Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls

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

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

William Setzer
University of Alabama at Birmingham
wjsetzer@uab.edu

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

Recent evidence has emerged linking coordinated campaigns by state-sponsored actors to manipulate public opinion on the Web. Campaigns revolving around major political events are enacted via mission-focused "trolls." While trolls are involved in spreading disinformation on social media, there is little understanding of how they operate, what type of content they disseminate, how their strategies evolve over time, and how they influence the Web's information ecosystem. In this paper, we begin to address this gap by analyzing 10M posts by 5.5K Twitter and Reddit users identified as Russian and Iranian state-sponsored trolls. We compare the behavior of each group of state-sponsored trolls with a focus on how their strategies change over time, the different campaigns they embark on, and differences between the trolls operated by Russia and Iran. Among other things, we find: 1) that Russian trolls were pro-Trump while Iranian trolls were anti-Trump; 2) evidence that campaigns undertaken by such actors are influenced by real-world events; and 3) that the behavior of such actors is not consistent over time, hence detection is not straightforward. Using Hawkes Processes, we quantify the influence these accounts have on pushing URLs on four platforms: Twitter, Reddit, 4chan's Politically Incorrect board (/pol/), and Gab. In general, Russian trolls were more influential and efficient in pushing URLs to all the other platforms with the exception of /pol/ where Iranians were more influential. Finally, we release our source code to ensure the reproducibility of our results and to encourage other researchers to work on understanding other emerging kinds of state-sponsored troll accounts on Twitter.

1 Introduction

Recent political events and elections have been increasingly accompanied by reports of disinformation campaigns attributed to state-sponsored actors [14]. 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].

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 [14, 19, 35]. However, automated content diffusion is only a part of the issue. In fact, recent work has shown that human actors are actually key in spreading false information on Twitter [42]. Overall, many aspects of state-sponsored disinformation remain unclear, e.g., how do state-sponsored trolls operate? What kind of content do they disseminate? How does their behavior change over time? And, 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 two different sources of ground truth data about state-sponsored actors. First, we use 10M tweets posted by Russian and Iranian trolls between 2012 and 2018 [18]. Second, we use a list of 944 Russian trolls, identified by Reddit, and find all their posts between 2015 and 2018 [36]. We analyze the two datasets across several axes in order to understand their behavior and how it changes over time, their targets, and the content they shared. For the latter, we leverage word embeddings to understand in what context specific words/hashtags are used and shed light to the ideology of the trolls. Also, we use Hawkes Processes [29] to model the influence that the Russian and Iranian trolls had over multiple Web communities; namely, Twitter, Reddit, 4chan's Politically Incorrect board (/pol/) [20], and Gab [52].

Main findings. Our study leads to several key observations:

(1) Our influence estimation results reveal that Russian trolls were extremely influential and efficient in spreading URLs on Twitter. Also, by comparing Russian to Iranian trolls, we find that Russian trolls were more efficient and influential in spreading URLs on Twitter, Reddit, Gab, but not on /pol/.

(2) By leveraging word embeddings, we find ideological differences between Russian and Iranian trolls (e.g., Russian trolls were pro-Trump, while Iranian trolls were anti-Trump).

(3) We find evidence that the Iranian campaigns were motivated by real-world events. Specifically, campaigns against France

and Saudi Arabia coincided with real-world events that affect the relations between these countries and Iran.

(4) We observe that the behavior of trolls varies over time. We find substantial changes in the use of language and Twitter clients over time for both Russian and Iranian trolls. These insights allow us to understand the targets of the orchestrated campaigns for each type of trolls over time.

(5) We find that the topics of discussion vary across Web communities. For example, we find that Russian trolls on Reddit were discussing about cryptocurrencies, while this does not apply in great extent for the Russian trolls on Twitter.

Finally, we make our source code publicly available [41] for reproducibility purposes and to encourage researchers to further work on understanding other types of state-sponsored trolls on Twitter (i.e., on January 31, 2019, Twitter released data related to trolls originating from Venezuela and Bangladesh [38]).

2 Related Work

Opinion manipulation. The practice of swaying opinion in Web communities has become a hot-button issue as malicious actors are intensifying their efforts to push their subversive agenda. Kumar et al. [27] study how users create multiple accounts, called *sockpuppets*, that actively participate in some communities with the goal to manipulate users' opinions. Mihaylov et al. [31] show that trolls can indeed manipulate users' opinions in online forums. In follow-up work, Mihaylov and Nakov [32] highlight two types of trolls: those paid to operate and those that are called out as such by other users. Then, Volkova and Bell [47] aim to predict the deletion of Twitter accounts because they are trolls, focusing on those that shared content related to the Russia-Ukraine crisis . Elyashar et al. [13] distinguish authentic discussions from campaigns to manipulate the public's opinion. Also, Steward et al. [43] focus on discussions related to the Black Lives Matter movement and how content from Russian trolls was retweeted by other users. Using community detection techniques, they unveil that Russian trolls infiltrated both left and right leaning communities, setting out to push specific narratives. Finally, Varol et al. [46] aim to identify memes (ideas) that become popular due to *coordinated* efforts.

False information on the political stage. Conover et al. [7] focus on Twitter activity during the 2010 US midterm elections and the interactions between right and left leaning communities. Ratkiewicz et al. [35] study political campaigns using controlled accounts to disseminate support for an individual or opinion. They use machine learning to detect the early stages of false political information spreading on Twitter. Wong et al. [50] aim to quantify the political leanings of users and news outlets during the 2012 US election on Twitter by considering tweeting and retweeting behavior of articles. Yang et al. [51] 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. [28] study 50M tweets related to the 2016 US election primaries and note the importance of three factors in political discussions on social media, namely the *party*, *policy considerations*, and *personality* of the candidates. Howard and Kollanyi [21] study the role of bots in Twitter conversations during the 2016 Brexit referendum. They find that most tweets are in favor of Brexit and that there are bots

Platform	Origin of trolls	# trolls	# trolls with tweets/posts	# of tweets/posts
Twitter	Russia	3,836	3,667	9,041,308
	Iran	770	660	1,122,936
Reddit	Russia	944	335	21,321

Table 1: Overview of Russian and Iranian trolls on Twitter and Reddit. We report the overall number of identified trolls, the trolls that had at least one tweet/post, and the overall number of tweets/posts.

with various levels of automation.Also, Hegelich and Janetzko [19] study whether bots on Twitter are used as political actors. By analyzing 1.7K bots on Twitter during the Russia-Ukraine conflict, they uncover their political agenda and show that bots exhibit various behaviors like trying to hide their identity and promoting topics through the use of hashtags.Badawy et al. [1] predict users that are likely to spread information from state-sponsored actors, while Dutt et al. [11] focus on the Facebook platform and analyze ads shared by Russian trolls in order to find the cues that make them effective. Finally, a large body of work focuses on social bots [3, 8, 14, 15, 45] and their role in spreading political disinformation, highlighting that they can manipulate the public's opinion at a large scale.

Remarks. In contrast, our study focuses on a set of Russian and Iranian trolls that were independently identified and suspended by Twitter and Reddit. To the best of our knowledge, this constitutes the first effort not only to characterize a ground truth of troll accounts, but also to quantify their influence on the greater Web.

3 Troll Datasets

Twitter. On October 17, 2018, Twitter released a large dataset of Russian and Iranian troll accounts [18]. Although the exact methodology used to determine that these accounts were state-sponsored trolls is unknown, based on the most recent Department of Justice indictment [9], the dataset appears to have been constructed in a manner that we can assume essentially no false positives, while we cannot make any postulation about false negatives. Table 1 summarizes the troll dataset.

Reddit. On April 10, 2018, Reddit released a list of 944 accounts which they determined were Russian state-sponsored trolls [36]. We recover the submissions, comments, and account details for these accounts using two mechanisms: 1) Reddit dumps provided by Pushshift [33]; and 2) crawling the user pages of those accounts. Although omitted for lack of space, we note that the union of these two datasets reveals some gaps in both, likely due to a combination of subreddit moderators removing posts or the troll users themselves deleting them, which would affect the two datasets in different ways. In any case, we merge the two datasets, with Table 1 describing the final dataset. Note that only 335 of the accounts released by Reddit had at least one submission or comment in our dataset. We suspect the rest were simply used as dedicated voting accounts used in an effort to push (or bury) specific content.

Ethics. We only work with anonymized public data and we follow standard ethical guidelines [37].

4 Analysis

4.1 Accounts Characteristics

Account Creation. Fig. 1 plots the Russian and Iranian troll accounts creation dates on Twitter and Reddit. We observe that the majority of Russian troll accounts were created around the time of

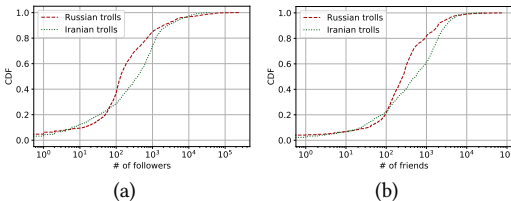Figure 1: Number of Russian and Iranian troll accounts created per week.

Figure 2: CDF of the number of a) followers and b) friends for the Russian and Iranian trolls on Twitter.

the Ukrainian conflict: 80% of them have an account creation date earlier than 2016. That said, there are some meaningful peaks in account creation during 2016 and 2017. 57 accounts were created between July 3-17, 2016, which was right before the Republican National Convention (July 18-21) where Donald Trump was named the Republican nominee for President [48] . Later, 190 accounts were created between July and August, 2017, during the run up to the infamous Unite the Right rally in Charlottesville [49]. Taken together, this might be evidence of coordinated activities aimed at manipulating users' opinions on Twitter with respect to specific events. This is further evidenced when examining the Russian trolls on Reddit: 75% of the accounts on Reddit were created in a single massive burst in the first half of 2015. Also, there are a few smaller spikes occurring just prior to the 2016 US election. For the Iranian trolls on Twitter we observe that they are much "younger," with the larger bursts of account creation *after* the 2016 US election.

Followers/Friends. Fig. 2 plots the CDF of the number of followers and friends for both Russian and Iranian trolls. 25% of Iranian trolls had more than 1k followers, while the same applies for only 15% of the Russian trolls. In general, Iranian trolls tend to have more followers than Russian trolls (median of 392 and 132, respectively). Both Russian and Iranian trolls tend to follow a large number of users, probably in an attempt to increase their follower count via reciprocal follows. Iranian trolls have a median followers to friends ratio of 0.51, while Russian trolls have a ratio of 0.74. This might indicate that Iranian trolls were more effective in acquiring followers without resorting in massive followings of other users, or perhaps that they used services that offer followers for sale [44].

4.2 Temporal Analysis

We next explore troll activity over time, looking for behavioral patterns. Fig. 3(a) plots the (normalized) volume of tweets/posts shared per week in our dataset. We observe that both Russian and Iranian trolls on Twitter became active during the Ukrainian conflict. Although lower in overall volume, there an increasing trend starts around August 2016 and continues through summer of 2017. We also see three major spikes in activity by Russian trolls on Reddit. The first is during the latter half of 2015, approximately around the time that Donald Trump announced his candidacy for

(a) Date

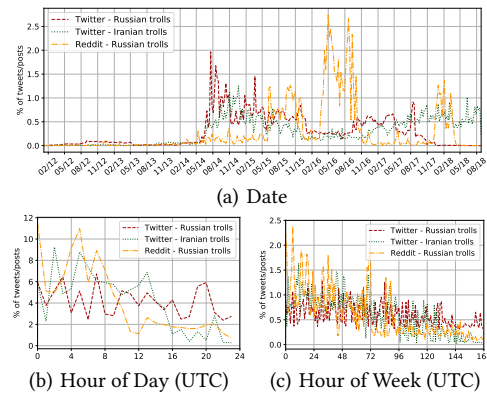(b) Hour of Day (UTC)　　(c) Hour of Week (UTC)

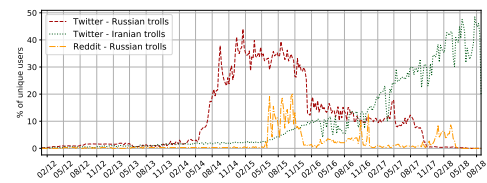Figure 3: Temporal characteristics of tweets from Russian and Iranian trolls.

Figure 4: Percentage of unique trolls that were active per week.
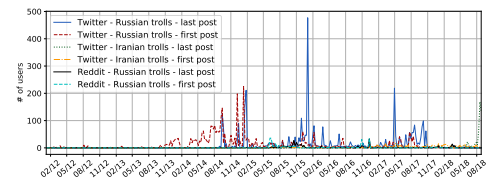

Figure 5: Number of trolls that posted their first/last tweet/post for each week in our dataset.

President. Next, we see solid activity through the middle of 2016, trailing off shortly before the election. Finally, we see another burst of activity in late 2017 through early 2018, at which point the trolls were detected and had their accounts locked by Reddit.

Furthermore, we examine the hour of day and week that the trolls post. Fig. 3(b) shows that Russian trolls on Twitter are active throughout the day, while on Reddit they are particularly active during the first hours of the day. Similarly, Iranian trolls on Twitter tend to be active from early morning until 13:00 UTC. When looking at the activity based on hour of the week (Fig. 3(c)), we find that Russian trolls on Twitter follow a diurnal pattern with slightly less activity during Sunday. In contrast, Russian trolls on Reddit and Iranian trolls on Twitter are particularly active during the first days of the week, while their activity decreases during the weekend. For Iranians this is likely due to the Iranian work week being from Sunday to Wednesday with a half day on Thursday.

But are *all* trolls in our dataset active throughout the span of our datasets? To answer this question, we plot the percentage of unique troll accounts that are active per week in Fig. 4 from which we draw the following observations. First, the Russian troll campaign on Twitter targeting Ukraine was much more diverse in terms of accounts when compared to later campaigns. There are several possible explanations for this. One explanation is that trolls learned from their Ukrainian campaign and became more efficient in later campaigns, perhaps relying on large networks of bots in their earlier
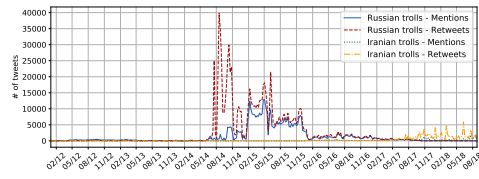
Figure 6: Number of tweets that contain mentions among Russian trolls and among Iranian trolls on Twitter.
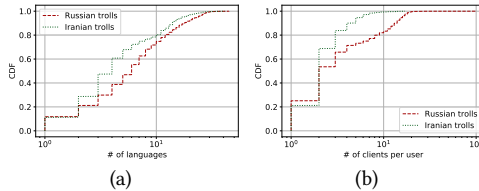

(a)   (b)

Figure 7: CDF of number of (a) languages used (b) clients used for Russian and Iranian trolls on Twitter.

campaigns which were later abandoned in favor of more focused campaigns like project Lakhta [10]. Another explanation could be that attacks on the US election might have required "better trained" trolls, perhaps those that could speak English more convincingly. The Iranians, on the other hand, seem to be slowly building their troll army over time. There is a steadily increasing number of active trolls posting per week over time. We speculate that this is due to their troll program coming online in a slow-but-steady manner, perhaps due to more effective training. Finally, on Reddit we see most Russian trolls posted irregularly, possibly performing other operations on the platform like manipulating votes on other posts.

Next, we investigate the point in time when each troll in our dataset made his first and last tweet. Fig. 5 shows the number of users that made their first/last post for each week in our dataset, which highlights when trolls became active as well as when they "retired." We see that Russian trolls on Twitter made their first posts during early 2014, almost certainly in response to the Ukrainian conflict. When looking at the last tweets of Russian trolls on Twitter we see that a substantial portion of the trolls "retired" by the end of 2015. This is likely because the Ukrainian conflict was over and Russia turned their attention to other targets (e.g., the USA, this is also aligned with the increase in the use of English language, see Section 4.3). When looking at Russian trolls on Reddit, we do not see a spike in first posts close to the time that the majority of the accounts were created (see Fig. 1). This indicates that the newly created Russian trolls on Reddit started posting gradually.

Finally, we assess whether Russian and Iranian trolls mention or retweet each other, and how this behavior occurs over time. Fig. 6 shows the number of tweets that were mentioning/retweeting other trolls' tweets. Russian trolls were particularly fond of this strategy during 2014 and 2015, while Iranian trolls started using this strategy after August, 2017. This again highlights how the strategies employed by trolls adapts and evolves to new campaigns.

4.3 Languages and Clients

Languages. First we study the languages used by trolls as it provides an indication of their targets. The language information is included in the datasets released by Twitter. Fig. 7(a) plots the CDF of the number of languages used by troll accounts. We find that
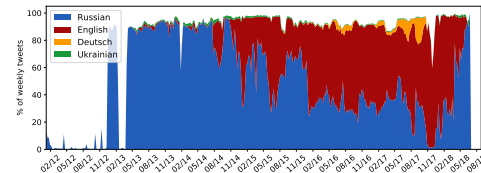
(a) Russians
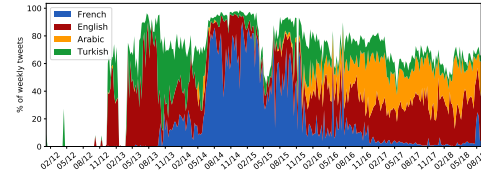
(b) Iranians
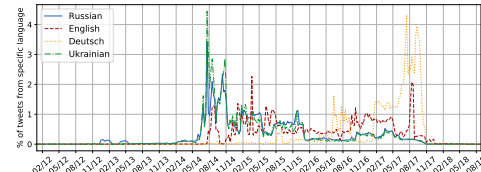
(c) Russians
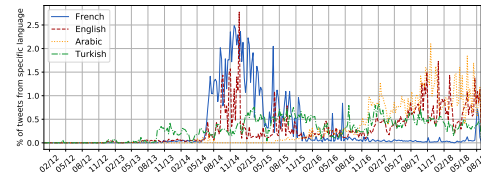
(d) Iranians

Figure 8: Use of the four most popular languages by Russian and Iranian trolls over time on Twitter. (a) and (b) show the percentage of weekly tweets in each language. (c) and (d) show the percentage of total tweets per language that occurred in a given week.

80% and 75% of the Russian and Iranian trolls, respectively, use more than 2 languages. Next, we note that in general, Iranian trolls tend to use fewer languages than Russian trolls. The most popular language for Russian trolls is Russian (53% of all tweets), followed by English (36%), German (1%), and Ukrainian (0.9%). For Iranian trolls we find that French is the most popular language (28% of tweets), followed by English (24%), Arabic (13%), and Turkish (8%).

Fig. 8 plots the use of different languages over time. Fig. 8(a) and Fig. 8(b) plot the percentage of tweets that were in a given language on a given week for Russian and Iranian trolls, respectively, in a stacked fashion, which lets us see how the usage of different languages changed over time relative to each other. Fig. 8(c) and Fig. 8(d) plot the language use from a different perspective: normalized to the overall number of tweets in a given language. This view gives us a better idea of how the use of each particular language changed over time. We make the following observations. First, there is a clear shift in targets based on the campaign. For example, Fig. 8(a) shows that the majority of early tweets by Russian trolls were in Russian, with English only reaching the volume of Russian language tweets in 2016. This coincides with the "retirement" of several Russian trolls on Twitter (see Fig 5). Next, we see evidence of other campaigns, for example German language tweets begin showing up in early to mid 2016, and reach their highest volume in the latter half of 2017, close to the 2017 German Federal

(a) Russians

(b) Iranians

Figure 9: Use of the eight most popular clients by Russian and Iranian trolls over time on Twitter.

elections. Additionally, we note that Russian language tweets have a huge drop off in activity the last two months of 2017.
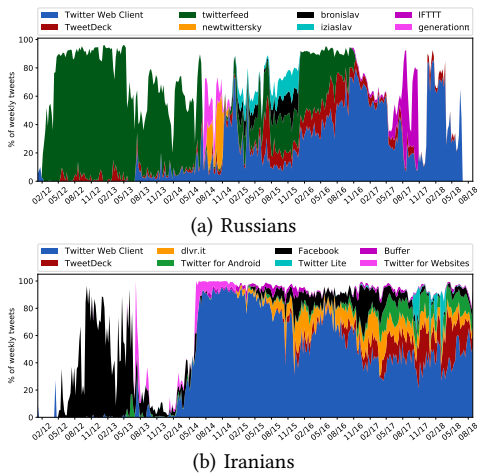
For the Iranians, we see more obvious evidence of multiple campaigns. For example, although Turkish and English are present for most of the timeline, French quickly becomes a commonly used language in the latter half of 2013, becoming the dominant language used from around May 2014 until the end of 2015. This is likely due to political events that happened during this time period. E.g., in November, 2013 France blocked a stopgap deal related to Iran's uranium enrichment program [26], leading to some fiery rhetoric from Iran's government. As tweets in French fall off, we also observe a dramatic increase in the use of Arabic in early 2016. This coincides with an attack on the Saudi embassy in Tehran [22], the primary reason the two countries ended diplomatic relations.

When looking at the language usage normalized by the total number of tweets in that language, we can get a more focused perspective. In particular, from Fig. 8(c) it becomes strikingly clear that the initial burst of Russian troll activity was targeted at Ukraine, with the majority of Ukrainian language tweets coinciding directly with the Crimean conflict [2]. From Fig. 8(d) we observe that English language tweets from Iranian trolls, while consistently present over time, have a relative peak corresponding with French language tweets, likely indicating an attempt to influence non-French speakers with respect to the campaign against French speakers.

Client usage. Finally, we analyze the clients used to post tweets. When looking at the most popular clients, we find that Russian and Iranian trolls use the main Twitter Web Client (28.5% for Russian trolls, and 62.2% for Iranian trolls). This is in contrast with what normal users use: using a random set of Twitter users, we find that mobile clients make up a large chunk of tweets (48%), followed by the TweetDeck dashboard (32%). We next look at how many different clients trolls use throughout our dataset: in Fig. 7(b), we plot the CDF of the number of clients used per user. 25% and 21% of the Russian and Iranian trolls, respectively, use only one client, while in general Russian trolls tend to use more clients than Iranians.

Fig. 9 plots the usage of clients over time in terms of weekly tweets by Russian and Iranian trolls. We observe that the Russians (Fig. 9(a)) started off with almost exclusive use of the "twitterfeed"
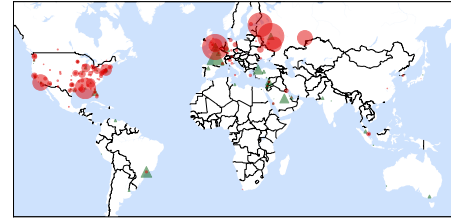

Figure 10: Distribution of reported locations for tweets by Russian (red circles) and Iranian trolls (green triangles) on Twitter.

client. Usage of this client drops off when it was shutdown in October, 2016. During the Ukrainian crisis, however, we see several new clients come into the mix. Iranians (Fig. 9(b)) started off almost exclusively using the "facebook" Twitter client, which automatically Tweets any posts you make on Facebook, indicating that Iranians likely started with a campaign on Facebook. At the beginning of 2014, we see a shift to using the Twitter Web Client, which only begins to decrease towards the end of 2015. Of particular note in Fig. 9(b) is the appearance of "dlvr.it," an automated social media manager, in the beginning of 2015. This corresponds with the creation of IUVM [24], which is a fabricated ecosystem of (fake) news outlets and social media accounts created by the Iranians, and might indicate that Iranian trolls stepped up their game around that time, starting using services that allowed them for better account orchestration to run their campaigns more effectively.

4.4 Geographical Analysis

We then study users' location, relying on the self-reported location field in their profiles, since only very few tweets have actual GPS coordinates. Note that the self-reported field is not required, and users are also able to change it whenever they like, so we look at locations for each tweet. We find that 16.8% and 20.9% of the tweets from Russian and Iranians trolls, respectively, do not include a self-reported location. To infer the geographical location from the self-reported text, we use pigeo [34], which provides geographical information (e.g., latitude, longitude, country, etc.) given the text that corresponds to a location. Specifically, we extract 626 self-reported locations for the Russian trolls and 201 locations for the Iranian trolls. Then, we use pigeo to obtain a geographical location (and its coordinates) for each text that corresponds to a location. Fig. 10 shows the locations inferred for Russian trolls (red circles) and Iranian trolls (green triangles). The size of the shapes on the map indicates the number of tweets that appear on each location. We observe that most of the tweets from Russian trolls come from Russia (34%), the USA (29%), and some from European countries, like United Kingdom (16%), Germany (0.8%), and Ukraine (0.6%). This suggests that Russian trolls may be pretending to be from certain countries, e.g., USA or United Kingdom, aiming to pose as locals and manipulate opinions. A similar pattern exists with Iranian trolls, which were active in France (26%), Brazil (9%), the USA (8%), Turkey (7%), and Saudi Arabia (7%). Also, Iranians trolls, unlike Russian trolls, did not report locations from their country, indicating that these trolls were primarily used for campaigns targeting foreign countries. Finally, we note that the location-based findings are in-line with the findings on the languages analysis (see Section 4.3), further evidencing that both Russian and Iranian trolls were specifically targeting different countries over time.
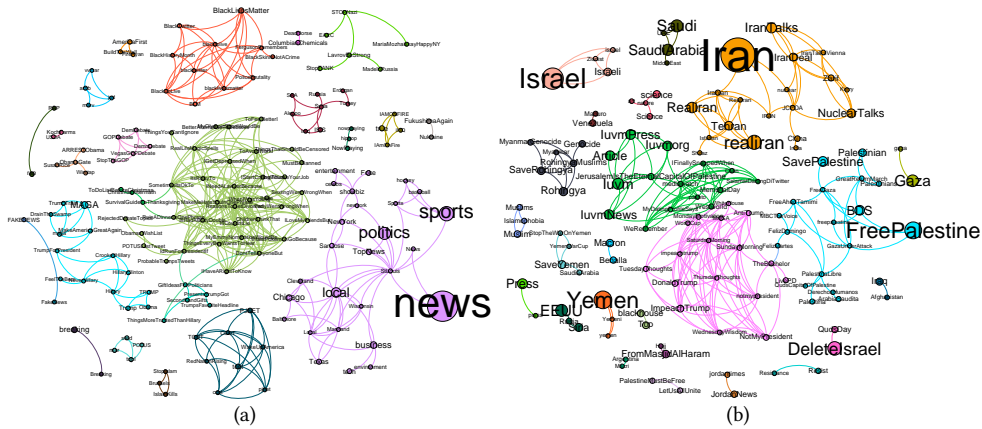
Figure 11: Visualization of the top hashtags used by a) Russian trolls and b) Iranian trolls on Twitter (see [40] and [39] for interactive versions).

Russian trolls on Twitter		Iranian trolls on Twitter	
Word	Cosine Similarity	Word	Cosine Similarity
trumparmi	0.68	impeachtrump	0.81
trumptrain	0.67	stoptrump	0.80
votetrump	0.65	fucktrump	0.79
makeamericagreatagain	0.65	trumpisamoron	0.79
draintheswamp	0.62	dumptrump	0.79
trumppenc	0.61	ivankatrump	0.77
@realdonaldtrump	0.59	theresist	0.76
wakeupamerica	0.58	trumpresign	0.76
thursdaythought	0.57	notmypresid	0.76
realdonaldtrump	0.57	worstpresidentev	0.75

Table 2: Top 10 similar words to "maga" and their respective cosine similarities (obtained from the word2vec models).

Russian trolls on Twitter				Iranian trolls on Twitter			
Hashtag	(%)	Hashtag	(%)	Hashtag	(%)	Hashtag	(%)
news	9.5%	USA	0.7%	Iran	1.8%	Palestine	0.6%
sports	3.8%	breaking	0.7%	Trump	1.4%	Syria	0.5%
politics	3.0%	TopNews	0.6%	Israel	1.1%	Saudi	0.5%
local	2.1%	BlackLivesMatter	0.6%	Yemen	0.9%	Gaza	0.5%
world	1.1%	true	0.5%	FreePalestine	0.8%	SaudiArabia	0.4%
MAGA	1.1%	Texas	0.5%	QudsDay4Return	0.8%	InternationalQudsDay2018	0.4%
business	1.0%	NewYork	0.4%	US	0.7%	Realiran	0.4%
Chicago	0.9%	Fukushima2015	0.4%	realiran	0.6%	Iuvm	0.4%
health	0.8%	quote	0.4%	ISIS	0.6%	News	0.4%
love	0.7%	Foke	0.4%	DeleteIsrael	0.6%		

Table 3: Top 20 (English) hashtags in tweets from Russian and Iranian trolls on Twitter.

4.5 Content Analysis

Word Embeddings. Recent indictments by the US Department of Justice have indicated that troll messaging was crafted, with certain phrases and terminology designated for use in certain contexts. To get a better handle on how this was expressed, we build two word2vec models on the tweets: one for the Russian trolls and one for the Iranian trolls. To train the models, we first extract the tweets posted in English, according to the data provided by Twitter. Then, we remove stop words, perform stemming, tokenize the tweets, and keep only words that appear at least 500 and 100 times for the Russian and Iranian trolls, respectively. Table 2 shows the top 10 most similar terms to "maga" for each model. "Maga" refers to Donald Trump's slogan and means "Make America Great Again". We see a marked difference between its usage by Russian and Iranian trolls. Russian trolls are clearly pushing heavily in favor of Donald Trump, while it is the exact opposite with Iranians.

Hashtags. Next, we study the use of hashtags with a focus on the ones written in English. In Table 3, we report the top 20 English hashtags for both Russian and Iranian trolls. Trolls appear to use hashtags to disseminate news (9.5%) and politics (3.0%) related

content, but also use several that might be indicators of propaganda or controversial topics, e.g., #BlackLivesMatter. For instance, one notable example is: "WATCH: Here is a typical #BlackLivesMatter protester: 'I hope I kill all white babes!' #BatonRouge".

Fig. 11 shows a visualization of hashtag usage built from the two word2vec models. Here, we show hashgtags used in a similar context, by constructing a graph where nodes are words that correspond to hashtags from the word2vec models, and edges are weighted by the cosine distances (as produced by the word2vec models) between the hashtags. After trimming out all edges between nodes with weight less than a threshold, based on methodology from [16], we run the community detection heuristic presented in [5], and mark each community with a different color. Finally, the graph is layed out with the ForceAtlas2 algorithm [25], which considers the weight of the edges when laying out the nodes. Note that the size of the nodes is proportional to the number of times the hashtag appeared in each dataset. We first observe that, in Fig. 11(a) there is a central mass of what we consider "general audience" hashtags (see green community on the center of the graph): hashtags related to a holiday or a specific trending topic (but non-political) hashtag. In the bottom right portion of the plot we observe "general news" related categories; in particular American sports related hashtags (e.g., "baseball"). Next, we see a community of hashtags (light blue, towards the bottom left of the graph) clearly related to Trump's attacks on Hillary Clinton. The Iranian trolls again show different behavior. There is a community of hashtags related to nuclear talks (orange), a community related to Palestine (light blue), and a community that is clearly anti-Trump (pink). The central green community exposes some of the ways they pushed the IUVM fake news network by using innocuous hashtags like "#MyDatingProfileSays" as well as politically motivated ones like "#JerusalemIsTheEternalCapitalOfPalestine."

We also study *when* these hashtags are used by the trolls, finding that most of them are well distributed over time. However we find some interesting exceptions. We highlight a few of these in Fig. 12, which plots the top ten hashtags that Russian and Iranian trolls posted with substantially different rates before and after the 2016 US election. The set of hashtags was determined by examining the relative change in posting volume before and after the election (a hashtag is selected when it has a ratio of appearance of before/after or after/before election of 0.5 or less, this is the reason that we can

(a) Russian trolls - Before election

(b) Russian trolls - After election

(c) Iranians trolls - Before election
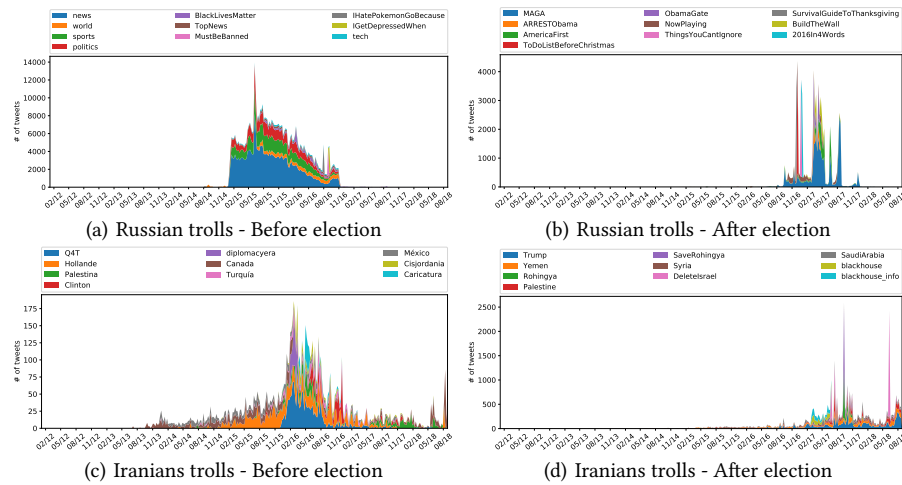
(d) Iranians trolls - After election

Figure 12: Top ten hashtags that appear a) c) substantially more times before the US elections rather than after the elections; and b) d) substantially more times after the elections rather than before.
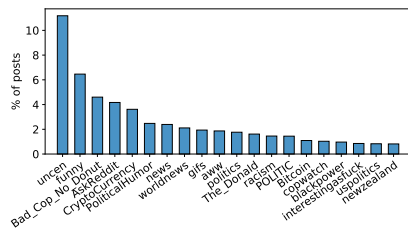

Figure 13: Top 20 subreddits that Russian trolls were active.

see hashtag usage both before and after election in Fig. 12(c). We make several observations. First, we note that more general audience hashtags remain a staple of Russian trolls before the election (the relative decrease corresponds to the overall relative decrease in troll activity following the Crimea conflict). They also use relatively innocuous/ephemeral hashtags like #IHatePokemonGoBeacause, likely in an attempt to hide the true nature of their accounts. That said, we also see them attaching to politically divisive hashtags like #BlackLivesMatters around the time that Donald Trump won the Republican Presidential primaries in June 2016. In the ramp up to the 2016 election, we see a variety of clearly political related hashtags, with #MAGA seeing peaks starting in early 2017 (higher than any peak during the 2016 Presidential campaigns). We also see a large number of politically ephemeral hashtags attacking Obama and a campaign to push the border wall between Mexico. In addition to these politically oriented hashtags, we again see the usage of ephemeral hashtags related to holidays. #SurvivalGuideToThanksgiving in late November 2016 is interesting as it was heavily used for discussing how to deal with interacting with family members with wildly different view points on the recent election results. This hashtag was exclusively used to give trolls a vector to sow discord. When it comes to Iranian trolls, we note that, prior to the 2016 election, they share many posts with hashtags related to Hillary Clinton (see Fig. 12(c)). After the election they shift to posting negatively about Donald Trump (see Fig. 12(d)).

LDA analysis. We also use the Latent Dirichlet Allocation (LDA) model [4] to analyze tweets' semantics. We train an LDA model for each of the datasets and extract ten distinct topics with ten words,

as reported in Table 4. While both Russian and Iranian trolls tweet about politics, for Iranian trolls, this seems to be focused more on regional, and possibly even internal issues. For example, "iran" itself is a common term in several of the topics, as is "israel," "saudi," "yemen," and "isis." While both sets of trolls discuss the proxy war in Syria, the Iranian trolls have topics pertaining to Russia and Putin, while the Russian trolls do not make any mention of Iran, instead focusing on more vague political topics like gun control and racism. For Russian trolls on Reddit (see Table 5) we again find topics related to politics and cryptocurrencies (e.g., topic 4).

Subreddits. Fig. 13 shows the top 20 subreddits that Russian trolls on Reddit exploited and their respective percentage of posts over our dataset. The most popular subreddit is /r/uncen (11% of posts), which is a subreddit created by a specific Russian troll and, via manual examination, appears to be mainly used to share news articles of questionable credibility. Other popular subreddits include general audience subreddits like /r/funny (6%) and /r/AskReddit (4%), likely in an attempt to hide the fact that they are state-sponsored trolls in the same way that innocuous hashtags were used on Twitter. Finally, it is worth noting that the Russian trolls were particularly active on communities related to cryptocurrencies like /r/CryptoCurrency (3.6%) and /r/Bitcoin (1%) possibly attempting to influence the prices of specific cryptocurrencies. This is particularly noteworthy considering cryptocurrencies have been reportedly used to launder money, evade capital controls, and perhaps used to evade sanctions [6].

URLs. We next analyze the URLs included in the tweets/posts. In Table 6, we report the top 20 domains for both Russian and Iranian trolls. Livejournal (5.4%) is the most popular domain in the Russian trolls dataset on Twitter, likely due the Ukrainian campaign. Overall, we can observe the impact of the Crimean conflict, with essentially all domains posted by the Russian trolls being Russian language or Russian oriented. One exception to Russian language sites is RT, the Russian-controlled propaganda outlet. The Iranian trolls similarly post more "localized" domains, for example, jordan-times, but we also see them pushing the IUVM fake news network. When it comes to Russian trolls on Reddit, we find that they were posting random images through Imgur (27.6% of the URLs), likely in an attempt to

Topic	Terms (Russian trolls on Twitter)	Topic	Terms (Iranian trolls on Twitter)
1	news, showbiz, photos, baltimore, local, weekend, stocks, friday, small, fatal	1	isis, first, american, young, siege, open, jihad, success, sydney, turkey
2	like, just, love, white, black, people, look, got, one, didn	2	can, people, just, don, will, know, president, putin, like, obama
3	day, will, life, today, good, best, one, usa, god, happy	3	trump, states, united, donald, racist, society, structurally, new, toonsonline, president
4	can, don, people, get, know, make, will, never, want, love	4	saudi, yemen, arabia, israel, war, isis, syria, oil, air, prince
5	trump, obama, president, politics, will, america, media, breaking, gop, video	5	iran, front, press, liberty, will, iranian, irantalks, realiran, tehran, nuclear
6	news, man, police, local, woman, year, old, killed, shooting, death	6	attack, usa, days, terrorist, cia, third, pakistan, predict, cfb, cfede
7	sports, news, game, win, words, nfl, chicago, star, new, beat	7	israeli, israel, palestinian, palestine, gaza, killed, palestinians, children, women, year
8	hillary, clinton, now, new, fbi, video, playing, russia, breaking, comey	8	state, fire, nation, muslim, muslims, rohingya, syrian, sets, ferguson, inferno
9	news, new, politics, state, business, health, world, says, bill, court	9	syria, isis, turkish, turkey, iraq, russian, president, video, girl, erdo
10	nyc, everything, tcot, miss, break, super, via, workout, hot, soon	10	iran, saudi, isis, new, russia, war, chief, israel, arabia, peace

Table 4: Terms extracted from LDA topics of tweets from Russian and Iranian trolls on Twitter.

Topic	Terms (Russian trolls on Reddit)
1	like, also, just, sure, korea, new, crypto, tokens, north, show
2	police, cops, man, officer, video, cop, cute, shooting, year, btc
3	old, news, matter, black, lives, days, year, girl, iota, post
4	tie, great, bitcoin, ties, now, just, hodl, buy, good, like
5	media, hahaha, thank, obama, mass, rights, use, know, war, case
6	man, black, cop, white, eth, cops, american, quite, recommend, years
7	clinton, hillary, one, will, can, definitely, another, job, two, state
8	trump, will, donald, even, well, can, yeah, true, poor, country
9	like, people, don, can, just, think, time, get, want, love
10	will, can, best, right, really, one, hope, now, something, good

Table 5: LDA topics of posts from Russian trolls on Reddit.

Domain (Russian trolls on Twitter)	(%)	Domain (Iranian trolls on Twitter)	(%)	Domain (Russian trolls on Reddit)	(%)
livejournal.com	5.4%	awdnews.com	29.3%	imgur.com	27.6%
riafan.ru	5.0%	dlvr.it	7.1%	blackmattersus.com	8.3%
twitter.com	2.5%	fb.me	4.8%	donotshoot.us	3.6%
ift.tt	1.8%	whatsupic.com	4.2%	reddit.com	1.9%
ria.ru	1.8%	googl.gl	3.9%	nytimes.com	1.5%
googl.gl	1.7%	realnienovosti.com	2.1%	theguardian.com	1.4%
dlvr.it	1.5%	twitter.com	1.7%	cnn.com	1.3%
gazeta.ru	1.4%	libertyfrontpress.com	1.6%	foxnews.com	1.2%
yandex.ru	1.2%	iuvmpress.com	1.5%	youtube.com	1.2%
j.mp	1.1%	buff.ly	1.4%	washingtonpost.com	1.2%
rt.com	0.8%	7sabah.com	1.3%	huffingtonpost.com	1.1%
nevnov.ru	0.7%	bit.ly	1.2%	photographyisnotacrime.com	1.0%
youtu.be	0.6%	documentinterdit.com	1.0%	butthis.com	1.0%
vesti.ru	0.5%	facebook.com	0.8%	thefreethoughtproject.com	0.9%
kievsmi.net	0.5%	al-hadath24.com	0.7%	dailymail.co.uk	0.7%
youtube.com	0.5%	jordan-times.com	0.7%	rt.com	0.7%
kiev-news.com	0.5%	iuvmonline.com	0.6%	politico.com	0.6%
inforeactor.ru	0.4%	youtu.be	0.6%	reuters.com	0.6%
lenta.ru	0.4%	alwaght.com	0.5%	youtu.be	0.6%
emaidan.com.ua	0.3%	ift.tt	0.5%	nbcnews.com	0.6 %

Table 6: Top 20 domains included in tweets/posts from Russian and Iranian trolls on Twitter and Reddit.

URLs shared by	Events per community							Total	
	/pol/	Reddit	Twitter	Gab	The_Donald	Iran	Russia	Events	URLs
Russians	76,155	366,319	1,225,550	254,016	61,968	0	151,222	2,135,230	48,497
Iranians	3,274	28,812	232,898	5,763	971	19,629	0	291,347	4,692
Both	331	2,060	85,467	962	283	334	565	90,002	153

Table 7: Number of events for URLs shared by a) Russian trolls; b) Iranian trolls; and c) Both Russian and Iranian trolls.

accumulate karma score. We also note a substantial portion URLs to (fake) news sites linked with the Internet Research Agency like blackmattersus.com (8.3%) and donotshootus.us (3.6%).

5 Influence Estimation

Thus far, we have analyzed the behavior of Russian and Iranian trolls on Twitter and Reddit, with a focus on how they evolved over time. Allegedly, one of their main goals is to manipulate the opinion of other users and extend the cascade of information that they share (e.g., lure other users into posting similar content) [12]. Therefore, we now set out to determine their impact in terms of the dissemination of information on Twitter, and on the greater Web.

To assess their influence, we look at three different groups of URLs: 1) URLs shared by Russian trolls on Twitter, 2) URLs shared by Iranian trolls on Twitter, and 3) URLs shared by both Russian *and* Iranian trolls on Twitter. We then find all posts that include any of these URLs in Reddit, Twitter (from the 1% Streaming API, with posts from confirmed Russian and Iranian trolls removed), Gab, and 4chan's Politically Incorrect board (/pol/). For Reddit and Twitter our dataset spans January 2016 to October 2018, for /pol/ it spans July 2016 to October 2018, and for Gab it spans August 2016 to October 2018. We select these communities as previous work shows they play an important and influential role on the dissemination of news [54] and memes [53]. Table 7 summarizes the

number of events (i.e., URL occurrences) for each community/group of users that we consider (Russia refers to Russian trolls on Twitter, while Iran refers to Iranian trolls on Twitter). Note that we decouple The_Donald from the rest of Reddit as previous work showed that it is quite efficient in pushing information in other communities [53]. From the table we observe: 1) Twitter has the largest number of events in all groups of URLs mainly because it is the largest community and 2) Gab has a considerably large number of events; more than /pol/ and The_Donald, which are bigger communities.

For each unique URL, we fit a Hawkes Processes model [29, 30], which allows us to estimate the strength of connections between each of these communities 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. We fit each Hawkes model using the methodology presented by [53]. In a nutshell, by fitting a Hawkes model we obtain all the necessary parameters that allow us to assess the root cause of each event (i.e., the community that is "responsible" for the creation of the event). By aggregating the root causes for all events we are able to measure the influence and efficiency of each Web community we considered. We show our results with two different metrics: 1) the absolute influence, or percentage of events on the destination community caused by events on the source community and 2) the influence relative to size, which shows the number of events caused on the destination platform as a percent of the number of events on the *source* platform. The latter can also be interpreted as a measure of how *efficient* a community is in pushing URLs to other communities.

Fig. 14 reports our results for the absolute influence for each group of URLs. When looking at the influence for the URLs from Russian trolls on Twitter (Fig. 14(a)), we find that Russian trolls were particularly influential to Gab (1.9%), the rest of Twitter (1.29%), and /pol/ (1.08%). When looking at the communities that influenced the Russian trolls we find the rest of Twitter (7%) and Reddit (4%). By looking at URLs shared by Iranian trolls on Twitter (Fig. 14(b)), we find that Iranian trolls were most successful in pushing URLs to The_Donald (1.52%), the rest of Reddit (1.39%), and Gab (1.05%). This is ironic considering The_Donald and Gab's zealous pro-Trump leanings and the Iranian trolls' clear anti-Trump leanings [17, 52]. Similarly to Russian trolls, the Iranian trolls were most influenced by Reddit (5.6%) and the rest of Twitter (4.6%). When looking at the URLs posted by both Russian and Iranian trolls we find that, overall, the Russian trolls were more influential in spreading URLs to the other Web communities with the exception of /pol/.

But how do these results change when we normalize the influence with respect to the number of events that each community creates (i.e., efficiency)? Fig. 15 shows the efficiency for each pair of communities/groups of users. For URLs shared by Russian trolls (Fig. 15(a)) we find that Russian trolls were particularly efficient in spreading the URLs to Twitter (10.4%)—which is not a surprise,
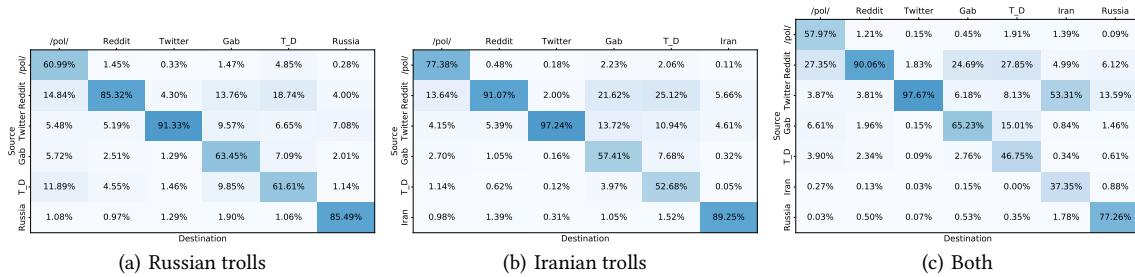
(a) Russian trolls

Source \ Destination	/pol/	Reddit	Twitter	Gab	T_D	Russia
/pol/	60.99%	1.45%	0.33%	1.47%	4.85%	0.28%
Reddit	14.84%	85.32%	4.30%	13.76%	18.74%	4.00%
Twitter	5.48%	5.19%	91.33%	9.57%	6.65%	7.08%
Gab	5.72%	2.51%	1.29%	63.45%	7.09%	2.01%
T_D	11.89%	4.55%	1.46%	9.85%	61.61%	1.14%
Russia	1.08%	0.97%	1.29%	1.90%	1.06%	85.49%

(b) Iranian trolls

Source \ Destination	/pol/	Reddit	Twitter	Gab	T_D	Iran
/pol/	77.38%	0.48%	0.18%	2.23%	2.06%	0.11%
Reddit	13.64%	91.07%	2.00%	21.62%	25.12%	5.66%
Twitter	4.15%	5.39%	97.24%	13.72%	10.94%	4.61%
Gab	2.70%	1.05%	0.16%	57.41%	7.68%	0.32%
T_D	1.14%	0.62%	0.12%	3.97%	52.68%	0.05%
Iran	0.98%	1.39%	0.31%	1.05%	1.52%	89.25%

(c) Both

Source \ Destination	/pol/	Reddit	Twitter	Gab	T_D	Iran	Russia
/pol/	57.97%	1.21%	0.15%	0.45%	1.91%	1.39%	0.09%
Reddit	27.35%	90.06%	1.83%	24.69%	27.85%	4.99%	6.12%
Twitter	3.87%	3.81%	97.67%	6.18%	8.13%	53.31%	13.59%
Gab	6.61%	1.96%	0.15%	65.23%	15.01%	0.84%	1.46%
T_D	3.90%	2.34%	0.09%	2.76%	46.75%	0.34%	0.61%
Iran	0.27%	0.13%	0.03%	0.15%	0.00%	37.35%	0.88%
Russia	0.03%	0.50%	0.07%	0.53%	0.35%	1.78%	77.26%

Figure 14: Percent of *destination* events caused by the source community to the destination community for URLs shared by a) Russian trolls; b) Iranian trolls; and c) both Russian and Iranian trolls.
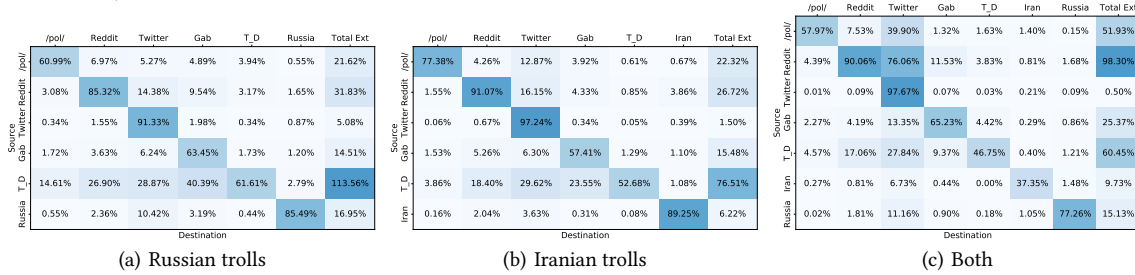
(a) Russian trolls

Source \ Destination	/pol/	Reddit	Twitter	Gab	T_D	Russia	Total Ext
/pol/	60.99%	6.97%	5.27%	4.89%	3.94%	0.55%	21.62%
Reddit	3.08%	85.32%	14.38%	9.54%	3.17%	1.65%	31.83%
Twitter	0.34%	1.55%	91.33%	1.98%	0.34%	0.87%	5.08%
Gab	1.72%	3.63%	6.24%	63.45%	1.73%	1.20%	14.51%
T_D	14.61%	26.90%	28.87%	40.39%	61.61%	2.79%	113.56%
Russia	0.55%	2.36%	10.42%	3.19%	0.44%	85.49%	16.95%

(b) Iranian trolls

Source \ Destination	/pol/	Reddit	Twitter	Gab	T_D	Iran	Total Ext
/pol/	77.38%	4.26%	12.87%	3.92%	0.61%	0.67%	22.32%
Reddit	1.55%	91.07%	16.15%	4.33%	0.85%	3.86%	26.72%
Twitter	0.06%	0.67%	97.24%	0.34%	0.05%	0.39%	1.50%
Gab	1.53%	5.26%	6.30%	57.41%	1.29%	1.10%	15.48%
T_D	3.86%	18.40%	29.62%	23.55%	52.68%	1.08%	76.51%
Iran	0.16%	2.04%	3.63%	0.31%	0.08%	89.25%	6.22%

(c) Both

Source \ Destination	/pol/	Reddit	Twitter	Gab	T_D	Iran	Russia	Total Ext
/pol/	57.97%	7.53%	39.90%	1.32%	1.63%	1.40%	0.15%	51.93%
Reddit	4.39%	90.06%	76.06%	11.53%	3.83%	0.81%	1.68%	98.30%
Twitter	0.01%	0.09%	97.67%	0.07%	0.03%	0.21%	0.09%	0.50%
Gab	2.27%	4.19%	13.35%	65.23%	4.42%	0.29%	0.86%	25.37%
T_D	4.57%	17.06%	27.84%	9.37%	46.75%	0.40%	1.21%	60.45%
Iran	0.27%	0.81%	6.73%	0.44%	0.00%	37.35%	1.48%	9.73%
Russia	0.02%	1.81%	11.16%	0.90%	0.18%	1.05%	77.26%	15.13%

Figure 15: Influence from source to destination community, normalized by the number of events in the *source* community for URLs shared by a) Russian trolls; b) Iranian trolls; and c) Both Russian and Iranian trolls. We also include the total external influence of each community.

given that the accounts operate directly on this platform—and Gab (3.19%). For the URLs shared by Iranian trolls, we again observe that they were most efficient in pushing the URLs to Twitter (3.6%), and the rest of Reddit (2.04%). Also, it is worth noting that in both groups of URLs The Donald had the highest external influence to the other platforms. This highlights that The Donald is an impactful actor in the information ecosystem and is quite possibly exploited by trolls as a vector to push specific information to other communities. Finally, when looking at the URLs shared by both groups of trolls, we find that Russian trolls were more efficient (greater impact relative to the number of URLs posted) at spreading URLs in all the communities except /pol/, where Iranians were more efficient.

6 Discussion & Conclusion

In this paper, we analyzed the behavior and evolution of Russian and Iranian trolls on Twitter and Reddit during several years. We shed light to the campaigns of each group of trolls, we examined how their behavior evolved over time, and what content they disseminated. Furthermore, we find some interesting differences between the trolls depending on their origin and the platform from which they operate. For instance, for the latter, we find discussions related to cryptocurrencies only on Reddit by Russian trolls, while for the former we find that Russian trolls were pro-Trump and Iranian trolls anti-Trump. Also, we quantify the influence that these state-sponsored trolls had on several Web communities (Twitter, Reddit, /pol/, and Gab), showing that Russian trolls were more efficient and influential in spreading URLs on other Web communities than Iranian trolls, with the exception of /pol/. In addition, we make our source code publicly available [41], which helps in reproducing our results and it is a leap towards understanding other types of trolls.

Our findings have serious implications for society at large. First, our analysis shows that while troll accounts use peculiar tactics and talking points to further their agendas, these are not completely disjoint from regular users, and therefore developing automated systems to identify and block such accounts remains an open challenge. Second, our results also indicate that automated systems to detect trolls are likely to be difficult to realize: trolls change their behavior over time, and thus even a classifier that works perfectly on one campaign might not catch future campaigns. Third, and perhaps most worrying, we find that state-sponsored trolls have a meaningful amount of influence on fringe communities like The Donald, 4chan's /pol/, and Gab, and that the topics pushed by the trolls resonate strongly with these communities. This might be due to users on these communities that sympathize with the views the trolls aim to share (i.e., "useful idiots" [55]) or to unidentified state-sponsored actors. In either case, considering recent tragic events like the Tree of Life Synagogue shootings, perpetrated by a Gab user seemingly influenced by content posted there, the potential for mass societal upheaval cannot be overstated. Due to this, we implore the research community, as well as governments and non-government organizations to expend whatever resources are at their disposal to develop technology and policy to address this new, and effective, form of digital warfare.

References

[1] A. Badawy, K. Lerman, and E. Ferrara. Who Falls for Online Political Manipulation? *ArXiv 1808.03281*, 2018.
[2] BBC. Ukraine crisis: Timeline. https://www.bbc.com/news/world-middle-east-26248275, 2014.
[3] A. Bessi and E. Ferrara. Social bots distort the 2016 US Presidential election online discussion. *First Monday*, 2016.
[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.
[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical*

mechanics: theory and experiment, 2008.

[6] Bloomberg. IRS Cops Are Scouring Crypto Accounts to Build Tax Evasion Cases. https://www.bloomberg.com/news/articles/2018-02-08/irs-cops-scouring-crypto-accounts-to-build-tax-evasion-cases, 2018.

[7] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonalves, F. Menczer, and A. Flammini. Political Polarization on Twitter. In *ICWSM*, 2011.

[8] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. BotOrNot: A System to Evaluate Social Bots. In *WWW*, 2016.

[9] Department of Justice. Grand Jury Indicts 12 Russian Intelligence Officers for Hacking Offenses Related to the 2016 Election. https://goo.gl/SCyrm6, 2018.

[10] Department of Justice. Russian National Charged with Interfering in U.S. Political System. https://goo.gl/HAUehB, 2018.

[11] R. Dutt, A. Deb, and E. Ferrara. 'Senator, We Sell Ads': Analysis of the 2016 Russian Facebook Ads Campaign. *ArXiv 1809.10158*, 2018.

[12] S. Earle. TROLLS, BOTS AND FAKE NEWS. https://goo.gl/nz7E8r, 2017.

[13] A. Elyashar, J. Bendahan, and R. Puzis. Is the Online Discussion Manipulated? Quantifying the Online Discussion Authenticity within Online Social Media. *CoRR*, abs/1708.02763, 2017.

[14] E. Ferrara. Disinformation and social bot operations in the run up to the 2017 French presidential election. *ArXiv 1707.00086*, 2017.

[15] E. Ferrara, O. Varol, C. A. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Commun. ACM*, 2016.

[16] J. Finkelstein, S. Zannettou, B. Bradlyn, and J. Blackburn. A Quantitative Approach to Understanding Online Antisemitism. *ArXiv 1809.01644*, 2018.

[17] C. Flores-Saviaga, B. C. Keegan, and S. Savage. Mobilizing the trump train: Understanding collective action in a political trolling community. In *ICWSM '18*, 2018.

[18] V. Gadde and Y. Roth. Enabling further research of information operations on Twitter. https://blog.twitter.com/official/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html, 2018.

[19] S. Hegelich and D. Janetzko. Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet. In *ICWSM*, 2016.

[20] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM '17*, 2017.

[21] P. N. Howard and B. Kollanyi. Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. *CoRR*, abs/1606.06356, 2016.

[22] B. Hubbard. Iranian Protesters Ransack Saudi Embassy After Execution of Shiite Cleric. https://nyti.ms/1P7RKUZ, 2016.

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

[24] IUVM. IUVM's About page. https://iuvm.org/en/about/, 2015.

[25] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS one*, 2014.

[26] Julian Borger and Saeed Dehghan. Geneva talks end without deal on Iran's nuclear programme. https://www.theguardian.com/world/2013/nov/10/iran-nuclear-deal-stalls-reactor-plutonium-france, 2013.

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

[28] H. T. Le, G. R. Boynton, Y. Mejova, Z. Shafiq, and P. Srinivasan. Revisiting The American Voter on Twitter. In *CHI*, 2017.

[29] S. W. Linderman and R. P. Adams. Discovering Latent Network Structure in Point Process Data. In *ICML*, 2014.

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

[31] T. Mihaylov, G. Georgiev, and P. Nakov. Finding Opinion Manipulation Trolls in News Community Forums. In *CoNLL*, 2015.

[32] T. Mihaylov and P. Nakov. Hunting for Troll Comments in News Community Forums. In *ACL*, 2016.

[33] Pushshift. Reddit Dumps. https://files.pushshift.io/reddit/, 2018.

[34] A. Rahimi, T. Cohn, and T. Baldwin. pigeo: A python geotagging tool. *ACL*, 2016.

[35] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Gonalves, A. Flammini, and F. Menczer. Detecting and Tracking Political Abuse in Social Media. In *ICWSM*, 2011.

[36] Reddit. Reddit's 2017 transparency report and suspect account findings. https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/, 2018.

[37] C. M. Rivers and B. L. Lewis. Ethical research standards in a world of big data. *F1000Research*, 3, 2014.

[38] Y. Roth. Empowering further research of potential information operations. https://blog.twitter.com/en_us/topics/company/2019/further_research_information_operations.html, 2019.

[39] S. Zannettou et al. Interactive Graph of Hashtags - Iranian trolls on Twitter. https://trollspaper2018.github.io/trollspaper.github.io/index.html#iranians_graph.gexf, 2018.

[40] S. Zannettou et al. Interactive Graph of Hashtags - Russian trolls on Twitter. https://trollspaper2018.github.io/trollspaper.github.io/index.html#russians_graph.gexf, 2018.

[41] S. Zannettou et al. Source code. https://github.com/zsavvas/trolls_analysis, 2019.

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

[43] L. Steward, A. Arif, and K. Starbird. Examining Trolls and Polarization with a Retweet Network. In *MIS2*, 2018.

[44] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao. Follow the green: growth and dynamics in twitter follower markets. In *IMC*, 2013.

[45] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *ICWSM*, 2017.

[46] O. Varol, E. Ferrara, F. Menczer, and A. Flammini. Early detection of promoted campaigns on social media. *EPJ Data Science*, 2017.

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

[48] Wikipedia. Republican National Convention. https://en.wikipedia.org/wiki/2016_Republican_National_Convention, 2016.

[49] Wikipedia. Unite the Right rally. https://en.wikipedia.org/wiki/Unite_the_Right_rally, 2017.

[50] F. M. F. Wong, C.-W. Tan, S. Sen, and M. Chiang. Quantifying Political Leaning from Tweets and Retweets. In *ICWSM*, 2013.

[51] X. Yang, B.-C. Chen, M. Maity, and E. Ferrara. Social Politics: Agenda Setting and Political Communication on Social Media. In *SocInfo*, 2016.

[52] S. Zannettou, B. Bradlyn, E. De Cristofaro, M. Sirivianos, G. Stringhini, H. Kwak, and J. Blackburn. What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber? In *WWW Companion*, 2018.

[53] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the Origins of Memes by Means of Fringe Web Communities. In *IMC*, 2018.

[54] S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *IMC*, 2017.

[55] S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *arXiv preprint arXiv:1804.03461*, 2018.