Uncovering Coordinated Networks on Social Media: Methods and Case Studies

Diogo Pacheco,[*1,2] **Pik-Mai Hui,**[*1] **Christopher Torres-Lugo,**[*1]
Bao Tran Truong,[1] **Alessandro Flammini,**[1] **Filippo Menczer**[1]
[1]Observatory on Social Media, Indiana University Bloomington, USA
[2]Department of Computer Science, University of Exeter, UK
d.pacheco@exeter.ac.uk,{huip,torresch,baotruon,aflammin,fil}@iu.edu

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, unsupervised network-based methodology to uncover groups of accounts that are likely coordinated. The proposed method constructs coordination networks based on arbitrary behavioral traces shared among accounts. We present five case studies of influence campaigns, four of which in the diverse contexts of U.S. elections, Hong Kong protests, the Syrian civil war, and cryptocurrency manipulation. In each of these cases, we detect networks of coordinated Twitter accounts by examining their identities, images, hashtag sequences, retweets, or temporal patterns. The proposed approach 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: 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; Pennycook et al. 2019).

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; 2020; Sayyadiharikandeh et al. 2020). 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.

Here we propose an approach to reveal coordinated behaviors among multiple actors, regardless of their automated/organic nature or malicious/benign intent. The idea is to extract features from social media data to build a coordination network, where two accounts have a strong tie if they display unexpectedly similar behaviors. 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 way to detect significant clusters of coordinated accounts. Our main contributions are:

- We present a general approach to detect coordination,

[*] Equal contributions.

which can in principle be applied to any social media platform where data is available. Since the method is completely unsupervised, no labeled training data is required.

- Using Twitter data, we present five case studies by instantiating the approach to detect different types of coordination based on (i) handle changes, (ii) image sharing, (iii) sequential use of hashtags, (iv) co-retweets, and (v) synchronization.
- The case studies illustrate the generality and effectiveness 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.
- We show that coordinated behavior does not necessarily imply automation. In the case studies, we detected a mix of likely bot and human accounts working together in malicious campaigns.
- Code and data are available at github.com/IUNetSci/coordination-detection to reproduce the present results and apply our methodology to other cases.

Related Work

Inauthentic coordination on social media can occur among social bots as well as human-controlled accounts. However, most research to date has focused on detecting social bots (Ferrara et al. 2016). Supervised machine learning models require labeled data describing how both humans and bots behave. Researchers created datasets using automated honeypot methods (Lee, Eoff, and Caverlee 2011), human annotation (Varol et al. 2017), or likely botnets (Echeverria, Besel, and Zhou 2017; Echeverria and Zhou 2017). These datasets have proven useful in training supervised models for bot detection (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 accounts. This limitation has been explored in the context of 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 (Al-khateeb 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). *SynchroTrap* (Cao et al. 2014) and *Debot* (Chavoshi, Hamooni, and Mueen 2016) leverage temporal information to identify clusters of accounts that tweet in synchrony. Content-based methods proposed by Chen and Subramanian (2018) and Giglietto et al. (2020) consider co-sharing of links on Twitter and Facebook, respectively. Timestamp and content similarity were both used to identify coordinated accounts during the 2012 election in South Korea (Keller et al. 2017; 2019).

While these approaches can work well, each is designed to consider only one of the many possible coordination dimensions. Furthermore, they are focused on coordination features that are likely observed among automated accounts; inauthentic coordination among human-controlled accounts is also an important challenge. The unsupervised approach proposed here is more general in allowing multiple similarity criteria that can detect human coordination in addition to bots. As we will show, several of the aforementioned unsupervised methods can be considered as special cases of the methodology proposed here.

Methods

The proposed approach 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 approach 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) Content: if the coordination is based on the *content* being shared, suspicious traces may include words, n-grams, hashtags, media, links, user mentions, etc.

   (b) Activity: coordination could be revealed by spatio-temporal patterns of *activity*. Examples of traces that can reveal suspicious behaviors are timestamps, places, and geo-coordinates.

   (c) Identity: 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, description, profile picture, homepage, account creation date, etc.

   (d) Combination: the detection of coordination might require a *combination* of multiple dimensions. For instance, instead of tracing only which hashtags were used or when accounts were active, as would be the case for a content- or activity-based suspicious trace, one can combine both these traces to have a temporal-content detection approach. The combined version is more restrictive and, therefore, can reduce the number of false positives.

   Once traces of interest are identified, we can build a network of accounts based on similar behavioral traces. Preliminary data cleaning may be applied, filtering nodes with lack of *support* — low activity or few interactions with the chosen traces — because of insufficient evidence to establish their coordination. For example, an account

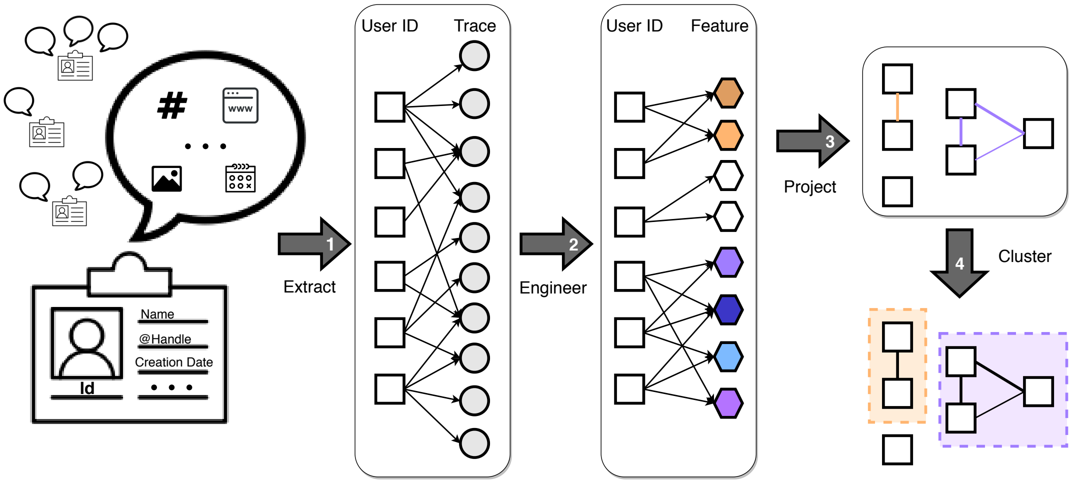


Figure 1: A chart of our proposed coordination detection approach. On the left we see behavioral traces that can be extracted from social media profiles and messages. Four steps described in the text lead to identification of suspicious clusters of accounts.

sharing few images will not allow a reliable calculation of image-based similarity.

2. **Bipartite network construction:** The next 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 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:** 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. The *weight* of an edge in the resulting undirected *coordination network* may be computed via simple co-occurrence, Jaccard coefficient, cosine similarity, or more sophisticated statistical metrics such as mutual information or χ^2. In some cases, every edge in the coordination network is suspicious by construction. In other cases, edges may provide noisy signals about coordination among accounts, leading to false positives. For example, accounts sharing several of the same memes are not necessarily suspicious if those memes are very popular. In these cases, manual curation may be needed to *filter out* low-weight edges in the coordination network to focus on the most suspicious interactions. One way to do this is to preserve edges with a top percentile of weights. The Discussion section presents edge weight distributions is some case studies, illustrating how aggressive filtering allows one to prioritize precision over recall, thus minimizing false positives.

4. **Cluster analysis:** The final step is to *find groups* of accounts whose actions are likely coordinated on the account network. Network community detection algorithms that can be used for this purpose include connected components, k-core, k-cliques, modularity maximization, and label propagation, among others (Fortunato 2010). In the case studies presented here we use connected components because we only consider suspicious edges (by design or by filtering).

In summary, the four phases of the proposed approach to detect coordination are translated into eight actionable steps: (i) formulate a conjecture for suspicious behavior; (ii) choose traces of such behavior, or (iii) engineer features if necessary; (iv) pre-filter the dataset based on support; choose (v) a weight for the bipartite network and (vi) a similarity measure as weight for the account coordination network; (vii) filter out low-weight edges; and finally, (viii) extract the coordinated groups. Although the proposed method is unsupervised and therefore does not required labeled training data, 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 sections we present five case studies, in which we implement the proposed approach to detect coordination through shared identities, images, hashtag sequences, co-retweets, and activity patterns.

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` 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 pictures 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 approach using identity traces, namely Twitter handles. We started from a log of requests to 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. We focus on users with at least ten entries (queries) such that multiple handle changes could be observed. This yielded 54 million records with 1.9 million handles. For further details see Table 1.

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. Therefore no edges are filtered. One could be more restrictive, for example by considering an account suspicious if it has taken more than one suspicious handle. To detect the suspicious clusters we project the network, connecting accounts based on the number of times they shared a handle. This is equivalent to using co-occurrence, the simplest similarity measure. Each connected component in the resulting network identifies a cluster of coordinated accounts as well as the set of handles they shared. Table 1 summarizes the method decisions.

Conjecture	Identities should not be shared
Support filter	Accounts with < 10 records
Trace	Screen name
Eng. trace	No
Bipartite weight	NA, the bipartite is unweighted
Proj. weight	Co-occurrence
Edge filter	No
Clustering	Connected components
Data source	Botometer (Yang et al. 2019)
Data period	Feb 2017–Apr 2019
No. accounts	1,545,892

Table 1: Case study 1 summary

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:

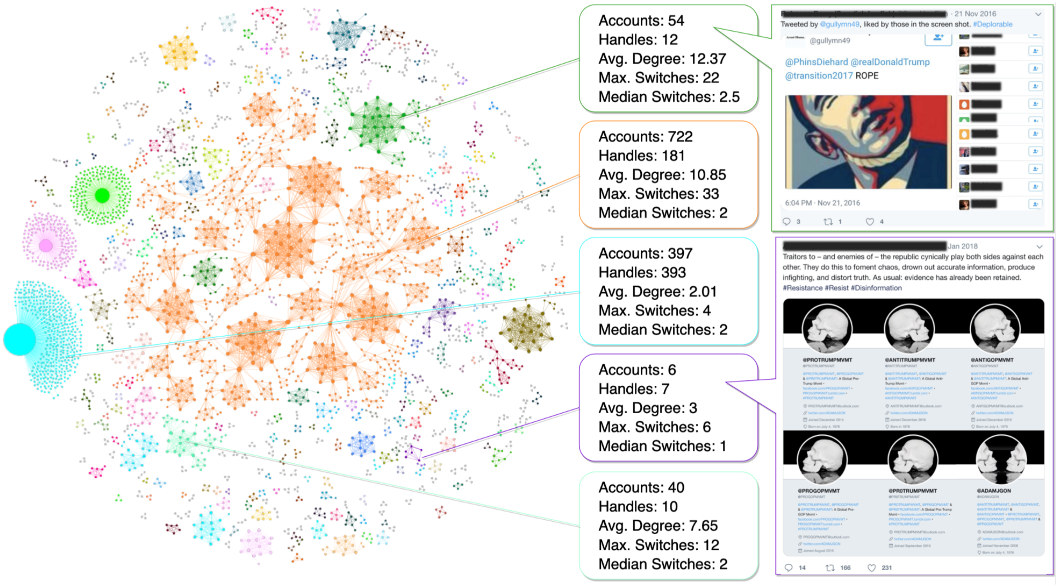
1. **Star-like components** capture the major accounts (hub nodes) practicing name squatting and/or hijacking. To confirm this, we analyzed the temporal sequence of handle switches involving star-like components. Typically, a handle switches from an account (presumably the victim) to the hub, and later (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** includes 722 accounts sharing 181 names (orange group in the center of Fig. 2). Using the Louvain community detection algorithm (Blondel et al. 2008), we further divide the giant component into 13 subgroups. 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 was 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. 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

[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

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



Accounts: 54
Handles: 12
Avg. Degree: 12.37
Max. Switches: 22
Median Switches: 2.5

Accounts: 722
Handles: 181
Avg. Degree: 10.85
Max. Switches: 33
Median Switches: 2

Accounts: 397
Handles: 393
Avg. Degree: 2.01
Max. Switches: 4
Median Switches: 2

Accounts: 6
Handles: 7
Avg. Degree: 3
Max. Switches: 6
Median Switches: 1

Accounts: 40
Handles: 10
Avg. Degree: 7.65
Max. Switches: 12
Median Switches: 2

[redacted] · 21 Nov 2016

Tweeted by @gullymn49, liked by those in the screen shot. #Deplorable

@gullymn49

@PhinsDiehard @realDonaldTrump @transition2017 ROPE

6:04 PM · Nov 21, 2016

3 1 4

[redacted] Jan 2018

Traitors to – and enemies of – the republic cynically play both sides against each other. They do this to foment chaos, drown out accurate information, produce infighting, and distort truth. As usual: evidence has already been retained. #Resistance #Resist #Disinformation

14 166 231

Figure 2: A 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 different colors. 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.

alive and evade suspension by Twitter following reports of abuse to the platform. Currently, the `@GullyMN49` handle is banned but 21 of the 23 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.

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. We used the BotSlayer tool (Hui et al. 2019) to collect tweets matching a couple dozen hashtags related to the protest in six languages, and subsequently downloaded all images and thumbnails in the collected tweets. We focus on 31,772 tweets that contain one or more images, and remove all retweets to avoid trivial replications of the same images. More on the data source can be found in Table 2.

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. While enlarging the bins would give more matches of variants, we want to ensure the space is sparse enough to retain high match precision.

We exclude accounts who tweeted less than five images to reduce noise from insufficient support. One could tune precision and recall by adjusting this support threshold. We set the threshold to maximize precision while maintaining reasonable recall. The sensitivity of precision to the support threshold parameter is analyzed in the Discussion section.

We then construct an unweighted bipartite network of accounts and image features by linking accounts with the vectors of their shared images. We project the bipartite network to obtain a weighted account coordination network, with edge weights computed by the Jaccard coefficient. We consider accounts that are highly similar in sharing the same images as coordinated. To this end, we retain the edges with the largest 1% of the weights (see Fig. 11). Excluding the singletons (accounts with no evidence of coordination), we rank the connected components of the network by size. Table 2 summarizes the method decisions in this case.

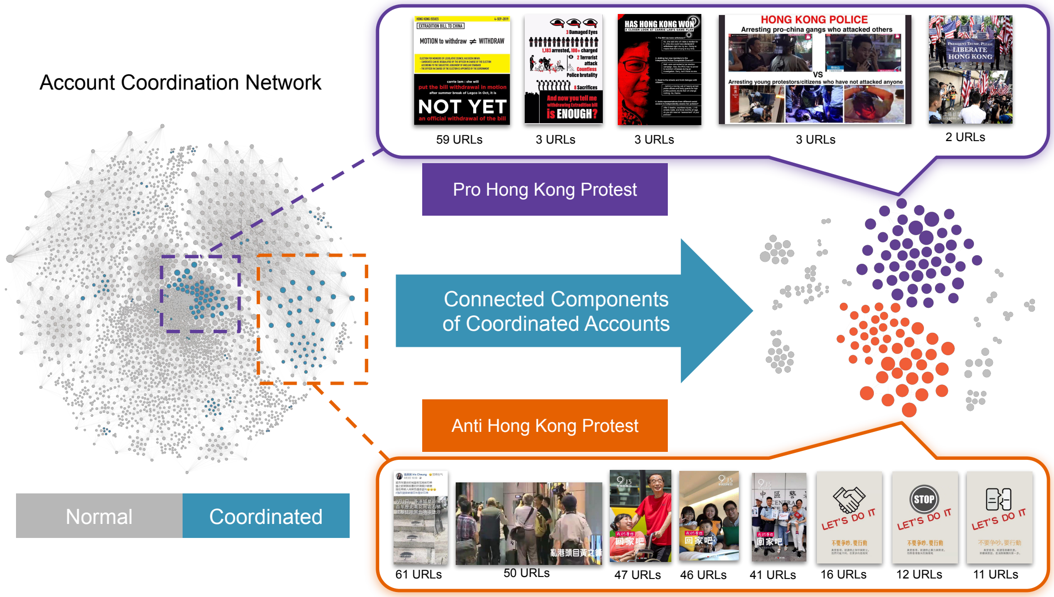


Figure 3: An account coordination network about Hong Kong protest on Twitter. Nodes represent accounts, whose sizes are proportional to their degrees. On the left-hand side, accounts are colored blue if they are likely coordinated, otherwise gray. On the right-hand side we focus on the connected components 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 purple and orange respectively. We show some exemplar images shared by these groups, along with the corresponding numbers of distinct URLs.

Conjecture	Unlike to upload duplicated images
Support filter	Accounts with <5 tweets w/image
Trace	Raw images
Eng. trace	RBG intervals (128 bins on each ch.)
Bipartite weight	NA, the bipartite is unweighted
Proj. weight	Jaccard similarity
Edge filter	Keep top 1% weights
Clustering	Connected components
Data source	BotSlayer (Hui et al. 2019)
Data period	Aug–Sep 2018
No. accounts	2,945

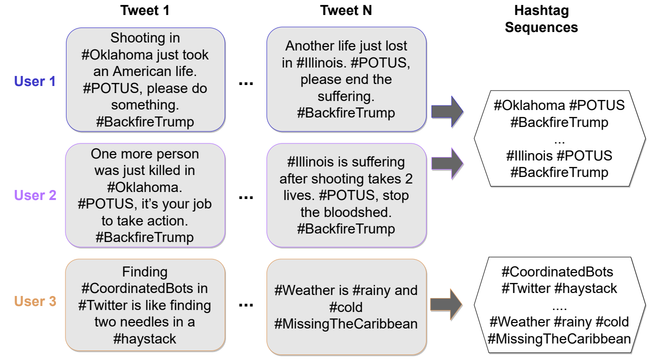
Table 2: Case study 2 summary

Analysis

Fig. 3 shows the account coordination network. We identify three suspicious clusters involving 315 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 humans (see Discussion). 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.



Shooting in #Oklahoma just took an American life. #POTUS, please do something. #BackfireTrump

Another life just lost in #Illinois. #POTUS, please end the suffering. #BackfireTrump

#Oklahoma #POTUS #BackfireTrump ... #Illinois #POTUS #BackfireTrump

One more person was just killed in #Oklahoma. #POTUS, it's your job to take action. #BackfireTrump

#Illinois is suffering after shooting takes 2 lives. #POTUS, stop the bloodshed. #BackfireTrump

Finding #CoordinatedBots in #Twitter is like finding two needles in a #haystack

#Weather is #rainy and #cold #MissingTheCaribbean

#CoordinatedBots #Twitter #haystack .... #Weather #rainy #cold #MissingTheCaribbean

Figure 4: A hashtag sequence features. Hashtags and their positions are extracted from tweet metadata. Accounts tweeting the same sequence of hashtags are easily identified.

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

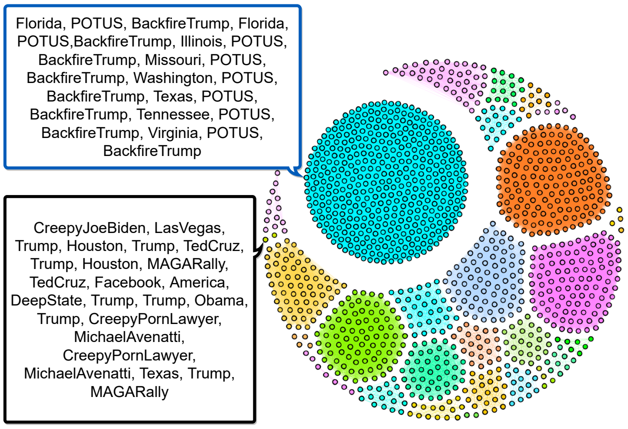

Florida, POTUS, BackfireTrump, Florida, POTUS,BackfireTrump, Illinois, POTUS, BackfireTrump, Missouri, POTUS, BackfireTrump, Washington, POTUS, BackfireTrump, Texas, POTUS, BackfireTrump, Tennessee, POTUS, BackfireTrump, Virginia, POTUS, BackfireTrump
CreepyJoeBiden, LasVegas, Trump, Houston, Trump, TedCruz, Trump, Houston, MAGARally, TedCruz, Facebook, America, DeepState, Trump, Trump, Obama, Trump, CreepyPornLawyer, MichaelAvenatti, CreepyPornLawyer, MichaelAvenatti, Texas, Trump, MAGARally

Figure 5: A 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.

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[4]) to paraphrase their messages. Detection could become harder due to apps that publish paraphrased text on behalf of a user. An example of this behavior is exhibited by the "Backfire Trump" Twitter app, which tweeted to President Trump whenever there was 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 a dataset of original tweets (no retweets) collected around the 2018 U.S. midterm election. More on the data source can be found in Table 3. Prior to applying our framework, we split the dataset into daily intervals to detect when pairs of accounts become coordinated.

Coordination Detection

A data preprocessing step filters out accounts with few tweets and hashtags. The thresholds depend on the time period under evaluation. In this case we use a minimum of five tweets and five unique hashtags over a period of 24 hours to ensure sufficient support for possible coordination. More stringent filtering could be applied to decrease the probability of two accounts producing similar sequences by chance.

Conjecture	Similar large sequence of hashtags
Support filter	At least 5 tweets, 5 hashtags per day
Trace	Hashtags in a tweet
Eng. trace	Ordered sequence of hashtags in a day
Bipartite weight	NA, the bipartite is unweighted
Proj. weight	Co-occurrence
Edge filter	No
Clustering	Connected components
Data source	BEV (Yang, Hui, and Menczer 2019)
Data period	Oct–Dec 2018
No. accounts	59,389,305

Table 3: Case study 3 summary

In this case we engineer features that combine content (hashtags) and activity (timestamps) traces. In particular, we use *ordered sequences* of hashtags for each user (Fig. 4). The bipartite network consists of accounts in one layer and hashtag sequences in the other. In the projection phase, we draw an edge between two accounts with identical hashtag sequences. These edges are unweighted and we do not apply any filtering, based on the assumption that two independent users are unlikely to post identical sequences of five or more hashtags on the same day. We also considered a fuzzy method to match accounts with slightly different sequences and found similar results.

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. Table 3 summarizes the method decisions.

Analysis

We identified 617 daily instances of coordination carried out by 1,809 unique accounts. Fig. 5 illustrates 32 suspicious groups identified on a single day. The largest component consists of 404 nodes that sent a series of tweets through the "Backfire Trump" Twitter application, advocating for stricter gun control laws. This application no longer works. Some of the claims in these tweets are inconsistent with reports by the non-profit Gun Violence Archive. The smallest groups consist of just pairs of accounts. One of these pairs tweeted a link to a now-defunct page promoting bonuses for an online casino. Another pair of accounts promoted a link to a list of candidates for elected office that had been endorsed by the Humane Society Legislative Fund. One could of course use longer time windows and potentially reveal larger coordinated networks. For example, the Backfire Trump cluster in Fig. 5 is part of a larger network of 1,175 accounts.

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. Therefore we focus on retweets in this case study.

We apply the proposed approach to detect coordinated accounts that amplify narratives related to the White Helmets, a volunteer organization that was targeted by disinforma-

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

Conjecture	High overlapping of retweets
Support filter	Accounts with <10 retweets
Trace	Retweeted tweet ID
Eng. trace	No
Bipartite weight	TF-IDF
Proj. weight	Cosine similarity
Edge filter	Keep top 0.5% weights
Clustering	Connected components
Data source	DARPA SocialSim
Data period	Apr 2018–Mar 2019
No. accounts	11,669

Table 4: Case study 4 summary

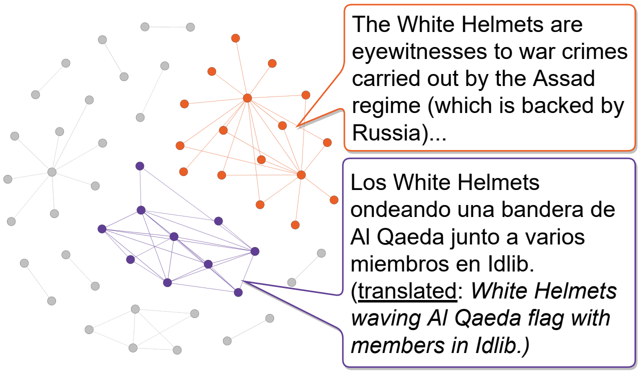

The White Helmets are eyewitnesses to war crimes carried out by the Assad regime (which is backed by Russia)...
Los White Helmets ondeando una bandera de Al Qaeda junto a varios miembros en Idlib. (translated: *White Helmets waving Al Qaeda flag with members in Idlib.)*

Figure 6: A co-retweet network. Two groups are highlighted with exemplar retweets. Singletons are omitted.

tion campaigns during the civil war in Syria.[5] Recent reports identify Russian sources behind these campaigns (Wilson and Starbird 2020). The data was collected from Twitter using English and Arabic keywords. More details about the data can be found in Table 4.

Coordination Detection

We construct the bipartite network between retweeting accounts and retweeted messages, excluding self-retweets and accounts having less than ten 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. Finally, to focus on evidence of potential coordination, we keep only the most suspicious 0.5% of the edges (see Fig. 11). These parameters can be tuned to trade off between precision and recall; the effect of the thresholds on the precision is analyzed in the Discussion section. Table 4 summarizes the method decisions.

Analysis

Fig. 6 shows the co-retweet network, and highlights two groups of coordinated accounts. Accounts in the orange and purple clusters retweet pro- and anti-White Helmets messages, respectively. The example tweets shown in the figure are no longer publicly available.

Case Study 5: Synchronized Actions

"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.

Social media provides fertile grounds for this type of scam (Mirtaheri et al. 2019). We investigate the effectiveness of our approach in detecting coordinated cryptocurrency pump & dump campaigns on Twitter. The data was collected using keywords and cashtags (e.g., $BTC) associated with 25 vulnerable cryptocoins as query terms. We consider both original tweets and retweets because they all add to the stream of information considered by potential buyers. More details on the dataset are found in Table 5.

Coordination Detection

We hypothesize that coordinated pump & dump campaigns use software to have multiple accounts post pump messages in close temporal proximity. Tweet timestamps are therefore used as the behavioral trace of the accounts. The shorter the time interval in which two tweets are posted, the less likely they are to be coincidental. However, short time intervals result in significantly fewer matches and increased computation time. On the other hand, longer (e.g., daily) intervals produce many false positive matches. To balance between these concerns, we use 30-minute time intervals.

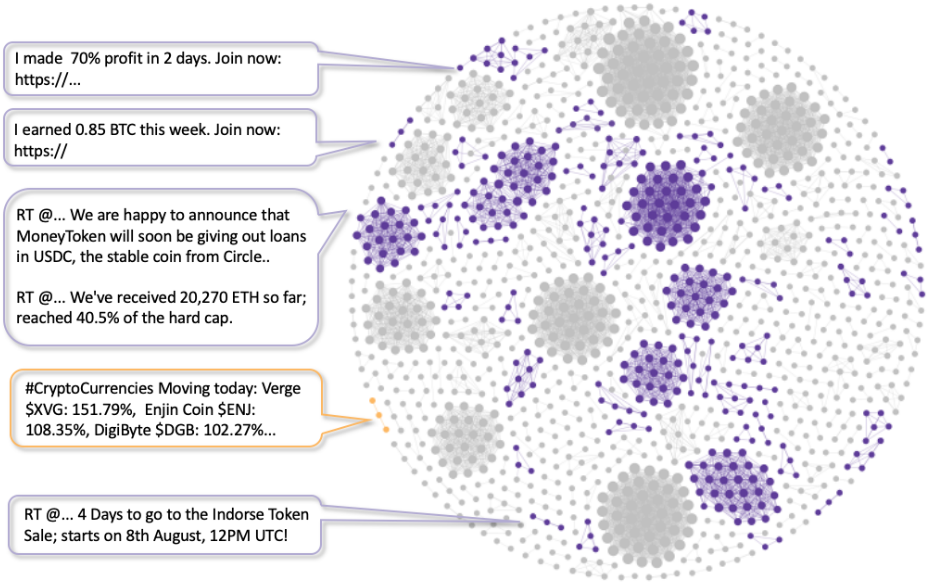
Intuitively, it is likely that any two users would post one or two tweets that fall within any time interval; however, the same is not true for a set of more tweets. To focus on accounts with sufficient support for coordination, we only keep those that post at least eight messages. This specific support threshold value is chosen to minimize false positive matches, as shown in the Discussion section.

The tweets are then binned based on the time interval in which they are posted. These time features are 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 weighted by the cosine similarity between the TF-IDF vectors. Upon manual inspection, we found that many of the tweets being shared in this network are not related to cryptocurrencies, while only a small percentage of edges are about this topic. These edges also have high similarity and yield a strong signal of coordination. Thus, we only preserve the 0.5% of edges with largest cosine similarity (see Fig. 11). Table 5 summarizes the method decisions.

Analysis

Fig. 7 shows the synchronized action network. The connected components in the network are qualitatively analyzed

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



I made 70% profit in 2 days. Join now: https://...

I earned 0.85 BTC this week. Join now: https://

RT @... We are happy to announce that MoneyToken will soon be giving out loans in USDC, the stable coin from Circle..

RT @... We've received 20,270 ETH so far; reached 40.5% of the hard cap.

#CryptoCurrencies Moving today: Verge $XVG: 151.79%, Enjin Coin $ENJ: 108.35%, DigiByte $DGB: 102.27%...

RT @... 4 Days to go to the Indorse Token Sale; starts on 8th August, 12PM UTC!

Figure 7: A time coordination network. Nodes (accounts) are connected if they post or retweet within the same 30-minute periods. Singletons are omitted. Accounts in the purple clusters and the small yellow cluster at 8 o'clock are highly suspicious of running pump & dump schemes. A few tweet excerpts are shown; these tweets are no longer publicly available.

Conjecture	Synchronous activities
Support filter	Accounts with <8 tweets
Trace	Tweet timestamp
Eng. trace	30-minute time intervals
Bipartite weight	TF-IDF
Proj. weight	Cosine similarity
Edge filter	Keep top 0.5% weights
Clustering	Connected components
Data source	DARPA SocialSim
Data period	Jan 2017–Jan 2019
No. accounts	887,239

Table 5: Case study 5 summary

to evaluate precision. The purple subgraphs flag clusters of coordinated accounts where suspicious pump & dump schemes are observed. We find different instances of this scheme for many cryptocurrencies. The excerpts included in Fig. 7 are from tweets pushing the Indorse Token and Bitcoin, respectively. These tweets allegedly state that the accounts have access to business intelligence and hint at the potential rise in coin price.

Changes in stock markets, especially those focusing on short-term trading such as cryptocurrencies, are hard to capture due to market volatility. Furthermore, it is difficult to attribute shifts in price to a single cause, such as pump & dump-related Twitter activities. This makes it difficult to quantitatively validate our results. However, in the week of Dec 15–21, 2017 there were daily uptrends for the coins Verge (XVG), Enjin (ENJ), and DigiByte (DGB). On each day, the prices spiked after large volumes of synchronized tweets commenting on their moving prices. These trends preceded the record price for these coins to date, which was on Dec 23, 2017 for XVG, and Jan 7, 2018 for both ENJ and DGB. The cluster of high-volume accounts pumping these three coins is highlighted in yellow in Fig. 7.

Inspection of the dense clusters shown in gray in Fig. 7 reveals they are composed of spam accounts or coordinated advertisement. Although not examples of pump & dump schemes, they do correctly reflect coordinated manipulation.

Discussion

The five case studies presented in this paper are merely illustrations of how our proposed methodology can be implemented to find coordination. The approach can in principle be applied to other social media platforms besides Twitter. For instance, the image coordination method can be applied on Instagram, and coordination among Facebook pages can be discovered via the content they share.

Several of the unsupervised methods discussed in the Related Work section, just like the five applications of our method presented here, focus on different types of coordination. These methods are therefore not directly comparable. A key contribution of this paper is to provide a flexible and general methodology to describe these different approaches in a unified scheme. For example, Debot (Chavoshi, Hamooni, and Mueen 2016) can be described as a special case of our approach based on a sophisticated temporal hashing scheme preserving dynamic time warping distance (Keogh and Ratanamahatana 2005), while SynchroTrap (Cao et al. 2014) exploits synchronization information by matching actions within time windows. The methods by Giglietto et al. (2020) and Chen and Subramanian (2018)

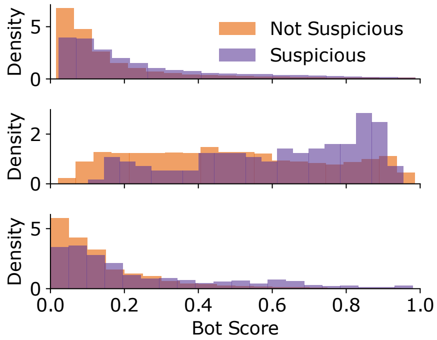


Figure 8: Bot scores of suspicious and non-suspicious accounts. Histograms of bot scores for the suspicious accounts identified by our methodology vs. other accounts. Top, center, and bottom panels represent account handle sharing (Case Study 1), image coordination (Case Study 2), and hashtag sequence (Case Study 3), respectively. Bot scores for Case Study 1 are obtained from version 3 of Botometer (Yang et al. 2019), collected between May 2018 and April 2019. For the other two cases, the bot scores are obtained from BotometerLite (Yang et al. 2020). The datasets may include multiple scores for the same account.

are special cases using similarity based on shared links. The method by Ahmed and Abulaish (2013) uses a contingency table of accounts by features equivalent to our bipartite network. Finally, we explored the use of similar text within short time windows to detect coordinated networks of websites pretending to be independent news sources (Pacheco, Flammini, and Menczer 2020).

Our approach aims to identify coordination between accounts, but it does not characterize the intent or authenticity of the coordination, nor does it allow to discover the underlying mechanisms. An example of malicious intent was highlighted in recent news reports about a coordinated network of teenagers posting false narratives about the election.[6] However, it is important to keep in mind that coordinated campaigns may be carried out by authentic users with benign intent. For instance, social movement participants use hashtags in a coordinated fashion to raise awareness of their cause.

Fig. 8 shows the distributions of bot scores in case studies 1–3. (We are unable to analyze bot scores in cases 4–5 due to anonymization in the datasets.) We observe that while coordinated accounts are more likely to have high bot scores, many coordinated accounts have low (human-like) scores — the majority in two of the three cases. Therefore, detecting social bots is not sufficient to detect coordinated campaigns.

Although the case studies presented here are based on data from diverse sources, they were not designed to inflate the effectiveness of the proposed method, nor to focus on malicious accounts. Fig. 9 shows that the sets of accounts analyzed in case studies 1 and 3 have bot score distributions consistent with those obtained from a random sample of tweets. We note this is not a random sample of accounts — it is biased by account activity. Case Study 2 is the exception; we conjecture that bots were used to post high volumes of images during the Hong Kong protest.

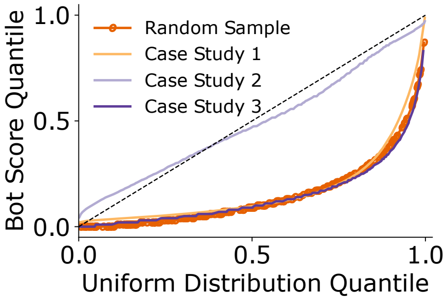


Figure 9: Bot score distributions. Q-Q plots comparing the distributions of bot scores in three case studies against that obtained from a 24-hour, 1% random sample of tweets. The sources of the bot scores are explained in Fig. 8. All distributions are heavily skewed towards lower bot score values (i.e., more humans than bots), except Case Study 2 in which bot scores are higher, with a near-uniform distribution.

While our methodology is very general, each implementation involves design decisions and related parameter settings. These parameters can be tuned to trade off between false positive and false negative errors. In this paper we focus on minimizing false positives — organic collective behaviors that may appear as coordinated. For instance, false positives could result from identical messages generated by social media share buttons on news websites; content similarity alone does not constitute evidence of coordination. One way to avert false positives is to engineer features that are suspicious by construction, as in case studies 1 and 3. Another way is to filter accounts based on support thresholds, filter edges in the coordination network based on similarity, and filter clusters based on characteristics such as density or size. Fig. 10 illustrates how the support threshold affects precision in case studies 2, 4, and 5, based on manual annotations. Case Study 2 shows that support can be selected to maximize precision. In Case Study 4, precision is high irrespective of support because all accounts co-retweeting a high number of tweets are suspicious in that context. In Case Study 5, on the other hand, precision is low because the dataset contains a large portion of tweets unrelated to cryptocurrencies — even though they too are coordinated. Fig. 11 illustrates the choices of edge filtering thresholds in the same case studies.

More rigorous methods could be investigated 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 formal 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.

With the exception of the hashtag sequence feature that

[6] www.washingtonpost.com/politics/turning-point-teens-disinformation-trump/2020/09/15/c84091ae-f20a-11ea-b796-2dd09962649c_story.html

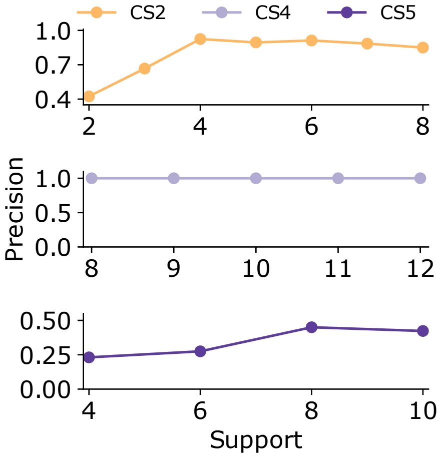


Figure 10: Precision vs support. In three of the case studies, we manually annotated accounts to calculate precision (fraction of accounts in suspicious clusters that are actually coordinated). Precision is plotted against support, namely, number of images (Case Study 2), number of retweets (Case Study 4), and number of tweets (Case Study 5).

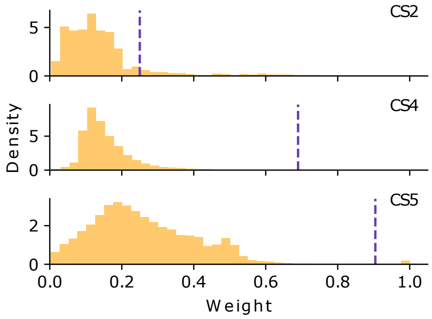


Figure 11: Weight distributions in coordination networks for three case studies. Dashed lines represent edge filters: we retain the edges with top 1% of weights in Case 2 and top 0.5% in Case 4 and 5.

combines content and activity traces, our case studies explore single behaviors in their distinct contexts. Considering multiple dimensions could yield larger groups if the different types of coordination are related. On the other hand, independent types of coordination could be revealed by separate clusters. To illustrate this, combining co-retweets and shared URLs in Case Study 4 yields separate clusters, suggesting distinct and independent coordination campaigns. More work can be done in considering multiple dimensions of coordination in specific scenarios. This presents the challenge of representing interactions through multiplex networks, and/or combining different similarity measures.

Conclusion

In this paper we proposed a network approach to identify coordinated accounts on social media. We presented five case studies demonstrating that our approach 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, our approach aims 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. Our method can also be leveraged to identify and characterize abusive accounts, which in turn can be used to train supervised learning algorithms.

The proposed approach 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.

BotSlayer (Hui et al. 2019) lets users track narratives that are potentially amplified by coordinated campaigns. We plan to incorporate the framework presented in this paper into the BotSlayer system. We believe that the framework's flexibility, combined with the user-friendliness of BotSlayer, will enable a broader community of users to join our efforts in countering disinformation on social media.

Acknowledgments

We thank Kaicheng Yang for helpful discussion. We gratefully acknowledge support from the Knight Foundation, Craig Newmark Philanthropies, DARPA (contract W911NF-17-C-0094), and NSF (NRT award 1735095).

References

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

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

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

Cao, Q.; Yang, X.; Yu, J.; and Palow, C. 2014. Uncovering large groups of active malicious accounts in online social networks. In *Proc. of the 2014 ACM SIGSAC Conf. on Computer and Commns. Security*, 477–488.

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

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, 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, 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 *Proc. of the 26th Intl. Conf. on World Wide Web companion*, 963–972.

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, 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 Proc. of The World Wide Web Conf.*, 237–247.

Echeverria, J., and Zhou, S. 2017. Discovery, retrieval, and analysis of the'star wars' botnet in twitter. In *Proc. of the 2017 IEEE/ACM Intl. Conf. on Adv. in Social Networks Anal. and Mining 2017*, 1–8.

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

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

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

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

Giglietto, F.; Righetti, N.; Rossi, L.; and Marino, G. 2020. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 italian elections. *Information, Communication & Society* 23(6):867–891.

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

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

Keller, F. B.; Schoch, D.; Stier, S.; and Yang, J. 2017. How to manipulate social media: Analyzing political astroturfing using ground truth data from south korea. In *Eleventh Intl. AAAI Conf. on Web and Social Media*.

Keller, F. B.; Schoch, D.; Stier, S.; and Yang, J. 2019. Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political Communication* 1–25.

Keogh, E., and Ratanamahatana, C. A. 2005. Exact indexing of dynamic time warping. *Knowledge and information systems* 7(3):358–386.

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, 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 Conf. on Weblogs and Social Media*.

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 Intl. World Wide Web Conf.*, 1161–1170.

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, 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*.

Pacheco, D.; Flammini, A.; and Menczer, F. 2020. Unveiling coordinated groups behind white helmets disinformation. In *CyberSafety 2020: The 5th Workshop on Computational Methods in Online Misbehavior. Companion Proc. Web Conf. (WWW)*.

Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. 2019. Understanding and reducing the spread of misinformation online. *PsyArXiv preprint: 3n9u8*.

Sayyadiharikandeh, M.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proc. 29th ACM International Conf. on Information & Knowledge Management (CIKM)*, 2725–2732.

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 Comms.* 9(1):4787.

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

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

Wilson, T., and Starbird, K. 2020. Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review* 1(1).

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. *Hum. Beh. and Emerging Techn.* 1(1):48–61.

Yang, K.-C.; Varol, O.; Hui, P.-M.; and Menczer, F. 2020. Scalable and generalizable social bot detection through data selection. In *Proc. 34th AAAI Conf. on Artificial Intelligence (AAAI)*.

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