Analysis of Strategy and Spread of Russia-sponsored Content in the US in 2017

Alexander Spangher[1]*, Gireeja Ranade[2,3], Besmira Nushi[3], Adam Fourney[3], and Eric Horvitz[3]
[1]Carnegie Mellon University
[2]University of California, Berkeley
[3]Microsoft Research

Abstract

The Russia-based Internet Research Agency (IRA) carried out a broad information campaign in the U.S. before and after the 2016 presidential election. The organization created an expansive set of internet properties: web domains, Facebook pages, and Twitter bots, which received traffic via purchased Facebook ads, tweets, and search engines indexing their domains. We investigate the scope of IRA activities in 2017, joining data from Facebook and Twitter with logs from the Internet Explorer 11 and Edge browsers and the Bing.com search engine. The studies demonstrate both the ease with which malicious actors can harness social media and search engines for propaganda campaigns, and the ability to track and understand such activities by fusing content and activity resources from multiple internet services. We show how cross-platform analyses can provide an unprecedented lens on attempts to manipulate opinions and elections in democracies.

The Internet Research Agency (IRA) has been identified as a Russia-based company focused on media and information propagation (Intelligence Community Assessment 2017). The organization was found to have spent at least 5.8 million rubles, or sixty-eight thousand USD, from June 8, 2015 to July 1, 2017, on disseminating information to the U.S. public via Facebook advertising. They fielded thousands of Facebook advertisements and sent millions of Tweets, promoting hundreds of web domains and Facebook groups that spanned the political spectrum (Schiff 2018a; 2018b; Darren L. Linvill 2018). According to an indictment made against IRA by the U.S. Special Counsel's Office on February 16, 2018, the purpose of IRA's expenditures of effort and capital was to "sow discord in the U.S. political system, including the 2016 U.S. presidential election" (Robert S. Mueller 2018).

In the process of investigating Russian election interference, the U.S. House Intelligence Committee released IRA-linked Facebook advertisements and Twitter accounts into the public domain (Schiff 2018a; 2018b). As stated in the release:

> Russia exploited real vulnerabilities that exist across online platforms and we must identify, expose, and defend ourselves against similar covert influence operations in the future.

We have pursued an understanding of *how IRA's campaign targeted these "vulnerabilities," in the hopes of contributing to an awareness of how malicious actors can exploit large-scale commercial internet services.* We merge Facebook and Twitter datasets with logs from Microsoft's web browsers and Bing search engine, thus fusing data from three major internet companies to provide a broader lens on the overall scope and influence of the IRA generated content (Fig. 1).

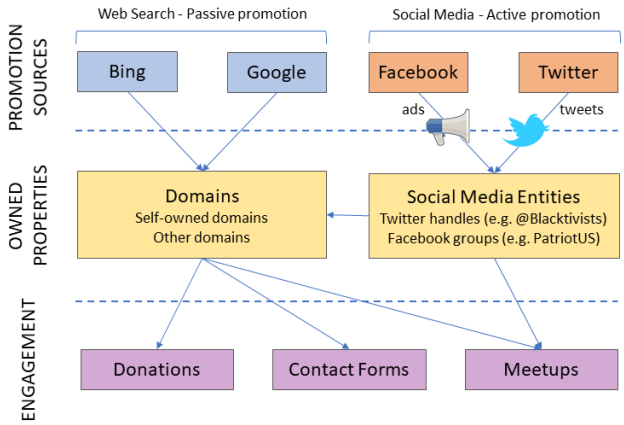


Figure 1: **IRA campaign structure.** Illustration of the structure of IRA sponsored content on the web. Content spread via a combination of paid promotions (Facebook ads), unpaid promotions (tweets and Facebook posts), and search referrals (organic search and recommendations). These pathways pointed to a combination of pages, some of which were created by the IRA (e.g. `blackmattersus.com`). IRA properties often contained engagement levers designed to maintain contact with users and push them towards events pages and donation pages.

We find that IRA-generated content spanned the political spectrum and covered a combination of apolitical, informative, local news and inflammatory articles on politically sensitive subjects. During the time frame of our studies, we found that the IRA invested more resources on Facebook in promoting left-leaning content than on right-leaning content. Right-leaning Twitter handles received more traffic than left-leaning handles on Twitter. We describe two case studies that show that the apolitical content and local news related IRA-properties reached users through search, and played a role in bringing traffic to IRA domains. We examine the IRA Facebook ad keyword targeting strategy and we investigate correlations between the regions where the ads received clicks, and the demographics of those regions.

Our specific findings are as follows:

1. *IRA Content:* The IRA produced and amplified a diverse array of left-leaning, right-leaning and apolitical content. We show two examples of Facebook advertisements in Fig. 2.

*Research performed during an internship at Microsoft Research.

Figure 2: Two example Facebook ads run by the IRA in 2017, out of 1,400 ads run post-election. The top ad is for the group "PatriotUs", while the bottom is for the group "Pan-African roots".

The advertisements and tweets ranged from emotionally charged to neutral (Fig. 3). IRA content drew traffic mainly from social media and search domains. While some IRA-established properties amplified politically sensitive topics (e.g. the Facebook group StopAllInvaders) others shared apolitical content[1]. One web property encouraged users to attend events, subscribe to mailing lists, and donate (see Tables 1, 4, Sections III, IV).

2. *IRA Strategies*: Through a study of traffic to the IRA properties we are able to make educated guesses about strategies that the IRA might have employed. For instance, we find that while there were similar amounts of left and right-leaning tweets on Twitter, the right-leaning tweets received more traffic in our datatset (Fig. 3). On the other hand, the IRA produced more left-leaning content on Facebook. Furthermore, they spent more money promoting left-leaning ads than right-leaning ads on Facebook.

   We use a basic model to estimate the effect of the paid promotions on traffic spikes. Under certain assumptions and for our basic model we find the paid promotions likely increased traffic spikes to the left-leaning properties (Fig. 8).

   We identify a strategy of employing apolitical content to garner credibility and engagement; such content received traffic via Bing on Microsoft promotional channels. One IRA article about Black Female Computer Scientists[2] appeared among the search results returned for a query that was promoted across Microsoft device lock-screens (Fig. 7 and Section V).

   Finally, we present a case study on a tactic employed on Twitter: rapid tweeting about evolving local news stories (Fig. 5). We believe a goal may have been to capture clicks before local news outlets were able to cover the story.

3. *IRA Outcome/Effects:* Despite the large volume of content generated by the IRA, on average only roughly 1 in 40,000 internet users in our dataset clicked on an IRA-property (Tweet, Facebook Group, URL) on a given day. Furthermore, we found only low volumes of traffic to meetup pages, donation pages and contact forms on the IRA domains (Table 4). Likewise, we found low correlation between geographic areas of heavy traffic to IRA domains and U.S. protests in 2017. We found no significant changes in traffic to news websites and to extreme news websites or to political donation websites (within a single browsing session, after a user was exposed to IRA content). This warrants further research that could take into account what we could not observe, such as traffic before and during the election, broader traffic from multiple browsers and search engines, traffic from mobile applications, long-term user behavioral change, real-world actions and posts made after exposure (see Section VI).

The remainder of the work is structured as follows: We begin by describing our datasets and methodology (Section II). We then describe the high-level structure of the IRA web properties and promotions (Section III), and follow by discussing the themes and messages that were ultimately promoted and clicked on (Section IV). We discuss the impact of paid promotions on Facebook as well as the role played by search through some case studies (Section V). Finally we discuss the implications and limitations of this research, and contextualize it with prior research in this space.

II. Data and Methodology

We have harnessed data from three large internet companies to provide a perspective on the strategy of the IRA, including views on the spread and access of IRA-generated content from January 2017 to August 2017. In addition, this work also leverages additional data sources including census information. We describe each data source in turn.

Primary datasets

Facebook ads: On May 10, 2018, the House of Representatives Permanent Select Committee on Intelligence released (as PDF documents) 3393 Facebook advertisements reported by Facebook as paid for by IRA-linked entities[3] (Schiff 2018a). We performed optical character recognition (OCR) and template-based information extraction, and this process yielded 3061 ads with actionable data.

IRA-linked Tweets: Alongside the Facebook ads data, the House Intelligence Committee also released a dataset on June 8th consisting of 3841 Twitter account handles believed to be associated with the IRA (Schiff 2018b). Researchers at Clemson University scraped over 2.9 million tweets from these handles and performed language and topic classification (Darren L. Linvill 2018).

[1]For example: "Black Female Computer Scientists. How Many Do You Know?" published at https://blackmattersus.com/31081-black-female-computer-scientists-how-many-do-you-know/

[2]https://blackmattersus.com/31081-black-female-computer-scientists-how-many-do-you-know/

[3]PDF documents for the individual ads included: text associated with the ad, the ad's start and end dates, targeting information, and the URL of the web property that the ad was promoting.

We leverage their work and focus our analysis on English-language tweets occurring from January 1st to August 1st, 2017 from the following categories they assign: "LeftTrolls", "RightTrolls", "Local" and "News" (See (Darren L. Linvill 2018) for more details). This resulted in a total of $471,000$ tweets and 320 handles.

Additionally, many of the tweets link to external domains. We augment the data by resolving Twitter's link-shorteners using historic instrumentation logs from Microsoft web browsers, described next.

Browsing data: We consider 212 days (January 1 to August 1, 2017) of instrumentation data collected by Microsoft Edge and Internet Explorer 11 desktop web browsers.[4] Data includes anonymized time-stamped records of page visits. Records are assembled into sessions of browsing activity: a "session" is defined as a sequence of page visits such that consecutive visits are less than 30 minutes apart.

We also used this data to identify upload dates for photos and videos in Facebook groups. Facebook assigns photos and videos unique ids visible in the url (i.e. `facebook.com/<IRAgroup>/photos/<id>` and `facebook.com/<IRAgroup>/videos/<id>`). We group the content by these ids and recover the earliest dates for each ID that we observe clicks. This gives us a sense of when the item was posted.

Web search data: Finally, we also analyze the logs of Bing, a major web search engine. This data reveals which URLs were presented as search results, even if such pages were not ultimately clicked by users – something that the browsing instrumentation data cannot provide. This dataset considers the same 212 days of data as the browser instrumentation logs.

Joins of primary datasets

We join the primary datasets to recover the scope of the IRA campaign. We associate IRA Facebook advertisements with browsing data by identifying clicks from `facebook.com` to a URL promoted by an IRA ad [5]. An equivalent approach was followed for joining browsing data with Twitter account handles and tweets [6].

Search engine logs were joined to the social media data by matching the URLs of search results to the URLs of pages actively promoted by Facebook ads and Tweets sent from IRA-linked twitter accounts. In search logs, we detect both when links are presented (impressions) and clicked.

Crowdsourced labeling of primary data

We used crowdworkers to label content[7] to ascertain information about the IRA content's political leaning, emotional intensity, and topic. We ran this study for Facebook ads, tweets, and search results related to IRA-owned web domains.

For Facebook ads, we rated all 1032 ads that ran after January 1, 2017. For Twitter we rated two subsets: a random subset of 500 IRA-linked tweets, and the top 500 most clicked tweets. For Search, we rated the 100 URLs with the most impressions (these 100 URLs captured 63% of search-result impressions to URLs in our datasets).

The task showed workers the text content of the promotion: the original Facebook ad, tweet text, or the snippet text shown in search results. Based on this information, the workers' task was to label the content with the most relevant option in each of the following categories: 1. *Political leaning*:{extreme-left, left, center, right, extreme-right, apolitical, local news}. 2. *Emotional intensity*:{neutral, low, medium, high, very high}. 3. *Discussed topic*: A predefined topic list.[8]

Each item was judged by five different crowd workers. We did quality control by excluding answers from workers with high disagreement rates (Inel et al. 2014). We performed a soft-assignment to label each item. For example, if a certain URL received 2 votes for extreme-right, 1 for left and 2 for extreme-left, then we counted $2/5, 1/5$ and $2/5$ clicks in each category respectively. To calculate the number of clicks on a certain label/topic, we used the soft assignment to proportionally divide the clicks across the assigned labels.

Secondary datasets

In addition to the primary datasets, we employ numerous external secondary datasets. We use Mediabias News and Media Categories (Zandt 2018) and SimilarWeb Domain Categories (Offer 2018), which provide URL-level categorizations, to characterize the content of URLs in our datasets. We use state-level popular vote counts for the 2016 presidential election (David Leip 2016), state-level voter registration data (U.S. Census Bureau 2016), state-level census demographic data (U.S. Census Bureau 2010), and state-level protest counts over time (Jeremy Pressman 2018) to characterize browsing data by geography. GDELT Global-Events data, which captures media publications over time, is used to characterize external news events (GDELT 2018). Finally, we use OpenSecrets.org (Center for Responsive Politics 2017) and Ballotpedia (BallotPedia 2017) to identify major political and donation sites over the period of interest.

III. Structure of the IRA Promotions

Before describing the traffic outcomes of the IRA campaign in Section IV, we present an overview of the scope and structure of the campaign to illustrate the breadth of the IRA's activity on different platforms.

Facebook Ads

We found that the Facebook advertisements paid for by the IRA promoted a smaller set of 350 distinct URLs. These URLs were distributed across 207 web properties. These properties include Facebook groups and profiles; Facebook or `meetup.com` event pages; news websites (including `CNN.com`); petitions (`whitehouse.gov`, `change.org`); and four domains identified as controlled

[4] Browser data is collected anonymously with user permission.

[5] After normalizing and reversing URL-transformations that Facebook applies to URLs (e.g. splicing `/pg/` or `/?/` from URLs). Such events are consistent with users clicking on ads, but are not sufficient to conclude that a particular ad was clicked because (a) a user may have arrived at a page from elsewhere in Facebook (e.g. from the newsfeed, or a notification) and (b) multiple ads promote a common URL.

[6] We searched for clicks with `twitter.com` and `<accounthandle>` as substrings, thus capturing clicks on tweets *and* account handles.

[7] We use Amazon Mechanical Turk. Workers were paid on average $12 per hour, above the national minimum wage at the time of publication.

[8] To extract the topic list we grouped Facebook ad keywords based on their co-occurrence across advertisements. These were then manually labeled (e.g. Black Lives Matter, Veterans, see Figure 4e for a partial list.) See Section IV for more discussion on targeting tags.

Promoted Content	# Properties	# Ads
Facebook groups	104	2674
Events & meetups	82	223
IRA domains	4	128
News organizations	4	4
Other	16	35
# Total	207	3061

Table 1: **Facebook Ad Categories.** Web properties promoted by the IRA Facebook campaign includes Facebook groups, events and domains suspected to be under the editorial control of the IRA (Schiff 2018a).

by the IRA (blackmattersus.com, dudeers.com, black4black.com, donotshoot.us)[9] (See Table 1).

The breadth of different content-types is notable (petitions, events, articles, Facebook groups), and raises questions about the intent of the IRA campaign. In the next section covering traffic patterns, we limit our analysis to traffic. However, we note that the reach of the IRA campaign did go beyond what can be measured by just traffic. For instance, one of the IRA's petitions[10] received 65,000 supporters, well within the range of typical mid-to-high performing change.org campaigns (according to top popular petitions listed at https://www.change.org/petitions).

Twitter

The 3,841 Twitter handles released by (Schiff 2018b) and 2.9 million tweets compiled by (Darren L. Linvill 2018) were filtered to 471,000 English-language tweets from 320 "LeftTrolls", "RightTrolls", "Local" and "News" accounts. Our analysis of links in these tweets revealed links to over 5,500 domains, tweets, and other properties. Over 95% of tweets are retweets, or reposting of other users' tweets, and many of the tweets point to well-known websites.

We found overlap in naming between the Facebook and Twitter campaigns: for instance, the Twitter campaign included a Twitter handle "blackmattersussoldier" and the Facebook campaign included a "BlackMattersUS" Facebook Group. However, we do not see the Facebook and Twitter campaigns sharing links or cross-promotions with each other. We observed only four URLs that were promoted both by Facebook ads and Tweets, all of them blackmattersus.com URLs with low-traffic.

We could only identify a small number of domains from the Twitter campaign as IRA-controlled (as compared to the Facebook campaign), and many of the links promoted by IRA tweets included major domains such as nytimes.com. As a result we focus on an analysis of the tweets themselves (for the subset outlined in Section II), and not the domains they promoted.

Web Search

According to the February 2018 indictment by the Special Counsel's office, the IRA was "*organized into departments, including: a graphics department; a data analysis department; a search-engine optimization (SEO) department; an information-technology (IT)...*" (Robert S. Mueller 2018). We find that IRA-promoted web properties were indexed by search engines, and surfaced to users in response to organic web search queries. Of the suspected IRA-owned domains we find one domain in particular, blackmattersus.com, received significant traffic via Bing (Fig. 6 and Fig. 7). We did not find evidence of paid advertising on Bing, but more research is needed for a comprehensive evaluation.

Targeting Decisions

The IRA used 890 different targeting keyword across the Facebook ads. The most frequent used keywords are all politically charged or demographically salient, with the top five being "african american civil rights movement", "african american history", "malcolm x" and "martin luther king jr". These five keywords are tagged to 32.3% of ads, which link to pages capturing 28% of traffic.

We identified pairs of keywords used together with correlation $c > .6$, ($p < .01$), then grouped these pairs together to identify clusters. We identified 19 such clusters using this method. When used together the keywords seem to indicate an attempt at demographic targeting. (We will analyze the demographic reach of the IRA campaign in Section V).

On Twitter, the most frequently used hashtags (of 44,700 total unique hashtags) are a combination of political and apolitical tags, with the top five being: "#MAGA[11]", "#NowPlaying", "#tcot[12]", "#top" and "#PJNET[13]". The top ten are tagged to 2.1% of tweets which themselves account for 1.3% of traffic: hashtags on Twitter were more diverse and did not repeat as frequently as the Facebook ads keywords.

IV. IRA Content and Traffic

Having described the breadth and size of the IRA's campaign in Section III, we now summarize the content of the campaign and provide an overview of the traffic to it.

Traffic Overview

As noted in the introduction, on average roughly 1 in 40,000 internet users was exposed to IRA ads on any given day in our dataset.

Perhaps because the platforms' audiences are different (Hughs 2012), or because there was little cross-promotion between the Twitter and Facebook IRA campaigns, we find very small overlap between users exposed to both campaigns. Cross-traffic between IRA Facebook properties and IRA tweets is also small, amounting to about 0.02% of the total users in our datasets clicking on content from both campaigns in the same day.

Table 2 shows the most trafficked Facebook groups and Twitter handles in each campaign, which account for more than 3/4th of total observed traffic in each case. The top-clicked Facebook groups seem to show a mix of topics and political leanings.

Figures 3a and 3b summarize the common referral pathways to IRA properties. Traffic to Facebook-advertised URLs and groups came mainly from Facebook and Search. The

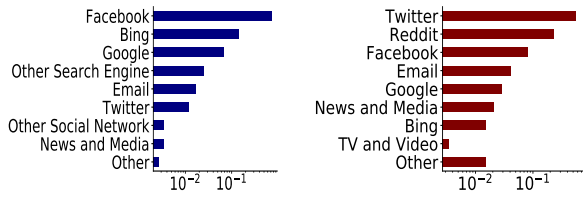
[9] See https://intelligence.senate.gov/sites/default/files/documents/exhibits-080118.pdf, https://euronews.com/2018/05/10/sean-hannity-black-lives-matter-among-targets-russian-influence-campaign-n872926, https://money.cnn.com/2017/10/12/media/dont-shoot-us-russia-pokemon-go/.

[10] https://www.change.org/p/barack-obama-u-s-house-of-representatives-u-s-senate-list-the-ku-klux-klan-as-an-official-terrorist-organization

[11] Make America Great Again

[12] Top conservatives on Twitter

[13] Patriot Journalist Network



(a) Referral pathways for clicks on URLs included in the IRA Facebook ads.

(b) Referral pathways for clicks on tweets linked to IRA accounts.

Figure 3: **Top traffic channels to IRA properties.** Top sources that brought users to IRA properties (Facebook-advertised URLs, Facebook groups and Tweets) from January 1, 2017 to August 1, 2017.

Top-Clicked Facebook Groups	Share of IRA Traffic
blacktivists	0.208
godblessthesouth	0.199
blackmattersus.mvmnt	0.169
brownunitedfront	0.116
patriototus	0.077
Other	0.231
Top-Clicked Twitter Handles	**Share of IRA Traffic**
ten_gop	0.462
pamela_moore13	0.245
crystal1johnson	0.076
southlonestar	0.048
jenn_abrams	0.045
Other	0.123

Table 2: **Top-performing IRA properties.** The top five Facebook groups (top) and Twitter handles (bottom) with most traffic. These Facebook groups collectively account for 77% of traffic to IRA-promoted Facebook properties, while these Twitter handles account for 88% of traffic.

Twitter campaign drew traffic mainly from Twitter, Reddit and Facebook, with Search playing a much smaller role.

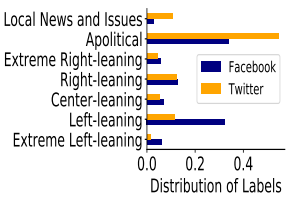
Traffic by Content-Type

Figure 4 summarizes results from our crowdsourcing study for both Facebook and Twitter. The left column summarizes the IRA generated content. The right column describes the traffic (volume of user cliks) to this content. The figure was created using the proportional assignment of clicks as described in the methodology section. The X-axis in all the histograms shows the distribution of content (ads or tweets) soft-assigned to each category on the Y-axis.

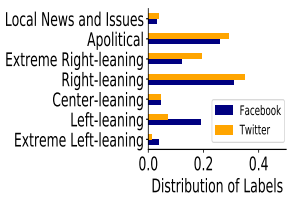
We see from Fig. 4a that there were more left-leaning and apolitical ads than right-leaning ads on Facebook. The Twitter content appears to be more balanced, with most tweets being apolitical and roughly equal numbers of left-leaning and right-leaning tweets. When we compare this to the traffic that was received in each category, we see that right-leaning and apolitical content got relatively more traffic in our dataset than left-leaning content on both Facebook and Twitter.

In addition to covering both sides of the political spectrum, the IRA content spanned a large topic space: Fig. 4e gives the distribution of content (ads, snippets, tweets) across categories, and Fig. 4f gives the distribution of traffic[14]. Especially prominent in the Facebook campaign were topics targeting African American and Mexican American interests. The Twitter campaign seems to have focused on general news, politics and other topics.

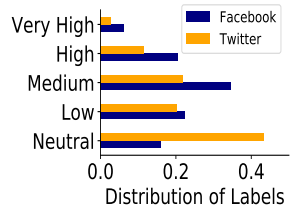
[14]Topic labels were chosen by manual assignment based on clusters among targeting tags in the Facebook ads. The assignment of ad/tweet to topic was done by crowdsourcing.



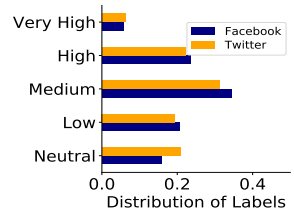
(a) Distribution of political leaning of IRA Facebook posts and tweets. We see more left-leaning and apolitical links than right leaning links on Facebook. Twitter has roughly equal amounts of right and left content.



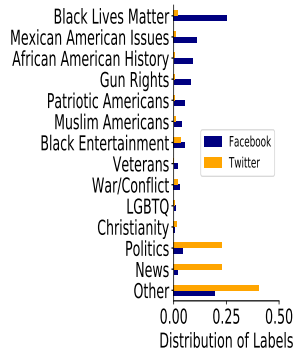
(b) Distribution of traffic to different categories of IRA Facebook posts and tweets, by political-leaning. Even though there was more left-leaning content on Facebook than right-leaning, the right-leaning content took a larger share of the traffic.



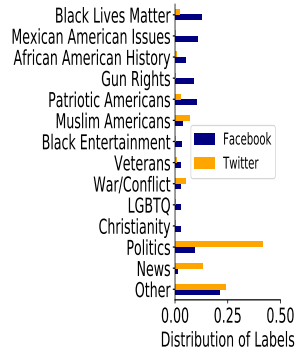
(c) Distribution of content across emotional intensity tags. A large fraction of the IRA promoted Facebook and Twitter content was neural and low emotional intensity. Especially pronounced is the Twitter channel, where nearly 43% of content is rated "Neutral".



(d) Distribution of traffic across the emotional valence of content. Relative to the proportions of overall content produced (c), content with medium and high emotional valence received more traffic on Twitter.



(e) Distribution of IRA content across topics (Facebook and Twitter).



(f) Distribution of traffic to IRA content across topics (Facebook and Twitter).

Figure 4: **Labels Summary: Content and Traffic.** This figure shows the distribution of the IRA content across political leaning, emotional intensity and specific politically charged topics using labels tagged in a crowdsourcing experiment. The assignment of content to categories was done using the soft-assignment of crowdsourcing labels as described in the methodology section.

V. IRA Strategy

Having summarized the IRA content and traffic to it in Section III, we now share inferences about IRA's strategies. We start with three case studies.

- **Case study 1:** Rapid-response tweeting to local news developments helped the IRA receive search engine impressions.
- **Case study 2:** Apolitical IRA-generated stories appeared among search results returned for search queries on Bing.com, including some Microsoft-promoted queries (e.g., "women scientists"). Overall, apolitical articles from organic and Microsoft-promoted queries received more traffic via the Bing.com search engine than other articles.
- **Case study 3:** Paid promotions were not the only traffic driving factor. In fact, we observe that for some of the groups, there exist other important unpaid actions (e.g. photo and video shares) that can predict high traffic.

We conclude this section by showing how the IRA reached various demographically and politically distinct geographic regions in the country and extensively used Facebook microtargeting tags.

Case study: Search and `twitter.com/TEN_GOP`

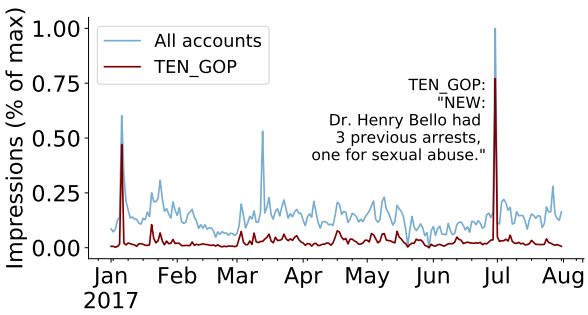
The IRA tweeted rapidly and in huge volumes across their accounts, tweeting in some cases thousands of times a day, and multiple times within seconds. Figure 4 shows that they principally tweeted about "News", "Politics", and "Entertainment". The high volume of rapid tweets allowed them to be unwittingly incorporated into legitimate news articles at major Western outlets (For example, an article published by `msn.com`[15] references a tweet by @SouthLoneStar, a confirmed IRA account).

In this example, we show that these tweets performed especially well at drawing unwitting traffic from users discovering a new news story. The IRA's TEN_GOP[16] Twitter handle, which drew 46% of traffic in the IRA's Twitter campaign, is a good case study of how an IRA Twitter account can capture significant traffic by being indexed by a search engine during the very early stages of a breaking news cycle.

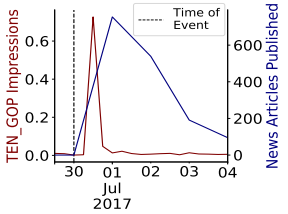
A news event about Dr. Henry Bello occurred on June 30, 2017. This turned out to be a major news story in New York City, receiving thousands of news articles of coverage (GDELT 2018). However, the bulk of these articles were published and indexed by the search engine on July 1, 2017. Fifteen of the stories published on July 1st were published by the Associated Press, however the remainder came from local domains, with the top domains being: nydailynews.com (12 stories), denik.cz (12 stories), heraldsun.com.au (11 stories) and news-sentinel.com (10 stories). Because the IRA tweeted about the event around noon, they got indexed on June 30, beating many of the articles published and receiving a surge of impressions (i.e. be displayed as a search result) for the story (see Figure Figure 5).

Case study: `blackmattersus.com`

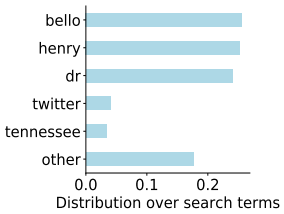
We showed earlier that a large volume of apolitical content was produced by the IRA. We examine the `blackmattersus.com` domain as a case study of how this



(a) Spike in search impressions (normalized as percent of max) that displayed the TENGOP tweet about Henry Bello on June 30, 2017.



(b) Distribution over impressions received by the TENGOP handle, compared with the number of news articles published on the subject on each day.



(c) Top search terms leading to impressions received in subfigure (a).

Figure 5: **Case study: Tweet impressions for the story of Henry Bello.** Subfigure (a) shows impressions on IRA tweets. The surge in impressions is likely linked to the early availability of news information that was unavailable elsewhere; most news outlets only published stories on this subject the following day (Subfigure (b)). Subfigure (c) shows the most common search terms that generated the impressions. Most of the stories were published by local news outlets.

content content played a role in a larger strategic aim: search engine traffic followed by an on-site engagement funnel.

Search Traffic `blackmattersus.com` contained a mix of content: content that was apolitical and emotionally neutral, as well as content that was political and emotionally intense (Table 3 and Figure 6). We found that content rated as "Apolitical" drew more traffic on Bing while "Left-leaning" drew more on Facebook. Additionally traffic to `blackmattersus.com` from Facebook was drawn more to pages that were "High" or "Medium" emotion-level, while traffic from Search centered on pages that were more likely to be rated "Neutral" or "Low" emotion.[17] We next explore the mechanisms *behind* this content performing well through search.

Beyond generating ranked lists of results in response to search queries input by end users, major search engines also power a host of company properties like news verticals, homepages and new-tab pages, which can link out to preformulated queries like, for example, "women scientists". Fig. 7 shows that, on several occasions, `blackmattersus.com` appeared in the results

[15] https://www.msn.com/en-gb/news/uknews/london-attack-woman-wearing-hijab-was-distressed-horrified-photographer-says/ar-BByFoiY

[16] TEN_GOP refers to Tennessee GOP. Many IRA accounts mimicked political accounts (Robert S. Mueller 2018).

[17] These findings provide support to the speculation voiced by journalists at (Mak 2018), that the IRA may have been generating apolitical content to serve a purpose: to gain Americans' trust in a network of seemingly local news handles, which could later be "operationalize[d] [to] significantly influence the narrative on a breaking news story".

Top Clicked URLs via facebook.com
/nj-police-officer-arrested-and-charged-f...
/her-provocative-ballet-dance-hurts-white...
/tamron-hall-has-been-replaced-by-megyn-k...
Top Clicked URLs via Search
/black-history-month-black-inventor...
/katherine-johnson-a-black-nasa-pioneer...
/black-female-computer-scientists...

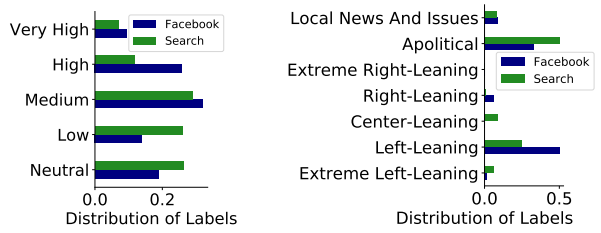
Table 3: **Case study: Blackmattersus.com, examples.** Top three most trafficked pages on blackmattersus.com from either Facebook (top) or Search (bottom).

	Share Button	Contact Page	Meetups	Donate
# links	3940	1	896	1
Click rate	2.64%	.28%	.11%	.056%

Table 4: **Blackmattersus.com Structure:** "Share Button" are links on each article that encourage sharing on social media, the "Contact Page" allows users to sign up for email correspondence, "Meetups" is a listing of local meetups and "Donate" is a page soliciting donations that leads to a paypal.com page. (We saw no evidence of clicks to paypal.com in the same session.)

returned for such preformulated queries, which were promoted across several Microsoft properties, including bing.com/news news verticals, new tab pages, and device lock screens.

By publishing neutral, entertaining content the IRA was able to draw more users to its pages.



(a) Emotion tags given by crowdsourced workers to blackmattersus.com URLs, weighed by traffic from either Facebook or search.

(b) Emotion tags given by crowdsourced workers to blackmattersus.com URLs, weighed by traffic from either Facebook or search.

Figure 6: **Case study: Blackmattersus.com, summary.** Blackmattersus.com was the most highly trafficked IRA-owned domain in our dataset. The emotional intensity of pages users reached from Facebook.com is higher than that for pages reached via search, as can be seen by looking at the top-trafficked pages on the domain, as well all trafficked pages as labeled by crowd workers.

Engagement Funnel While gaining a wide reach through various promotional channels on the internet, blackmatters.com was structured to potentially engage users beyond reading articles.

blackmattersus.com is a particularly expansive site with 4,608 pages: articles, pages linking to Facebook event pages, a donations page, and an input field soliciting users to join the site's mailing list. Shortly upon arrival, a pop-up asks users if they would like to receive notifications.

We performed a site-wide scrape to categorize pages hosted on the domain. Table 4 gives our overview of this domain. On the first row, we show the raw number of links we counted devoted to each of four action types: "Share Button", "Meetups", "Contact Page", and "Donate". Interestingly, the "Meetups" are pages onsite that link out to meetup.com and Facebook events, many of which appear to be legitimate events, and are still active on Facebook. The "Donate" button links to a paypal.com page[18].

Each of these link-types is accessible from any article. The number of users we see engaging with this funnel is very low. The second row of Table 4 shows the conversion rate, or the percentage of users who clicked on one of the links *after* clicking on another blackmattersus.com page. These numbers are only one 1/10th typical conversion rates observed across e-commerce sites industry-wide (Saleh 2017).

Case study: Did promotions matter?

The IRA's ad-promotion strategy on Facebook involved roughly 531 ads and nearly 1 million rubles (or $15,000.00) on left-leaning content, compared with 184 ads and 620 thousand rubles (or $9,112.00) on right-leaning content. The IRA spent more on left-leaning content during our observation period.

Clearly, paid advertisements were part of the IRA strategy to generate traffic to their groups and properties. However, paid promotions are only part of the story. There are other important known and unknown factors not to be ignored. Among the known ones: historical traffic, photo and video shares are the most correlated ones with traffic spikes.

For example, Fig. 8(a)-(b) illustrates the traffic to different Facebook groups over time along with (1) photo and video uploads and (2) paid promotions. For the StopAllInvaders Facebook group, it seems some traffic spikes co-occur with events like photo and video uploads, but do not co-occur with paid promotions. For another example group, BrownUnitedFront, spikes in traffic seem more aligned with paid promotions than with media uploads. To isolate the influence of paid promotions from other unpaid factors, we trained a model with the goal of predicting the changes in spikes of high-traffic Facebook groups in the absence of paid promotions.

We used our model to examine six Facebook groups chosen for having high traffic as well as a large number of promotions. For these groups, we infer that without paid promotion, the left-leaning groups would have received fewer traffic spikes, while traffic spikes for right-leaning groups would have been largely unchanged. However, an important take away here is that paid promotions were not the only factor that lead to traffic spikes, which is something to keep in mind for future work.

A more comprehensive discussion of this is included in Appendix 2.

Geographic patterns in traffic

87% of the traffic to Facebook-promoted properties was U.S. based, but we observed clicks to IRA's properties from 176 countries.

[18]The paypal.com page is operated by a merchant with the email address *xtimwalterssx@gmail.com*, which was one of the emails mentioned in the Special Counsel's Indictment (Robert S. Mueller 2018).

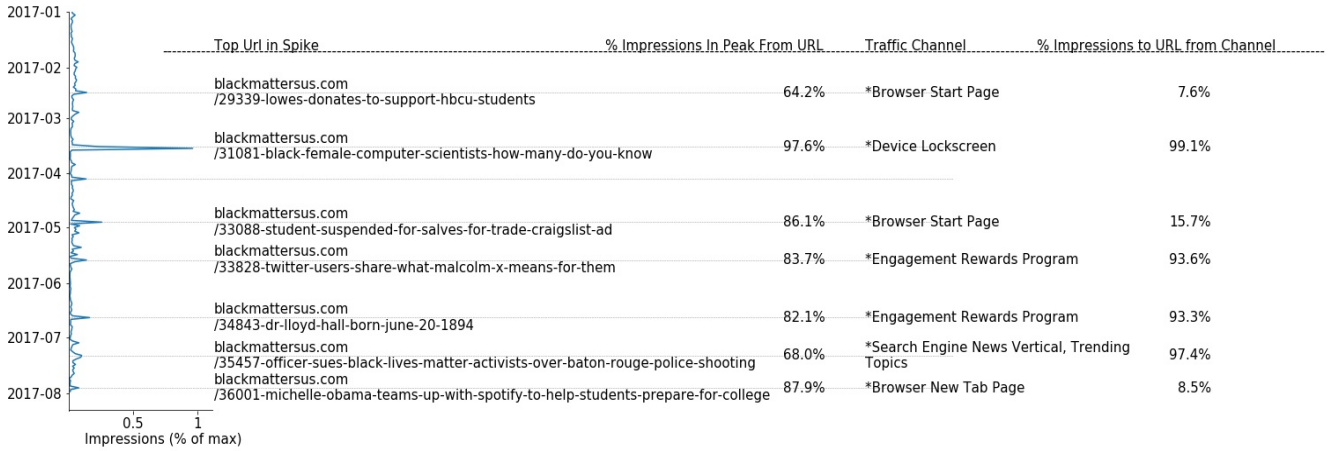
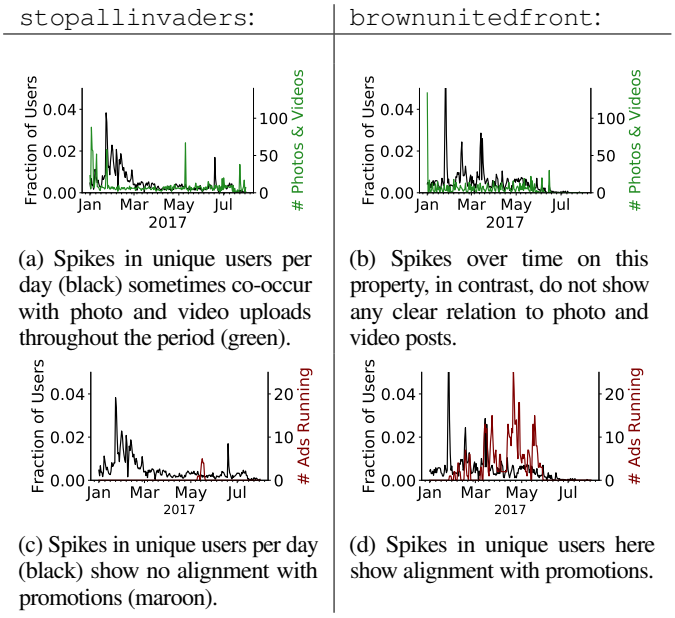


Figure 7: **Instances of Facebook-Promoted Domains Appearing in Search Results**: The domains most associated with the Facebook-ad campaign often appeared as results to search queries, with surges of search query impressions occurring from time to time. Some of these spikes can be associated with the promotion of various topics. For instance, the largest spike corresponds to the page `blackmattersus.com/31081-black-female-computer-scientists-how-many-do-you-know` appearing among the results returned for the query, "women scientists" – a query which was briefly promoted on the lock screen of some Microsoft devices.



(a) Spikes in unique users per day (black) sometimes co-occur with photo and video uploads throughout the period (green).

(b) Spikes over time on this property, in contrast, do not show any clear relation to photo and video posts.

(c) Spikes in unique users per day (black) show no alignment with promotions (maroon).

(d) Spikes in unique users here show alignment with promotions.

Figure 8: **Drivers of Spikes.** Many promoted web-properties (shown above) seem to have click-spikes associated with factors like paid actions (Facebook-ads) or unpaid actions (photo or video uploads). The Y-axis in each of the graphs represents the fraction of total unique users to the group over the time period that visited on a given day. We quantify the effect of a paid Facebook promotion on the day of promotion by predicting spikes the property would have received had it not been promoted (see Appendix 2). This figure shows that in addition to paid promotions, unpaid actions (such as photo/video uploads) also impacted the traffic to the Facebook properties.

We consider a broad array of geographic features to describe the areas contributing traffic including race-based, rural/urban, education-based, employment-based and other demographic features from the 2010 census (U.S. Census Bureau 2010). We look at voting patterns from the 2016 presidential election (David Leip 2016). We also examine voting as a percentage of registered voters, a metric that we use to proxy voter enthusiasm (U.S. Census Bureau 2016).

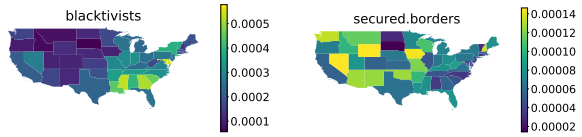
Fig. 9 shows state-level correlations between census features and traffic to IRA properties promoted via Facebook. The IRA properties are clustered by targeting keywords and topics as per the results of the crowdsourcing study. We only show a subset of features, and only display correlations with $p < .1$.

Traffic to IRA content in certain categories is positively correlated with demographic features. For example, traffic to IRA content labeled "Black Lives Matter" is significantly correlated with the percentage of census respondents self-identifying as African American in a state ($c = .88$, $p < .001$). On the other hand, this content is negatively correlated with the percentage of census respondents self-identifying as White in a state.

Moreover, there is a significant negative correlation ($c = -.29, p < .001$) between the percentage of of Registered Voters who Voted (in 2016) in a state with traffic across most ad groups. Thus, there was more traffic to ads in states with lower voter turnout. This may support the claim that the IRA operation was seeking to target areas of voter discontent (Shane and Mazzetti 2018), but is subject to many confounders. Furthermore, it is important to note that our work is looking at data collected in 2017, two months after the 2016 U.S. Presidential election, and thus cannot infer anything about the pre-election timeframe.

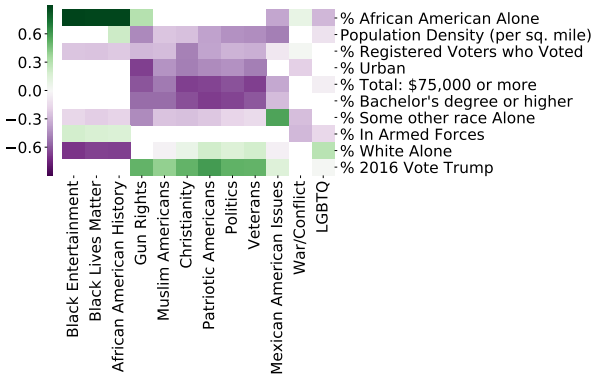
VI. Effects on exposed users

Having described case studies of the IRA strategy, we now close with a view of the effects on users. Although our main conclusion here is that few effects were observed, more research is needed due to the following limitations. First, this analysis cannot observe the long-term behavior of internet users online due to extensive anonymization techniques applied in the data. Second, browser and query logs give us only a limited view into users' online actions (e.g. we cannot see posts, tweets and emails), and most importantly we cannot observe real-world



(a) Traffic to Blacktivists Facebook group normalized by average daily active users by state.

(b) Traffic to Secured Borders Facebook group normalized average daily active users by state.



(c) Correlation coefficients across demographic census features with categories of the Facebook advertisements ($p<0.1$ shown)

Figure 9: **Correlation with state-level census features.** The Facebook-ad campaign was targeted primarily through interest-tags. We grouped them into clusters by asking crowd workers to label ad content. The top row shows distributions over traffic-normalized click counts by state for two sample groups. The bottom figure shows correlations between census features and normalized clicks by state ($p<0.1$ shown).

behaviors. Third, a complete analysis on these effects calls for data from potentially multiple browsers and search engines to better cover the online user base, as well as data from a broader range of mobile devices and applications. For completeness, we summarize all attempts we made to better understand impact on exposed users and hope that these attempts will seed further investigation on the topic.

Our study explored online news consumption, online political donations, and real-world protests. To categorize online news consumption, we used data from MediabiasFactcheck.com to label domains known for publishing various news-types (Zandt 2018). Using these labels, we examined the browsing behaviors of users exposed to an IRA ads before versus after exposure to the ad, while controlling for session length. We did not see significant changes in the total volume of news consumed, or in the extremism of news sites visited before and after exposure to the ad.

To understand online political donations, we catalogued the websites registered for candidates, politicians and major recipients of political donations, using data from OpenSecrets.org (Center for Responsive Politics 2017) and BallotPedia.org (BallotPedia 2017). Again, we failed to see any significant changes in pages visits before and after exposure to an IRA ad.

In addition, we also explored the notion that the IRA was attempting to draw more people into protests. For example, Robert S. Mueller's indictment points to anecdotes of Americans being driven to attend real-world protests as a result of misinformation (Robert S. Mueller 2018). Using Crowd Counting Consortium's numbers tracking protests in 2017, we examined whether there was a correlation between where protests occurred in the U.S. and the locations of ad-clicks by users (Jeremy Pressman 2018). After normalizing for state-level populations and state-level browsing behavior, we found no significant correlation.

VII. Related Work

Social media and politics. The dataset generated by this paper was further used in followup work by Boyd, Spangher et al. (Boyd et al. 2018). This work examined language differences in the IRA tweets and those posted by a general U.S. population during the same time. Additionally, Boyd et al., report that the majority of Facebook ads were posted during 9am and 6pm, Moscow Standard Time. As such, (Boyd et al. 2018) provides further analysis for the point-of-origin of IRA tweets and advertisements.

The impact of social media on the political ambit has attracted significant attention more broadly in both political science and social media analysis studies. Social networks have facilitated faster and targeted user communication, which has created new and amplified effects due to high information diffusion (Bakshy et al. 2012). From a political perspective, such effects can often be positive and increase political participation awareness (Gil de Zúñiga, Jung, and Valenzuela 2012). However, as any powerful means of communication, social platforms can also be exploited by external entities to produce negative effects in the society such as drastic political division (Sunstein 2018) as well as manipulation and misinformation (Marwick and Lewis 2017; Fourney et al. 2017). This paper specifically focuses on extracting data-analytic insights that can reveal information regarding the extreme and potentially divisive characteristics of the IRA campaign. Being aware of the nature and the scale of these characteristics is crucial for increasing general awareness and developing robust protective mechanisms in the future.

Division and polarization. Political division has been investigated using the notion of "information bubbles" (Resnick et al. 2013) and "echo chambers" (Bessi 2016; Garimella et al. 2018). Authors study the phenomenon of political and intellectual isolation that can occur when users are only exposed to information that confirms their own beliefs. While some of these effects can exist due to organic user behavior, recent research has shown that they have been leveraged to increase polarization and division (Stewart, Arif, and Starbird 2018; Conover et al. 2011; Garimella and Weber 2017; Bessi 2016). In this context, various tools have been proposed to foster diverse information consumption either via visualization (Gillani et al. 2018) or exploration strategies that go beyond the personal network bubble (Resnick et al. 2013).

Manipulation and misinformation. Prior research—including our own—has found social media can be used to to spread and amplify propaganda at a global scale (Menczer 2016; Fourney et al. 2017; Silverman 2016). Combined with the fact that humans are inherently bad at detecting deception (Mihalcea, Pérez-Rosas, and Burzo 2013; Abouelenien et al. 2014; Pérez-Rosas et al. 2015), decades of algorithmic advances in targeted advertisement have laid the groundwork for large-scale, general-purpose mass manipulation.

Manipulation and misinformation strategies transform information in a form that best propagates ideas or agendas of specific groups of interest. A comprehensive summary of these techniques (Marwick and Lewis 2017) shows several examples where social media has shown to be vulnerable to such attacks

especially from extremist political groups and Internet trolling entities. In the most harmful form, manipulation can be encountered as jointly combined with misinformation, where information is intentionally twisted or even fabricated. Social media has been a lucrative target of misinformation (Allcott and Gentzkow 2017; Fourney et al. 2017; Faris et al. 2017) in the recent years. Social bots on Twitter have played a role in influencing the spread of information in a network (Varol et al. 2017a; 2017b; Bessi and Ferrara 2016). Various approaches have been proposed to isolate and limit the harmful effects of misinformation by harnessing linguistic features, semantic analysis of the content, and network topology properties (Shu et al. 2017; Budak, Agrawal, and El Abbadi 2011; Conroy, Rubin, and Chen 2015).

The 2016 US election. The events of the IRA ad campaigns in the US presidential election in 2016 are not unique. Similar developments have been noticed in the last five years in the case of the Brexit referendum in the UK in 2016 (Howard and Kollanyi 2016), German Federal election in 2017 (Morstatter et al. 2018), and elections in Pakistan in 2013 (Younus et al. 2014). This work was motivated by the need to better understand the IRA campaign in 2016 and was enabled by the release of the recent datasets from Facebook and Twitter on this matter during the respective testimonies at the United States House of Representatives (Schiff 2018a; 2018b). Guided by the same motivation, there is recent and ongoing work tackling important questions related to the structure of the retweet network (Stewart, Arif, and Starbird 2018) and the characteristics of IRA-promoted Twitter handles (Darren L. Linvill 2018). Multiple journalistic efforts (Shane 2017; Mak 2018) published illustrative examples of the promoted content and informed the broader audience. This paper aims at providing a data-oriented analysis, joining multiple sources of information funneling from the promoted ads and content, to post-election web traffic, to observable user effects.

VIII. Discussion and Conclusion

While our studies present a set of insights on the strategy, structure, and scope of IRA-related web activities, our analyses have limitations. We highlight these as caveats and as guides for future work in pursuit of confirmation of the results and inferences shared in this paper. First, much of our analysis relies on browsing instrumentation data on desktop computers and excludes mobile devices. The released Facebook ads data set and the web search data set we use, on the other hand, do include mobile traffic. Secondly, because the browsing instrumentation collects only URLs, our browser logs cannot directly observe interactions that occur within a page (e.g., enabled by AJAX or JavaScript). We inspected many of the web properties and, while we found that most activities of interest result in changes to the URL, some (e.g., donations brokered by Facebook) could not be detected. Also, as with all studies of this nature, we cannot observe actions taken in the physical world. For example, we cannot be sure if a person attended a protest that they read about. Likewise, we restrict our analysis to short-term trends because our browsing instrumentation data set does not maintain long-term histories. Our data is also collected in 2017, after the 2016 election, which was one of the targets of the IRA campaign.

Nevertheless, we have demonstrated multiple ways by which malicious agents can manipulate the digital landscape showcasing the large attack surface susceptible to malevolent interventions. We highlight how Facebook's fine-grained interest-based targeting keywords can yield emergent demographic-based information flow. We show how search platforms, and their agnostic content-promotion strategies can also be leveraged – especially when agents are able to get their late-breaking content indexed before more reputable sources. We hope this work will further stimulate the development of new approaches to monitoring, understanding, and uncovering propaganda campaigns as well as technology and policy-based solutions that help avoid political manipulation. Our findings suggest that cross-organization collaboration will be valuable in this endeavor.

Circumstances around the 2016 U.S. presidential elections and rising concerns about the influence of propaganda campaigns led to the public availability of valuable datasets for understanding IRA activities. Weaving together the public datasets with proprietary data on search and browsing activity provides a previously unavailable lens on the workings of the IRA campaign. We consider this work a preface to numerous opportunities ahead and to the many directions that remain to be explored. We see a dual moving forward, with the Web jointly holding great promise for strengthening liberal democracies while also serving as a platform that can be harnessed by those who seek to manipulate and disrupt. As the digital world continues evolve, and risks to democracy continue to emerge, openness and cooperation among major stakeholders will be essential to understand and counter malevolent threats.

References

Abouelenien, M.; Pérez-Rosas, V.; Mihalcea, R.; and Burzo, M. 2014. Deception detection using a multimodal approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, 58–65. ACM.

Allcott, H., and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*.

Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The Role of Social Networks in Information Diffusion. In *WWW*.

BallotPedia. 2017. "BallotPedia: The Encyclopedia of American Politics". `https://ballotpedia.org/`.

Bessi, A., and Ferrara, E. 2016. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* 21(11).

Bessi, A. 2016. Personality Traits and Echo Chambers on Facebook. *Computers in Human Behavior*.

Bloomberg. 2018. Company Overview of BuzzFeed, Inc. `https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId=46607363`.

Boyd, R. L.; Spangher, A.; Fourney, A.; Nushi, B.; Ranade, G.; Pennebaker, J. W.; and Horvitz, E. 2018. Characterizing the Internet Research Agency's Social Media Operations During the 2016 U.S. Presidential Election using Linguistic Analyses. `https://psyarxiv.com/ajh2q`.

Budak, C.; Agrawal, D.; and El Abbadi, A. 2011. Limiting the Spread of Misinformation in Social Networks. In *WWW*.

Center for Responsive Politics. 2017. Opensecrets.org. `https://www.opensecrets.org/`.

Conover, M.; Ratkiewicz, J.; Francisco, M. R.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political Polarization on Twitter. *ICWSM*.

Conroy, N. J.; Rubin, V. L.; and Chen, Y. 2015. Automatic deception detection: Methods for finding fake news. In *ASISST Annual Meeting*. American Society for Information Science.

Darren L. Linvill, P. L. W. 2018. "The Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building". `http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf`.

David Leip. 2016. "2016 Presidential General Election Data - Popular Vote by State.". `https://uselectionatlas.org/`.

Faris, R.; Roberts, H.; Etling, B.; Bourassa, N.; Zuckerman, E.; and Benkler, Y. 2017. Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election. (ID 3019414).

Fourney, A.; Racz, M. Z.; Ranade, G.; Mobius, M.; and Horvitz, E. 2017. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *CIKM*. ACM.

Garimella, V. R. K., and Weber, I. 2017. A Long-Term Analysis of Polarization on Twitter. In *ICWSM*.

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *WWW*.

GDELT. 2018. Global events database. https://www.gdeltproject.org.

Gil de Zúñiga, H.; Jung, N.; and Valenzuela, S. 2012. Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*.

Gillani, N.; Yuan, A.; Saveski, M.; Vosoughi, S.; and Roy, D. 2018. Me, my echo chamber, and I: introspection on social media polarization. In *WWW*.

Howard, P. N., and Kollanyi, B. 2016. Bots,# StrongerIn, and# Brexit: Computational Propaganda during the UK-EU Referendum. https://papers.ssrn.com/abstract=2798311.

Hughs, D. 2012. A Tale of Two Sites: Twitter vs. Facebook and the Personality Predictors of Social Media Usage.

Inel, O.; Khamkham, K.; Cristea, T.; Dumitrache, A.; Rutjes, A.; van der Ploeg, J.; Romaszko, L.; Aroyo, L.; and Sips, R.-J. 2014. Crowdtruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In *ISWC*, 486–504. Springer.

Intelligence Community Assessment. 2017. Background to "Assessing Russian Activities and Intentions in Recent US Elections": The Analytic Process and Cyber Incident Attribution. Technical Report ICA 2017-01D, Office of the Director of National Intelligence.

Irwin, N., and Mui, Y. Q. 2013. Washington Post Sale: Details of Bezos Deal. *The Washington Post*.

Jeremy Pressman, E. C. 2018. https://sites.google.com/view/crowdcountingconsortium.

Mak, T. 2018. Russian Influence Campaign Sought To Exploit Americans' Trust In Local News. *National Public Radio*.

Marwick, A., and Lewis, R. 2017. Media Manipulation and Disinformation Online. *New York: Data & Society Research Institute*.

Menczer, F. 2016. The Spread of Misinformation in Social Media. In *WWW*.

Meyer, R. 2016. How Many Stories Do Newspapers Publish Per Day. *The Atlantic*.

Mihalcea, R.; Pérez-Rosas, V.; and Burzo, M. 2013. Automatic detection of deceit in verbal communication. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 131–134. ACM.

Morstatter, F.; Shao, Y.; Galstyan, A.; and Karunasekera, S. 2018. From alt-right to alt-rechts: Twitter analysis of the 2017 german federal election. In *WWW*.

Offer, O. 2018. https://www.similarweb.com/.

Pompeo, J. 2014. Taking Stock of Newsroom Head Counts. *Politico*.

Pérez-Rosas, V.; Abouelenien, M.; Mihalcea, R.; and Burzo, M. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 59–66. ACM.

Resnick, P.; Garrett, R. K.; Kriplean, T.; Munson, S. A.; and Stroud, N. J. 2013. Bursting Your (Filter) Bubble: Strategies for Promoting Diverse Exposure. *CSCW*.

Robert S. Mueller, I. 2018. Internet research agency indictment. https://www.justice.gov/file/1035477/download.

Saleh, K. 2017. The Average Website Conversion Rate by Industry. https://www.invespcro.com/blog/the-average-website-conversion-rate-by-industry/.

Scheines, R. 1997. An Introduction to Causal Inference. In *Causality in Crisis? University of Notre Dame*, 185–200. Press.

Schiff, A. 2018a. Statement on release of facebook advertisements. https://democrats-intelligence.house.gov/news/documentsingle.aspx?DocumentID=379.

Schiff, A. 2018b. Statement on release of twitter ads, accounts and data. https://democrats-intelligence.house.gov/news/documentsingle.aspx?DocumentID=396.

S.E.C. 2012. The Washington Post Company Form 10K. https://www.sec.gov/Archives/edgar/data/104889/000010488913000009/d10k.htm.

S.E.C. 2017. The New York Times Company 2017 Annual Report. https://s1.q4cdn.com/156149269/files/doc_financials/annual/2017/Final-2017-Annual-Report.pdf.

Shane, S., and Mazzetti, M. 2018. Inside a 3-Year Russian Campaign to Influence U.S. Voters. *The New York Times*.

Shane, S. 2017. These Are the Ads Russia Bought on Facebook in 2016. *The New York Times*.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *KDD*.

Silverman, C. 2016. This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. *Buzzfeed*.

Spangher, A. 2015. Building the Next New York Times Recommendation Engine. *The New York Times*.

Stewart, L. G.; Arif, A.; and Starbird, K. 2018. Examining trolls and polarization with a retweet network. In *WSDM Workshop on Misinformation and Misbehavior Mining on the Web*.

Sunstein, C. R. 2018. *# Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

U.S. Census Bureau. 2010. American community survey. Prepared by Social Explorer.

U.S. Census Bureau. 2016. Voting and registration. https://www.census.gov/topics/public-sector/voting/data/tables.html.

Varol, O.; Ferrara, E.; Davis, C. A.; Menczer, F.; and Flammini, A. 2017a. Online Human-Bot Interactions: Detection, Estimation, and Characterization. *arXiv:1703.03107 [cs]*. arXiv: 1703.03107.

Varol, O.; Ferrara, E.; Menczer, F.; and Flammini, A. 2017b. Early detection of promoted campaigns on social media. *EPJ Data Science* 6(1):13.

Younus, A.; Qureshi, M. A.; Saeed, M.; Touheed, N.; O'Riordan, C.; and Pasi, G. 2014. Election Trolling: Analyzing Sentiment in Tweets During Pakistan Elections 2013. In *WWW*.

Zandt, D. V. 2018. https://mediabiasfactcheck.com/.

Appendices

Appendix 1: Campaign details

Campaign Structure and Content

In this appendix, we describe the campaign in more detail and discuss how our observations match with statistics released by Congress ((Schiff 2018a), (Schiff 2018b)).

In the post accompanying the release of Facebook and Twitter data by the U.S. House of Representatives Permanent Select Committee on Intelligence[19], Representative Adam Schiff notes:

> "The Russians ... weav[ed] together fake accounts, pages, and communities to push politicized content and videos, and to mobilize real Americans to sign online petitions and join rallies and protests."

We presented evidence of examples of the IRA promoting protests, meetups, and numerous engagement techniques with their ads, as discussed above. However, we did not find evidence of extensive user engagement. On the other hand, the content that we analyze in our paper contains some gaps and discrepancies with the content Schiff describes, which we explore and justify here.

Facebook Rep. Schiff reports identifying the following content for the IRA's Facebook campaign (Schiff 2018a):

- 3,519 Facebook ads.
- 470 IRA-created Facebook pages.
- More than 80,000 pieces of organic content posted to Facebook accounts.

We observe 3,519 released PDF documents, each describing a Facebook ad, but as noted in Section II, we were only able to extract and process usable content from 3,061 of these, i.e. 84% of the posted ads as a result of blank ad-description fields, missing URL fields, blank documents, and OCR errors.

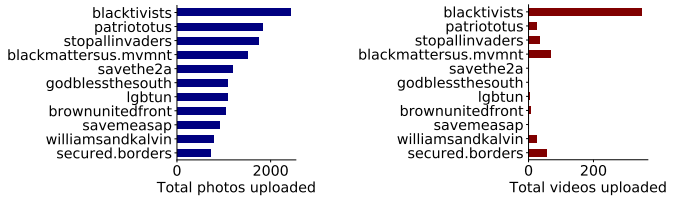
As noted in Section III, we only find 104 IRA-linked Facebook groups being promoted by the ads, in contrast to (Schiff 2018a)'s stated 470 groups. Formal communication and fact-checking with researchers at Facebook have confirmed that only 107 of the groups identified were advertised. The missing 3 groups in our data is likely due to OCR errors introduced when parsing released Facebook-ad PDFs.

We seek to verify the volume of content reported by Rep. Schiff by deduplicating unique content URLs. We find a total of 14,860 photo and video posts to IRA-linked accounts, shown in Figure 10. We were not able to successfully track other post-types due to the nature of our instrumentation logs, and the manner by which other content appears on Facebook. It is likely that out of the 80,000 pieces of content reported by (Schiff 2018a), the over 65,000 pieces that we did not observe included both: organic posts without photos and videos posted to the 104 IRA-controlled groups included in our study, as well as content posted to the 366 IRA-controlled Facebook groups not included in our analysis (as detailed above). Discussions with both Facebook representatives and Congressman Adam Schiff's office for the purposes of fact-checking have confirmed that this interpretation is reasonable, and there are ongoing efforts to further clarify.

Twitter Schiff reports finding the following content for the IRA's Facebook campaign:

- More than 36,000 Russian-linked bot accounts.

[19] https://democrats-intelligence.house.gov/social-media-content/



(a) Total photos posted by account. A photo upload is determined by the minimum timestamp of a click on it's unique URL key.

(b) Total videos posted by account. A unique video upload is determined the same way a photo upload is.

Figure 10: **Facebook photos and video uploads:** Metrics capturing the volume of organic (unpaid) posts involving photos and videos uploaded to IRA Facebook groups.

- 3,841 Twitter accounts affiliated with the IRA.
- More than 130,000 tweets by accounts linked to the IRA.

We had more difficulty verifying Twitter data because of ambiguity between the definitions of "Russian-linked bot accounts" and "accounts affiliated or linked to the IRA". Further the number 130,000 was smaller than what was observed by other researchers (Darren L. Linvill 2018). We attempted to fact-check with Representative Schiff's office and representatives from Twitter to resolve these queries but were unable to come to a resolution. In this paper, we therefore rely on data provided by (Darren L. Linvill 2018), which aggregates historical tweets produced over time by the aforementioned IRA affiliated or linked accounts.

After excluding non-English tweets and restricting to the categories as described in the methodology section, we were left with 1,746,000 tweets from 917 accounts believed to be IRA-affiliated. Further restricting to our time-period of observation, we are left with 471,000 from 360 accounts.

Campaign Size and Discussion

In this discussion section, we seek to give a comparative analysis of the size of the IRA's operation in terms of staff, budget and content output. In his indictment of the IRA (Robert S. Mueller 2018), Special Counsel Robert Mueller states:

> The ORGANIZATION [the IRA] employed hundreds of individuals for its online operations, ranging from creators of fictitious personas to technical and administrative support. The ORGANIZATION's annual budget totaled the equivalent of millions of U.S. dollars.

As a rough comparison, the *New York Times* newsroom, according to public information, employs about 1,300 journalists, writers, copy-editors, social media producers and others (Pompeo 2014). According to the *Times'* 10-K report, its operating cost in 2017 was 1.48 billion USD (S.E.C. 2017).

The *Times* uploads between 200-300 articles, blogs, and interactives a day (Spangher 2015). From public browsing, it appears the *Times* tweets from 5-10 accounts, about 50-100 times a day. They post on Facebook from 5-10 accounts, posting 50-100 times a day as well.

The *Washington Post* likewise has a newsroom of more than 700 staff. According to public statements, they produce an average of 500 pieces of content a day (videos, photos, articles and interactive pieces). The last available public data for the

Post shows operating costs in the range of 1.8 billion USD, in 2012[20](S.E.C. 2012). According to a brief survey of their social media accounts, they too produce between 50-100 posts on their Facebook accounts and 50-100 tweets from their Twitter accounts a day.

Buzzfeed, according to public reporting, has a newsroom of around 460 staff. They produced roughly 200 articles a day in 2016 (Meyer 2016). A brief survey on social media reveals that *BuzzFeed* posts to Facebook between 100-200 times a day and tweets roughly the same amount.[21]

Based on material revealed from congressional testimony, the IRA generated over 2.9 million tweets from over $2,800$ active accounts[22]. During our observation period, they tweeted at least, on average, $2,000$ times a day. They fielded $3,519$ Facebook ads, roughly 5 ads a day during our period of interest. According to Facebook photo/video URL scraping, discussed earlier, they additionally produced at least $14{,}285$ photos and 575 videos over the period of interest, or roughly 70 photos and 3 videos a day (Figure 12).

This does not include content on their owned domains. The domains we have examined show abundant material being produced. We find `blackmattersus.com` to be particularly expansive. By performing a site-wide scrape, we find over 3,900 articles published, and roughly 5 per day during our observation period. They logged more than 3 new meetup links a day during our observation period.

An apples-to-apples comparison based on raw volume of output is doubtlessly flawed, since we can not compare the production time each piece of content requires at the *Times*, the *Post* or *BuzzFeed* with the time required for content produced by the IRA. Much of the IRA's content is likely derivative or simply copied: many articles on `blackmattersus.com` appear duplicative, and 95% of the tweets recorded from accounts of interest are retweets. (Less than 1% of tweets from major news organizations are retweets). However, even with very conservative estimates, assuming their staff spent fractions of the time per piece of content as *BuzzFeed*, we add to Mueller's estimate (Robert S. Mueller 2018) to project that just their *content-focused* staff still numbered in the low hundreds. Their budget in the "millions of USD" (Robert S. Mueller 2018) means they spent far capital per piece of content than the *Times* or the *Post*, but we emphasize that nevertheless, this was an effort that approached the scope of a large newsroom.

An important question is whether the IRA campaign was worthwhile. As presented in Table 5 and Figure 11, the campaign and the agency received a significant volume of coverage from Western mainstream outlets at hundreds of articles per outlet and thousands of articles a day. These facts speak to the scale of such campaigns and the role that journalists and scientists – present authors included – should play in these ongoing attempts to manipulate the online information landscape.

Appendix 2: Effect of paid promotions

As noted in the paper, the IRA spent more money on left-leaning content than on right-leaning content on Facebook during our observation period. Furthermore, traffic spikes to IRA Facebook groups were also influenced by events such as photo and video uploads in addition to paid promotions. To understand if paid promotions mattered more for some groups than for other groups, we asked: Would these groups still have had high traffic spikes had they not been advertised on Facebook?

We are limited in our ability to answer this question given that we cannot conduct a controlled study of traffic with and without paid promotions. Hence we decided to build a simple model to predict traffic to a group in the absence of a paid promotion, and use this as a comparison baseline.

We trained a model with the goal of predicting how each Facebook group would have performed *had it not been promoted using a paid ad*. For this purpose we trained the model to predict the probability of a spike (a spike only using (`group url`,`date`) tuples for which the url was not promoted on the given date and we omitted the promotion treatment from the input feature set. We then evaluate the predicted performance on days that articles *did* receive promotions and compared our predictions.

We note that our causal analysis here is restricted. The question we ask is "Would these groups still have had high

Media Outlet Mentioning IRA	Count of articles
yahoo.com	830
iheart.com	550
nbcnews.com	306
msn.com	256
washingtonpost.com	210
reuters.com	210
enterprise-security-today.com	168
dailymail.co.uk	153

Table 5: **IRA mentions in Western media, top outlets.** Top outlets publishing articles mentioning "Russian Troll", "IRA", "internet-research-agency", "blacktivist", "black-matters-us" in URLs (calculated using (GDELT 2018)).

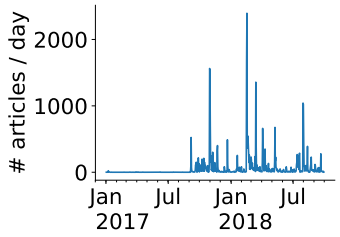


Figure 11: **IRA mentions in Western media, over time.** Number of articles in western media mentioning "Russian Troll", "IRA", "internet-research-agency", "blacktivist", "black-matters-us" in URLs(calculated using (GDELT 2018)).

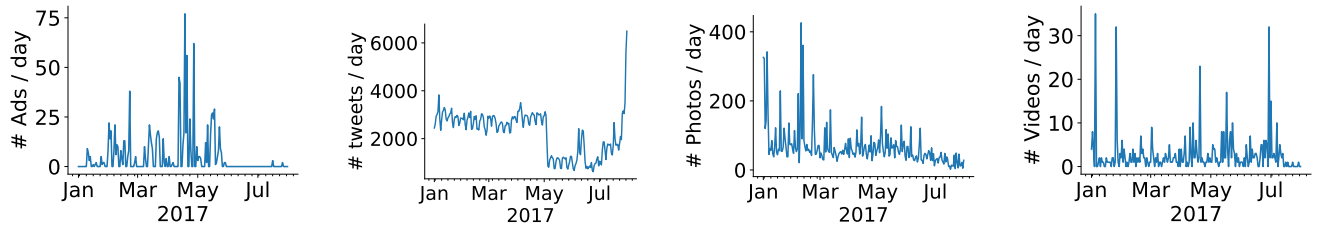
Feature type (X)	**Feature name**
historical	click volume (7-day window excluding the last 2 days)
unpaid action	#posted photos #posted videos
contextual	#clicks on external similar urls #search queries on similar topics
content	topics present in the url (binary)

Table 6: Feature types used in counterfactual modeling.

[20]The *Post* was bought in 2013 by Jeffrey Preston Bezos, and is now held by a privately owned company, Nash Holdings LLC. (Irwin and Mui 2013)

[21]*BuzzFeed* has always been a privately held, and does not disclose costs (Bloomberg 2018).

[22]Of the 3841 Twitter handles in the dataset, 1034 handles posted no content during the period of observation (Darren L. Linvill 2018).



(a) Number of Facebook ads posted over time by IRA accounts.

(b) Number of tweets tweeted over time by IRA handles.

(c) Photos posted on Facebook over time to IRA accounts.

(d) Videos posted on Facebook over time to IRA accounts.

Figure 12: Metrics capturing the volume of unpaid photos and videos uploaded to IRA Facebook groups.

Facebook Group	(a) AUC *no prom.*	(b) AUC *prom.*
blacktivists	0.965	0.698
patriototus	0.876	0.703
brownunitedfront	0.961	0.792
williamsandkalvin	0.808	0.771
secured.borders	0.792	0.760
stopallinvaders	0.712	0.724

Table 7: Area under ROC curve (AUC) of prediction models tested on time-window holdout data on days (a) without promotions or (b) with promotions, ordered by the difference between the two columns.

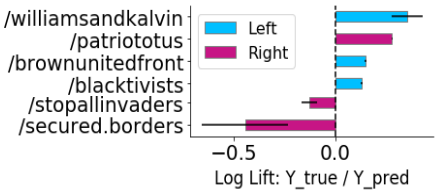


Figure 13: **Effect of Promotions.** The average log lift of click percentile for each group on days with a paid promotion, compared the probability we predict a spike would have occurred without promotion. Positive values indicate higher importance of promotions with respect to traffic spikes.

traffic spikes had they not been advertised on Facebook?" This is different from the question: "What is the impact of a promotion?". In fact, the question we ask is closer to: "When did the promotions not matter for traffic spikes?"

Data and Methodology. For this case study, we consider a specific set of Facebook groups that received both large amounts of traffic and many promotions during the train and test periods. These constraints yield six groups[23]. We limit our study to these groups since it is difficult to make any statements on the effect of promotions for groups that did not receive many promotions.

We formatted our dataset as (`group url`, `date`) pairs. We trained separate ℓ_2-constrained Logistic Regression models for each group on all (`group url`,`date`) pairs with *no* promotions to predict the likelihood of a traffic spike to that group on the given date. Here, a "spike" is defined as click-counts larger than the 50th percentile of all daily click-counts for the Facebook group being modeled. We train only on dates *without* promotions, to predict the counterfactual traffic outcome on dates *with* promotions. Our training data came from 1/1 to 5/1 and our test data was that collected from 5/2 to 8/1.

Table 6 summarizes all features used for these models. *Unpaid action* features count the number of photos and videos posted by each group on the given date. *Content* features are binary indicators marking topic-relevance based on our crowdsourcing task. *Contextual* features track traffic to non-IRA news websites properties and volumes of search queries that were topically relevant to any of the topics of any of the groups being studied. These features capture the possible influence of external news factors.

Additionally, instead of using traffic immediately preceding each date to calculate historical features, a buffer of 2 days is used. We purposely exclude a feature indicating whether the property had been promoted during the given day. And we took steps to ensure that no other model features would leak this information (e.g. the click volume feature only counts clicks up to two days before the spike date). We experimented with adding a feature counting past-promotions, but found little change.

We interrogated the independence relationships in our observed data to test whether our assumptions hold. We test for conditional independence between treatment and all other variables, given outcome (using $corr(x_i,x_j|y)=0$ for all values of $y\in\{0,1\}$, and for all x_i, x_j pairs). We observe independence between all treatment and outcome variables ($correlation > .1, p < .1$). (We apply significance testing with Bonferroni corrections to critique our *faithfulness* assumption.)[24]

These independence observations give us confidence that our assumptions are being met.

The modeling decisions we made improve the adherence of our data to the causal structure we infer.

1. Treatment assignment: We do not find strong evidence for a dependence relationship between treatment assignment and outcome, mediated through confounding factors. Based on a Causal Markov assumption, this implies that we can safely model using these features as independent (Scheines 1997).
2. Historical treatments: We also did not find significant relationships between historical promotions and current

[23]These 6 groups together account for 64% of traffic and 36% of ads ran (15% of ad spend) over the observation period.

[24]Note: we continued to see correlations between individual features in X, but according to the *Causal Markov* assumption, this does not impact our causal claims regarding the promotion's effect on the outcome.

promotions, strengthening our belief that independence across different dates and urls is maintained.

Additionally, we took steps to limit overfitting. As mentioned above, we model each Facebook Group separately. We tuned each model's ℓ_2 hyperparameter using holdout validation data during training. We also iterated to include external news and traffic features. We choose to measure the Area Under the Receiver Operating Characteristic Curve (AUC) as a performance measure in order to handle class imbalance and also avoid picking a probability decision threshold for when a prediction should be leaning towards a spike.

Findings. The most important features across our models, in terms of average coefficient values across all models, were *historical clicks*, *# posted photos*, *# posted videos*. This suggests a role for *unpaid actions* in the IRA strategy.

Table 7 shows the AUCs for each group calculated for days with more than one promotion running (*promo.* column) or with no running promotion (*no prom.* column) in the test period. Again, since the models were trained only on *no prom.* days in the training period, the *no prom.* column is a validation test to empirically check how well the model generalizes to previously unseen data with no promotions, while the *prom.* column tests our counterfactual extrapolation.

For groups with a *prom.* AUC closer to a *no-prom.* AUC (e.g. "StopAllInvaders"), promotional signal is not necessary for the model to predict a spike: i.e., a variable capturing promotion-information would not have added signal and other factors, encoded in the model features, had a higher impact. For cases where the AUC score is divergent (e.g. "Blacktivists"), model-features were not sufficient, meaning that other factors unknown to this model, including promotions, played a significant role. Again, the reason why we look at the difference between these two columns and not only at the absolute value of each, is to decouple the ability of the model to generalize to unseen no promotion data from the ability of the model to extrapolate traffic spike predictions on dates with promotions.

Figure 8 shows another view of these results. In Figure 8, we consider only the days *with* promotions, and compare the observed traffic patterns (showing *prom.* traffic) with our model predictions (modeling *no prom.*, or the counterfactual). We calculate the logarithmic ratio between the observed traffic (in percentile) and the predicted probability of a spike on that date, as assigned by the model.