further research on information operations on twitter.


Figure 2: Word Clouds of Account Descriptions from IRA Accounts and External Accounts They Touched

IRA Account Descriptions

2019 Descriptions of Accounts Retweeted by IRA Accounts

2019 Descriptions of Accounts Replied to by IRA Accounts

28
Figure 3: Internal and External Networked Weekly Output by Troll Type
Figure 4: Follower Counts (natural log) of Accounts Retweeted by Trolls as Compared to Random Political Accounts (Black)
Figure 5: Status Counts (natural log) of Accounts Retweeted by Trolls as Compared to Random Political Accounts (Black)
Figure 6: External Activity Weekly Shares
Figure 7: Troll Account Birth and Death Dates

First Tweet Date for IRA Accounts of Indicated Type

Last Tweet Date for IRA accounts of indicated Type
Figure 8: Followers per Account by Troll Type by Day
Figure 9: Internal Networked Behavior Shares over Account Life

(a) Internal Replies

(b) Internal Retweets
Figure 10: External Networked Behavior Shares over Account Life

(a) External Replies

(b) External Retweets
Figure 11: Size of External Networked Accounts over Account Life

(a) External Replies

(b) External Retweets
Figure 12: Diff-in-Diff Estimates as a Function of the Size of the Retweeting Accounts

(a) Followers

(b) Following

(c) Updates