Uncovering Coordinated Networks on Social Media

Diogo Pacheco*, **Pik-Mai Hui***, **Christopher Torres-Lugo***,
Bao Tran Truong, Alessandro Flammini, Filippo Menczer
Center for Complex Networks and Systems Research
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington

Abstract

Coordinated campaigns are used to influence and manipulate social media platforms and their users, a critical challenge to the free exchange of information online. Here we introduce a general network-based framework to uncover groups of accounts that are likely coordinated. The proposed method construct coordination networks based on arbitrary behavioral traces shared among accounts. We present five case studies of influence campaigns in the diverse contexts of U.S. elections, Hong Kong protests, the Syrian civil war, and cryptocurrencies. In each of these cases, we detect networks of coordinated Twitter accounts by examining their identities, images, hashtag sequences, retweets, and temporal patterns. The proposed framework proves to be broadly applicable to uncover different kinds of coordination across information warfare scenarios.

Introduction

Online social media have revolutionized how people access news and information, and form opinions. By enabling exchanges that are unhindered by geographical barriers, and by lowering the cost of information production and consumption, social media have enormously broadened participation in civil and political discourse. Although this could potentially strengthen democratic processes, there is increasing evidence of malicious actors polluting the information ecosystem with disinformation and manipulation campaigns (Lazer et al. 2018; Vosoughi, Roy, and Aral 2018; Bessi and Ferrara 2016; Shao et al. 2018; Ferrara 2017; Stella, Ferrara, and De Domenico 2018; Deb et al. 2019; Bovet and Makse 2019; Grinberg et al. 2019).

While influence campaigns, misinformation, and propaganda have always existed (Jowett and O'Donnell 2018), social media have created new vulnerabilities and abuse opportunities. Just as easily as like-minded users can connect in support of legitimate causes, so can groups with fringe, conspiratorial, or extremist beliefs reach critical mass and become impervious to expert or moderating views. Platform APIs and commoditized fake accounts make it simple to develop software to impersonate users and hide the identity of those who control these social bots — whether they are fraudsters pushing spam, political operatives amplifying misleading narratives, or nation-states waging online warfare (Ferrara et al. 2016). Cognitive and social biases make us even more vulnerable to manipulation by social bots: our limited attention facilitates the spread of unchecked claims, confirmation bias makes us disregard facts, group-think and echo chambers distort perceptions of norms, and the bandwagon effect makes us pay attention to bot-amplified memes (Weng et al. 2012; Hills 2019; Ciampaglia et al. 2018; Lazer et al. 2018).

Despite advances in countermeasures such as machine learning algorithms and human fact-checkers employed by social media platforms to detect misinformation and inauthentic accounts, malicious actors continue to effectively deceive the public, amplify misinformation, and drive polarization (Barrett 2019). We observe an arms race in which the sophistication of attacks evolves to evade detection.

Most machine learning tools to combat online abuse target the detection of social bots, and mainly use methods that focus on individual accounts (Davis et al. 2016; Varol et al. 2017; Yang et al. 2019). However, malicious groups may employ *coordination* tactics that appear innocuous at the individual level, and whose suspicious behaviors can be detected only when observing networks of interactions among accounts. For instance, an account changing its handle might be normal, but a group of accounts switching their names in rotation is unlikely to be coincidental.

In this paper we propose an approach that considers coordination of multiple actors to reveal suspicious behaviors, regardless of their automated/organic nature and malicious/benign intent. The idea is to use features extracted from social media data to build a coordination network, where two accounts have a strong tie if they display unexpectedly similar behavioral traces. These similarities can stem from any metadata, such as content entities and profile features. Networks provide an efficient representation for sparse similarity matrices, and a natural framework to detect significant clusters of coordinated accounts. We demonstrate the effectiveness of the proposed framework on Twitter, but the method can in principle be applied to any social media platform where data is available.

*Equal contributions.

After describing our methodology, we present five case studies by instantiating the framework to detect different types of coordination: (i) handle changes, (ii) image sharing, (iii) sequential use of hashtags, (iv) co-retweets, and (v) synchronization. These examples illustrate the generality of our approach: we are able to detect coordinated campaigns based on what is presented as identity, shown in pictures, written in text, retweeted, or when these actions are taken.

Methods

The proposed framework to detect accounts acting in coordination on social media is illustrated in Fig. 1. It can be described by four phases:

1. **Behavioral trace extraction:** The starting point of coordination detection should be a conjecture about suspicious behavior. Assuming that authentic users are somewhat independent of each other, we consider a surprising lack of independence as evidence of coordination. The implementation of the framework is guided by a choice of traces that capture such suspicious behavior. For example, if we conjecture that accounts are controlled by an entity with the goal of amplifying the exposure of a disinformation source, we could extract shared URLs as traces. Coordination scenarios may be associated with a few broad categories of suspicious traces:
   (a) If the coordination is based on the *content* being shared, suspicious traces may include words, n-grams, hashtags, media, links, user mentions, etc.
   (b) Coordination could be revealed by spatiotemporal patterns of *activity*. Examples of traces that can reveal suspicious behaviors are timestamps, places, and geocoordinates.
   (c) Accounts could coordinate on the basis of personas or groups. Traces of *identity* descriptors could be used to detect these kinds of coordination: name, handle, profile picture, homepage, account creation date, etc.
   (d) The detection of coordination might require a *combination* of multiple dimensions. For instance, the number of false positive cases could be reduced by a combination of hashtags and temporal signals.
2. **Bipartite network construction:** Once traces of interest are identified, we can build a network of users based on similar behavioral traces. The first step is to build a bipartite network connecting accounts and features extracted from their profiles and messages. In this phase, we may use the behavioral traces as features, or engineer new features derived from the traces. For example, content analysis may yield features based on sentiment, stance, and narrative frames. Temporal features such as hour-of-day and day-of-week could be extrapolated from timestamp metadata. Features could be engineered by aggregating traces, for example by conflating locations into countries or images into color profiles. More complex features could be engineered by considering sets or sequences of traces. The bipartite network may be weighted based on the strength of association between an account and a feature — sharing the same image many times is a stronger signal than sharing it just once. Weights may also incorporate normalization such as IDF to account for popular features; it is not suspicious if many accounts mention the same celebrity.
3. **Projection onto account network:** Depending on the features, the bipartite network may be more or less sparse. If the features are common, it may be useful to prune some low-weight edges that provide noisy signals about coordination among accounts — we cannot state with confidence that an account tweeting twice has been coordinating with anyone due to lack of data. At this stage, the bipartite network is projected onto a network where the account nodes are preserved, and edges are added between nodes based on some similarity measure over the features. This may be done via simple co-occurrence, Jaccard coefficient, cosine similarity, or more sophisticated statistical metrics such as mutual information or χ^2. The edges in the resulting undirected network are weighted by the similarity measure and reveal interaction patterns among the accounts.
4. **Cluster analysis:** Low-weight edges in the account network may be filtered out to focus on the most suspicious interactions. One way to do this is to preserve edges with weight above a threshold, another is to preserve some top percentile of weights. More advanced methods like multi-scaling backbone (Serrano, Boguná, and Vespignani 2009) can also be considered. The final step is to perform cluster analysis on the account network. Many network community detection algorithms can be used, such as connected components, k-core, k-cliques, modularity maximization, and label propagation, among others (Fortunato 2010). The clusters obtained in this final step represent groups of accounts whose actions are likely to be coordinated.

We recommend a manual inspection of the suspicious clusters and their content. Such analysis will provide validation of the method and evidence of whether the coordinated groups are malicious and/or automated.

In the following section we present five case studies, in which we implement the proposed framework to detect coordination. First, we use a dataset of profiles submitted to a social bot classifier to identify coordination based on sharing identities. Then, we leverage a tweet dataset tracking the 2019 Hong Kong protests to detect content (images) coordination. Third, we use a database of tweets related to the 2018 US midterm elections to uncover temporal-content coordination from the usage of hashtags. Fourth, we look at co-retweets in a dataset of content about the White Helmets. And finally, we consider the synchronization of messages to spot pump & dump schemes to manipulate cryptocurrencies.

Case Study 1: Account Handle Sharing

On Twitter and some other social media platforms, although each user account has an immutable ID, many relationships are based on an account handle (called `screen_name`

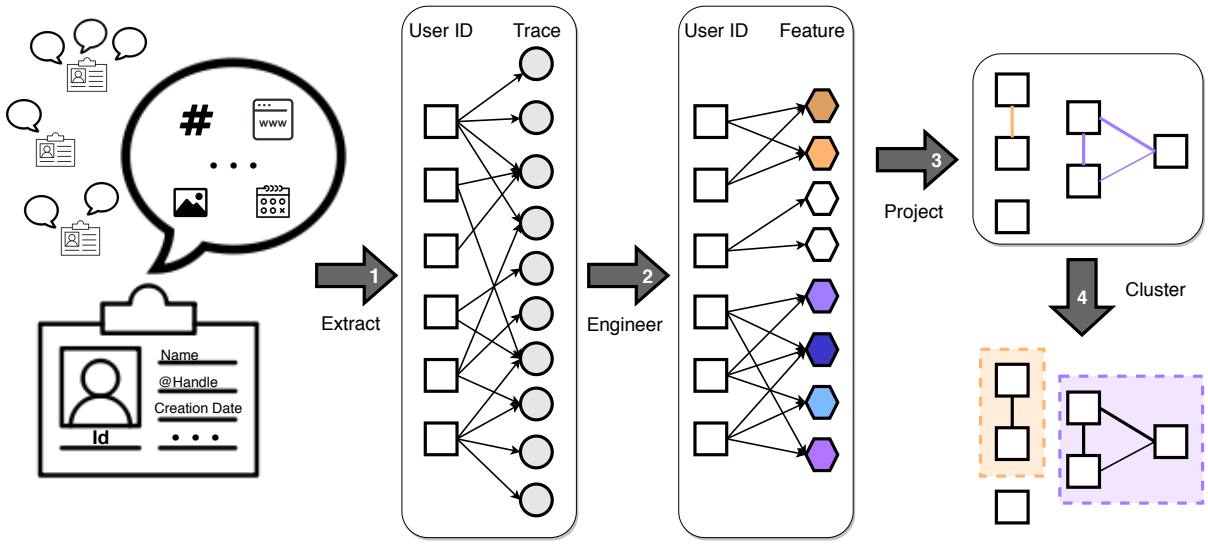

Figure 1: **Coordination Detection Framework**. On the left we see behavioral traces that can be extracted from social media profiles and messages. Four steps described in the text lead to the identification of suspicious clusters of accounts.

on Twitter) that is changeable and in general reusable. An exception is that handles of suspended accounts are not reusable on Twitter. Users may have legitimate reasons for changing handles. However, the possibility of changing and reusing handles exposes users to abuse such as username squatting[1] and impersonation (Mariconti et al. 2017). In a recent example, multiple Twitter handles associated with different personas were used by the same Twitter account to spread the name of the Ukraine whistleblower in the US presidential impeachment case.[2]

For a concrete example of how handle changes can be exploited, consider the following chronological events:

1. `user_1` (named `@super_cat`) follows `user_2` (named `@kittie`) who posts pictures of felines.
2. `user_3` (named `@super_dog`) post pictured of canines.
3. `user_1` tweets mentioning `user_2`: "I love `@kittie`". A mention on Twitter creates a link to the mentioned account profile. Therefore, at time step 3, `user_1`'s tweet is linked to `user_2`'s profile page.
4. `user_2` renames its handle to `@tiger`.
5. `user_3` renames its handle to `@kittie`, reusing `user_2`'s handle.

Even though `user_1`'s social network is unaltered regardless of the name change (`user_1` still follows `user_2`), name changes are not reflected in previous posts, so anyone who clicks on the link at step 3 will be redirected to `user_3`'s profile instead of to `user_2` as originally intended by `user_1`. This type of user squatting, in coordination with multiple accounts, can be used to promote entities, run "follow-back" campaigns, infiltrate communities, or even promote polarization (Mariconti et al. 2017). Since social media posts are often indexed by search engines, these manipulations can be used to promote content beyond social media boundaries.

To detect this kind of coordination on Twitter, we applied our framework using identity traces, namely Twitter handles. We started from a log of requests to Botometer (botometer.org), a social bot detection service of the Indiana University Observatory on Social Media (Yang et al. 2019). Each log record consists of a timestamp, the Twitter `user_id` and handle, and the bot score. For this case study, we analyzed 54 million records from February 2017 to April 2019, containing 1.8 million unique accounts (each with a distinct `user_id`) and 1.9 million handles.

Coordination Detection

We create a bipartite network of suspicious handles and accounts. We consider a handle suspicious if it is shared by at least two accounts, and an account suspicious when it has taken at least one suspicious handle. To detect the suspicious groups we project the network, connecting accounts based on the number of times they shared a handle.

Each connected component in the resulting network identifies a cluster of coordinated accounts as well as the set of handles used by them.

Analysis

Fig. 2 shows the handle sharing network. It is a weighted, undirected network with 7,879 nodes (Twitter accounts). We can classify the components into three classes:

1. **Star-like components** capture the major accounts (hub nodes) practicing name squatting and/or hijacking. To confirm this, we analyzed the temporal sequence of handle switches involving star-like components. Typically, a

[1] help.twitter.com/en/rules-and-policies/twitter-username-squatting

[2] www.bloomberg.com/news/articles/2019-12-28/trump-names-ukraine-whistle-blower-in-a-retweet-he-later-deleted

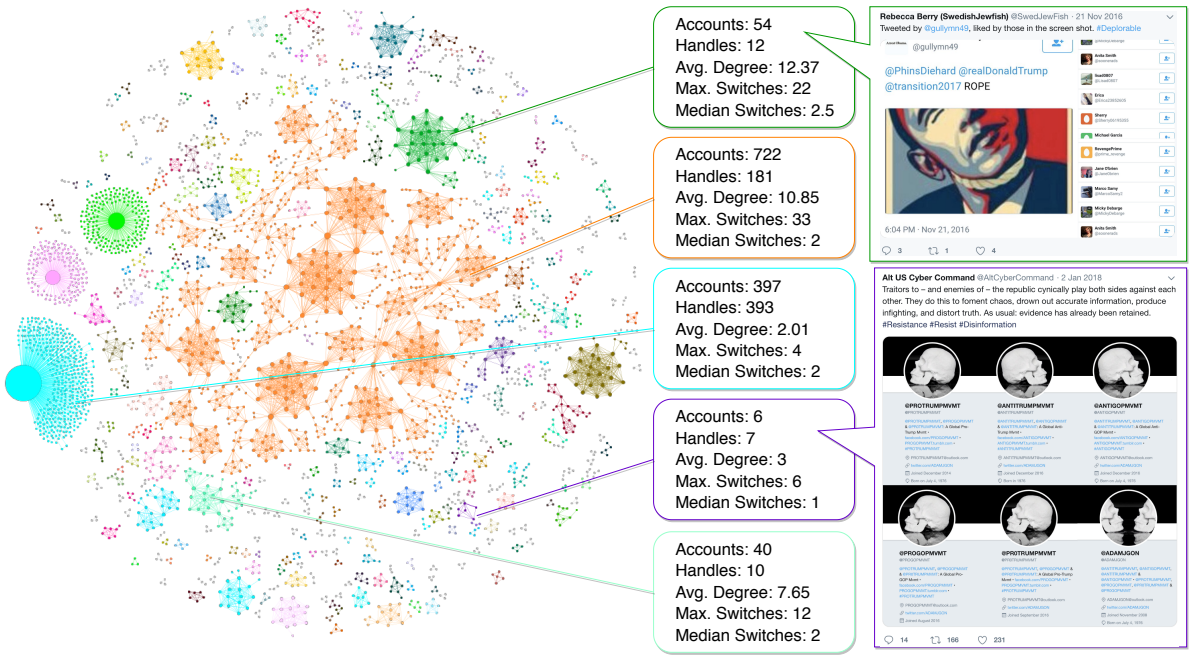

Figure 2: **Handle sharing network.** A node represents a Twitter account and its size is proportional to the number of accounts with which it shares handles. The weight of an edge is the number of unique handles shared by two accounts. Suspicious coordinated groups are identified by connected components, each with a different color. We illustrate the characteristics of a few coordinated groups, namely the number of accounts, number of shared handles, average number of accounts with which handles are shared, and the maximum and median number of times that a handle is switched among accounts. The number of switches is a lower-bound estimate on the basis of our data sample. We also show tweets by independent parties who uncovered the malicious activity of a couple of the coordinated groups, discussed in the main text.

handle switches from an account (presumably the victim) to the hub, and later (presumably after some form of ransom is paid) it switches back from the hub to the original account. These kinds of reciprocal switches occur 12 times more often in stars than any other components.

2. **The giant component** is composed by 722 accounts sharing 181 names (orange group in the center of Fig. 2). Using the Louvain community detection algorithm (Blondel et al. 2008), we can further divide the giant coponent into 13 sub-groups. We suspect they represent temporal clusters corresponding to different coordinated campaigns by the same group. This investigation is left for future study.
3. **Other components** can represent different cases requiring further investigation, as discussed next.

Fig. 2 illustrates a couple of stories about malicious behaviors corresponding to two of the coordinated handle sharing groups, which had previously been uncovered by others. In June 2015, the handle `@GullyMN49` was reported in the news due to an offensive tweet against President Obama.[3] More than one year later, the same handle was still posting similar content.[4] In March 2017, we observed 23 different accounts taking the handle in a 5-day interval. We conjecture that this may have been an attempt to keep the persona created back in 2015 alive and evade suspension by Twitter following reports of abuse to the platform. Currently, the `@GullyMN49` handle is banned but 21 of the accounts are still active.

The second example in Fig. 2 shows a cluster of six accounts sharing seven handles. They have all been suspended since. Interestingly, the cluster was sharing handles that appeared to belong to conflicting political groups, e.g., `@ProTrumpMvmt` and `@AntiTrumpMvmt`. Some of the suspicious accounts kept changing sides over time. Further investigation revealed that these accounts were heavily active; they created the appearance of political fundraising campaigns in an attempt to take money from both sides.

Fig. 3 shows the distributions of bot scores for the suspicious accounts identified by our analysis and an equally-sized sample of non-suspicious accounts from the same dataset. The two distributions can be discriminated by a Kolmogorov-Smirnov test, showing that coordinated accounts are more likely to have higher bot scores ($p < 0.01$). However, most coordinated accounts have low (human-like) scores. Moreover, Fig. 3 shows that as the coordination level increases, the average bot score decreases. These results highlight that in general, bot detection tools are not sufficient to detect this kind of coordination.

[3] minnesota.cbslocal.com/2015/06/03/obama-tweeter-says-posts-cost-him-his-job-2/

[4] twitter.com/SwedJewFish/status/800946662386503680

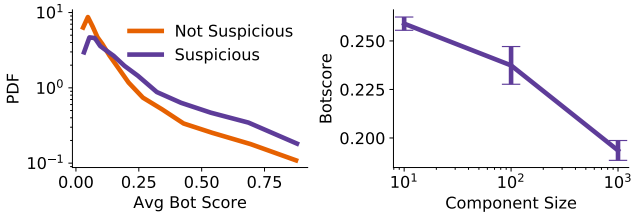

Figure 3: **Bot scores of handle sharing accounts.** Left: Probability distributions of average bot scores for 6,061 coordinated accounts and a sample of the same number of non-suspicious accounts. The number derives from our use of bot scores from the current version of Botometer, collected since May 2018. The dataset may include multiple scores for the same account. Right: Relationship between coordination level (inferred from the size of connected components corresponding to coordination groups) and average bot score. Logarithmic bins are used for component sizes. Error bars are standard errors across bot score measurements in each bin.

Case Study 2: Image Coordination

Images constitute a large portion of the content on social media. A group of accounts posting many of the same or similar images may reveal suspicious coordinated behavior. In this case study, we identify such groups on Twitter in the context of the 2019 Hong Kong protest movement by leveraging media images as content traces.

The dataset used in this case study was collected between August 31 and September 30, 2019 using the BotSlayer tool developed at the Indiana University Observatory on Social Media (Hui et al. 2019). We focus on tweets that contain one or more images, and remove all retweets to avoid trivial replications of the same images. We further exclude users who tweeted less than three images to minimize noise in the network projection phase. The remaining 2,945 users generated 31,772 tweets with images.

Coordination Detection

Every time an image is posted, it is assigned a different URL. Therefore detecting identical or similar images is not as simple as comparing URLs; it is necessary to analyze the actual image content. We represent each image by its RGB color histogram, binning each channel into 128 intervals and resulting in a 384-dimensional vector. The binned histograms allow for matching variants: images with the same vector are either identical or similar, and correspond to the same feature. We then construct a bipartite network of accounts and image features by linking accounts with the vectors of their shared images. The network edges are weighted by their corresponding tweet frequencies.

We perform a projection of the bipartite network to obtain a weighted account coordination network. The weights of the edges can be derived from any similarity measure calculated using the bipartite network weights; in this case study we adopt the Jaccard similarity.

We select the edges with the largest 1% of the weights to focus on reliable coordination relations. Excluding the singletons (accounts with insufficient evidence of coordination), we rank the connected components of the network by size. Next, let us focus on the largest components.

Analysis

Fig. 4 shows the account coordination network. We identify three suspicious clusters involving 317 accounts, posting pro- or anti-protest images. The anti-protest group shares images with Chinese text, targeting Chinese-speaking audiences, while the pro-protest group shares images with English text.

We observe that some of the shared image features correspond to the exact same image, others are slight variants. For example, the 59 image URLs corresponding to the same feature in the pro-protest cluster include slight variations with different brightness and cropping. The same is true for 61 corresponding anti-protest images.

Although this method identifies coordination of accounts, it does not characterize the coordination as malicious or benign, nor as automated or organic. In fact, many of these coordinated accounts behave like authentic users according to Botometer (Yang et al. 2019). These groups are identified because their constituent accounts have circulated the same sets of pictorial content significantly more often than the rest of the population.

Case Study 3: Hashtag Sequences

A key element of a disinformation campaign is an ample audience to influence. To spread beyond one's followers, a malicious actor can use hashtags to target other users who are interested in a topic and may search for related tweets.

If a set of automated accounts were to publish messages using identical text, it would look suspicious and would be easily detected by a platform's anti-spam measures. To minimize the chances of detection, it is easy to imagine a malicious user leveraging a language model (e.g., GPT-2[5]) to paraphrase their messages. Detection could become even harder due to legitimate apps that publish paraphrased text on behalf of a user. An example of this behavior is exhibited by the "Backfire Trump" Twitter app, which tweets to President Trump whenever there is a fatality resulting from gun violence.

However, we conjecture that even paraphrased text is likely to include the same hashtags based on the targets of a coordinated campaign. Therefore, in this case study we explore how to identify coordinated accounts that post highly similar sequences of hashtags across multiple tweets.

We evaluated this approach on the dataset generated for the Bot Electioneering Volume project (Yang, Hui, and Menczer 2019). The dataset consist of tweets containing election-related hashtags, collected using Twitter's filtering API between October and December 2018, around the U.S. midterm election. Prior to applying our framework, we split the dataset into daily intervals to detect when pairs of accounts become coordinated. We filter out accounts with fewer than five tweets or fewer than five unique hashtags.

[5] openai.com/blog/better-language-models/

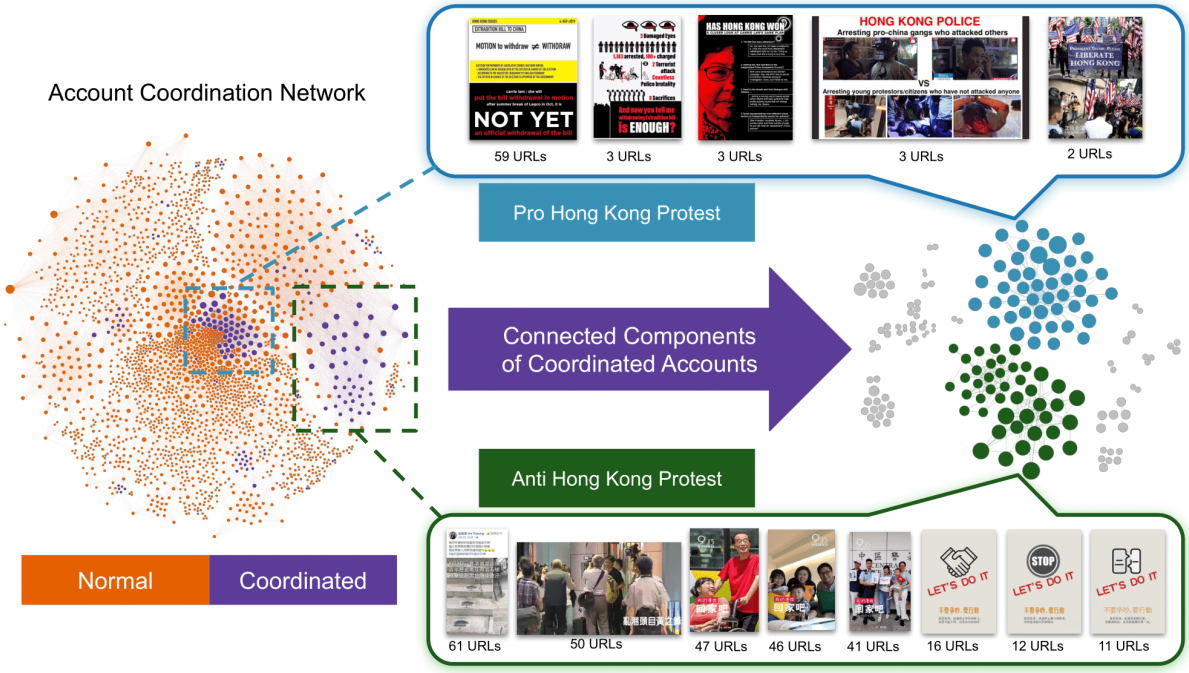

Figure 4: **Account coordination network about Hong Kong protest on Twitter.** Nodes represent user accounts, whose sizes are proportional to their degrees. If two nodes share any image in common, they are connected by an edge weighted by the Jaccard similarity between their features. On the left-hand side, accounts are colored purple if they are in likely coordinated groups, otherwise orange. On the right-hand side we focus on the connected components corresponding to the likely coordinated groups. The three largest components are colored according to the content of their images — one pro- and two anti-protest clusters, in blue and green respectively. We show some exemplar images shared by these groups, along with the corresponding numbers of distinct URLs.

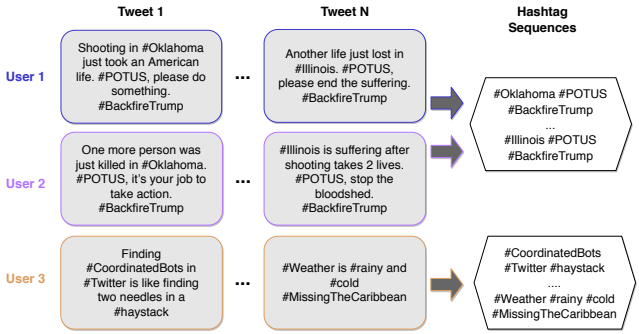

Figure 5: **Hashtag sequence features.** Hashtags and their positions are extracted from tweet metadata. Accounts tweeting the same sequence of hashtags are easily identified.

Coordination Detection

The bipartite network in the first phase consists of accounts in one layer and hashtag traces in the other. This graph is transformed to extract our features of interest, consisting of an ordered sequence of hashtags for each user (Fig. 5). In practice, we can process these bipartite networks using one of two approaches:

1. Any duplicate detection method, such as a dictionary where the keys are ordered sequences of hashtags, allows for accounts with identical sequences to be grouped together, but will not match accounts with similar sequences.
2. A locality-sensitive hashing method, such as MinHash (Broder 1997), allows us the flexibility to query similar users based on the Jaccard similarity between sets of hashtag n-grams.

The duplicate approach might cause us to miss some coordinated accounts due to the arbitrary time cutoff, or occasional mismatches. To construct the inputs for the MinHash method,[6] we use hashtag n-gram strings. We achieve similar results using MinHash with a Jaccard similarity threshold of 0.8 and using a duplicate method; in the following we present results based on the latter.

To project each daily bipartite network onto a hashtag coordination network, we draw an edge between two accounts with matched hashtag sequences. We identify suspicious groups of accounts by removing singleton nodes and then extracting the connected components of the network. Large components are more suspicious, as it is less likely that many accounts post the same hashtag sequences by chance. Fig. 6 illustrates 32 suspicious groups identified on a single day.

[6] github.com/ekzhu/datasketch

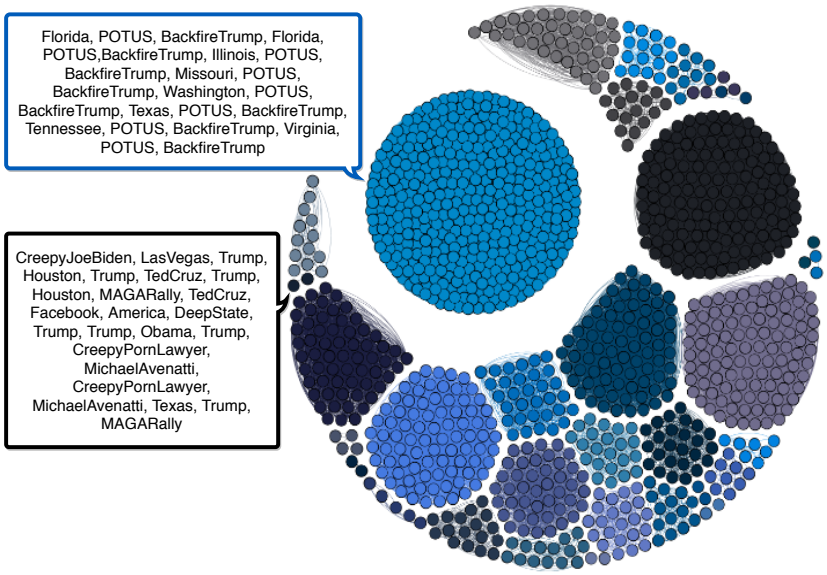

Figure 6: **Hashtag coordination network.** Accounts are represented as nodes, with edges connecting accounts that tweeted the same sequences of hashtags. There are 32 connected components, identified by different colors. The hashtag sequences shared by two of the coordinated groups (the smallest and largest) are shown. This network is based on tweets from October 22, 2018.

Analysis

Using our approach, we identified 617 daily instances of coordination carried out by 1,809 unique accounts across the 3-month period. The largest coordinated group on a single day consisted of 1,175 accounts — the largest component shown in Fig. 6 is a subset of this group. The smallest group consisted of just a pair of accounts.

We observe that many of the coordinated accounts hijack hashtags to amplify their reach or use Twitter apps that tweet on behalf of an account. The latter is the case of the largest coordination group.

Given that coordination can occur over multiple groups of accounts and that these groups can evolve over time, it may be desirable to merge the daily networks to reveal more complex types of coordination among different groups. One could also use different time resolutions to build the network; shorter intervals enable more matches but also increase the noise, whereas longer intervals will yield fewer matches and fewer false positives.

Case Study 4: Co-Retweets

Amplification of information sources is perhaps the most common form of manipulation. On Twitter, a group of accounts retweeting the same tweets or the same set of accounts may signal coordinated behavior.

We apply the proposed framework to detect coordinated accounts that amplify narratives related to the White Helmets, a volunteer organisation that was targeted by disinformation campaigns during the civil war in Syria.[7] The anonymized dataset in this case study is provided by the DARPA SocialSim project; it includes over 800 thousand retweets by approximately 42 thousand accounts, collected between April 2018 and March 2019.

Coordination Detection

We construct the bipartite network between retweeting accounts and retweeted messages, excluding self-retweets. This network is weighted using TF-IDF to discount the contributions of popular tweets. Each account is therefore represented as a TF-IDF vector of retweeted tweet IDs. The projected co-retweet network is then weighted by the cosine similarity between the account vectors. We finally apply two filters: we retain edges with similarity above 0.9, and with both accounts having at least ten retweets.

Analysis

Fig. 7 shows the co-retweet network, and highlights two groups of coordinated accounts. Accounts in the green and blue subgraphs retweet pro- and anti-White Helmets messages, respectively. The example tweets shown in the figure are no longer publicly available.

Case Study 5: Synchronized Action

"Pump & dump" is a shady scheme where the price of a stock is inflated by simulating a surge in buyer interest through false statements (pump) to sell the cheaply purchased stock at a higher price (dump). Investors are vulnerable to this kind of manipulation because they want to act quickly when acquiring stocks that seem to promise high future profits. By exposing investors to information seemingly from different sources in a short period of time, fraudsters create a false sense of urgency that prompts victims to act.

[7] www.theguardian.com/world/2017/dec/18/syria-white-helmets-conspiracy-theories

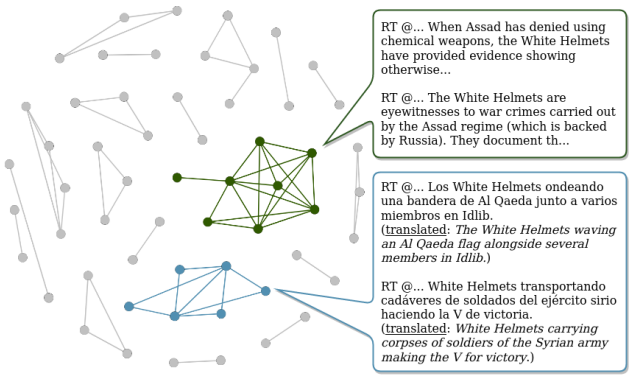

Figure 7: **Co-retweet network.** Two connected components are highlighted with exemplar retweets. Singleton nodes are omitted.

Social media provide fertile ground for this type of scam (Mirtaheri et al. 2019). We investigate the effectiveness of our framework in detecting coordinated cryptocurrency pump & dump campaigns on Twitter. Our analysis leverages an anonymized dataset provided by the DARPA SocialSim project; it includes 3.2 million posts by approximately 900 thousand accounts, collected between January 2017 and January 2019. Here we treat tweets and retweets the same since they all add to the stream of information considered by potential buyers.

Coordination Detection

We hypothesize that coordinated pump & dump campaigns use software to have multiple accounts post pump messages in close temporal proximity. We therefore use tweet timestamps as the behavioral traces of the accounts. These are binned into 30-minute time intervals to obtain features, used to construct the bipartite network of accounts and tweet times. Edges are weighted using TF-IDF. Similar to the previous case, the projected account coordination network is therefore weighted by the cosine similarity between the TF-IDF vectors. We only keep edges with both nodes producing at least three messages and a cosine similarity above 0.9.

Analysis

Fig. 8 shows the synchronized action network. The green subgraphs flag suspicious pump & dump schemes pushing the Indorse Token cryptocurrency, and the blue subgraph corresponds to accounts pumping Bitcoin. Tweet excerpts allegedly state that the accounts have access to business intelligence and hint at the potential rise in coin price. The tweets shown in the figure are no longer publicly available.

We observe other dense clusters in Fig. 8. Inspection reveals that these groups are composed of spam accounts — although they are not examples of pump & dump schemes, they do flag coordinated manipulation.

Discussion

The five case studies presented in this paper are merely illustrations of how our proposed framework can be implemented to find coordination. While our examples are based on Twitter, the framework can in principle be applied to other social media platforms, such as Facebook and Instagram, using any behavioral traces in the data. For instance, the image coordination method can be applied on Instagram, and coordination among Facebook pages can be discovered via the content they share.

While our framework is very general, each implementation involves design decisions and related parameter settings. For example, in our case studies, suspicious groups are identified by simply considering connected components in the account coordination network. This may suffice if the network is sparse, whereas additional filtering (e.g., k-core or clique discovery) may help obtain better results in denser networks.

Our framework aims to identify coordination among user accounts, but it does not characterize the intent or authenticity of the coordination. In some cases, traces and features may be chosen to target specific suspicious behaviors. For example, it is reasonable to assume that large groups of accounts sharing handles have malicious intent. In other cases, coordinated campaigns may be carried out by authentic (human) users with benign intent. For instance, social movement participants use hashtags in a coordinated fashion to raise awareness of their causes.

It is important to minimize false positive errors — spontaneous, organic collective behaviors that may appear as coordinated. For instance, false positives would result by considering identical messages generated by social media share buttons on news websites. Such content similarity alone does not constitute evidence of coordination.

While we use filters based on link weights, feature support, and component size in our case studies, more rigorous and general methods are needed to exclude spurious links that can be attributed to chance. One approach we plan to explore in future work is to design null models for the observed behaviors, which in turn would enable a statistical test to identify meaningful coordination links. For example, one could apply Monte Carlo shuffling of the bipartite network before projection to calculate the p-values associated with each similarity link.

Each of our case studies explores a single behavior in a distinct context. One may wish to consider multiple dimensions of coordination in a single scenario. This would present the challenge of representing interactions through multiplex networks, and/or combining different similarity measures.

Related Work

Academic efforts to detect social bots on social media have been a topic of research since at least 2012. At that time, a platform was developed with the goal of crowd-sourcing their detection (Wang et al. 2012). However, the low cost of deploying social bots meant that automated approaches had to be developed to keep pace with the ever-increasing amount of bots. This led to the development of machine learning models, initially based on the supervised learning paradigm. These models require the use of labeled data describing how both humans and bots behave. Given the

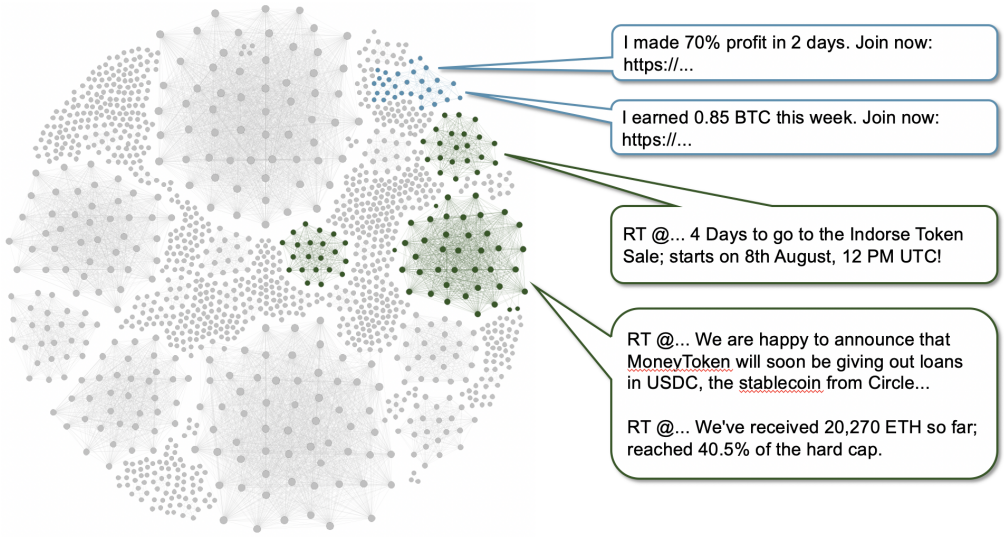

Figure 8: **Time coordination network.** Nodes represent accounts, and two nodes are connected if they post or retweet within the same 30-minute periods. Singletons and dyads are omitted. Four connected components are highlighted with excerpts of a few exemplar suspicious retweets.

lack of ground truth data, researchers created datasets using automated honeypot methods (Lee, Eoff, and Caverlee 2011), human annotation (Varol et al. 2017), or identified suspicious groups of accounts that seemed to be botnets (Echeverria, Besel, and Zhou 2017; Echeverria and Zhou 2017). These datasets have proven to be a good approximation of ground truth, upon which successful detection tools have been built (Davis et al. 2016; Varol et al. 2017; Yang et al. 2019).

One downside of supervised detection methods is that by relying on features from a single account or tweet, they are not as effective at detecting coordinated social bots (Chen and Subramanian 2018; Cresci et al. 2017; Grimme, Assenmacher, and Adam 2018). The detection of coordinated accounts requires a shift toward the unsupervised learning paradigm. Initial applications focused on clustering or community detection algorithms in an attempt to identify similar features among pairs of accounts (Ahmed and Abulaish 2013; Miller et al. 2014). Recent applications look at specific coordination dimensions, such as content or time (Alkhateeb and Agarwal 2019). A method named *Digital DNA* proposed to encode the tweet type or content as a string, which was then used to identify the longest common substring between accounts (Cresci et al. 2016). *Debot* is a time-based method that compares the time series of accounts with the purpose of identifying accounts that tweet in synchrony (Chavoshi, Hamooni, and Mueen 2016). A content-based method proposed by Chen and Subramanian (2018) searches for accounts tweeting similar content. While these approaches work well, they have the disadvantage of considering only one of the many possible coordination dimensions, some of which in the worst-case scenario requires the comparison of a quadratic number of accounts.

The framework proposed here is also unsupervised, but more general in allowing multiple similarity criteria. The quadratic complexity problem is mitigated by imposing sparse network representations for the bipartite relationships between accounts and traces, and consequently the derived feature and projection networks.

Conclusion

In this paper we proposed a network framework to identify coordinated accounts on social media. We presented five case studies demonstrating that our framework can be applied to detect multiple types of coordination on Twitter.

Unlike supervised methods that evaluate the features of individual accounts to estimate the likelihood that an account belongs to some class, say a bot or troll, the objective of our framework is to detect coordinated behaviors at the group level. Therefore, the proposal is intended to complement rather than replace individual-level approaches to counter social media manipulation.

The proposed framework provides a unified way of tackling the detection of coordinated campaigns on social media. As such, it may help advance research in this area by highlighting the similarities and differences between approaches.

We hope that this work will shed light on new techniques that social media companies may use to combat malicious actors, and also empower the general public to become more aware of the threats of modern information ecosystems.

As part of our future work, we intend to incorporate this framework onto BotSlayer (Hui et al. 2019). We believe that the flexibility afforded by this framework, along with the user-friendliness of BotSlayer, will enable an ample audience of users to join our efforts to counter dis- and misinformation on social media.

Acknowledgments.

We thank Kaicheng Yang for helpful discussion. We gratefully acknowledge support from the Knight Foundation, Craig Newmark Philanthropies, and DARPA (contract W911NF-17-C-0094).

References

[Ahmed and Abulaish 2013] Ahmed, F., and Abulaish, M. 2013. A generic statistical approach for spam detection in online social networks. *Computer Communications* 36(10-11):1120–1129.

[Al-khateeb and Agarwal 2019] Al-khateeb, S., and Agarwal, N. 2019. *Deviance in Social Media and Social Cyber Forensics: Uncovering Hidden Relations Using Open Source Information (OSINF)*. Springer.

[Barrett 2019] Barrett, P. M. 2019. Disinformation and the 2020 Election: How the social media industry should prepare. White paper, Center for Business and Human Rights, New York University.

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

[Blondel et al. 2008] 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):P10008.

[Bovet and Makse 2019] Bovet, A., and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10(1):7.

[Broder 1997] Broder, A. Z. 1997. On the resemblance and containment of documents. In *Proc. Compression and Complexity of Sequences (IEEE SEQUENCES)*, 21–29.

[Chavoshi, Hamooni, and Mueen 2016] Chavoshi, N.; Hamooni, H.; and Mueen, A. 2016. Debot: Twitter bot detection via warped correlation. In *Proc. Intl. Conf. on Data Mining (ICDM)*, 817–822.

[Chen and Subramanian 2018] Chen, Z., and Subramanian, D. 2018. An unsupervised approach to detect spam campaigns that use botnets on twitter. *arXiv preprint arXiv:1804.05232*.

[Ciampaglia et al. 2018] Ciampaglia, G. L.; Mantzarlis, A.; Maus, G.; and Menczer, F. 2018. Research challenges of digital misinformation: Toward a trustworthy web. *AI Magazine* 39(1):65–74.

[Cresci et al. 2016] Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2016. Dna-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems* 31(5):58–64.

[Cresci et al. 2017] Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on World Wide Web companion*, 963–972.

[Davis et al. 2016] Davis, C. A.; Varol, O.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. Botornot: A system to evaluate social bots. In *Proc. 25th Intl. Conf. Companion on World Wide Web*, 273–274.

[Deb et al. 2019] Deb, A.; Luceri, L.; Badaway, A.; and Ferrara, E. 2019. Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms. In *Companion Proceedings of The World Wide Web Conference*, 237–247.

[Echeverria and Zhou 2017] Echeverria, J., and Zhou, S. 2017. Discovery, retrieval, and analysis of the'star wars' botnet in twitter. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, 1–8. ACM.

[Echeverria, Besel, and Zhou 2017] Echeverria, J.; Besel, C.; and Zhou, S. 2017. Discovery of the twitter bursty botnet. *Data Science for Cyber-Security*.

[Ferrara et al. 2016] Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM* 59(7):96–104.

[Ferrara 2017] Ferrara, E. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22(8).

[Fortunato 2010] Fortunato, S. 2010. Community detection in graphs. *Physics reports* 486(3-5):75–174.

[Grimme, Assenmacher, and Adam 2018] Grimme, C.; Assenmacher, D.; and Adam, L. 2018. Changing perspectives: Is it sufficient to detect social bots? In *Proc. Intl. Conf. on Social Computing and Social Media*, 445–461.

[Grinberg et al. 2019] Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363(6425):374–378.

[Hills 2019] Hills, T. T. 2019. The dark side of information proliferation. *Perspectives on Psychological Science* 14(3):323–330.

[Hui et al. 2019] Hui, P.-M.; Yang, K.-C.; Torres-Lugo, C.; Monroe, Z.; McCarty, M.; Serrette, B.; Pentchev, V.; and Menczer, F. 2019. Botslayer: real-time detection of bot amplification on twitter. *Journal of Open Source Software* 4(42):1706.

[Jowett and O'Donnell 2018] Jowett, G., and O'Donnell, V. 2018. *Propaganda & persuasion*. SAGE Publications, seventh edition.

[Lazer et al. 2018] Lazer, D.; Baum, M.; Benkler, Y.; Berinsky, A.; Greenhill, K.; Menczer, F.; Metzger, M.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S.; Sunstein, C.; Thorson, E.; Watts, D.; and Zittrain, J. 2018. The science of fake news. *Science* 359(6380):1094–1096.

[Lee, Eoff, and Caverlee 2011] Lee, K.; Eoff, B. D.; and Caverlee, J. 2011. Seven months with the devils: A long-term study of content polluters on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.

[Mariconti et al. 2017] Mariconti, E.; Onaolapo, J.; Ahmad, S. S.; Nikiforou, N.; Egele, M.; Nikiforakis, N.; and Stringhini, G. 2017. What's in a name? Understanding profile name reuse on Twitter. In *Proc. 26th International World Wide Web Conference*, 1161–1170.

[Miller et al. 2014] Miller, Z.; Dickinson, B.; Deitrick, W.; Hu, W.; and Wang, A. H. 2014. Twitter spammer detection using data stream clustering. *Information Sciences* 260:64–73.

[Mirtaheri et al. 2019] Mirtaheri, M.; Abu-El-Haija, S.; Morstatter, F.; Steeg, G. V.; and Galstyan, A. 2019. Identifying and analyzing cryptocurrency manipulations in social media. *arXiv preprint arXiv:1902.03110*.

[Serrano, Boguná, and Vespignani 2009] Serrano, M. Á.; Boguná, M.; and Vespignani, A. 2009. Extracting the multiscale backbone of complex weighted networks. *PNAS* 106(16):6483–6488.

[Shao et al. 2018] Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K. C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature Communications* 9(1):4787.

[Stella, Ferrara, and De Domenico 2018] Stella, M.; Ferrara, E.; and De Domenico, M. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *PNAS* 115(49):12435–12440.

[Varol et al. 2017] Varol, O.; Ferrara, E.; Davis, C. A.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection,

estimation, and characterization. In *Proc. 11th Intl. AAAI Conf. on Web and Social Media*.

[Vosoughi, Roy, and Aral 2018] Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

[Wang et al. 2012] Wang, G.; Mohanlal, M.; Wilson, C.; Wang, X.; Metzger, M.; Zheng, H.; and Zhao, B. Y. 2012. Social turing tests: Crowdsourcing sybil detection. *arXiv preprint arXiv:1205.3856*.

[Weng et al. 2012] Weng, L.; Flammini, A.; Vespignani, A.; and Menczer, F. 2012. Competition among memes in a world with limited attention. *Scientific Reports* 2(1):335.

[Yang et al. 2019] Yang, K.-C.; Varol, O.; Davis, C. A.; Ferrara, E.; Flammini, A.; and Menczer, F. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies* 1(1):48–61.

[Yang, Hui, and Menczer 2019] Yang, K.-C.; Hui, P.-M.; and Menczer, F. 2019. Bot electioneering volume: Visualizing social bot activity during elections. In *Companion Proceedings of The 2019 World Wide Web Conference*, 214–217. ACM.