Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Characterizing the Use of Images in State-Sponsored Information Warfare Operations by Russian Trolls on Twitter

Savvas Zannettou,[1] **Tristan Caulfield,**[2] **Barry Bradlyn,**[3]
Emiliano De Cristofaro,[2] **Gianluca Stringhini,**[4] **Jeremy Blackburn**[5]
[1]Max-Planck-Institut für Informatik, [2]University College London, [3]University of Illinois at Urbana-Champaign,
[4]Boston University, [5]Binghamton University
szannett@mpi-inf.mpg.de, {t.caulfield, e.decristofaro}@ucl.ac.uk, bbradlyn@illinois.edu,
gian@bu.edu, jblackbu@binghamton.edu

Abstract

State-sponsored organizations are increasingly linked to efforts aimed to exploit social media for information warfare and manipulating public opinion. Typically, their activities rely on a number of social network accounts they control, aka trolls, that post and interact with other users disguised as "regular" users. These accounts often use images and memes, along with textual content, in order to increase the engagement and the credibility of their posts.

In this paper, we present the first study of images shared by state-sponsored accounts by analyzing a ground truth dataset of 1.8M images posted to Twitter by accounts controlled by the Russian Internet Research Agency. First, we analyze the content of the images as well as their posting activity. Then, using Hawkes Processes, we quantify their influence on popular Web communities like Twitter, Reddit, 4chan's Politically Incorrect board (/pol/), and Gab, with respect to the dissemination of images. We find that the extensive image posting activity of Russian trolls coincides with real-world events (e.g., the Unite the Right rally in Charlottesville), and shed light on their targets as well as the content disseminated via images. Finally, we show that the trolls were more effective in disseminating politics-related imagery than other images.

1 Introduction

Social network users are constantly bombarded with digital content. While the sheer amount of information users have access to was unthinkable just a couple of decades ago, the way in which people process that information has also evolved drastically. Social networks have become a battlefield for *information warfare*, with different entities attempting to disseminate content to achieve strategic goals, push agendas, or fight ideological battles (Rowett 2018; Denning 1999).

As part of this tactic, governments often employ "armies" of actors, operating from believable accounts and posting content that aims to manipulate opinion or sow public discord by actively participating in online discussions. Previous work has studied the involvement of state-sponsored accounts in divisive events, e.g., the Black Lives Matter movement (Stewart, Arif, and Starbird 2018) or the 2016 US elections (Badawy, Ferrara, and Lerman 2018; Boyd et al. 2018), highlighting how these entities can be impactful both on the information ecosystem and in the real world.

In today's information-saturated society, the effective use of images when sharing online content can have a strong influence in whether content will catch people's attention and go viral (Berger and Milkman 2012; Jenders, Kasneci, and Naumann 2013; Khosla, Das Sarma, and Hamid 2014). Users often feel overwhelmed with how much content they are exposed to (Koroleva, Krasnova, and Günther 2010), and pay attention to each piece of information for short amounts of time, with repercussion to their attention span (Wrzus et al. 2013). In fact, previous research showed that 60% of social network users re-share articles on social media without reading them, basing their decision on limited cues such as the title of the article or the thumbnail image associated with it (Gabielkov et al. 2016).

Therefore, as part of the efforts aimed to actively push agendas, state-sponsored accounts do not only use textual content, but also take advantage of the expressive power of images and pictures, e.g., using politically and ideologically charged memes (Rowett 2018). In Figure 1, we report some (self-explanatory) examples of images pushed by state-sponsored accounts on Twitter, showcasing their unequivocally political nature and how they can be used to push agendas. Nonetheless, the role of images in information diffusion on the Web has attracted limited attention from the research community, which has thus far mainly focused on textual content (Badawy, Ferrara, and Lerman 2018).

In this paper, we begin filling this gap by studying the use of images by state-sponsored accounts, aka Russian trolls (Gadde and Roth 2018). In particular, we focus on the following research questions:

1. What content is disseminated via images by state-sponsored accounts?
2. Can we identify the target audience of Russian state-sponsored accounts by studying the images they share?
3. How influential are these accounts in making images go viral on the Web? How does this influence results compare to previous characterizations that look into the spread of news by these accounts?

Aiming to address these questions, we use an image-

Figure 1: Examples of politically-charged images posted by Russian trolls.

processing pipeline, expanding that presented by (Zannettou et al. 2018), to study images shared by state-sponsored trolls on Twitter. More precisely, we implement a custom annotation module that uses Google's Cloud Vision API to annotate images in the absence of high-quality ground truth data, or for images that are not bounded to a specific domain (e.g., memes). We then run the new pipeline on a dataset of 1.8M images from the 9M tweets released by Twitter in October 2018 as part of their effort to curb state-sponsored propaganda (Gadde and Roth 2018). These tweets were posted by 3.6K accounts identified as being controlled by the Russian Internet Research Agency (IRA). Finally, we quantify the influence that state-sponsored trolls had on other mainstream and alternative Web communities: namely, Twitter, Reddit, Gab, and 4chan's Politically Incorrect board (/pol/). To do this, we use Hawkes Processes (Linderman and Adams 2014; 2015), which allow us to model the spread of the images across multiple Web communities and assess the root cause of the image appearances.

Main Findings. Along with a first-of-its-kind characterization of how images are used by state-sponsored actors, our work yields a number of interesting findings:

1. The sharing of images by the trolls coincides with real-world events. For instance, we find a peak in activity that is clearly in close temporal proximity with the Unite the Right rally in Charlottesville (Spencer 2017), likely suggesting their use to sow discord during dividing events.
2. Our analysis provides evidence of their general themes and targets. For instance, we find that Russian trolls were mainly posting about Russia, Ukraine, and the USA.
3. By studying the co-occurrence of these images across the Web, we show that the same images appeared in many popular social networks, as well as mainstream and alternative news outlets. Moreover, we highlight interesting differences in popular websites for each of the detected entities: for instance, troll-produced images related to US matters were mostly co-appearing on mainstream English-posting news sites.
4. Our influence estimation results highlight that the Russian state-sponsored trolls, despite their relatively small size, are particularly influential and efficient in pushing images related to politics to other Web communities. In particular, we find that Russian state-sponsored trolls were more influential in spreading political imagery when compared to other images. Finally, by comparing these results to previous analysis focused on news (Zannettou et al. 2019b), we find that trolls were slightly more influential in spreading news via URLs than images.

2 Related Work

Trolls and politics. Previous work has focused on understanding the behavior, role, and impact of state-sponsored accounts on the US political scene. (Boyd et al. 2018) perform linguistic analysis on posts by Russian state-sponsored accounts over the course of the 2016 US election; they find that right- and left-leaning communities are targeted differently to maximize hostility across the political spectrum in the USA. (Stewart, Arif, and Starbird 2018) investigate the behavior of state-sponsored accounts around the BlackLivesMatter movement, finding that they infiltrated both right- and left-leaning political communities to participate in both sides of the discussions. (Jensen 2018) finds that, during the 2016 US election, Russian trolls were mainly interested in defining the identity of political individuals rather than particular information claims.

Trolls in social networks. (Dutt, Deb, and Ferrara 2018) analyze the advertisements purchased by Russian accounts on Facebook. By performing clustering and semantic analysis, they identify their targeted campaigns over time, concluding that their main goal is to sway division on the community, and also that the most effective campaigns share similar characteristics. (Zannettou et al. 2019a) compare a set of Russian troll accounts against a random set of Twitter users, showing that Russian troll accounts exhibit different behaviors in the use of the Twitter platform when compared to random users. In follow up work, (Zannettou et al. 2019b) analyze the activities of Russian and Iranian trolls on Twitter and Reddit, finding substantial differences between them (e.g., Russian trolls were pro-Trump, Iranian ones anti-Trump), that their behavior and targets vary greatly over time, and that Russian trolls discuss different topics across Web communities (e.g., they discuss about cryptocurrencies on Reddit but not on Twitter). Also, (Spangher et al. 2018) examine the exploitation of various Web platforms (e.g., social networks and search engines), showing that state-sponsored accounts use them to advance their propaganda by promoting content and their own controlled domains. Finally, (Broniatowski et al. 2018) focus on the vaccine debate and study Twitter discussions by Russian trolls, bots, and regular users. They find that the trolls amplified both sides of the debate, while at the same time their messages were more political and divisive in nature when compared to messages from bots and regular users.

Detection & Classification. (Badawy, Lerman, and Ferrara 2019) use machine learning to detect Twitter users that are likely to share content that originates from Russian state-sponsored accounts, while (Im et al. 2019) detect Russian trolls using machine learning techniques, finding that these

accounts are still very active on the Web. Also, (Kim et al. 2019) classify Russian state-sponsored trolls into various roles: left- or right-leaning or accounts that pose as news outlets. By applying their technique on 3M tweets posted by Russian trolls on Twitter, they find that despite the fact that trolls had multiple roles, they worked together, while for trolls that pose as news outlets, they find that they had multiple agendas. For instance, some were posting about violent news to create an atmosphere of fear, while others focused on posting highly biased political news.

Remarks. Overall, unlike previous work, we focus on content shared via *images* by state-sponsored accounts. Indeed, to the best of our knowledge, ours is the first study performing a large-scale image analysis on a ground truth dataset of images shared by Russian trolls on Twitter. Previous research (Gabielkov et al. 2016) has showed that social network users usually decide what to share and consume content based on visual cues; thus, as state-sponsored accounts tend to post disinformation (Mejias and Vokuev 2017), studying the images they share provides an important tool to understand and counter disinformation.

3 Methodology

We now present our dataset and our methodology for analyzing images posted by state-sponsored trolls on Twitter.

Dataset. We use a ground truth dataset of tweets posted by Russian trolls released by Twitter in October 2018 (Gadde and Roth 2018). The dataset includes over 9M tweets posted by 3.6K Russian state-sponsored accounts, and their associated metadata and media (1.8M images). Note that the methodology employed by Twitter for detecting/labeling these state-sponsored accounts is not publicly available. That said, to the best of our knowledge, this is the most up-to-date and the largest ground truth dataset of state-sponsored accounts and their activities on Twitter.

Ethics. We only work with publicly available data, which was anonymized by Twitter, and follow standard ethical guidelines (Rivers and Lewis 2014)—e.g., we do not try to de-anonymize users based on their tweets.

Image analysis pipeline. To analyze the images posted by these state-sponsored accounts, we build on the image processing pipeline presented by (Zannettou et al. 2018). This relies on Perceptual Hashing, or pHash (Monga and Evans 2006), and clustering techniques (Ester et al. 1996) to group similar images according to their visual peculiarities, yielding clusters of visually similar images. Then, clusters are annotated based on the similarity between a ground truth dataset and each cluster's medoid (i.e., the representative image in the cluster). For this process, (Zannettou et al. 2018) use crowdsourced meme metadata obtained from Know Your Meme. Our effort, however, has a broader scope as the images shared by state-sponsored accounts are not limited to memes. Consequently, we use a different annotation approach, relying on Google's Cloud Vision API[1], a state-of-the-art solution in Computer Vision tasks to gather useful insights from open-domain images (i.e., not bounded to a specific domain like Internet memes).

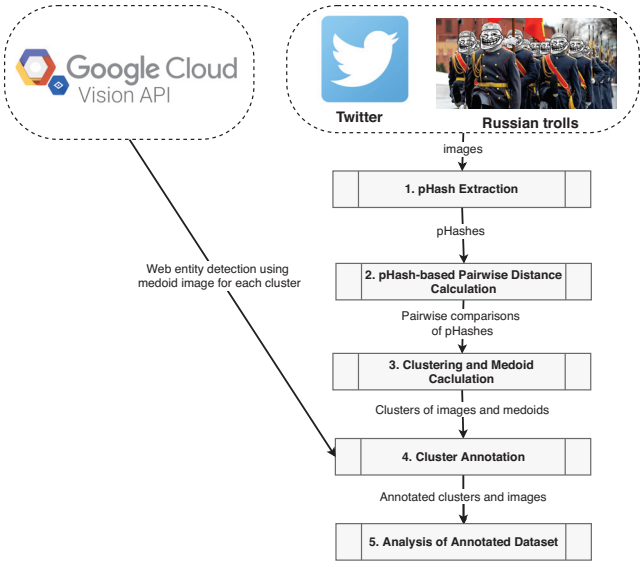


Figure 2: Overview of our image processing pipeline.

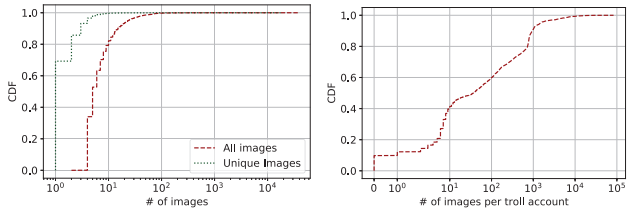
Figure 2 shows the extended pipeline. We perform the "Web Detection" task using Cloud Vision API, which provides us with two very useful pieces of information for each image: 1) a set of *entities*, and their associated confidence scores, that best describe the image (e.g., an image showing Donald Trump yields an entity called "Donald Trump"); and 2) a set of *URLs* on the Web that the same image appeared. To extract this information, the API leverages Google's image search functionality to find URLs to identical and similar images. Furthermore, by extracting data from the text of these URLs, the API provides a set of entities that are related to the image. These two pieces of information are crucial for our analysis as they allow us to understand the context of the images and their appearance across the Web.

Running the pipeline. First, we extract a pHash for each image using the ImageHash library.[2] This reveals that there is a substantial percentage of images that are either visually identical or extremely similar as they have the same pHashes (43% of the images). Next, we cluster the images by calculating all the pairwise comparisons of all the pHashes. This results in 78,624 clusters containing 753,634 images. Then, for each cluster, we extract the medoid, which is the image that has the minimum average Hamming distance between all the images in the cluster. Then, using each medoid, we perform "Web Detection" using the Cloud Vision API, which provides us with a set of entities and URLs, which we assign for each image in the cluster. This is doable since the average number of unique images per cluster is 1.8 with a median of 1 unique image per cluster (see Figure 3(a)).

Pipeline Evaluation. To evaluate the performance of our pipeline, we manually annotate a random sample of 500

[1]https://cloud.google.com/vision/

[2]https://github.com/JohannesBuchner/imagehash



(a) (b)

Figure 3: CDF of a) #images per cluster (image uniqueness is based on their pHash); and b) #images per troll account.

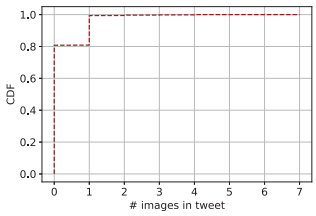


Figure 4: CDF of the number of images per tweet in our dataset.

clusters. Specifically, the first author of this paper manually checked the 500 random clusters and the corresponding Cloud Vision entity with the highest confidence score to assess whether the entity is "appropriate" with respect to the images in the cluster. We find that the Cloud Vision API-based annotation provides an appropriate entity in 83.7% of the clusters in the random sample. Thus, we argue this is a reasonable performance for the purposes of our study.

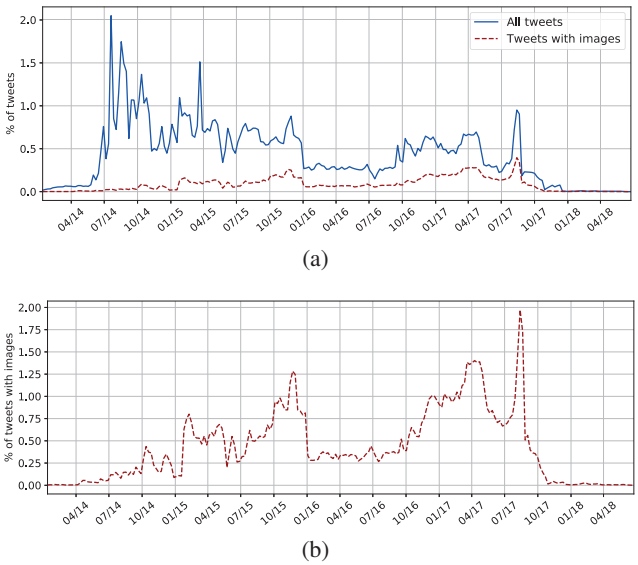
4 Image Analysis

We now present the results of our analysis. First, we perform a general characterization of the images posted by state-sponsored accounts on Twitter and then an analysis of the content of the images. Also, we study the occurrence of the images across the Web.

4.1 General Characterization

We begin by looking at the prevalence of images in tweets by state-sponsored trolls. In Figure 3(b), we plot the CDF of the number of images posted per confirmed state-sponsored account that had at least one tweet (4.5% of the identified trolls never tweeted). We find that only a small percentage of these accounts do not share images (9.7% of the Russian troll accounts). Also, some accounts shared an extremely large number of images, 8% of the Russian trolls posted over 1K images. Furthermore, we find an average of 502.2 images per account with a median number of images of 37.

Then, in Figure 4, we report the CDF of the number of images per tweet; we find that 19% of tweets posted by Russian trolls include at least one image. One explanation for this relatively large fraction is that Twitter automatically generates a preview/thumbnail image when you post a URL. Indeed, by inspecting the URLs in the tweets, we find that out of the 19% of the tweets that contained images, 11.8% of them



(a)

(b)

Figure 5: Temporal overview of: a) all tweets and tweets with images as a percentage of all tweets; and b) all tweets with images as a percentage of all tweets that contained at least one image.

contained automatically generated one, while the rest (7.2%) include images that are explicitly posted (i.e., not generated based on a posted URL). That said, we include *all* images in our dataset and analysis, as generated images too provide insight into the content posted by the state-sponsored accounts, especially considering their proclivity to post "fake news" (Mejias and Vokuev 2017) and the role images might play in catching people's attention.

Temporal Analysis. Next, we look into how the tweets from Russian trolls are shared over time with a particular focus on the tweets that contain images. Figure 5(a) reports the percentage of tweets shared each week normalized by the number of all tweets, while Figure 5(b) the percentages normalized by the number of tweets that contained at least one image. The former shows that, in the early stages of their operations (before 2016), Russian trolls were posting tweets mostly without images, whereas, after 2016 it seems that they started posting more tweets containing images. This indicates that they started using more images in their tweets after 2016, likely because they started targeting specific foreign countries (e.g., the US (Mueller 2019)), suggesting the Russian trolls might believe the use of images can be better for pushing specific narratives.

Figure 5(b) reveals an overall increase in the use of images after October 2016 with a peak of activity in use of images during the week leading to the Charlottesville rally in August 2017 (Spencer 2017), which led to the death of one counter protester (Caron 2017) and was a significant turning point in the use of online hate speech and anti-Semitism in fringe Web communities (Zannettou et al. 2020). This peak likely indicates that the use of images is an effective tactic used by Russian trolls to sow discord on social networks

Top entity	#clusters (%)	Top entity	#images (%)
Russia	2,783 (3.5%)	Russia	30,426 (4.0%)
Vladimir Putin	1,377 (1.7%)	Vladimir Putin	15,718 (2.0%)
Donald Trump	1,281 (1.6%)	Breaking news	15,071 (2.0%)
Car	1,262 (1.6%)	Donald Trump	13,807 (1.8%)
Ukraine	1,031 (1.3%)	Car	10,236 (1.3%)
U.S.A.	907 (1.1%)	Ukraine	10,169 (1.3%)
Barack Obama	823 (1.0%)	U.S.A.	8,638 (1.1%)
Petro Poroshenko	621 (0.8%)	Barack Obama	8,380 (1.1%)
Document	530 (0.6%)	Petro Poroshenko	6,654 (0.9%)
Moscow	495 (0.6%)	Logo	6.017 (0.8%)
Hillary Clinton	479 (0.6%)	Moscow	5,524 (0.7%)
Meme	461 (0.6%)	Syria	4,540 (0.6%)
Logo	456 (0.6%)	Public Relations	4,459 (0.6%)
Product	422 (0.5%)	Police	4,301 (0.6%)
Public Relations	416 (0.5%)	Hillary Clinton	4,167 (0.5%)
Illustration	393 (0.5%)	Document	4,060 (0.5%)
Syria	372 (0.5%)	Meme	3,886 (0.4%)
Web page	310 (0.4%)	Product	3,256 (0.4%)
Advertising	295 (0.3%)	Saint Petersburg	2,870 (0.4%)
Police	290 (0.3%)	Illustration	2,862 (0.4%)

Table 1: Top 20 entities found in images shared by Russian troll accounts. We report the top entities both in terms of the number of clusters and of images.

with respect to events related to politics, the alt-right, and white supremacists.

4.2 Entity Analysis

We now explore the content of images with a special focus on the *entities* they contain, which allows us to better understand what "messages" images were used to convey. To do so, we use the image processing pipeline presented in (Zannettou et al. 2018) to create clusters of visually similar images but leverage Google's Cloud Vision API to annotate each cluster (as discussed in the Methodology section).

Then, for each image, we assign the entity with the highest confidence score as returned by the Cloud Vision API. We also associate the tweet metadata to each image (i.e., which image appears in which tweet). The final annotated dataset allows us to study the popularity of entities in images posted by state-sponsored accounts on Twitter.

Popular Entities. We first look at the popularity of entities for the trolls: Table 1 reports the top 20 entities that appear in our image dataset both in terms of the number of clusters, as well as the number of images within the clusters. We observe that the two most popular entities for Russian trolls are referring to Russia itself (i.e., "Russia" and "Vladimir Putin" entities). Also, trolls are mainly focused on events related to Russia, Ukraine, USA, and Syria (their top entities correspond to these countries). Moreover, several images include screenshots of news articles (see entity "Web page") as well as logos of news sites (see entity "Logo"), hence indicating that these accounts were sharing news articles via images. This is because the state-sponsored accounts shared URLs of news articles, which do not include images, hence Twitter automatically adds the logo of the news site to the tweet. Finally, we find a non-negligible percentage of images and clusters that show memes, highlighting that memes are exploited by such accounts to disseminate their ideology and probably weaponized information via memes.

Graph Visualization. To get a better picture of the spectrum of entities and the interplay between them, we also build a graph, reported in Figure 6, where nodes correspond to clusters of images and each edge to the similarity of the entities between the clusters. For each cluster, we use the set of entities from the Google Cloud Vision API and calculate the Jaccard similarity between each cluster. Jaccard similarity is useful here, because it exposes meta relationships between clusters. While images that appears within the same cluster are visually similar, there are likely to be other clusters that represent the same subjects, but from a different visual perspective. Then, we create an edge between clusters (weighted by their Jaccard similarity) with similarities below a pre-defined threshold. We set this threshold to 0.4, i.e., we discard all edges between clusters that have a Jaccard similarity less than 0.4, because we want to 1) capture the main connections between the clusters and 2) increase the readability of the graph. We then perform the following operations: 1) we run a community detection algorithm using the Louvain method (Blondel et al. 2008) and paint each community with a different color; 2) we lay out the graph with the Force Atlas2 layout (Jacomy et al. 2014), which takes into account weights of edges (i.e., clusters with higher similarity will be positioned closer in the graph); 3) for readability purposes, we show the top 30% of nodes according to their degree in the graph; and 4) we manually annotate the graph with representative images for each community, allowing us to understand the content within each community. In a nutshell, this graph allows us to understand the main communities of entities pushed by the state-sponsored accounts and how they are connected.

Main Communities. From Figure 6, we observe a large community (sapphire) corresponding to clusters related to Vladimir Putin and Russia. This community is tightly connected with communities related to Donald Trump/Hillary Clinton/USA (green), Ukraine/Petro Poroshenko (light blue), and Sergey Lavrov (gray). Also, we observe that other big communities include logos from news outlets (pink) that are connected with communities including screenshots of articles (brown), images of documents (light green), and various other screenshots (emerald). Other communities worth noting are those including comics and illustrations (yellow) as well as images of products and advertisements (orange). Overall, these findings highlight that state-sponsored troll accounts shared many images with a wide variety of themes, ranging from memes to news via screenshots.

4.3 Images Occurrence across the Web

Our next set of measurements analyze the co-occurrence of the images posted by Russian state-sponsored accounts across the greater Web. Recall that the Cloud Vision API also provides details about the appearance of an image across the Web. This is useful when studying the behavior of state-sponsored accounts, as it either denotes that they posted the images on other domains too, or they obtained the image from a different domain, or that other users on the Web posted them on other domains too. Thus, studying the domains that shared the same images as state-sponsored

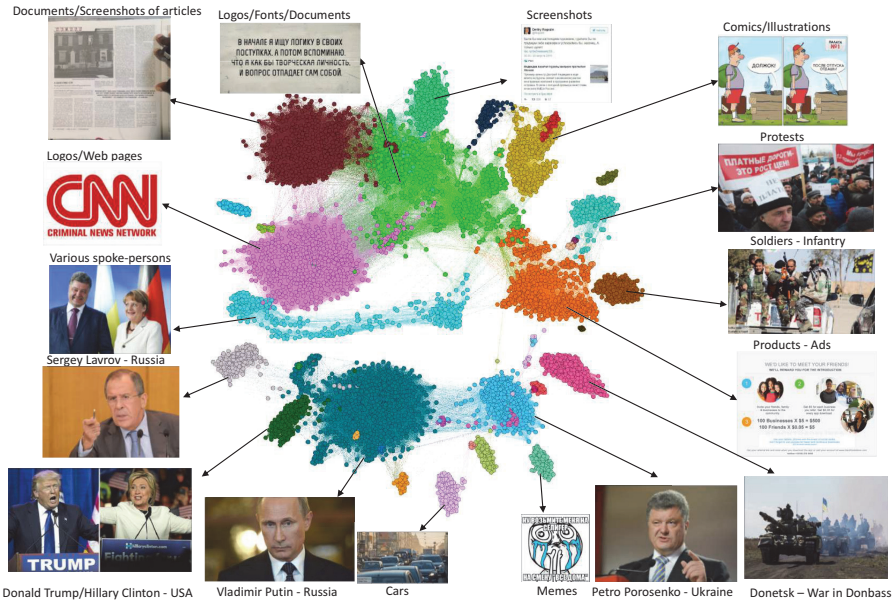


Figure 6: Overview of a subset of the clusters obtained from images shared by the troll accounts.

Domain	#clusters (%)	Domain	#images (%)
pinterest.com	9,433 (12.0%)	pinterest.com	76,231 (10.1%)
twitter.com	5,481 (7.0%)	twitter.com	46,609 (6.1%)
youtube.com	4,132 (5.2%)	youtube.com	40,540 (5.4%)
wordpress.com	3,329 (4.2%)	riafan.ru	35,497 (4.7%)
ria.ru	3,260 (4.1%)	ria.ru	31.153 (4.1%)
riafan.ru	2,734 (3.4%)	wordpress.com	30,464 (4.0%)
blogspot.com	2,432 (3.0%)	blogspot.com	20,890 (2.7%)
livejournal.com	2,381 (3.0%)	sputniknews.com	20,558 (2.7%)
pikabu.ru	2,073 (2.6%)	livejournal.com	20,227 (2.6%)
me.me	1,984 (2.5%)	pikabu.ru	17,250 (2.2%)
sputniknews.com	1,943 (2.4%)	rambler.ru	15,227 (2.0%)
reddit.com	1,826 (2.3%)	me.me	14,675 (1.9%)
theguardian.com	1,527 (1.9%)	theguardian.com	14,111 (1.9%)
rambler.ru	1,524 (1.9%)	reddit.com	14,025 (1.8%)
facebook.com	1,336 (1.7%)	wikimedia.org	12,897 (1.7%)
dailymail.co.uk	1,271 (1.6%)	wikipedia.org	12,081 (1.6%)
imgur.com	1,210 (1.5%)	facebook.com	12,012 (1.6%)
wikimedia.org	1,051 (1.3%)	dailymail.co.uk	9,854 (1.3%)
pinterest.co.uk	1,027 (1.3%)	imgur.com	9,381 (1.2%)
wikipedia.org	996 (1.2%)	cnn.com	8,606 (1.1%)

Table 2: Top 20 domains that shared the same images as the trolls. We report the top domains both in terms of number of clusters and number of images within the clusters.

accounts allows us to understand their behavior and potential impact on the greater Web. For instance, this information can be used to detect domains that are exclusively controlled by state-sponsored actors to spread disinformation.

In Table 2, we report the top domains, both in terms of number of clusters and number images within the clusters, that shared the same images as the state-sponsored accounts. Unsurprisingly, the most popular domains are actually mainstream social networking sites (e.g., Pinterest, Twitter, YouTube, and Facebook). Also, among the popular domains we find popular Russian news outlets like ria.ru and riafan.ru, as well as Russian-owned social networking sites like livejournal.com and pikabu.ru. This highlights the efforts by Russian trolls to sway public opinion about public matters related to Russia. We further find both mainstream and alternative news outlets like theguardian.com and sputniknews.com, respectively (we use the list provided by (Zannettou et al. 2017) to distinguish mainstream and alternative news outlets). This provides evidence that the efforts of Russian trolls had an impact on, or were inspired by, content shared on a wide variety of important sites in the information ecosystem on the Web.

Next, we aim to provide a holistic view of the domains while considering the interplay between the entities of the images and the domains that they also shared them. To do this, we create a graph where nodes are either entities or domains that were returned from the Cloud Vision API. An edge exists between a domain node and an entity node if an image appearing on the domain contained the given entity. Then, we perform the operations (1) and (2) as described in the entities analysis section (i.e., community detection and layout algorithm). We do this for the images posted by the trolls and present the resulting graph in Figure 7. This graph allows us to understand which domains shared images pertaining to various semantic entities. We find popular Web communities like Twitter, Pinterest, Facebook and YouTube in the middle of the graph, constituting a separate community (light blue), i.e., they are used for sharing images across all entities. Entities mainly related to Russia are shared via Russian state-sponsored outlets like sputniknews.com (see orange community). Entities that are related to the USA and political persons like Donald Trump, Barack Obama, and Hillary Clinton are part of a separate community (pink) with popular news outlets like washingtonpost.com and nytimes.com. Finally, for matters

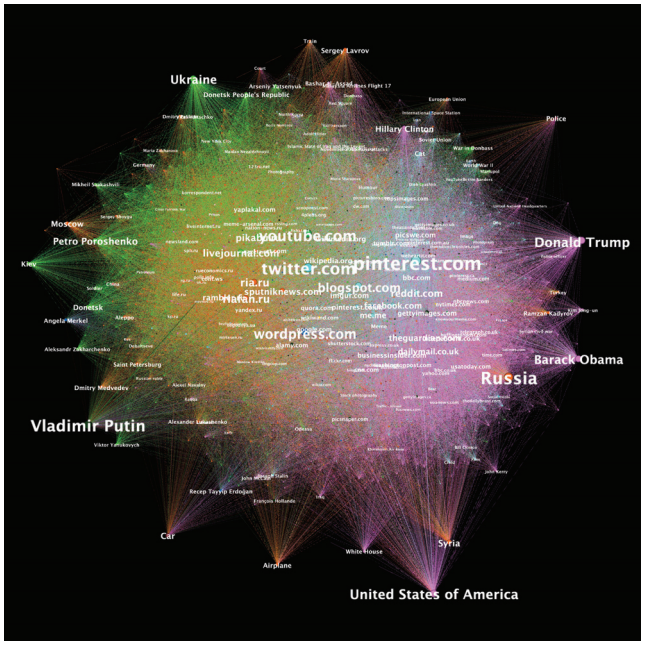


Figure 7: Visualization of the interplay between entities and domains that shared matching images as the ones shared by the trolls.

related to Ukraine (green community) most of the images co-appeared on popular Russian-owned social networks like livejournal.com and pikabu.ru.

Overall, our findings indicate that the same images often appear on both their feeds and specific domains. Thus, state-sponsored trolls might be trying to make their accounts look more credible and push their agenda by targeting unwitting users on popular Web communities like Twitter.

5 Cross-Platform Influence

Our analysis above studies the occurrence of images shared by Russian state-sponsored accounts on other domains, but does not encapsulate the interplay between multiple communities. In reality, the Web consists of a large number of communities that are not exclusively independent of each other, as communities naturally influence each other. For instance, a Twitter user might share an image on Twitter that she previously saw on Reddit: in this case, we see that the Reddit community has "influenced" the Twitter community with respect to the sharing of that particular image.

In this section, we model and measure the interplay and the influence across Web communities with respect to the dissemination of images that were also shared by Russian state-sponsored accounts on Twitter. In other words, we aim to understand how *influential* the trolls were in spreading images to other communities. To do so, we rely on Hawkes Processes (Linderman and Adams 2014; 2015), which allow us to estimate the probabilities that an appearance of an image on one community is due to that image previously occurring on the same or on another Web community.

5.1 Hawkes Processes

Hawkes Processes are self-exciting temporal point processes (Hawkes 1971) that describe how *events* occur on a set of *processes*. In our setting, events are the posting of an image, and processes are Web communities. Generally, a Hawkes model consists of K processes; each process has a *rate* of events that dictates the frequency of the creation of events in the specific process. The occurrence of an event on one process, causes *impulses* to the rest of the processes, temporarily increasing the rate of events in the other processes. The impulses comprise two useful pieces of information: the intensity of the increase in the rate, and how it is distributed and decays over time.

By fitting a Hawkes model using Bayesian inference to data that describes the appearances of events in the processes, we obtain the parameter values for the impulses. This lets us quantify the overall rate of events in each process, as well as how much previous events contribute to the rate, at a given point in time. Naturally, we cannot possibly know what exactly triggered the creation of an event on a process, however, we can use Hawkes Processes to calculate the *probability* that the cause of an event is another process in the model, as also done by previous work (Zannettou et al. 2017; 2018). Note that the background rate of the Hawkes Processes allow us to capture and model the interplay of external sources (i.e., platforms that we do not use in our analysis), hence the resulting probabilities encapsulate the influence of the greater Web via the background rates.

5.2 Datasets

Cross-Platform Dataset. We use a publicly available dataset consisting of 160M pHashes and image URLs for all the images posted on Twitter (using 1% Streaming API), Reddit, 4chan's /pol/, and Gab, between July 2016 and July 2017.[3] Then, we select the images that have the same pHashes with the ones shared by Russian state-sponsored accounts on Twitter. For each one of these images, we find all their occurrences on Reddit, /pol/, Gab, and Twitter. Next, we omit images that appear less than five times across all Web communities we study, ultimately obtaining a set of 90K pHashes. Finally, we annotate each pHash using the Web entities obtained from the Cloud Vision API.

Since our dataset focuses primarily on the year before and after the 2016 US elections, we concentrate our analysis around this major event. Specifically, we want to assess whether Russian state-sponsored accounts were more effective in pushing images related to the Democratic Party or Republican Party. To do so, we select all the pHashes that have a Cloud Vision Web entity corresponding to "Democratic Party," "Hillary Clinton," and "Barack Obama" for the Democratic Party, and "Republican Party" and "Donald Trump" for the Republican Party. Using these entities, we find 9.9K images related to the Republican Party and 6K images related to the Democratic Party.

Examples. To provide an intuition on what some of the politically charged images look like, we provide some ex-

[3]https://zenodo.org/record/1451841.

	/pol/	Reddit	Twitter	Gab	T_D	Russia	Total Events	pHashes
Republican Party-related Images	96,569	85,457	145,372	18,496	21,733	18,332	385,959	9,947
Democratic Party-related Images	64,282	38,602	96,082	13,485	17,797	12,465	242,713	6,043
All images	409,026	421,115	1,904,570	75,361	72,679	231,730	3,114,481	90,299

Table 3: Number of events for images shared by Russian state-sponsored accounts. We report the number of events on Twitter (other users), Russian state-sponsored accounts on Twitter (Russia), Gab, /pol/, Reddit, and The_Donald subreddit.

Figure 8: Example images in the Democratic Party sample.

amples in Figure 8 and Figure 9 for the Democratic and the Republican party, respectively. These illustrate how Russian trolls use images to spread disinformation: for instance, Figure 8(b) shows Senator Robert Byrd meeting with Hillary Clinton and, in another occasion, wearing a Ku Klux Klan robe. The image with the robe is known to be fake as reported later by Snopes (Snopes 2016). We can also observe how state-sponsored accounts rely on edited/photoshopped images to make specific personalities look bad: e.g., Figure 8(a) is an edited image aimed at reinforcing the idea that Hillary Clinton was involved in the Pizzagate conspiracy theory (her face was edited and a baby was added in the picture). Finally, we find several memes that are meant to be funny; however they have a strong political nature and can effectively disseminate ideology. For instance, Figure 8(c) makes fun of Hillary Clinton, while Figure 9(a) and Figure 9(b) are clearly pro-Trump and celebrate him winning the 2016 US elections.

Events. Table 3 summarizes the number of events for our dataset. Note that we elect to decouple The_Donald subreddit from the rest of Reddit mainly because of its strong political nature and support towards Donald Trump (Flores-Saviaga, Keegan, and Savage 2018). By looking at the raw numbers of events per category, we note that in general Russian state-sponsored accounts shared more content related to the Republican Party when compared to the Democratic Party. The same applies for all the other communities we study: in general we find 1.59 times more events for the Republican Party than the Democratic Party (385K vs 242K events). This indicates that content related to the Republican Party was more popular in all Web communities during this time period and that Russian state-sponsored accounts pushed more content related to the Republican Party, likely in favor of Donald Trump as previous research show (Zannettou et al. 2019b).

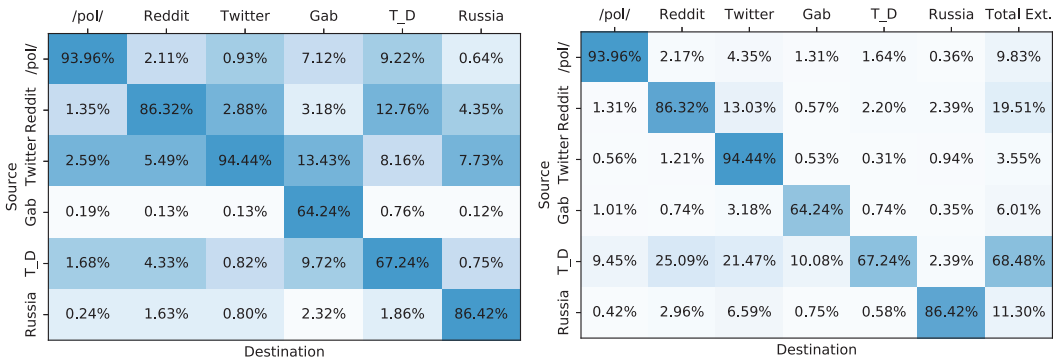
Figure 9: Example images in the Republicans Party sample.

5.3 Results

We create a Hawkes model for each pHash. Each model consists of six processes, one for each of Reddit, The_Donald subreddit, Gab, Russian state-sponsored accounts on Twitter, and other Twitter users. Then, we fit a Hawkes model using Gibbs sampling as described in (Linderman and Adams 2014) for each of the 90K pHashes.

Metrics. After fitting the models and obtaining all the parameters for the models, following the methodology presented in (Zannettou et al. 2018), we calculate the *influence* and *efficiency* that each community had to each other. The former denotes the percentage of events (i.e., image appearances) on a specific community that appear because of previous events on another community, while the latter is a normalized influence metric that denotes how efficient a community is in spreading images to the other communities irrespectively to the number of events that are created within the community. In other words, efficiency describes how influential the posting of a single event to a particular community is, with respect to how it spreads to the other communities.

Overall Influence & Efficiency. Figure 10 reports the influence estimation results for all the events (i.e., all the images that were shared by Russian state-sponsored accounts and have at least five occurrences across all the Web communities we study). When looking at the raw influence results (Figure 10(a)), we observe that Russian state-sponsored accounts had the most influence towards Gab (2.3%), followed by The_Donald subreddit (1.8%), and the rest of Reddit (1.6%), while they had the least influence to 4chan's /pol/ (0.2%). By comparing the influence of regular Twitter users,



(a) (b)

Figure 10: Influence estimation for all images shared by Russian state-sponsored accounts on Twitter: a) Raw *influence* between source and destination Web communities; and b) Normalized influence (*efficiency*) of each Web community as the results are normalized by the numbers of events created on the source community. The numbers in the cell can be interpreted as the expected percentage of events created on the destination community because of previously occurring events on the source community.

with respect to the dissemination of memes, to the influence of the state-sponsored actors (see Figure 11 in extended version of (Zannettou et al. 2018)[4]), we observe similar patterns. That is, regular Twitter users were more influential on Gab (8%), followed by The_Donald (3.6%), and the rest of Reddit (2.8%), while they had the least influence on /pol/ (0.7%). This comparison indicates that Russian trolls influenced other platforms similarly to regular Twitter users with the difference that the intensity of their influence is substantially lower (between 3.5x-1.5x times lower), mainly due to the fact that Russian trolls consist of a few thousands accounts. Furthermore, when comparing the results for Twitter against previous characterizations of Russian trolls on news URLs (see Figure 14 (a) in (Zannettou et al. 2019b)), we find that actually Russian trolls were more influential in spreading news URLs compared to images (1.29% for news URLs and 0.8% for images).

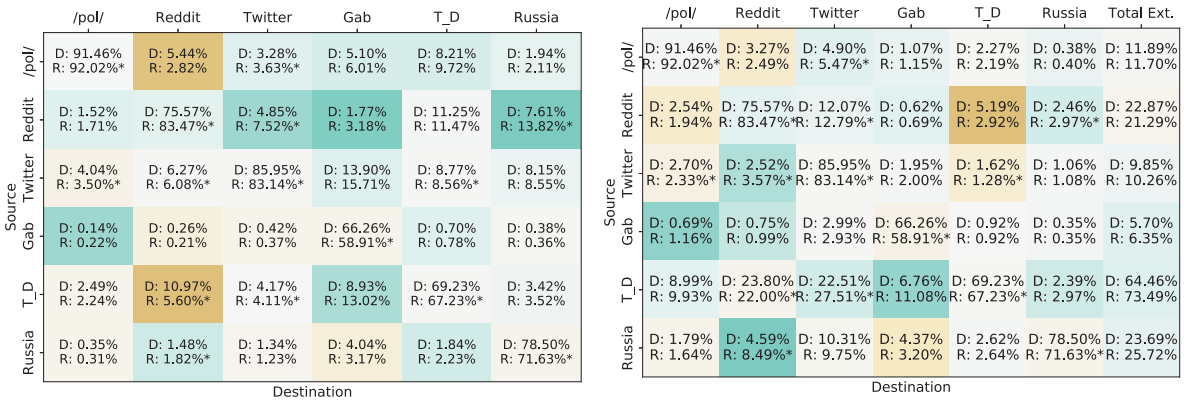
When looking at the efficiency of Russian state-sponsored accounts (last row in Figure 10(b)), we find that they were most efficient in pushing the images on Twitter (6.5%) likely because it is the same social network. Also, they were particularly efficient in pushing images towards the rest of Reddit (2.9%), while again we find that they were not very effective towards 4chan's /pol/ (0.4%). Furthermore, we report the overall external efficiency of each community towards all the other communities (right-most column in Figure 10(b)). We find that the most efficient platform in the ones that we study is The_Donald subreddit (68.4%), followed by the rest of Reddit (19.5%) and the Russian state-sponsored accounts on Twitter (11.3%). Again, by looking at previous results based on news (see Figure 15 (a) in (Zannettou et al. 2019b)), we observe that Russian trolls were more efficient in spreading news URLs compared to images (16.95% external influence for news, while for images we find 11.3%).

[4]Available via https://arxiv.org/abs/1805.12512

Politics-related Images. Next, we investigate how our influence estimation results change when considering only the politics-related images, and in particular the differences between the images pertaining to the Republican and Democratic Parties. Figure 11 reports our influence and efficiency estimation results for the images related to the Republican Party (R) and Democratic Party (D). *NB:* To assess the statistical significance of these results, we perform a two-sample Kolmogorov-Smirnov test to the influence distributions of the two samples and annotate the figures with an * for cases where $p < 0.01$. We make the following observations. First, Russian state-sponsored accounts were most influential in pushing both Democratic and Republican Party-related images to Gab, The_Donald subreddit, and the rest of the Reddit, while again were the least influential in spreading these images in 4chan's /pol/ (see last row in Figure 11(a)).

Second, when comparing the results for both parties, we observe that on Twitter they have more or less the same influence for both Republicans and Democratic parties (1.3% vs 1.2%), on Gab they were more influential in spreading Democratic Party images when compared to Republican party (4.0% vs 3.1%). For The_Donald and the rest of Reddit we observe the opposite: they were more influential in spreading Republican Party related images when compared to the Democratic Party (see last row in Figure 11(a)).

Third, by looking at the efficiency results (Figure 11(b)), we find that again that Russian state-sponsored accounts were most efficient in spreading political images to big mainstream communities like Twitter and Reddit (see last row in Figure 11(b)). Fourth, by looking at the overall external influence of the communities (right-most column in Figure 11(b)), we observe that again The_Donald subreddit had the bigger efficiency (over 60% for both parties), followed by the Russian state-sponsored accounts on Twitter and the rest of Reddit. Finally, by comparing the efficiency of state-sponsored trolls on all images vs the political-related images (cf. Figure 10(b) and Figure 11(b)), we find that Rus-



(a) (b)

Figure 11: Influence estimation for images shared by Russian state-sponsored accounts on Twitter related to the Republican party (R) and the Democratic Party (D): a) Raw *influence* between source and destination Web communities; and b) Normalized influence (*efficiency*) of each Web community as the results are normalized by the numbers of events created on the source community.

sian state-sponsored trolls were over 2 times more efficient in spreading political-related imagery when compared to all the images in our dataset (11.3% vs 23.6% and 25.7%).

Most Influential Images. Since our influence estimation experiments are done with the granularity of specific pHashes, we can also assess *which* images the Russian state-sponsored accounts were more influential in spreading. To do so, we sort the influence results for the Democratic and Republican Parties according to the external influence that Russian state-sponsored accounts had to all the other Web communities, and report the top three images with the most influence. Figure 12 and Figure 13 show the three most influential images shared by Russian state-sponsored accounts for the Democratic and Republicans party, respectively.

Evidently, Russian state-sponsored accounts were particularly influential in spreading images "against" the Democratic Party: for instance, Figure 12(a) is an image that trolls Nancy Pelosi, currently serving as speaker of the US House of Representatives, while Figure 12(b) shares a political message against Hillary Clinton's chances during the 2016 US elections. On the other hand, the most influential images related to the Republican Party (Figure 13) are neutral and likely aim to disseminate pro-Trump messages and imagery.

6 Discussion & Conclusion

This paper presented a large-scale quantitative analysis of 1.8M images shared by Russian state-sponsored accounts ("Russian trolls") on Twitter. Our work is motivated, among other things, by the fact that social network users tend to put little effort into verifying information and they are often driven by visual cues, e.g., images, for re-sharing content (Gabielkov et al. 2016). Therefore, as state-sponsored accounts tend to post disinformation (Mejias and Vokuev 2017), analyzing the images they share represents a crucial step toward understanding and countering the spread of false information on the Web, and its impact on society.

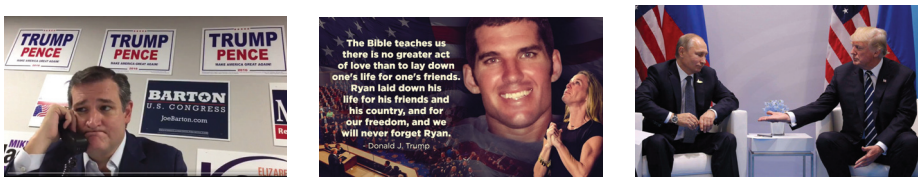
By extending the image processing pipeline presented in (Zannettou et al. 2018), we clustered the images and annotated them using Google's Cloud Vision API. Our analysis shed light on the content and targets of these images, finding that Russian trolls had multiple targets: mainly the USA, Ukraine, and Russia. Furthermore, we found an overall increase in image use after 2016 with a peak in activity during divisive real-world events like the Charlottesville rally. Finally, by leveraging Hawkes Processes, we quantified the influence that Russian state-sponsored accounts had with respect to the dissemination of images on the Web, finding that these accounts were particularly influential in spreading politics-related images. Also, by comparing our results to previous analysis made on news URLs, we find that these actors were more influential and efficient in spreading news via URLs when compared to images.

Our findings demonstrate that state-sponsored accounts pursued a political agenda, aimed at influencing users on Web communities w.r.t. specific world events and individuals (e.g., politicians). Some of our findings confirm previous analysis performed on the text of the tweets posted by these accounts (Zannettou et al. 2019b), highlighting how state-sponsored actors post images that are conceptually similar to their text, possibly in an attempt to make their content look more credible. Our influence estimation also demonstrated that Russian state-sponsored accounts were particularly influential and efficient in spreading political images to a handful of Web communities. Also considering the relatively small number of Russian state-sponsored accounts that were actually identified by Twitter, our analysis suggests that these actors need to be taken very seriously in order to tackle online manipulation and spread of disinformation.

Naturally, our study is not without limitations. First, our pipeline relies on a closed-system (i.e., Cloud Vision API) with a relatively unknown methodology for extracting entities. However, our small-scale manual evaluation showed that the API provides an acceptable performance for our

(a) (b) (c)

Figure 12: Top three most influential images related to the Democratic Party shared by Russian state-sponsored accounts.



(a) (b) (c)

Figure 13: Top three most influential images related to the Republican Party shared by Russian state-sponsored accounts.

needs. Second, we study the images in isolation, without considering other features of the tweets like shared text, thus we may lose important knowledge that exists in the text like sentiment, entities that are referenced, toxicity, etc. Finally, our study relies on a dataset that is independently identified and released by Twitter, and the methodology for identifying these accounts is unknown and it is unclear on whether there are false positives within the dataset.

Implications of our work. Overall, our study has several implications related to the exploitation of social media by Russian state-sponsored actors, who share weaponized information on divisive matters with the ultimate goal of sowing discord and influencing online discussions. As such, their activities should be considered as having broader impact than "just" political campaigns, rather, as direct attacks against individuals and communities, since they can lead to erratic real-world behavior outside the scope of any particular election—e.g., disease epidemics as parents are not vaccinating kids because of disinformation (Broniatowski et al. 2018; Mejias and Vokuev 2017). We also argue that the public should be adequately informed about the existence and the strategies of these actors, particularly their use of weaponized information beyond just "fake news," as a necessary step toward educating users in how to process and digest information on the Web.

Our analysis also complements, to some extent, the Mueller Report (Mueller 2019). Although it represents the first comprehensive investigation of large-scale state-sponsored "information warfare," much of the Report currently remains redacted. Even if it is eventually released in its entirety, it is unlikely to contain a quantitative understanding of how these state-sponsored actors behaved and what kind of influence they had. Furthermore, state-sponsored attacks are reportedly *still on going* (Barnes and Goldman 2019). While still awaiting scientific study, new campaigns, including for instance the Qanon conspiracy theory, have been launched by Russian trolls, and at least partially supported by the use of images initially appearing on imageboards like 4chan and 8chan (Collins and Murphy 2019). Overall, our work can be beneficial to policy makers, law enforcement, and military personnel, as well as political and social scientists, historians, and psychologists who will be studying the events surrounding the 2016 US Presidential Elections for years to come. Our scientific study of how state sponsored actors used images in their attacks can serve to inform this type of interdisciplinary work by providing, at minimum, a data-backed dissection of the most notable and effective information warfare campaign to date.

Finally, the research community can re-use the tools and techniques presented in this paper to study image sharing by various teams or communities on the Web, e.g., state-sponsored accounts from other countries, bots, or any coordinated campaign. In fact, Twitter recently released new datasets for state-sponsored trolls that originate from Venezuela and Bangladesh (Roth 2019); our techniques can be immediately be applied on this data.

Future Work. As part of future work, we plan to study the use of news articles and social network posts from state-sponsored accounts with a particular focus on detecting possibly doctored images. Finally, we aim to build on top of our work to detect domains that are controlled by state-sponsored actors and aim to push specific (disinformation) narratives on the Web.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie EN-

CASE project (GA No. 691025). Also, this work was partially supported by a Content Policy Research on Social Media Platforms award from Facebook Research.

References

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *ASONAM*.

Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Who Falls for Online Political Manipulation? In *WWW Companion*.

Barnes, J., and Goldman, A. 2019. F.B.I. Warns of Russian Interference in 2020 Race and Boosts Counterintelligence Operations. https://nyti.ms/33MC6P2.

Berger, J., and Milkman, K. L. 2012. What makes online content viral? *Journal of marketing research*.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10).

Boyd, R. L.; Spangher, A.; Fourney, A.; Nushi, B.; Ranade, G.; Pennebaker, J.; and Horvitz, E. 2018. Characterizing the Internet Research Agency's Social Media Operations During the 2016 US Presidential Election using Linguistic Analyses. *PsyArXiv*.

Broniatowski, D. A.; Jamison, A. M.; Qi, S.; AlKulaib, L.; Chen, T.; Benton, A.; Quinn, S. C.; and Dredze, M. 2018. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health* 108(10).

Caron, C. 2017. Heather Heyer, Charlottesville Victim, Is Recalled as "a Strong Woman". https://nyti.ms/2vuxFZx.

Collins, B., and Murphy. 2019. Russian troll accounts purged by Twitter pushed Qanon and other conspiracy theories . https://nbcnews.to/3bqeXEE.

Denning, D. E. R. 1999. *Information Warfare and Security*.

Dutt, R.; Deb, A.; and Ferrara, E. 2018. "Senator, We Sell Ads": Analysis of the 2016 Russian Facebook Ads Campaign. In *International Conference on Intelligent Information Technologies*.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*.

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

Gabielkov, M.; Ramachandran, A.; Chaintreau, A.; and Legout, A. 2016. Social clicks: What and who gets read on Twitter? *ACM SIGMETRICS Performance Evaluation Review*.

Gadde, V., and Roth, Y. 2018. Enabling further research of information operations on Twitter. https://bit.ly/2wE1daF.

Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58(1).

Im, J.; Chandrasekharan, E.; Sargent, J.; Lighthammer, P.; Denby, T.; Bhargava, A.; Hemphill, L.; Jurgens, D.; and Gilbert, E. 2019. Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. *arXiv:1901.11162*.

Jacomy, M.; Venturini, T.; Heymann, S.; and Bastian, M. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9(6).

Jenders, M.; Kasneci, G.; and Naumann, F. 2013. Analyzing and predicting viral tweets. In *WWW*.

Jensen, M. 2018. Russian Trolls and Fake News: Information or Identity Logics? *Journal of International Affairs* 71(1.5).

Khosla, A.; Das Sarma, A.; and Hamid, R. 2014. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*.

Kim, D.; Graham, T.; Wan, Z.; and Rizoiu, M.-A. 2019. Tracking the Digital Traces of Russian Trolls: Distinguishing the Roles and Strategy of Trolls On Twitter. *arXiv:1901.05228*.

Koroleva, K.; Krasnova, H.; and Günther, O. 2010. Stop spamming me!: Exploring information overload on Facebook. In *Americas Conference on Information Systems*.

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

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

Mejias, U. A., and Vokuev, N. E. 2017. Disinformation and the media: the case of Russia and Ukraine. *Media, Culture & Society* 39(7).

Monga, V., and Evans, B. L. 2006. Perceptual Image Hashing Via Feature Points: Performance Evaluation and Tradeoffs. *IEEE Transactions on Image Processing*.

Mueller, R. S. 2019. Report On The Investigation Into Russian Interference In The 2016 Presidential Election. US Department of Justice.

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

Roth, Y. 2019. Empowering further research of potential information operations. https://bit.ly/2WJO9uV.

Rowett, G. 2018. The Strategic Need to Understand Online Memes and Modern Information Warfare Theory. In *IEEE Big Data*.

Snopes. 2016. Senator Robert Byrd in Ku Klux Klan Garb. https://www.snopes.com/fact-check/robert-byrd-kkk-photo/.

Spangher, A.; Ranade, G.; Nushi, B.; Fourney, A.; and Horvitz, E. 2018. Analysis of Strategy and Spread of Russia-sponsored Content in the US in 2017. *arXiv:1810.10033*.

Spencer, H. 2017. A Far-Right Gathering Bursts Into Brawls. https://nyti.ms/2uTmIgV.

Stewart, L. G.; Arif, A.; and Starbird, K. 2018. Examining Trolls and Polarization with a Retweet Network. In *WSDM*.

Wrzus, C.; Hänel, M.; Wagner, J.; and Neyer, F. J. 2013. Social network changes and life events across the life span: a meta-analysis. *Psychological bulletin*.

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

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

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019a. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *WWW Companion*.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019b. Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls. In *WebSci*.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *ICWSM*.