Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web

Savvas Zannettou
Cyprus University of Technology
sa.zannettou@edu.cut.ac.cy

Tristan Caulfield
University College London
t.caulfield@ucl.ac.uk

Emiliano De Cristofaro
University College London
e.decristofaro@ucl.ac.uk

Michael Sirivianos
Cyprus University of Technology
michael.sirivianos@cut.ac.cy

Gianluca Stringhini
Boston University
gian@bu.edu

Jeremy Blackburn
University of Alabama at Birmingham
blackburn@uab.edu

ABSTRACT

Over the past couple of years, anecdotal evidence has emerged linking coordinated campaigns by state-sponsored actors with efforts to manipulate public opinion on the Web, often around major political events, through dedicated accounts, or "trolls." Although they are often involved in spreading disinformation on social media, there is little understanding of how these trolls operate, what type of content they disseminate, and most importantly their influence on the information ecosystem.

In this paper, we shed light on these questions by analyzing 27K tweets posted by 1K Twitter users identified as having ties with Russia's Internet Research Agency and thus likely state-sponsored trolls. We compare their behavior to a random set of Twitter users, finding interesting differences in terms of the content they disseminate, the evolution of their account, as well as their general behavior and use of Twitter. Then, using Hawkes Processes, we quantify the influence that trolls had on the dissemination of news on social platforms like Twitter, Reddit, and 4chan. Overall, our findings indicate that Russian trolls managed to stay active for long periods of time and to reach a substantial number of Twitter users with their tweets. When looking at their ability of spreading news content and making it viral, however, we find that their effect on social platforms was minor, with the significant exception of news published by the Russian state-sponsored news outlet RT (Russia Today).

CCS CONCEPTS

• **General and reference → General conference proceedings**; **Measurement**; • **Mathematics of computing → Probabilistic algorithms**; • **Networks → Social media networks**; **Online social networks**.

KEYWORDS

disinformation, trolls, social networks, twitter, reddit, 4chan

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316495

ACM Reference Format:
Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3316495

1 INTRODUCTION

Recent political events and elections have been increasingly accompanied by reports of disinformation campaigns attributed to state-sponsored actors [7]. In particular, "troll farms," allegedly employed by Russian state agencies, have been actively commenting and posting content on social media to further the Kremlin's political agenda [23]. In late 2017, the US Congress started an investigation on Russian interference in the 2016 US Presidential Election, releasing the IDs of 2.7K Twitter accounts identified as Russian trolls.

Despite the growing relevance of state-sponsored disinformation, the activity of accounts linked to such efforts has not been thoroughly studied. Previous work has mostly looked at campaigns run by bots [7, 10, 20]; however, automated content diffusion is only a part of the issue, and in fact recent research has shown that human actors are actually key in spreading false information on Twitter [21]. Overall, many aspects of state-sponsored disinformation remain unclear, e.g., how do state-sponsored trolls operate? What kind of content do they disseminate? And, perhaps more importantly, is it possible to quantify the influence they have on the overall information ecosystem on the Web?

In this paper, we aim to address these questions, by relying on the set of 2.7K accounts released by the US Congress as ground truth for Russian state-sponsored trolls. From a dataset containing all tweets released by the 1% Twitter Streaming API, we search and retrieve 27K tweets posted by 1K Russian trolls between January 2016 and September 2017. We characterize their activity by comparing to a random sample of Twitter users. Then, we quantify the influence of these trolls on the greater Web, looking at occurrences of URLs posted by them on Twitter, 4chan [11], and Reddit, which we choose since they are impactful actors of the information ecosystem [29]. Finally, we use Hawkes Processes [15] to model the influence of each Web community (i.e., Russian trolls on Twitter, overall Twitter, Reddit, and 4chan) on each other.

Main findings. Our study leads to several key observations:

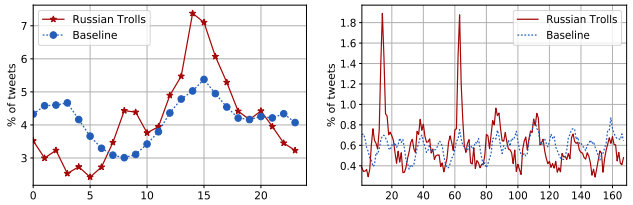
(1) Trolls actually bear very small influence in making news go viral on Twitter and other social platforms alike. A noteworthy exception are links to news originating from RT (Russia Today), a state-funded news outlet: indeed, Russian trolls are quite effective in "pushing" these URLs on Twitter and other social networks.
(2) The main topics discussed by Russian trolls target very specific world events (e.g., Charlottesville protests) and organizations (such as ISIS), and political threads related to Donald Trump and Hillary Clinton.
(3) Trolls adopt different identities over time, i.e., they "reset" their profile by deleting their previous tweets and changing their screen name/information.
(4) Trolls exhibit significantly different behaviors compared to other (random) Twitter accounts. For instance, the locations they report concentrate in a few countries like the USA, Germany, and Russia, perhaps in an attempt to appear "local" and more effectively manipulate opinions of users from those countries. Also, while random Twitter users mainly tweet from mobile versions of the platform, the majority of the Russian trolls do so via the Web Client.

2 DATASETS

Russian trolls. We start from the 2.7K Twitter accounts suspended by Twitter because of connections to Russia's Internet Research Agency. The list of these accounts was released by the US Congress as part of their investigation of the alleged Russian interference in the 2016 US presidential election, and includes both Twitter's *user ID* (which is a numeric unique identifier associated to the account) and the *screen name*.[1] From a dataset storing all tweets released by the 1% Twitter Streaming API, we search for tweets posted between January 2016 and September 2017 by the user IDs of the trolls. Overall, we obtain 27K tweets from 1K out of the 2.7K Russian trolls. Note that the criteria used by Twitter to identify these troll accounts are not public. What we do know is that this is not the complete set of active Russian trolls, because 6 days prior to this writing Twitter announced they have discovered over 1K more troll accounts.[2] Nonetheless, it constitutes an invaluable "ground truth" dataset enabling efforts to shed light on the behavior of state-sponsored troll accounts.

Baseline dataset. We also compile a list of random Twitter users, while ensuring that the distribution of the average number of tweets per day posted by the random users is similar to the one by trolls. To calculate the average number of tweets posted by an account, we find the first tweet posted after January 1, 2016 and retrieve the overall tweet count. This number is then divided by the number of days since account creation. Having selected a set of 1K random users, we then collect all their tweets between January 2016 and September 2017, obtaining a total of 96K tweets. We follow this approach as it gives a good approximation of posting behavior, even though it might not be perfect, since (1) Twitter accounts can become more or less active over time, and (2) our datasets are based on the 1% Streaming API, thus, we are unable to control the number of tweets we obtain for each account.

[1] See https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf
[2] https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html



(a) Hour of Day (b) Hour of Week

Figure 1: Temporal characteristics of tweets from Russian trolls and random Twitter users.

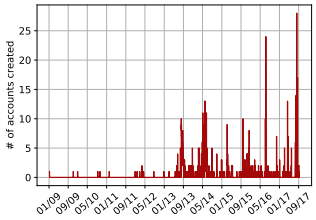


Figure 2: Number of Russian troll accounts created per day.

3 ANALYSIS

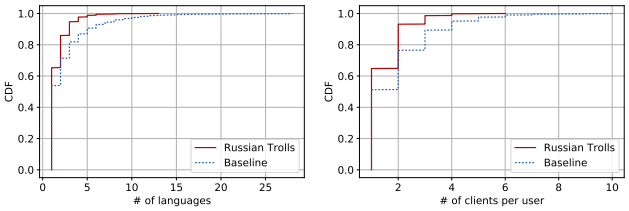
Temporal analysis. We observe that Russian trolls are continuously active on Twitter between January, 2016 and September, 2017, with a peak of activity just before the second US presidential debate (October 9, 2016). Fig. 1(a) shows that most tweets from the trolls are posted between 14:00 and 15:00 UTC. In Fig. 1(b), we also report temporal characteristics based on hour of the week, finding that both datasets follow a diurnal pattern, while trolls' activity peaks around 14:00 and 15:00 UTC on Mondays and Wednesdays. Considering that Moscow is three hours ahead UTC, this distribution does not rule out that tweets might actually be posted from Russia.

Account creation. Next, we examine the dates when the trolls infiltrated Twitter, by looking at the account creation dates. From Fig. 2, we observe that 71% of them are actually created before 2016. There are some interesting peaks, during 2016 and 2017: for instance, 24 accounts are created on July 12, 2016, approx. a week before the Republican National Convention (when Donald Trump received the nomination), while 28 appear on August 8, 2017, a few days before the infamous Unite the Right rally in Charlottesville. Taken together, this might be evidence of coordinated activities aimed at manipulating users' opinions with respect to specific events.

Account characteristics. We also shed light on the troll account profile information. In Table 1, we report the top ten words appearing in the screen names and the descriptions of Russian trolls, as well as character 4-grams for screen names and word bigrams for profile descriptions. Interestingly, a substantial number of Russian trolls pose as news outlets, evident from the use of the term "news" in both the screen name (1.3%) and the description (10.7%). Also, it seems they attempt to increase the number of their followers, thus their reach of Twitter users, by nudging users to follow them (see, e.g., "follow me" appearing in almost 8% of the accounts). Finally,

Table 1: Top 10 words found in Russian troll screen names and account descriptions. We also report character 4-grams for the screen names and word bigrams for the description.

Screen Name				Description			
Word	(%)	4-gram	(%)	Word	(%)	Word bigram	(%)
news	1.3%	news	1.5%	news	10.7%	follow me	7.8%
bote	1.2%	line	1.5%	follow	10.7%	breaking news	2.6%
online	1.1%	blac	1.3%	conservative	8.1%	news aus	2.1%
daily	0.8%	bote	1.3%	trump	7.8%	uns in	2.1%
today	0.6%	rist	1.1%	und	6.2%	deiner stdt	2.1%
ezekiel2517	0.6%	nlin	1.1%	maga	5.9%	die news	2.1%
maria	0.5%	onli	1.0%	love	5.8%	wichtige und	2.1%
black	0.5%	lack	1.0%	us	5.3%	nachrichten aus	2.1%
voice	0.4%	bert	1.0%	die	5.0%	aus deiner	2.1%
martin	0.4%	poli	1.0%	nachrichten	4.3%	die dn	2.1%



(a) (b)

Figure 3: CDF of number of (a) languages used (b) clients used per user.

Table 2: Top 10 Twitter clients (as % of tweets).

Client (Trolls)	(%)	Client (Baseline)	(%)
Twitter Web Client	50.1%	TweetDeck	32.6%
twitterfeed	13.4%	Twitter for iPhone	26.2%
Twibble.io	9.0%	Twitter for Android	22.6%
IFTTT	8.6%	Twitter Web Client	6.1%
TweetDeck	8.3%	GrabInbox	2.0%
NovaPress	4.6%	Twitter for iPad	1.4%
dlvr.it	2.3%	IFTTT	1.0%
Twitter for iPhone	0.8%	twittbot.net	0.9%
Zapier.com	0.6%	Twitter for BlackBerry	0.6%
Twitter for Android	0.6%	Mobile Web (M2)	0.4%

10.3% of the Russian trolls describe themselves as Trump supporters: "trump" and "maga" (Make America Great Again, one of Trump campaign's main slogans).

Language. Looking at the language (as provided via the Twitter API) of the tweets posted by Russian trolls, we find that most of them (61%) are in English, although a substantial portion are in Russian (27%), and to a lesser extent in German (3.5%). In Fig. 3(a), we plot the cumulative distribution function (CDF) of the number of different languages for each user: 64% of the Russian trolls post all their tweets in only one language, compared to only 54% for random users. Overall, by comparing the two distributions, we observe that random Twitter users tend to use more languages in their tweets compared to the trolls.

Client. Finally, we analyze the clients used to post tweets. We do so since previous work [5] shows that the client used by official or professional accounts are quite different that the ones used by regular users. Table 2 reports the top 10 clients for both Russian trolls and baseline users. We find the latter prefer to use Twitter

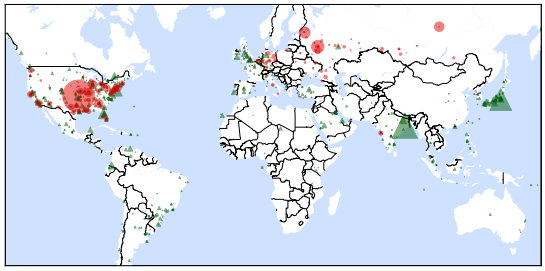


Figure 4: Distribution of reported locations for tweets by Russian trolls (red circles) and baseline (green triangles).

clients for mobile devices (48%) and the TweetDeck dashboard (32%), whereas, the former mainly use the Web client (50%). We also assess how many different clients Russian trolls use throughout our dataset: in Fig. 3(b), we plot the CDF of the number of clients used per user. We find that 65% of the Russian trolls use only one client, 28% of them two different clients, and the rest more than three, which is overall less than the random baseline users.

Location. We then study users' location, relying on the self-reported location field in their profiles. Note that users not only may leave it empty, but also change it any time they like, so we look at locations for each tweet. We retrieve it for 75% of the tweets by Russian trolls, gathering 261 different entries, which we convert to a physical location using the Google Maps Geocoding API. In the end, we obtain 178 unique locations for the trolls, as depicted in Fig. 4 (red circles). The size of the circles on the map indicates the number of tweets that appear at each location. We do the same for the baseline, getting 2,037 different entries, converted by the API to 894 unique locations. We observe that most of the tweets from Russian trolls come from locations within the USA and Russia, and some from European countries, like Germany, Belgium, and Italy. On the other hand, tweets in our baseline are more uniformly distributed across the globe, with many tweets from North and South America, Europe, and Asia. This suggests that Russian trolls may be pretending to be from certain countries, e.g., USA or Germany, aiming to pose as locals and better manipulate opinions. This explanation becomes more plausible when we consider that a plurality of trolls' tweets have their location set as a generic form of "US," as opposed to a specific city, state, or even region. Interestingly, the 2nd, 3rd, and 4th most popular location for trolls to tweet from are Moscow, St. Petersburg, and a generic form of "Russia." We also assess whether users change their country of origin based on the self-reported location: only a negligible percentage (1%) of trolls change their country, while for the baseline the percentage is 16%.

Media. We then assess whether Russian trolls use images and videos in a different way than random baseline users. For Russian trolls (resp., baseline accounts), 66% (resp., 73%) of the tweets include no images, 32% (resp., 18%) exactly one image, and 2% (resp., 9%) more than one. This suggests that Russian trolls disseminate a considerable amount of information via single-image tweets. As for videos, only 1.5% of the tweets from Russian trolls includes a video, as opposed to 6.4% for baseline accounts.

Table 3: Top 20 hashtags in tweets from Russian trolls and baseline users.

Trolls				Baseline			
Hashtag	(%)	Hashtag	(%)	Hashtag	(%)	Hashtag	(%)
news	7.2%	US	0.7%	iHeartAwards	1.8%	UrbanAttires	0.6%
politics	2.6%	tcot	0.6%	BestFanArmy	1.6%	Vacature	0.6%
sports	2.1%	PJNET	0.6%	Harmonizers	1.0%	mPlusPlaces	0.6%
business	1.4%	entertainment	0.5%	iOSApp	0.9%	job	0.5%
money	1.3%	top	0.5%	JouwBaan	0.9%	Directioners	0.5%
world	1.2%	topNews	0.5%	vacature	0.9%	JIMIN	0.5%
MAGA	0.8%	ISIS	0.4%	KCA	0.9%	PRODUCE101	0.5%
health	0.8%	Merkelmussbleiben	0.4%	Psychic	0.8%	VoteMainFPP	0.5%
local	0.7%	IslamKills	0.4%	RT	0.8%	Werk	0.4%
BlackLivesMatter	0.7%	breaking	0.4%	Libertad2016	0.6%	dts	0.4%

Hashtags. Our next step is to study the use of hashtags in tweets. Russian trolls use at least one hashtag in 32% of their tweets, compared to 10% for the baseline. Overall, we find 4.3K and 7.1K unique hashtags for trolls and random users, respectively, with 74% and 78% of them only appearing once. In Table 3, we report the top 20 hashtags for both datasets. Trolls appear to use hashtags to disseminate news (7.2%) and politics (2.6%) related content, but also use several that might be indicators of propaganda and/or controversial topics, e.g., #ISIS, #IslamKills, and #BlackLivesMatter. For instance, we find some notable examples including: "We just have to close the borders, 'refugees' are simple terrorists #IslamKills" on March 22, 2016, "#SyrianRefugees ARE TERRORISTS from #ISIS #IslamKills" on March 22, 2016, and "WATCH: Here is a typical #BlackLivesMatter protester: 'I hope I kill all white babes!' #BatonRouge <url>" on July 17, 2016.

We also study when these hashtags are used by the trolls, finding that most of them are well distributed over time. However, there are some interesting exceptions, e.g., with #Merkelmussbleiben (a hashtag seemingly supporting Angela Merkel) and #IslamKills. Specifically, tweets with the former appear exclusively on July 21, 2016, while the latter on March 22, 2016, when a terrorist attack took place at Brussels airport. These two examples illustrate how the trolls may be coordinating to push specific narratives on Twitter.

Mentions. We find that 46% of trolls' tweets include *mentions* to 8.5K unique Twitter users. This percentage is much higher for the random baseline users (80%, to 41K users). Table 4 reports the 20 top mentions we find in tweets from Russian trolls and baseline users. We find several Russian accounts, like 'leprasorium' (a popular Russian account that mainly posts memes) in 2% of the mentions, as well as popular politicians like 'realDonaldTrump' (0.6%). The practice of mentioning politicians on Twitter may reflect an underlying strategy to mutate users' opinions regarding a particular political topic, which has been also studied in previous work [2].

URLs. We then analyze the URLs included in the tweets. First of all, we note that 53% of the trolls' tweets include at least a URL, compared to only 27% for the random baseline. There is an extensive presence of URL shorteners for both datasets, e.g., bit.ly (12% for trolls and 26% for the baseline) and ift.tt (10% for trolls and 2% for the baseline), therefore, in November 2017, we visit each URL to obtain the final URL after all redirections. In Fig. 5, we plot the CDF of the number of URLs per unique domain. We observe that Russian trolls disseminate more URLs in their tweets compared to the baseline. This might indicate that Russian trolls include URLs to increase their credibility and positive user perception; indeed, [9] show that adding a URL in a tweet correlates with higher credibility scores. Also, in Table 6, we report the top 20 domains for both Russian trolls and the baseline. Most URLs point to content within Twitter itself; 13% and 35%, respectively. Links to a number of popular social networks like YouTube (1.8% and 4.2%, respectively) and Instagram (1.5% and 1.9%) appear in both datasets. We also note that among the top 20 domains, there are also a number of news outlets linked from trolls' tweets, e.g., Washington Post (0.7%), Seattle Times (0.7%), and state-sponsored news outlets like RT (0.8%) in trolls' tweets, but much less so from the baseline.

Table 4: Top 20 mentions in tweets from trolls and baseline.

Trolls				Baseline			
Mention	(%)	Mention	(%)	Mention	(%)	Mention	(%)
leprasorium	2.1%	postsovet	0.4%	TasbihIstighfar	0.3%	RasSpotlights	0.1%
zubovnik	0.8%	DLGreez	0.4%	raspotlights	0.2%	GenderReveals	0.1%
realDonaldTrump	0.6%	DanaGeezus	0.4%	FunnyBrawls	0.2%	TattedChanel	0.1%
midnight	0.6%	ruopentwit	0.3%	YouTube	0.2%	gemvius	0.1%
blicqer	0.6%	Spoontamer	0.3%	Harry_Styles	0.2%	DrizzyNYC__	0.1%
gloed_up	0.6%	YouTube	0.3%	shortdancevids	0.2%	August_Alsina_	0.1%
wylsacom	0.5%	ChrixMorgan	0.3%	UrbanAttires	0.2%	RihannaBITCH_	0.1%
TalibKweli	0.4%	sergeylazarev	0.3%	BTS_twt	0.2%	sexualfeed	0.1%
zvezdanews	0.4%	RT_com	0.3%	KylieJenner_NYC	0.2%	PetsEvery30	0.1%
GiselleEvns	0.4%	kozheed	0.3%	BaddiessNation	0.2%	IGGYAZALEAoO	0.1%

Table 5: Top 10 domains in tweets from trolls and the baseline.

Domain (Trolls)	(%)	Domain (Baseline)	(%)
twitter.com	12.81%	twitter.com	35.51%
reportsecret.com	7.02%	youtube.com	4.21%
riafan.ru	3.42%	vine.co	3.94%
politexpert.net	2.10%	factissues.com	3.24%
youtube.com	1.88%	blogspot.com.cy	1.92%
vk.com	1.58%	instagram.com	1.90%
instagram.com	1.53%	facebook.com	1.68%
yandex.ru	1.50%	worldstarr.info	1.47%
infreactor.org	1.36%	trendytopic.info	1.39%
cbslocal.com	1.35%	minibird.jp	1.25%

Table 6: Top 20 domains included in tweets from Russian trolls and baselines users.

Domain (Trolls)	(%)	Domain (Baseline)	(%)
twitter.com	12.81%	twitter.com	35.51%
reportsecret.com	7.02%	youtube.com	4.21%
riafan.ru	3.42%	vine.co	3.94%
politexpert.net	2.10%	factissues.com	3.24%
youtube.com	1.88%	blogspot.com.cy	1.92%
vk.com	1.58%	instagram.com	1.90%
instagram.com	1.53%	facebook.com	1.68%
yandex.ru	1.50%	worldstarr.info	1.47%
infreactor.org	1.36%	trendytopic.info	1.39%
cbslocal.com	1.35%	minibird.jp	1.25%
livejournal	1.35%	yaadlinksradio.com	1.24%
nevnov.ru	1.07%	soundcloud.com	1.24%
ksnt.com	1.01%	linklist.me	1.15%
kron4.com	0.93%	twimg.com	1.09%
viid.me	0.93%	appparse.com	1.08%
newinform.com	0.89%	cargobayy.net	0.88%
inforeactor.ru	0.84%	virralclub.com	0.84%
rt.com	0.81%	tistory.com	0.50%
washigntonpost.com	0.75%	twitcasting.tv	0.49%
seattletimes.com	0.73%	nytimes.com	0.48%

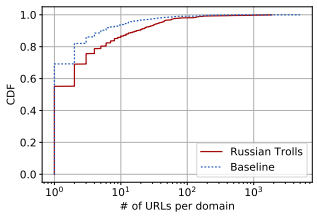
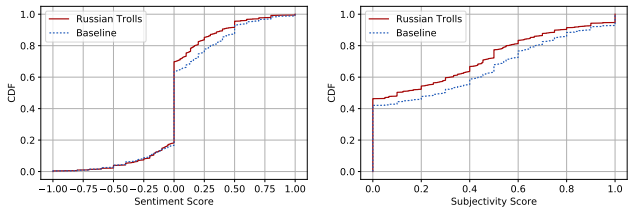


Figure 5: CDF of number of URLs per domain.



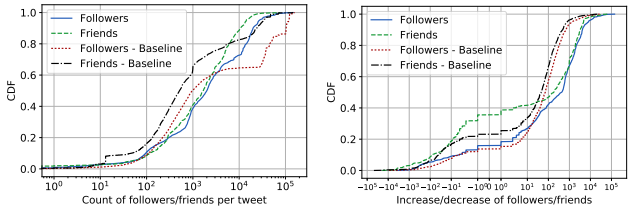
(a) (b)

Figure 6: CDF of sentiment and subjectivity scores for tweets of Russian trolls and random users.

Sentiment analysis. Next, we assess the sentiment and subjectivity of each tweet for both datasets. Fig. 6(a) plots the CDF of the sentiment scores of tweets posted by Russian trolls and our baseline users. We observe that 30% of the tweets from Russian trolls have a positive sentiment, and 18% negative. These scores are not too distant from those of random users where 36% are positive and 16% negative, however, Russian trolls exhibit a unique behavior in terms of sentiment, as a two-sample Kolmogorov-Smirnov test unveils significant differences between the distributions ($p < 0.01$). Overall, we observe that Russian trolls tend to be more negative/neutral, while our baseline is more positive. We also compare subjectivity scores (Fig. 6(b)), finding that tweets from trolls tend to be more subjective; again, we perform significance tests revealing differences between the two distributions ($p < 0.01$).

LDA analysis. We also use the Latent Dirichlet Allocation (LDA) model to analyze tweets' semantics. We train an LDA model for each of the datasets and extract 10 distinct topics with 10 words, as reported in Table 7. Overall, topics from Russian trolls refer to specific world events (e.g., Charlottesville) as well as specific news related to politics (e.g., North Korea and Donald Trump). By contrast, topics extracted from the random sample are more general (e.g., tweets regarding birthdays).

Screen name changes. Previous work [17] has shown that malicious accounts often change their screen name in order to assume different identifies. Therefore, we investigate whether trolls show a similar behavior, as they might change the narrative with which they are attempting to influence public opinion. Indeed, we find that 9% of the accounts operated by trolls change their screen name, up to 4 times during the course of our dataset. Some examples include changing screen names from "OnlineHouston" to "HoustonTopNews," or "Jesus Quintin Perez" to "WorldNewsPolitics," in a clear attempt to pose as news-related accounts. In our baseline, we find that 19% of the accounts changed their Twitter screen names, up



(a) (b)

Figure 7: CDF of the number of (a) followers/friends for each tweet and (b) increase in followers/friends for each user from the first to the last tweet.

to 11 times during our dataset; highlighting that changing screen names is a common behavior of Twitter users in general.

Followers/Friends. Next, we look at the number of followers and friends (i.e., the accounts one follows) of the Russian trolls, as this is an indication of the overall impact of a tweet. In Fig. 7(a), we plot the CDF of the number of followers per tweet measured at the time of that tweet. On average, Russian trolls have 7K followers and 3K friends, while our baseline has 25K followers and 6K friends. We also note that in both samples, tweets reached a large number of Twitter users; at least 1K followers, with peaks up to 145K followers. These results highlight that Russian trolls have a non-negligible number of followers, which can assist in pushing specific narratives to a much greater number of Twitter users. We also assess the evolution of the Russian trolls in terms of the number of their followers and friends. To this end, we get the follower and friend count for each user on their first and last tweet and calculate the difference. Fig. 7(b) plots the CDF of the increase/decrease of the followers and friends for each troll as well as random user in our baseline. We observe that, on average, Russian trolls increase their number of followers and friends by 2,065 and 1,225, respectively, whereas for the baseline we observe an increase of 425 and 133 for followers and friends, respectively. This suggests that Russian trolls work hard to increase their reachability within Twitter.

Tweet Deletion. Arguably, a reasonable strategy to avoid detection after posting tweets that aim to manipulate other users might be to delete them. This is particularly useful when troll accounts change their identity and need to modify the narrative that they use to influence public opinion. With each tweet, the Streaming API returns the total number of available tweets a user has up to that time. Retrieving this count allows us to observe if a user has deleted a tweet, and around what period; we call this an "observed deletion." Recall that our dataset is based on the 1% sample of Twitter, thus, we can only estimate, in a conservative way, how many tweets are deleted; specifically, in between subsequent tweets, a user may have deleted and posted tweets that we do not observe. In Fig. 8, we plot the CDF of the number of deleted tweets per observed deletion. We observe that 13% of the Russian trolls delete some of their tweets, with a median percentage of tweet deletion equal to 9.7%. Whereas, for the baseline set, 27% of the accounts delete at least one tweet, but the median percentage is 0.1%. This means that the trolls delete their tweets in batches, possibly trying to cover their tracks or get a clean slate, while random users make a larger number of deletions but only a small percentage of their overall tweets, possibly

Table 7: Terms extracted from LDA topics of tweets from Russian trolls and baseline users.

Topic	Terms (Trolls)	Topic	Terms (Baseline)
1	trump, black, people, really, one, enlist, truth, work, can, get	1	want, can, just, follow, now, get, see, don, love, will
2	trump, year, old, just, run, obama, breaking, will, news, police	2	2016, july, come, https, trump, social, just, media, jabberduck, get
3	new, trump, just, breaking, obamacare, one, sessions, senate, politics, york	3	happy, best, make, birthday, video, days, come, back, still, little
4	man, police, news, killed, shot, shooting, woman, dead, breaking, death	4	know, never, get, love, just, night, one, give, time, can
5	trump, media, tcot, just, pjnet, war, like, video, post, hillary	5	just, can, everyone, think, get, white, fifth, veranomtv2016, harmony, friends
6	sports, video, game, music, isis, charlottesville, will, new, health, amb	6	good, like, people, lol, don, just, look, today, said, keep
7	can, don, people, want, know, see, black, get, just, like	7	summer, seconds, team, people, miss, don, will, photo, veranomtv2016, new
8	trump, clinton, politics, hillary, video, white, donald, president, house, calls	8	like, twitter, https, first, can, get, music, better, wait, really
9	news, world, money, business, new, one, says, state, 2016, peace	9	dallas, right, fuck, vote, police, via, just, killed, teenchoice, aldubmainecelebration
10	now, trump, north, korea, people, right, will, check, just, playing	10	day, black, love, thank, great, new, now, matter, can, much

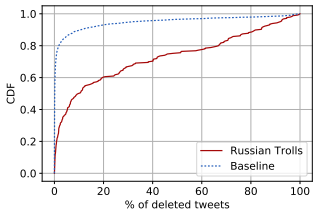


Figure 8: CDF of the number of deleted tweets per observe deletion.

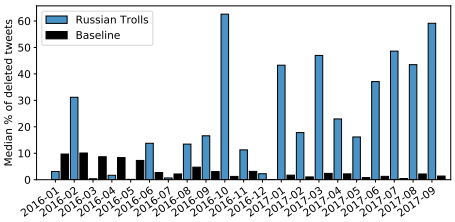


Figure 9: Average percentage of observed deletions per month.

because of typos. We also report the distribution, over each month, of tweet deletions in Fig. 9. Specifically, we report the mean of the percentages for all observed deletions in our datasets. Most of the tweets from Russian trolls are deleted in October 2016, suggesting that these accounts attempted to get a clean slate a few months before the 2016 US elections.

Case Study. While the previous results provide a quantitative characterization of Russian trolls behavior, we believe there is value showing a concrete example of the behavior exhibited and how techniques played out. We start on May 15, 2016, where the troll with screen name 'Pen_Air', was posing as a news account via its profile description: "National American news." On September 8, 2016 as the US presidential elections approached, 'Pen_Air' became a Trump supporter, changing its screen name to 'Blacks4DTrump' with a profile description of "African-Americans stand with Trump to make America Great Again!" Over the next 11 months, the account's tweet count grew from 49 to 642 while its follower count rose from 1.2K to *nearly* 9K. Then, around August 18, 2017, the account was seemingly repurposed. Almost all of its previous tweets were deleted (the account's tweet count dropped to 35), it gained a new screen name ('southlonestar2'), and was now posing as a "Proud American and TEXAN patriot! Stop ISLAM and PC. Don't mess with Texas" according to its profile description. When examining the accounts tweets, we see that most are clearly related

Table 8: Total URLs with at least one event in Twitter, /pol/, Reddit, and Russian trolls on Twitter; total events for Russian state-sponsored news URLs, other news URLs and all the URLs; and mean background rate (λ_0) for each platform.

		/pol/	Reddit	Twitter	Trolls
URLs	Russian state-sponsored	6	13	19	19
	Other news sources	47	168	159	192
	All	127	482	861	989
Events	Russian state-sponsored	19	42	118	19
	Other news sources	720	3,055	2,930	195
	All	1,685	9,531	1,537,612	1,461
Mean λ_0	Russian state-sponsored	0.0824	0.1865	0.2264	0.1228
	Other news sources	0.0421	0.1447	0.1544	0.0663
	All	0.0324	0.1557	0.1553	0.0753

to politics, featuring blunt right-wing attacks and "talking points." For example, "Mr. Obama! Maybe you bring your girls and leave them in the bathroom with a grown man! #bathroombill #NObama <url>" on May 15, 2016, "#HiLIARy has only two faces! And I hate both! #NeverHillary #Hillaryliesmatter <url>" on May 19, 2016, and "RT @TEN_GOP: WikiLeaks #DNCLeaks confirms something we all know: system is totally rigged! #NeverHillary <url>." on July 22, 2016.

Take-aways. In summary, our analysis leads to the following observations. First, we find evidence that trolls were actively involved in the dissemination of content related to world news and politics, as well as propaganda content regarding various topics such as ISIS and Islam. Moreover, several Russian trolls were created or repurposed a few weeks before notable world events, including the Republican National Convention meeting or the Charlottesville rally. We also find that the trolls deleted a substantial amount of tweets in batches and overall made substantial changes to their accounts during the course of their lifespan. Specifically, they changed their screen names aiming to pose as news outlets, experienced significant rises in the numbers of followers and friends, etc. Furthermore, our location analysis shows that Russian trolls might have tried to manipulate users located in the USA, Germany, and possibly in their own country (i.e., Russia), by appearing to be located in those countries. Finally, the fact that these accounts were active up until their recent suspension also highlights the need to develop more effective tools to detect such actors.

4 INFLUENCE ESTIMATION

Thus far, we have analyzed the behavior of the Russian trolls on the Twitter platform, and how this differs from that of a baseline of random users. Allegedly, their main goal is to ultimately manipulate the opinion of other users and extend the cascade of disinformation

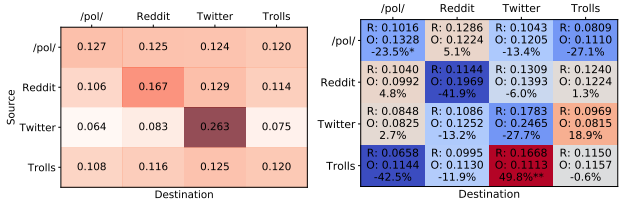
they share (e.g., other users post similar content) [4]. Therefore, we now set out to shed light on their impact, in terms of the dissemination of disinformation, on Twitter and on the greater Web.

To assess their influence, we look at the URLs posted by four groups of users: Russian trolls on Twitter, "normal" accounts on Twitter, Reddit users, and 4chan users (/pol/ board). For each unique URL, we fit a statistical model known as Hawkes Processes [15, 16], which allows us to estimate the strength of connections between each of these four groups in terms of how likely an event – the URL being posted by either trolls or normal users to a particular platform – is to cause subsequent events in each of the groups. For example, a strong connection from Reddit to /pol/ would mean that a URL that appears on Reddit is likely to be seen and then re-posted on /pol/; whereas, a weak connection from trolls to normal users on Twitter indicates that a URL posted by trolls is less likely to be re-tweeted or re-posted by the latter. We fit the Hawkes Processes using the methodology presented by [29].

To study the dissemination of different types of content, we look at three different sets of URLs: 1) The complete set of all URLs posted by Russian trolls; 2) The subset of URLs for Russian state-sponsored news, namely, RT (Russia Today); and 3) The subset of URLs from other mainstream and alternative news sources using the list provided by [29]. Table 8 summarizes the number of URLs, number of events (i.e., occurrences of a given URL) as well as the mean background rate for each category and social network. The background rate defines the rate at which events occur excluding the influence of the platforms included in the model; the background rate includes events created spontaneously on each platform, such as by a user sharing the article from the original source, or those generated by another platform not monitored by us like Facebook. The number of events for Russian state-sponsored news sources is substantially lower than the number of events from other news sources. This is expected since the former only includes one news source (RT), however, it is interesting that the background rates for these URLs are higher than for other news sources, meaning that events from Russian state-sponsored news are more likely to occur spontaneously.

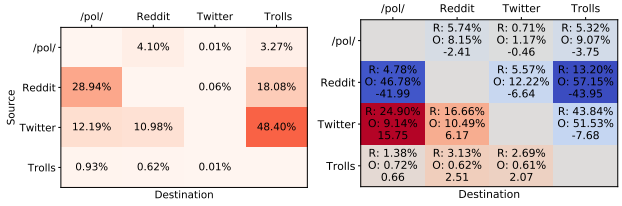
Fitting a Hawkes model yields a weight matrix, which characterizes the strength of the connections between the groups we study. Each weight value represents the connection strength from one group to another and can be interpreted as the expected number of subsequent events that will occur on the second group after each event on the first. The mean weight values over all URLs, as well as for the URLs from RT and other news URLs, are presented in Fig. 10. We observe that for /pol/, Reddit, and normal users on Twitter, the greatest weights are from each group to itself, meaning that reposts/retweets on the same site are more common than sharing the URL to the other platforms (Fig. 10(a)). For the Russian Trolls on Twitter, however, the weight is greater from the trolls to Twitter than from the trolls to themselves, perhaps reflecting their use as an avenue for *disseminating* information to normal Twitter users (Fig. 10(b)). Also, we observe that, in most cases, the connections are stronger for non-Russia state-sponsored news, indicating that regular users are more inclined to share news articles from mainstream and alternative news sources.

Looking at the Russian trolls and normal Twitter users, we see that the trolls are more likely to retweet Russian state-sponsored



(a) All URLs (b) News URLs

Figure 10: Mean weights for (a) all URLs in our dataset and (b) news URLs categorized as Russian state-sponsored (R) and other mainstream and alternative news URLs (O). We also show the percent of increase/decrease between the two categories. Note that * and ** refer to statistical significance with, resp., $p < 0.05$ and $p < 0.01$.



(a) All URLs (b) News URLs

Figure 11: Estimated mean percentages of events created because of other events for (a) all URLs and (b) Russian state-sponsored URLs (R) and other mainstream and alternative news URLs (O). We also show the difference between the two categories of news.

URLs from normal Twitter users than other news sources; conversely, normal Twitter users are more likely to retweet Russian state-sponsored URLs from the troll accounts. In order to assess the significance of our results, we perform two-sample Kolmogorov-Smirnov tests on the weight distributions for the RT URLs and the other news URLs for each source-destination platform pair (depicted as stars in the Fig. 10(b)). Small p value means there is a statistically significant difference in the way that RT URLs propagate from the source to the destination platform. Most of the source-destination pairs have no statistical significance, however for the Russian trolls–Twitter users pair, we find significance difference with $p < 0.01$.

In Fig. 11, we report the estimated total impact for each pair of platforms, for both Russian state-sponsored news, other news sources as well as all the observed URLs. We determine the impact by calculating, based on the estimated weights and the number of events, the percentage of events on a destination platform caused by events on a source platform, following the same methodology as [29]. For all URLs (Fig. 11(a)), we find that the influence of Russian trolls is negligible on Twitter (0.01%), while for /pol/ and Reddit it is slightly higher (0.93% and 0.62%, respectively). For other pairs, the larger impacts are between Reddit–/pol/ and Twitter-Russian trolls, mainly due to the larger population of users. Looking at the estimated impact for RT and other news sources (Fig. 11(b)), we note that the trolls influenced the other platforms approximately the same for alternative and mainstream news sources (0.72%, 0.62%, and 0.61 for /pol/, Reddit, and Twitter, respectively). Interestingly,

Russian trolls have a much larger impact on all the other platforms for the RT news when compared to the other news sources: approximately 2 times more on /pol/, 5 times more on Reddit, and 4 times more on Twitter.

Take-aways. Using Hawkes processes, we were able to assess the degree of influence Russian trolls had on Twitter, Reddit, and /pol/ by examining the diffusion of information via URLs to news. Our results indicate that their influence is actually quite limited. With the exception of news originating from the Russian state-sponsored news outlet RT, the troll accounts were generally less influential than other users on Reddit, Twitter, and 4chan. However, our analysis is based only on 1K troll accounts found in Twitter's 1% stream, and, as mentioned previously, Twitter recently announced they had discovered over 1K more trolls and more than 50K automated accounts. With that in mind, there are several potential explanations behind this limited influence. For example, there might be a lot of influence attributed to regular Twitter users that belongs to newly announced troll accounts. Considering that Twitter announced the discovery of "only" 1K more troll accounts, we suspect that this is not really the case. Another, more plausible explanation is that the troll accounts are just not terribly efficient at spreading news, and instead are more concerned with causing havoc by pushing ideas, engaging other users, or even taking both sides of controversial online discussions [22]. This scenario makes more sense considering that, along with 1K new troll accounts, Twitter also announced discovering over 50K *automated* accounts that might be more efficient in terms of driving traffic to specific URLs.

5 RELATED WORK

Opinion manipulation. The practice of swaying opinion on the Web is a long-standing issue as malicious actors attempt to push their agenda. Kumar et al. [13] study how users create multiple accounts, called *sockpuppets*, that participate in Web communities to manipulate users' opinions. They find that sockpuppets exhibit different posting behavior when compared to benign users. Mihaylov et al. [18] show that trolls can manipulate users' opinions in forums, while in their follow-up work [19] they highlight the two types of manipulation trolls: those paid to operate and those that are called out as such by other users. Then, Volkova and Bell [26] predict the deletion of Twitter accounts because they are trolls, focusing on those that shared content related to the Russian-Ukraine crisis.Elyashar et al. [6] distinguish authentic discussions from campaigns to manipulate the public's opinion, using a set of similarity functions alongside historical data. Also, Steward et al. [22] focus on discussions related to the Black Lives Matter movement and how content from Russian trolls was retweeted by other users. Using the retweet network, they find the existence of two communities; one left- and one right-leaning communities. Also, they note that trolls infiltrated both communities, setting out to push specific narratives. Finally, Varol et al. [25] aim to identify memes that become popular due to *coordinated* efforts, and achieve 75% AUC score before memes become trending and 95% AUC score afterwards.

False information on the political stage. Conover et al. [2] study the interactions of right- and left-leaning communities on Twitter during the 2010 US midterm elections: finding that the retweet network has limited connectivity between the two communities, which does not happen in the mentions network; mainly because users engage others users with different ideologies and expose them to different opinions. Ratkiewicz et al. [20] use machine learning to detect the early stages of false political information spreading on Twitter and introduce a framework that considers topological, content-based, and crowdsourced features of the information diffusion. Wong et al. [27] quantify political leaning of users and news outlets during the 2012 US presidential election on Twitter by using an inference engine that considers tweeting and retweeting behavior of articles. Yang et al. [28] investigate the topics of discussions on Twitter for 51 US political persons showing that Democrats and Republicans are active in a similar way on Twitter. Le et al. [14] study 50M tweets regarding the 2016 US election primaries and highlight the importance of three factors in political discussions on social media, namely the *party*, *policy considerations*, and *personality* of the candidates. Howard and Kollanyi [12] study the role of bots in Twitter conversations during the 2016 Brexit referendum. By analyzing 1.5M tweets, they find that most tweets are in favor of Brexit and that there are bots with various levels of automation.Also, Hegelich and Janetzko [10] highlight that bots have a political agenda and that they exhibit various behaviors, e.g., trying to hide their identity and promoting topics through hashtags and retweets. Finally, a large body of work focuses on social bots [1, 3, 7, 8, 24] and their role in spreading disinformation, highlighting that they can manipulate the public's opinion at large scale, thus potentially affecting the outcome of political elections.

Remarks. Unlike previous work, our study focuses on the set of Russian troll accounts that were suspended by Twitter and released by the US congress. To the best of our knowledge, this constitutes the first effort not only to characterize a ground truth of troll accounts independently identified by Twitter, but also to quantify their influence on the greater Web, specifically, on Twitter as well as on Reddit and 4chan.

6 CONCLUSION

In this paper, we analyzed the behavior and use of the Twitter platform by Russian trolls during the course of 21 months. We showed that Russian trolls exhibited interesting differences when compared with a set of random users, actively disseminated politics-related content, adopted multiple identities during their account's lifespan, and that they aimed to increase their impact on Twitter by increasing their followers. Also, we quantified the influence that Russian trolls have on Twitter, Reddit, and /pol/ using a statistical model known as Hawkes Processes. Our findings show that trolls' influence was not substantial with respect to the other platforms, with the significant exception of news published by the Russian state-sponsored news outlet RT.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska Curie ENCASE project (GA No. 691025) and under the CYBERSECURITY CONCORDIA project (GA No. 830927).

REFERENCES

[1] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016).

[2] Michael Conover, Jacob Ratkiewicz, Matthew R. Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *ICWSM*.

[3] Clayton A. Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot: A System to Evaluate Social Bots. In *WWW*.

[4] Samuel Earle. 2017. Trolls, Bots and Fake News: The Mysterious World of Social Media Manipulation. https://goo.gl/nz7E8r.

[5] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2017. Towards detecting compromised accounts on social networks. *IEEE TDSC* (2017).

[6] Aviad Elyashar, Jorge Bendahan, and Rami Puzis. 2017. Is the Online Discussion Manipulated? Quantifying the Online Discussion Authenticity within Online Social Media. *CoRR* abs/1708.02763 (2017).

[7] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *ArXiv 1707.00086* (2017).

[8] Emilio Ferrara, Onur Varol, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* (2016).

[9] Aditi Gupta and Ponnurangam Kumaraguru. 2012. Credibility ranking of tweets during high impact events. In *PSOSM '12*.

[10] Simon Hegelich and Dietmar Janetzko. 2016. Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet. In *ICWSM*.

[11] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*.

[12] Philip N. Howard and Bence Kollanyi. 2016. Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. *Arxiv 1606.06356* (2016).

[13] Srijan Kumar, Justin Cheng, Jure Leskovec, and V. S. Subrahmanian. 2017. An Army of Me: Sockpuppets in Online Discussion Communities. In *WWW*.

[14] Huyen T. Le, G. R. Boynton, Yelena Mejova, Zubair Shafiq, and Padmini Srinivasan. 2017. Revisiting The American Voter on Twitter. In *CHI*.

[15] Scott W. Linderman and Ryan P. Adams. 2014. Discovering Latent Network Structure in Point Process Data. In *ICML*.

[16] S. W. Linderman and R. P. Adams. 2015. Scalable Bayesian Inference for Excitatory Point Process Networks. *ArXiv 1507.03228* (2015).

[17] Enrico Mariconti, Jeremiah Onaolapo, Syed Sharique Ahmad, Nicolas Nikiforou, Manuel Egele, Nick Nikiforakis, and Gianluca Stringhini. 2017. What's in a Name?: Understanding Profile Name Reuse on Twitter. In *WWW*.

[18] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015. Finding Opinion Manipulation Trolls in News Community Forums. In *CoNLL*.

[19] Todor Mihaylov and Preslav Nakov. 2016. Hunting for Troll Comments in News Community Forums. In *ACL*.

[20] Jacob Ratkiewicz, Michael Conover, Mark R. Meiss, Bruno GonÃğalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and Tracking Political Abuse in Social Media. In *ICWSM*.

[21] Kate Starbird. 2017. Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter. In *ICWSM*.

[22] Leo Steward, Ahmer Arif, and Kate Starbird. 2018. Examining Trolls and Polarization with a Retweet Network. In *MIS2*.

[23] The Independent. 2017. St Petersburg 'troll farm' had 90 dedicated staff working to influence US election campaign. https://ind.pn/2yuCQdy.

[24] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *ICWSM*.

[25] Onur Varol, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2017. Early detection of promoted campaigns on social media. *EPJ Data Science* (2017).

[26] Svitlana Volkova and Eric Bell. 2016. Account Deletion Prediction on RuNet: A Case Study of Suspicious Twitter Accounts Active During the Russian-Ukrainian Crisis. In *NAACL-HLT*.

[27] Felix Ming Fai Wong, Chee-Wei Tan, Soumya Sen, and Mung Chiang. 2013. Quantifying Political Leaning from Tweets and Retweets. In *ICWSM*.

[28] Xinxin Yang, Bo-Chiuan Chen, Mrinmoy Maity, and Emilio Ferrara. 2016. Social Politics: Agenda Setting and Political Communication on Social Media. In *SocInfo*.

[29] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *ACM IMC*.