







# Did State-Sponsored Trolls Shape the 2016 US Presidential Election Discourse? Quantifying Influence on Twitter

Nikos Salamanos<sup>1</sup>(✉) , Michael J. Jensen<sup>2</sup>, Costas Iordanou<sup>1</sup>,  
and Michael Sirivianos<sup>1</sup>

<sup>1</sup> Department of Electrical Engineering, Computer Engineering and Informatics,  
Cyprus University of Technology, Limassol 3036, Cyprus  
{nik.salaman,michael.sirivianos}@cut.ac.cy,  
costas.iordanou@eecei.cut.ac.cy

<sup>2</sup> Institute for Governance and Policy Analysis, University of Canberra,  
Canberra 2601, Australia  
Michael.Jensen@canberra.edu.au

**Abstract.** It is a widely accepted fact that state-sponsored Twitter accounts operated during the 2016 US presidential election, spreading millions of tweets with misinformation and inflammatory political content. Whether these social media campaigns of the so-called “troll” accounts were able to manipulate public opinion is still in question. Here, we quantify the influence of troll accounts on Twitter by analyzing 152.5 million tweets (by 9.9 million users) from that period. The data contain original tweets from 822 troll accounts identified as such by Twitter. We construct and analyze a very large interaction graph of 9.3 million nodes and 169.9 million edges using graph analysis techniques and a game-theoretic centrality measure. Then, we quantify the influence of all Twitter accounts on the overall information exchange as defined by the retweet cascades. We provide a global influence ranking of all Twitter accounts, and we find that one troll account appears in the top-100 and four in the top-1000. This, combined with other findings presented in this paper, constitute evidence that the driving force of virality and influence in the network came from regular users - users who have not been classified as trolls by Twitter. On the other hand, we find that, on average, troll accounts were tens of times more influential than regular users were. Moreover, 23% and 22% of regular accounts in the top-100 and top-1000, respectively, have now been suspended by Twitter. This raises questions about their authenticity and practices during the 2016 US presidential election.

**Keywords:** Disinformation · Information Diffusion · Twitter Trolls · Political Trolls

## 1 Introduction

The Russian efforts to manipulate the outcome of the 2016 US presidential election were unprecedented in terms of the size and scope of the operation. Millions

of posts across multiple social media platforms gave rise to hundreds of millions of impressions targeting specific segments of the population in an effort to mobilize, suppress, or shift votes [11]. Trolls were particularly focused on the promotion of identity narratives [12], though that does not distinguish them from many other actors during the election [22]. The Special Counsel’s report described this interference as “sweeping and systematic” [18, vol 1, 1]. Russian efforts focused on inflicting significant damage to the integrity of the communication spaces where Americans became informed and discussed their political choices during the election [15]. Therefore, the question of whether these disinformation campaigns had a significantly real impact on social media is of paramount importance [5, 11, 22].

In this paper, we address this question by measuring the influence of the so-called “troll” accounts together with the virality of information that they spread on Twitter during the period of 2016 US Presidential election. Let us note that a “troll” is any Twitter account that deliberately spreads disinformation, tries to inflict conflict, or causes extreme emotional reactions. A troll account could be human or operated automatically. An automated operated account is called a “bot” and is controlled by an algorithm that autonomously performs actions on Twitter. The term “bot” is not synonymous with “troll” as benign bots do operate and have a positive impact on users<sup>1</sup>. In fact, Twitter has set specific rules for acceptable automated behavior<sup>2</sup>.

There are several obstacles to any empirical study on this subject: (i) the lack of complete and unbiased Twitter data – the Twitter API returns only a small sample of the users’ daily activity; (ii) Tweets from deactivated profiles are not available; (iii) The followers and followees lists are not always accessible (i.e., the social graph is unknown). Having that in mind, we collected 152.5 million election-related tweets during the period of the 2016 US presidential election, using the Twitter API along with a set of track terms related to political content. The data contain original troll tweets from that period which later on were deleted by Twitter. Then, based on the ground-truth data released by Twitter regarding state-sponsored accounts linked to Russia, Iran, Venezuela, and Bangladesh states, we identified 822 trolls in our data. Finally, we constructed a very large *interaction-graph* of 9.3 million nodes/users and 169.9 million edges. Using graph analysis techniques and Shapley Value-based centrality, we analyze (i) the graph structure; (ii) the diffusion of potential political content as represented by the retweet cascades of tweets with at least one web or media URL embedded in the text.

Our approach is agnostic with respect to the actual political content of the tweets. The goal is to measure the impact of all users on the overall diffusion of information and consequently estimate the impact of ground-truth trolls. For the rest of the paper, we call “*regular*” the users that have not been classified as trolls by Twitter; they are just the rest of the population and might not always represent benign accounts.

---

<sup>1</sup> <https://blog.mozilla.org/internetcitizen/2018/01/19/10-twitter-bots-actually-make-internet-better-place/>.

<sup>2</sup> <https://help.twitter.com/en/rules-and-policies/twitter-automation>.

**Research Questions (RQ):** We address the following RQ:

**RQ1:** Who are the most influential trolls and regular users? Can we rank them in order of contribution (impact) to the overall information diffusion?

**RQ2:** Which are the viral retweet cascades initiated by regular users and specific troll accounts?

**RQ3:** What is the proximity of top-k influential regular users to bot accounts and how many of them have been suspended by Twitter later on?

**Contributions:** Our primary contributions are as follows:

**C1:** We construct one of the largest graphs representing the interactions between state-sponsored troll accounts and regular users on Twitter during the 2016 US Presidential election. This counts as an approximation of the original social graph.

**C2:** We introduce the notion of *flow graphs* – a natural representation of the information diffusion that takes place in the Twitter platform during the retweeting process. This formulation allows us to apply a Shapley Value-based centrality measure for a fair estimation of users’ contribution to the information shared without imposing assumptions on the users’ behavior. Moreover, we estimate the virality of retweet cascades by the *structural virality* along with the influence each user has on them by the *influence-degree*.

**C3:** We present strong evidence that troll activity was not the main cause of viral cascades of web and media URLs on Twitter. Our measurements show that the regular users were generally the most active and influential part of the population, and their activity was the driving force of the viral cascades. At the same time, we find that, on average, trolls were tens of times more influential than regular users – an indicator of the effectiveness of their strategies to attract attention. These findings further substantiate previously reported insights [26, 28, 29]. Furthermore, more than 20% of the top-100 as well as the top-1000 regular users, have now been suspended by Twitter. This sets their authenticity in question, as well as their activity during that period.

**Data Availability:** Part of the dataset is available under proper restrictions for compliance with Twitter’s ToS and the GDPR<sup>3</sup>. The ground truth data are provided by Twitter<sup>4</sup>.

## 2 Related Work

In a seminal work on the general problem of disinformation on Twitter [26], the authors investigated the diffusion cascades of true and false rumors disseminated from 2006 to 2017 – approximately 126K rumor cascades spread by 3 million people. The main findings are (i) false news diffused faster and more broadly than true ones; (ii) human behavior contributes more to the spread of falsity than trolls. These findings are in line with our main result, that is, the regular users had the dominant role in the viral cascades. Moreover, part of our methodology for the construction of the retweet trees has been inspired by this work.

<sup>3</sup> <https://doi.org/10.5281/zenodo.6526783>.

<sup>4</sup> <https://about.twitter.com/en/our-priorities/civic-integrity>.

In [6], the authors analyzed 171 million tweets by 11 million users – collected five months prior to the 2016 US presidential election. They examined 30 million tweets that contained at least one web URL pointing to a news outlet website. 25% of the news was either fake or biased, representing the spreading of misinformation. Then, they investigated the flow of information by constructing retweet networks for each news category. Furthermore, they estimated the most influential users in the retweet networks using the Collective Influence (CI) algorithm [17]. One of their findings is that Trump supporters were the main group of users that spread fake news, although it was not the dominant one in the whole network. We note that in [6], the overall retweet graph is directly constructed by the data as they were provided by the Twitter API. In our study, we enrich the raw Twitter data by considering all the possible information paths, and at the same time, we provide an estimation of the retweet trees.

Grinberg et al. [10] investigates the extent to which Twitter users were exposed to fake news during the 2016 US presidential election. Their data consists of tweets from 16.4K Twitter accounts that were active during the 2016 US election season, along with their list of followers. They restrict their analysis to tweets containing a URL from a website outside Twitter. One of their main findings is that although a large part of the population had been exposed to fake news, only a small fraction (1%) was responsible for the diffusion of 80% of fake news. The authors introduce the notion of users’ “exposures”, i.e., tweets from a user to his followers. This approach is roughly in line with the *flow graphs* that we introduce in Sect. 4.2.

In [28, 29], the authors analyzed the characteristics and strategies of 5.5K Russian and Iranian troll accounts on Twitter and Reddit. Using *Hawkes Processes*, they compute an overall statistical measure of influence that quantifies the effect these accounts had on social media platforms, such as Twitter, Reddit, 4chan, and Gab. One of their main results is that even though the troll accounts reach a considerably large number of Twitter users and effectively spread URLs on Twitter, their overall effect on the social platforms is not dominant. Our findings verify these results and support the fact that some trolls have above-average influence.

In [3, 4], the authors examined the Russian disinformation campaigns on Twitter in 2016, based on 43M tweets shared by 5.7M users and 221 trolls. They focused on the characteristics of *spreaders*, namely the users that had been exposed to and shared content previously published by Russian trolls. They constructed the retweet graph by mapping retweet actions to edges. Then, they applied the label propagation algorithm to classify Twitter accounts as conservative or liberal. Finally, they used the *Botometer* [8] to determine whether spreaders and non-spreaders can be labeled as bots. We also apply this technique in order to examine whether the top-k influential users exhibit bot behavior.

In [14], a postmortem analysis is conducted on one million Twitter accounts, which although active during the 2016 US election period, later on, were suspended by Twitter. The authors focused on the community-level activities of the suspended accounts, and for that purpose, they clustered them into communities.

Then, they compared the characteristics of suspended account communities with the not suspended ones and found significant differences in their characteristics, especially in their posting behavior.

Finally, Bovet et al. [7] developed a method to infer the political opinion of Twitter users during the 2016 US presidential election. For that purpose, they constructed a directed social graph based on the users’ actions (replies, mentions, retweets) between them – a similar graph formulation technique to ours in this paper. Then, they monitored the evolution of three structural graph properties, the Strongly Connected Giant Component, the Weakly Connected Giant Component, and the Corona.

### 3 Datasets

**Ground-Truth Twitter Data:** Twitter has released a large collection of state-sponsored trolls activities as part of Twitter’s election integrity efforts (see footnote 4). This is ongoing work where the list of malicious accounts is constantly updated. We requested the unhashed version, which consists of 8,275 troll accounts information affiliated with Russia (3,838), Iran (2,861), Venezuela (1,565), and Bangladesh states (11), along with 25,076,853 tweets shared by them. In this study, we leverage only the troll IDs which served as ground-truth identifiers of the trolls in the tweets collection we presented next.

**Our Twitter Dataset:** Our analysis is based on 152,479,440 tweets from 9,939,698 users. We downloaded the data using the Twitter streaming (1%) and Tweepy<sup>5</sup> Python library, in the period before and up to the 2016 US presidential election – from September 21 to November 7, 2016 (47 days; we did not collect data on 02/10/2016). The tweets’ track terms<sup>6</sup> were related to political content such as “hillary2016”, “clinton2016”, “trump2016” and “donaldtrump2016” – namely, a list of phrases used to determine which Tweets are delivered by the stream. In addition to the *tweet text*, *user screen name*, and *user ID*, we also collected metadata, including the *hashtags*, the URLs, and *mentions* that were included in the tweet text, as well as information on the *account creation*, *user timezone*, and *user location*. Based on the ground-truth Twitter data, we identified 35,489 tweets from 822 troll accounts.

**Retweet Cascades:** When a user retweets, usually, he/she agrees with the context of the original tweet (root-tweet) that has been retweeted. For this reason, the analysis of the retweet cascades – i.e., a series of retweets upon the same root-tweet – is important for the identification of the viral cascades as well as the influential users in them. We analyze only the retweet cascades where the root tweet-text contains at least one URL and has been retweeted by at least 100 distinct retweeters (excluding the root-user since he/she may have retweeted his/her own tweet). This process resulted in 46.4K retweet cascades consisting of 19.6M tweets (see Table 1). In a retweet cascade, it is not only the actual tweet

<sup>5</sup> <https://www.tweepy.org/>.

<sup>6</sup> <https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters.html>.

**Table 1.** Retweet cascades with minimum 100 unique retweeters

	Regular Users	Trolls
Total users	3,633,457	233
Root users	8,192	12
Root tweets	45,986	423
Retweeters	3,630,764	228
Total retweets	19,588,072	
Total URLs	43,989	

that has been diffused but mainly the information it contains. So, the web or media URLs (i.e., videos and photos) that are embedded in the tweets serve as “anchors” by which we connect distinct retweet cascades, considering that they are referring to the same information. For example, in Sect. 5.3 we analyze the cascades that refer to URLs that have been spread by trolls.

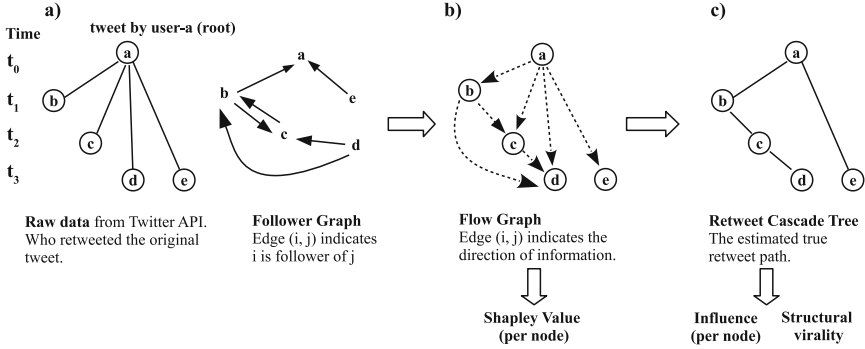
## 4 Methodology

### 4.1 The Social Network

We leverage the users’ activity as is recorded in the data (152.5M tweets) to construct an approximation of the follower-graph – the social network, which is not publicly available to a large extent. Specifically:

**Interaction-Graph:** In short, we map users to nodes and interactions between users to directed edges. In Twitter, the interactions between users belong to three categories: (i) *replies*; (ii) *retweets* or *quotes* – a special form of retweet; (iii) *mentions*. We define the directed edge  $(i, j)$ , from user  $i$  to user  $j$ , for every action of  $i$  on tweets of  $j$ . For example, if  $i$  had replied to a tweet of  $j$ . The direction of the edge implies that  $i$  is a *follower* of  $j$ , while the reverse direction represents the information flow from  $j$  to  $i$ . This process outputs a directed *multigraph*, where many edges may connect the same pair of users. It consists of 169,921,912 edges, 9,321,061 regular users, and 821 trolls. Even though the number of troll accounts is small, there are some indications that some troll accounts might have substantial activity, which is worth further investigation. For instance, we have 671K edges that point to 285 trolls. The total number of nodes is not equal to the total number of users who appear in the initial dataset because the isolated nodes have been discarded – i.e., users who, although tweeted, neither performed an action to other accounts nor received actions from others.

**Follower-graph:** Finally, we construct the follower-graph by discarding the duplicate edges and keeping only the earliest ones. It is a directed graph with 9.32M users/nodes and 84,1M edges, representing an approximation of the true follower-graph.



**Fig. 1.** Toy example of retweet analysis. (a) The raw data provided by Twitter API along with the follower-graph. (b) The flow graph shows the full information flow according to Twitter functionality and the follower-graph. The edges present the path of information that appears on the users’ timeline prior to their retweets. For instance, user  $c$  has retweeted on date  $t_2$ . At the same time user  $b$  – whom user  $c$  follows – has retweeted on date  $t_1 < t_2$ . Note that a given retweet contains both the name of the user who retweeted and the name of the root user who posted the original tweet. Hence, we have an edge from the root to any retweeter because the users have retweeted the root tweet even if they did not follow the root user. (c) The time-inferred cascade tree is constructed from the flow graph by assuming (see Sect. 4.2) that each retweeter has been influenced by the friend who very recently retweeted the original tweet.

### 4.2 Retweet Cascade Tree and Flow Graph

Generally, the retweet data returned by the Twitter API have, by design, limited information regarding the true chain of retweet events. For a given retweet, the information provided is the retweeter ID as well as the root-user ID. Hence, in terms of influence, this corresponds to the case where all the retweeters have been influenced by the root-user. In Fig. 1a, we present an example of the raw data. This star-like cascade structure does not always depict the true chain of retweet events. For example, a user may have retweeted a friend’s retweet and not the original one.

**Retweet Cascade Tree:** A widely used method for the reconstruction of the true retweet path is the time-inferred diffusion process [9, 25, 26]. It is based on the causality assumption that a given user, before retweeting, has been influenced by his “friend” who has recently retweeted the same original tweet. Moreover, since a user can retweet a tweet more than once, we assume that he has been influenced by another user on his first action only. Hence, the final retweet path (see Fig. 1c) is constructed by the raw data provided by Twitter in conjunction with the follower-graph (Fig. 1a). Thus, we have two rather extreme cases; one is the star tree that we take from Twitter API, where no real diffusion structure is present, and the other one is the cascade tree, where a specific hypothesis has been applied with respect to who was influenced by whom. The latter emphasizes

the most recent friend, whereas the former is always the root user. In order to define an intermediate case, we introduce the notion of *flow graph*.

**Flow Graph:** We introduce the concept of *flow graph*, which presents the direction of all possible influence between the retweeters that may have taken place by the information-diffusion in the Twitter platform. Let us consider the toy example in Fig. 1. Before constructing the retweet cascade tree in Fig. 1c, we first have to identify all the time-inferred edges from the users that retweeted in time  $t$  to the users who will retweet in  $t + 1$ . The edges direction indicates the information flow on the Twitter platform and is based on the fact that when a user retweets a given tweet, his action appears on his followers' timeline. For instance, when user  $b$  retweets the root tweet in  $t_1$ , he is transmitting this information to his followers  $c$  and  $d$ . Finally, we add an edge from the root user to any of the retweeters because, in any given retweet, the author's screen name is always visible. The construction of the flow graphs is based on the follower-graph, where the edges are time inferred. So, in a given time  $t_i$ , a given user  $i$  receives information from the users he had already started following at a certain time  $t < t_i$ .

The flow graph, together with the retweet tree, are the two graph structures we leverage to evaluate users' impact on the overall information exchange. Specifically, (a) Flow graph: we measure the contribution of the users to the overall diffusion of information by the Shapley Value-based centrality (see Sect. 4.3). (b) Retweet cascade tree: we measure the influence of every user in a given retweet tree by the influence-degree and the overall virality of the tree by the structural virality (see Sect. 4.4).

### 4.3 Shapley Value-Based Centrality

Towards evaluating the users in terms of the influence/impact they had on the retweet cascades, we have to create a consistent ranking where the top-k users are the most influential ones. One way to do so is to use a centrality measure that fits well in our problem. Here, we apply the Shapley Value-based degree centrality [1, 2, 16] one of the game-theory inspired methods of identifying influential nodes in networks [19, 20, 23]. These methods are based on the Shapley Value [21], a division scheme for the fair distribution of gains or costs in each player of a cooperative game. The Shapley Value of each player in the game is the average weighted marginal contribution of the player over all possible coalitions. Hence, the problem of computing the Shapley Value in a  $N$  player game has, in most cases, exponential complexity since the possible coalitions are  $2^N$ .

We apply the Shapley Value-based degree centrality introduced in [1, 16], which is further refined in [2]. First, in [1, 16], the authors provide a linear time algorithm for the exact computation of the Shapley Value in the following game. Given a directed graph  $G(V, E)$ , with  $V$  nodes and  $E$  edges, the set of players are the nodes in  $V$ , and each coalition is a subset of  $V$ . The value of a coalition  $C$  is defined by the size of the set *fringe*( $C$ ), i.e., the set that consists of the members of  $C$  along with their out-neighbors. This set represents the sphere of



influence of the coalition  $C$ . Moreover, we define that the value of the empty coalition is always zero. The exact closed-form solution of the Shapley Value of

$$\text{a node } u_i \text{ is } V(u_i) = \sum_{u_j \in \{u_i\} \cup N_{out}(u_i)} \frac{1}{1 + \text{indegree}_G(u_j)}.$$

Hence, the algorithm for computing the Shapley Values has running time  $O(|V| + |E|)$  (see Algorithm 1 in [1, 16]). In fact, the Shapley Value is the sum of probabilities that the node contributes to each of its neighbors and also itself.

This formulation is very similar to what we want to measure in the flow graphs. In our case, the value of a coalition is the set of users that have been informed by the members of the coalition about a given root-tweet. Having said that, we cannot directly apply the above formulation since a node cannot inform itself. This very problem has been addressed by the authors in [2] to solve the influence maximization problem. They refined the previous formulation so that the value of a coalition  $C$  is the size of the out-neighbors of the member in  $C$ , i.e., the number of nodes that can be directly influenced by  $C$ . In conclusion, we compute the Shapley Value for all nodes in any flow graph using the following formula (see [2]):

$$SV(u_i) = \sum_{u_j \in N_{out}(u_i)} \frac{1}{1 + \text{indegree}_G(u_j)} \quad (1)$$

In this way, the “leaf” nodes always have zero Shapley Value since they did not inform anyone in the flow graph. The advantage of this approach is that it provides a linear time computation of Shapley Values and also works for disconnected graphs. In fact, this is the case that we face here since the overall information flow is represented by the flow graphs, i.e., a set of disjoint graphs. Moreover, we can compute the overall Shapley Value for any subset of retweet cascades that a user is a part of. We will use this property in order to evaluate trolls and regular users together only in a subset of retweet cascades. The intuition of this approach is that Eq. 1 computes in a fair way the users’ contribution in informing the other members of the graph for a given piece of information, which in our case is the original root-tweet and the URL it contains. We note that from the method in [2], we use only the part that computes the Shapley Values and not the whole process (influence maximization). Our goal is to compute the users’ contribution without assumptions regarding the influence process.

Finally, the global Shapley Value of a user in the overall information exchange is the summation of his Shapley Values in the flow graphs (FG) the user participates in. Hence:

$$SV_{global}(u) = \sum_{FG \in \{FG\}_u} SV(u, FG) \quad (2)$$

#### 4.4 Structural Virality and Influence-Degree

Structural virality evaluates how viral a retweet cascade tree is [9]. The structural virality of a cascade tree  $T$  with  $n > 1$  nodes is the average distance between all pairs of nodes in a cascade. That is:

$$\nu(T) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n d_{ij} \quad (3)$$

where  $d_{ij}$  is the shortest path between the nodes  $i$  and  $j$ . The  $\nu(T)$  represents the average depth of nodes when we consider all nodes as the root of the cascade.

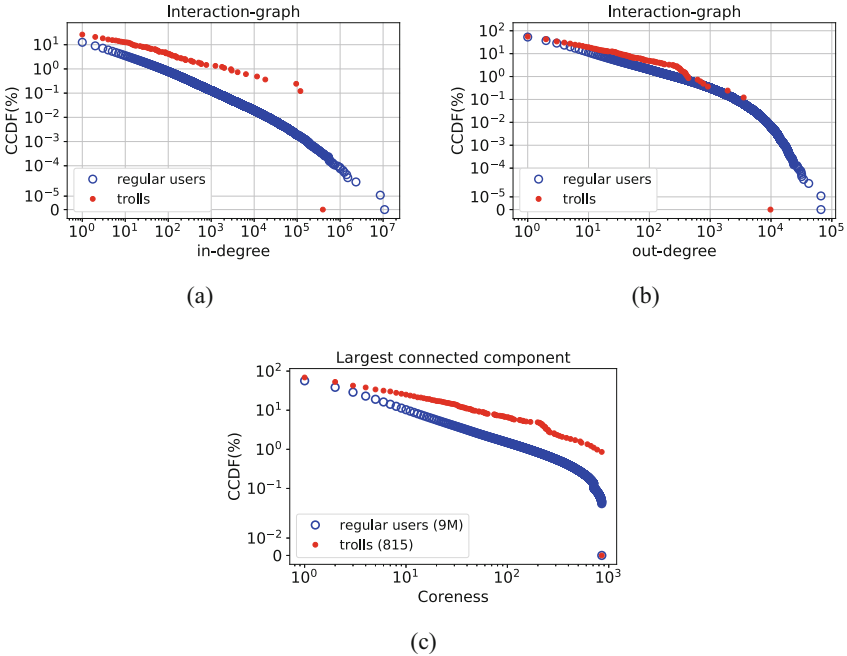
We expect the tree of a viral cascade will have many sub-trees, representing many generations of a viral diffusion process on a smaller scale. On the other hand, a cascade tree with many leaves directly connected with the root represents a “broadcast” – where in a single diffusion process, the material has been transmitted to many nodes (see an example of a broadcast in Fig. 1a). Even though the structural virality is a measure for the cascade tree, it also reflects the collective influence of the nodes in the tree, meaning that not only the root but also other intermediate nodes should have been influential since the material has been transmitted in several regions of the network. So, we expect to find influential nodes in cascades with large structural virality. Hence, in order to measure the influence on an individual level, we define the *influence-degree*. The influence-degree measures the direct influence a node had on a cascade tree. It is defined as the number of users that have been influenced by a given user  $i$  in the cascade tree. For instance, in Fig. 1c, the influence-degree of node  $a$  is two because he has influenced both  $b$  and  $e$ . The global *influence-degree* is the total number of users that have been influenced by  $i$  in all the cascade trees that  $i$  has participated in.

## 5 Results

The analysis is based on comparing the influence of two groups of users; the trolls and the regular users. First, we provide general topological features of the interaction-graph, as well as the follower-graph. Next, we focus on the retweet cascades. We compute the users’ Shapley Value and influence-degree along with the Structural Virality of the cascade trees. Finally, we provide global rankings where we identify the top-k influential users.

### 5.1 Graph Topology

**Degree Distribution.** In both interaction-graph and follower-graph, the in-degree represents the user’s popularity, i.e., the overall activity of his followers on his posts. On the other hand, the out-degree is a measure of a user’s sociability/extroversion, i.e., how active a given user is by interacting with other Twitter accounts. We compare the degree distributions of both graphs, since users with a high degree in the interaction-graph do not necessarily have a large degree in the follower-graph; for instance, users who are highly popular in a small group of followers. The results show that the degree distributions for both graphs are very similar; thus, we discuss the findings only for the interaction-graph which depicts the overall users’ activity. Figure 2 presents the empirical complementary cumulative distribution (CCDF) of in-degree and out-degree for regular



**Fig. 2.** (a) & (b) CCDF of in-degree and out-degree for trolls and regular users; (c) CCDF of coreness for the nodes in the largest connected component.

users and trolls. In summary: (i) 285 trolls and 2.3M regular users have non-zero in-degree; (ii) 675 trolls and 8.5M regular users have non-zero out-degree.

**In-degree** (Fig. 2a): (i) 12 troll accounts have in-degree larger than 1K. The top-3 trolls have 396K, 119K, and 95K in-degrees. On the other hand, we have 12K and 1.8K regular users with in-degrees larger than 1K and 10K, respectively. The top-3 regular users have 10.8M, 8.6M, and 2.3M in-degrees.

**Out-degree** (Fig. 2b): (i) the troll activity is not substantial, i.e., three accounts have out-degree larger than 1K, and the top-3 trolls have 9.8K, 3.5K, and 1.9K out-degrees; (ii) the regular users appear to be considerably more active, i.e., 29.6K and 594 accounts have out-degree larger than 1K and 10K, respectively. In conclusion, it seems that in our dataset the troll activity is not dominant compared to the activity of regular users.

Finally, Table 2 presents the average values for in-degree and out-degree for trolls and regular users in interaction-graph and follower-graph. Even though regular users are the dominant part of the population, the trolls attracted on average, a considerably large amount of traffic. For instance, the trolls’ average in-degree is 45 times higher than the regular users’ average in-degree.

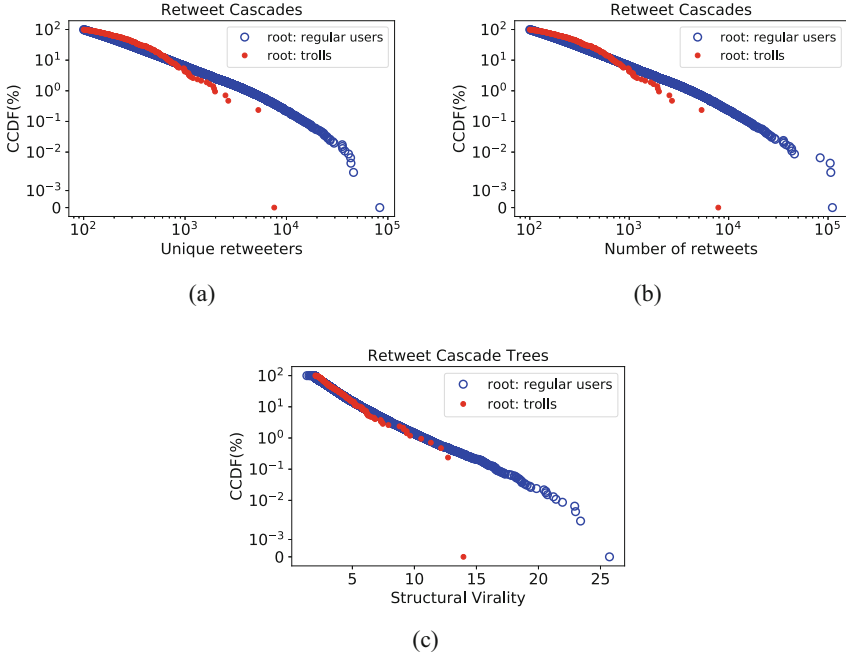
**Table 2.** Average values: Regular Users vs Trolls

		Regular Users	Trolls
Interaction-graph	In-degree	18.16	821.22
	Out-degree	18.23	38.97
Follower-graph	In-degree	8.99	258.63
	Out-degree	9.02	22.48
Largest Comp.	Coreness	9.22	31.75
RT Cascades	Shapley Value	3.21	269.02
	Infl. Degree	5.35	382.71
	Ranking by Shapley	$1, 82 \cdot 10^6$	$1, 61 \cdot 10^6$

**K-Core Decomposition.** First, we identify the connected components of the undirected version of the follower-graph – 9.32M nodes and 82.8M reciprocal edges. We identified 104,954 connected components. The largest connected component consists of 9M nodes and 82,7M edges while the second largest has only 223 nodes. Hence, the largest part of the graph is well-connected. Then, we compute the *k-core decomposition* of the nodes in the largest connected component. The *k-core decomposition* is the process of computing the cores of a graph  $G$ . The *k-core* is the maximal subgraph of  $G$  where each node has a degree of at least  $k$ . The *k-shell* is the subgraph of  $G$  that consists of the nodes that belong to *k-core* but not to  $(k + 1)$ -core. A node has *coreness* (or core number)  $k$  if it belongs to the *k-shell*. In other words, each node is assigned to a shell layer of the graph  $G$ . The graph *k-core number* is the maximum value of  $k$  where the *k-core* is not empty. Coreness is one of the most effective centrality measures for identifying the influential nodes in a complex network [13].

Figure 2c presents the CCDF of the coreness values for trolls and regular users. The graph *k-core number* is 854. The majority of nodes in the larger *k*-shells are the users since their population is larger than that of the troll accounts. There are only eight trolls with large coreness; seven accounts are part of the largest 854-shell, and one account is part of the second-largest 853-shell. This is an indication that these accounts were probably influential. Regarding the regular users, 3,710 and 250 of them belong to the largest and second-largest *k*-shell, respectively. Finally, from Table 2, we observe that the average coreness of trolls is three times larger than the coreness of regular users.

**Summary of Results.** Few trolls have a substantial number of followers (in-degree), activity on other accounts (out-degree), and structural position in the network (coreness). Generally, the dominant part of the population is the regular users. On the other hand, on average, the trolls attracted tens of times more traffic than the regular users.



**Fig. 3.** (a) and (b) CCDF of retweet cascades in terms of the unique number of retweeters and the total number of retweets. The retweeters might have retweeted the same tweet more than once; hence the number of retweets is larger than the number of retweeters. (c) Structural Virality of the retweet cascade trees.

## 5.2 Retweet Cascades and Structural Virality

We now turn our attention to the retweet cascades and provide general statistics about the popularity of the root tweets posted by regular users and trolls. In Fig. 3 we present the CCDF of the number of unique retweeters and the CCDF of the total number of retweets per retweet cascade. From the 423 retweet cascades that have been initiated by troll accounts, 18 of them have more than 1K retweeters. In addition, the two largest cascades have 5.2K and 7.5K retweeters (Fig. 3a). Regarding the cascades that were initiated by regular users, in 2,890 of them the number of retweeters is larger than 1K; 101 cascades have more than 10K retweeters, and the top-5 have between 40K to 83.2K. Regarding the number of retweets per cascade, the findings are similar to the previous ones. The most popular root tweets have been posted by regular users instead of trolls (Fig. 3b). Moreover, in the largest four cascades, the number of retweets is between 83K to 111K, which renders them considerably larger than the number of unique retweeters. This indicates that the root tweets of these four cascades were very popular and they have been retweeted multiple times by the same users.

**Structural Virality.** The previous results depict that the cascades initiated by trolls were not very large. However, the results are based on the unstructured raw data provided by Twitter API, where all the retweets point to the original tweet (see the example in Fig. 1a). Here, we aim to measure how viral the cascades were by using the measure of structural virality (see Sect. 4.4). For the computation of Eq. 3, we use the networkx<sup>7</sup> Python package (Dijkstra’s algorithm). In Fig. 3c, we compare the structural virality of cascade trees for: (i) the cascades initiated by trolls (423 root-tweets, see Table 1); and (ii) the 45,986 cascades initiated by regular users. We can see that regular users were the source of the most viral cascades. The top troll cascade has 13.95 structural virality. On the other hand, 138 user cascades have structural virality larger than 13.95.

**Summary of Results.** The vast majority of viral cascades were initiated by regular users and very few by troll accounts. Moreover, retweet cascades with thousands of retweets have very small structural virality, which indicates that their root users were the main source of influence.

### 5.3 Top-k Influential Users

We conclude the analysis by identifying the most influential Twitter accounts based on two measures, the Shapley Value-based centrality and the influence-degree. We produce the global ranking of all accounts (trolls and regular users) that are part of the retweet cascades (see Table 1; 233 trolls; 3.63M regular users). In addition, we measure how close to a Twitter bot the profiles of the top-1000 regular users are. Our goal is to examine whether the behavior of top-ranked accounts deviates from a human-operated account. As we mentioned in Sect. 1, an account can be automated (having a high Botometer score) and, at the same time, can be benign. On the other hand, a high bot-score raises questions about the authenticity of an account.

**Shapley Value-Based Centrality and Influence-Degree.** Here, based on the flow graphs and the Eqs. 1 and 2, we compute the global Shapley Value of every user who participated in the retweet cascades. Moreover, having the URLs that are embedded in the root-tweets as identifiers of the web and media material that has been diffused in the network, we collect only the cascades that refer to URLs that have been spread by trolls – either by posting an original root tweet or by retweeting. For simplicity, we call these URLs as *URLs-troll*.

In Fig. 4a, we plot the CCDF of the global Shapley Values. We have 27 out of 233 trolls and 161,513 out of 3.6 million regular users with non-zero Shapley Value. In other words, only 27 trolls have a non-zero contribution to the diffusion of information by the retweet cascades. Subsequently, based on the global Shapley Values, we get the global ranking, where the rank for the trolls is [27, 150, 181, 769, 1649, 1797, 2202, 3273, 3964, 4424, 10017, 12263,

<sup>7</sup> <https://networkx.github.io/>.

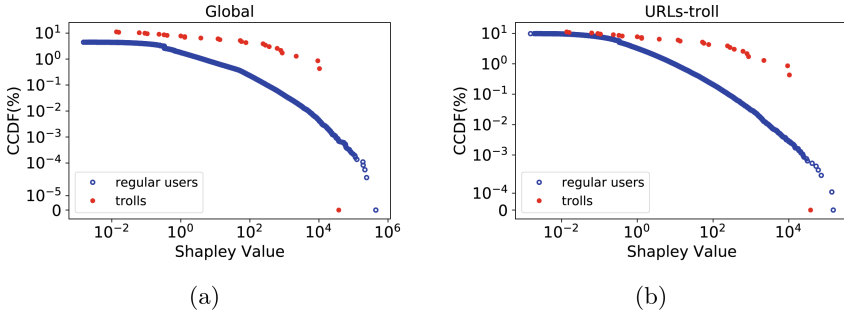


Fig. 4. CCDF of the Shapley Values for trolls and regular users.eps

12939, 22706, 23858, 38246, 58516, 58524, 64181, 90589, 114414, 124387, 139794, 142181, 146944, 158378, 158960]. Hence, only four troll accounts are in the top-1000, and one is in the top-100 (see Table 3). Moreover, the average ranking of trolls is not significantly larger than the rest of the population (see Table 2). At the same time, the average Shapley Value (global) for troll accounts is 83.8 times larger than the regular users’ Shapley Value, which indicates that the troll accounts were quite effective in spreading information. Furthermore, Fig. 4b reports the Shapley Values only for the retweet cascades of *URLs-troll*. We have 2,723 URLs that appear in 3,924 cascades of 934K regular users and 233 trolls. Twenty-seven trolls and 91,572 regular users have non-zero Shapley Value. The distribution for the trolls is the same as the global one since the retweet cascades of *URLs-troll* are the only ones with troll accounts present. Regarding the regular users, we recompute their total Shapley Value by Eq. 2 and only for the subset of retweet cascades that correspond to *URLs-troll*. Again, we reach a final ranking, where the ranking of trolls in the top-1000 is [7, 28, 32, 125, 335, 361, 444, 697, 864, 981]; namely, only ten trolls appear in the top-1000 and three of them in the top-100.

Finally, we use the influence-degree as a measure to rank regular users and trolls according to the effect they have on the retweets cascade trees (see Sect. 4.4). In summary, we have 21 trolls and 118,960 regular users with non-zero influence (we omit the plot). We found four troll accounts in the top-1000 with rankings [34, 201, 241, 899] and one of them in the top-100 (see also Table 3). On the other hand, the influence-degree of trolls is more than 71.5 times larger than regular users’ influence, on average, a similar result to the one for the Shapley Value (see Table 2).

**Bots and Suspended Accounts.** How similar to bot accounts are the top-k users? In order to estimate this, we use the Botometer scores for the top-1000 regular users (ranking by Shapley Values). Botometer<sup>8</sup> classifies Twitter accounts as a bot or human with 0.95 AUC classification performance [24,27]. It uses

<sup>8</sup> <https://botometer.iuni.iu.edu>.

**Table 3.** Top-k Twitter accounts. (We use bold for the suspended accounts)

Account-info User screen-name (User-ID)	Ranking (by-Shapley, by-Infl.)	Coreness	CAP (eng., univ.)
<b>Top-10 influential accounts</b>			
HillaryClinton (1339835893)	(1, 1)	854	(0.0015, 0.0019)
LindaSuhler (347627434)	(2, 2)	854	(0.0068, 0.019)
realDonaldTrump (25073877)	(3, 4)	854	(0.0015, 0.0022)
TeamTrump (729676086632656900)	(4, 5)	854	(0.0014, 0.0019)
wikileaks (16589206)	(5, 6)	854	(0.0013, 0.0019)
WDFx2EU7 ( <b>779739206339928064</b> )	(6, 13)	854	N/A
PrisonPlanet (18643437)	(7, 9)	854	(0.0012, 0.0020)
FoxNews (1367531)	(8, 7)	854	(0.0028, 0.0026)
magnifier661 ( <b>431917957</b> )	(9, 11)	854	N/A
CNN (759251)	(10, 8)	854	(0.0031, 0.0027)
ChristiChat ( <b>732980827</b> )	(11, 10)	854	N/A
StylishRentals ( <b>355355420</b> )	(13, 3)	96	N/A
<b>Troll accounts in top-1000</b>			
TEN_GOP ( <b>4224729994</b> )	(27, 34)	854	N/A
Pamela_Moore13 ( <b>4272870988</b> )	(150, 201)	854	N/A
America_1st_( <b>4218156466</b> )	(181, 241)	854	N/A
tpartynews ( <b>3990577513</b> )	(769, 899)	854	N/A
<b>Potentially Bot accounts in top-1000</b>			
resultzba (3248410062)	(275, 412)	854	(0.565, 0.297)
TrumpLadyFran (717627639159128064)	(311, 355)	854	(0.847, 0.446)
edebblazim (429229693)	(531, 571)	854	(0.892, 0.812)
WORIDSTARHIPHOP (2913627307)	(643, 552)	46	(0.511, 0.385)

various machine-learning models and more than a thousand features that have been extracted from the publicly available data of the account in question. For a given account, the Botometer API returns various scores where the more general one is the *Complete Automation Probability* (CAP) – the probability that a given account is completely automated. Two CAP scores are provided, one based on its English language tweets and one for universal features. Generally, CAP scores above 0.5 indicate a bot account [7].

In top-1000, four regular users have either CAP(english) or CAP(universal) score larger than 0.5, so they are potentially bots (see Table 3). On the other hand, only 22 and 21 users have CAP scores larger than 0.2. Moreover, 263 accounts were inactive. In order to verify the reasons for inactivity, we get the account information of the regular users in the top-10000, using Tweepy. When an account is not accessible, then Tweepy returns an error message<sup>9</sup> either “User not found” (code 50; corresponds with HTTP 404; deleted account by the user

<sup>9</sup> <https://developer.twitter.com/en/support/twitter-api/error-troubleshooting>.



itself) or “User has been suspended” (code 63; corresponds with HTTP 403; suspended account by Twitter due to violation of Twitter Rules<sup>10</sup>). In summary, we found that (i) in top-100: 23 suspended accounts out of the 26 inactive ones; (ii) in top-1000: 220 suspended out of the 263 inactive; (iii) in top-10000: 1,836 suspended out of the 2,508 inactive.

Lastly, Table 3 shows the account information for the top-10 influential users based on the Shapley Value. We also present the corresponding rankings in terms of influence-degree and coreness along with the Botometer scores. Two accounts in top-10 are suspended, which raises serious doubts about the authenticity of these users. The top-10 users are part of the largest 854-shell. In addition, we report the four trolls in top-1000 along with their rankings and coreness. All four of them are part of the largest 854-shell. Moreover, in retweet cascades initiated by them, more than 1.1% of the retweets were from regular users belonging to the top-1000 group.

**Summary of Results.** Four troll accounts were amongst the most influential users. Their tweets have been retweeted tens of times by top-1000 influential regular users. Four regular users in the top-1000 exhibit bot behavior. In addition, 23% and 22% of regular accounts in the top-100 and top-1000 respectively, have been suspended by Twitter, something that raises questions about their authenticity and practices overall.

## 6 Conclusion

In this paper, we have extensively studied the influence that state-sponsored trolls had during the 2016 US presidential election by analyzing millions of tweets from that period. We first constructed the interaction-graph between trolls and regular users, and then we concentrated our analysis on the retweet cascades. In order to measure the users’ impact on the diffusion of information, we introduce the notion of *flow graph*, where we apply a game theoretic-based centrality measure. Moreover, we estimate the retweet paths by constructing the retweet cascade trees where we measure the users’ direct influence. The results indicate that although the trolls initiated some viral cascades, their role was not dominant and the source of influence was mainly the regular users. On the other hand, the average influence of trolls was considerably larger than the average influence of regular users. This indicates that the strategies these trolls followed in order to attract and engage regular users were sufficiently effective. Furthermore, 23% and 22% of regular accounts in the top-100 and top-1000, respectively, have now been suspended by Twitter. This raises questions about the authenticity of these accounts.

**Acknowledgement.** We are grateful to Twitter for providing access to the trolls’ ground truth dataset. We thank Nikolaos Laoutaris for his insightful comments about

<sup>10</sup> <https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts>.

the Shapley Value. This project has received funding from the European Union's Horizon 2020 Research and Innovation program under the Cybersecurity CONCORDIA project (Grant Agreement No. 830927) and under the Marie Skłodowska-Curie INCOGNITO project (Grant Agreement No. 824015).

## References

1. Aadithya, K.V., Ravindran, B., Michalak, T.P., Jennings, N.R.: Efficient Computation of the Shapley Value for Centrality in Networks. In: Saberi, A. (ed.) WINE 2010. LNCS, vol. 6484, pp. 1–13. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-17572-5\\_1](https://doi.org/10.1007/978-3-642-17572-5_1)
2. Adamczewski, K., Matejczyk, S., Michalak, T.: How good is the shapley value-based approach to the influence maximization problem? *Front. Artif. Intell. Appl.* **263** (2014)
3. Badawy, A., Ferrara, E., Lerman, K.: Analyzing the digital traces of political manipulation: The 2016 Russian interference twitter campaign. In: Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 258–265. ASONAM 2018 (2018)
4. Badawy, A., Lerman, K., Ferrara, E.: Who falls for online political manipulation? In: Companion Proceedings of The 2019 World Wide Web Conference. pp. 162–168. WWW 2019, ACM (2019)
5. Benkler, Y., Faris, R., Roberts, H.: *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press (2018)
6. Bovet, A., Makse, H.A.: Influence of fake news in twitter during the 2016 us presidential election. *Nat. Commun.* **10**(7) (2019)
7. Bovet, A., Morone, F., Makse, H.A.: Validation of twitter opinion trends with national polling aggregates: Hillary clinton vs donald trump. *Sci. Rep.* **8**(1) (2018)
8. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: Botornot: A system to evaluate social bots. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 273–274. WWW 2016 Companion (2016)
9. Goel, S., Anderson, A., Hofman, J., Watts, D.J.: The structural virality of online diffusion. *Manag. Sci.* **2**(1) (2015)
10. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., Lazer, D.: Fake news on twitter during the 2016 u.s. presidential election. *Science* **363**(6425), 374–378 (2019)
11. Jamieson, K.H.: *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don't, Can't, and Do Know*. Oxford University Press (2018)
12. Jensen, M.: Russian trolls and fake news: Information or identity logics? *J. Int. Affairs* **71**(1.5), 115–124 (2018)
13. Kitsak, M., et al.: Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893 (2010)
14. Le, H., Boynton, G.R., Shafiq, Z., Srinivasan, P.: A postmortem of suspended twitter accounts in the 2016 U.S. presidential election. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 258–265. ASONAM 2019, ACM (2019)
15. Mazarr, M., et al.: *Hostile Social Manipulation: Present Realities and Emerging Trends*. Santa Monica: Rand Corporation (March 2019). [https://www.rand.org/pubs/research\\_reports/RR2713.html](https://www.rand.org/pubs/research_reports/RR2713.html)

16. Michalak, T., Aadithya, K., Szczepański, P., Ravindran, B., Jennings, N.: Efficient computation of the shapley value for game-theoretic network centrality. *J. Artif. Intell. Res* **46**(2014)
17. Morone, F., Makse, H.A.: Influence maximization in complex networks through optimal percolation. *Nature* **524**(7563), 65–68 (2015)
18. Mueller, R.S.: Report on the Investigation into Russian Interference in the 2016 Presidential Election. Washington, DC: Department of Justice (2019). <https://www.justice.gov/storage/report.pdf>
19. Narayanam, R., Narahari, Y.: A shapley value-based approach to discover influential nodes in social networks. *IEEE Trans. Autom. Sci. Eng.* **8**(1), 130–147 (2011)
20. Papapetrou, P., Gionis, A., Mannila, H.: A shapley value approach for influence attribution. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) *ECML PKDD 2011. LNCS (LNAI)*, vol. 6912, pp. 549–564. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-23783-6\\_35](https://doi.org/10.1007/978-3-642-23783-6_35)
21. Shapley, L.S.: A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (eds.) *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, Princeton (1953)
22. Sides, J., Tesler, M., Vavreck, L.: *Identity Crisis: The 2016 Presidential Campaign and the Battle for the Meaning of America*. Princeton University Press (2018)
23. Suri, N.R., Narahari, Y.: Determining the top-k nodes in social networks using the shapley value. In: *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3*. pp. 1509–1512. *AAMAS 20'08* (2008)
24. Varol, O., Ferrara, E., Davis, C., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. *Proceedings of the International AAAI Conference on Web and Social Media* **11**(1), 280–289 (2017). <https://doi.org/10.1609/icwsm.v11i1.14871>
25. Vosoughi, S., Mohsenvand, M.N., Roy, D.: Rumor gauge: predicting the veracity of rumors on twitter. *ACM Trans. Knowl. Discov. Data* **11**(4) (2017)
26. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. *Science* **359**(6380), 1146–1151 (2018)
27. Yang, K.C., Varol, O., Hui, P.M., Menczer, F.: Scalable and generalizable social bot detection through data selection (2019)
28. Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., Blackburn, J.: Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In: *Workshop on Computational Methods in Online Misbehavior*, pp. 218–226. ACM (2019)
29. Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., Blackburn, J.: Who let the trolls out?: Towards understanding state-sponsored trolls. In: *Proceedings of the 10th ACM Conference on Web Science*, pp. 353–362. *WebSci '19*, ACM (2019)