Detection of Coordination Between State-Linked Actors

Keeley Erhardt(✉) and Alex Pentland

MIT Media Lab, Cambridge, MA 02139, USA
{keeley,pentland}@mit.edu

Abstract. Powerful actors have engaged in information control for centuries, restricting, promoting, or influencing the information environment as it suits their evolving agendas. In the Digital Age, information control has moved online, and information operations now target the online platforms that play a critical role in news engagement and civic debate. In this paper, we use a discrete-time stochastic model to analyze coordinated activity in an online social network, representing the behaviors of accounts as interacting Markov chains. From a dataset of 31,521 tweets posted by 206 accounts, half of which were identified by Twitter as participating in a state-linked information operation, we evaluate the coordination, measured by the apparent influence, between pairs of state-linked compared to unaffiliated accounts. We find that the state-linked actors exhibit more coordination amongst themselves than with the unaffiliated accounts. The degree of coordination between the state-linked accounts is also much higher than the observed coordination between the unaffiliated accounts. Additionally, we find that the account that represented the most coordinated activity in the network *had no followers*, demonstrating the power of our modeling approach to unearth hidden connections even in the absence of explicit network structure.

Keywords: Coordinated activity · Influence modeling · Markov chains

1 Introduction

The rate of online media consumption has dramatically increased and individuals' online social networks (OSNs) are an ever more popular source for news content. State and non-state actors desiring to manipulate the information environment have adapted to this trend, launching information operations targeted at a range of online platforms. Since October 2018, Twitter has publicly identified more than 40 state-linked information operations attributed to over 20 countries targeted at its platform [15]. From 2017 through mid-2021, Facebook similarly took down and reported over 150 information operations originating from more than 50 countries [7]. An information operation can be characterized as coordinated activity aimed at a strategic objective that is fundamentally deceptive in nature [6]. This deception may not necessarily imply explicitly false information (e.g., out-of-context images, agenda-setting, or flooding the information environment with superfluous messaging to confuse and distract [8,14]).

R. Thomson et al. (Eds.): SBP-BRiMS 2022, LNCS 13558, pp. 144–154, 2022.
https://doi.org/10.1007/978-3-031-17114-7_14

Much of the literature in this space focuses on detecting information operations through content-based features [1,13], or network-based approaches [16]. Other studies examine the temporal patterns of post activity [9,10]. In this paper, we choose to instead revisit the "Influence Model", first proposed in [2]. This model is most similar to the temporal approach in [10] but has the advantage of being able to distinguish the directionality of apparent influence rather than producing an undirected account to account coordination graph. The influence model describes the dynamics of networked, interacting Markov chains. A Markov chain is a method for generating a sequence of random variables in which the current value is always probabilistically dependent on only the most recent past value.

In this context, we choose to model individual social accounts as Markov chains with random variables representing post activity for a given user. With the influence model, we can measure the coordination between pairs of accounts based on post activity alone. From these coordination measures, it is possible to quantify hidden connections between accounts and, potentially, inauthentic activity. We focus on the coordination aspect of information operations for a few reasons. First, it alleviates some privacy and bias concerns associated with moderation. Second, an influence modeling approach is more language and media agnostic than content-based alternatives. Third, unlike network-based methods, this approach does not require access to the underlying network structure.

Our contributions are as follows. First, we present a novel application of the influence model for detecting accounts engaged in an information operation. Second, we demonstrate how state-linked accounts can be distinguished from other accounts in a network based on their coordinated post activity alone. And third, we have published an open-source Python library that efficiently implements the influence model and supports the learning of its parameters from sequences of observations.

2 The Influence Model

The "Influence Model" describes the relationships between networked Markov chains in terms of the "influence" chains have on one another. The model is made up of a network of interacting Markov chains each associated with a node in a network. At the network level, nodes are referred to as sites and their connections are described by the stochastic network matrix D. At the local level, each site has an internal Markov chain $\Gamma(A)$ and assumes one of the statuses of $\Gamma(A)$ at any given discrete-time instant. These statuses are represented by a length-m status vector $\boldsymbol{s}$, an indicator vector containing a single 1 in the position corresponding to the present status and 0 everywhere else:

$$\boldsymbol{s}_i'[k] = [0...010...1]. \tag{1}$$

Each chain evolves according to its own status and the statuses of its neighbors. Updating the status of the ith site in the influence model takes place in three stages:

1. The ith site, site$_i$, randomly selects one of its neighbors to be its determining site; site$_j$ is selected with probability d_{ij}.
2. The status of site$_j$ at time k, $\boldsymbol{s}_j[k]$, fixes the probability vector $\boldsymbol{p}_i[k+1]$ that is used in (3) to randomly select the next status of site$_i$.
3. The next status $\boldsymbol{s}_i[k+1]$ is realized according to $\boldsymbol{p}_i[k+1]$.

A state-transition matrix A_{ij} describes how the state-transition probabilities of site$_j$ depend on the previous status of site$_i$. A_{ij} is an $m_i \times m_j$ non-negative matrix with rows summing to 1. A is a matrix with A_{ij} in its (i, j)th block. From the stochastic network matrix D and the state-transition matrix A, one can compute the influence matrix H that describes the “influence" exerted by and on each site in the network. H is given by the generalized Kronecker product of D' and $\{A_{ij}\}$:

$$H = D' \otimes \{A_{ij}\}. \tag{2}$$

The influence model has been applied to a number of problems, ranging from modeling failures in a power grid to recognizing functional roles in meetings [3,5] For more detail on the model, its properties, and applications, we refer readers to [3] and [11].

2.1 The influence Library

In conjunction with this paper, we have published an open-source Python library that provides an efficient implementation of the influence model. The library supports defining new influence models and generating observations through applying the model's evolution equations. We also implement methods to reconstruct an influence model from observations, learning the parameters D, A, and H. Additionally, the project implements the basic, simulated example presented in [4] to familiarize new users with the core concepts of the model.

```python
import numpy as np

leader = Site("leader", np.array([[1], [0]]))
follower = Site("follower", np.array([[0], [1]]))
D = np.array([
    [1, 0],
    [1, 0],
])
A = np.array([
    [.5, .5, 1., 0.],
    [.5, .5, 0., 1.],
    [.5, .5, .5, .5],
    [.5, .5, .5, .5],
])
model = InfluenceModel([leader, follower], D, A)
initial_state = model.get_state_vector()
next(model)
next_state = model.get_state_vector()
```

3 Data

In this paper, we analyze an information operation targeted at Twitter and attributed to the People's Republic of China (PRC). The operation focused on promoting Chinese Communist Party (CCP) narratives related to the treatment of the Uyghur population in Xinjiang. In December 2021, Twitter published a representative sample of accounts and tweets associated with this state-linked information operation, including 31,269 tweets from 2,016 unique accounts [15]. The tweets begin April 20, 2019 and end April 5, 2021. We augment this dataset with "unaffiliated" accounts and tweets, defined as accounts and tweets still permitted on the Twitter platform as of March 2022. Tweets from unaffiliated accounts were collected using the Twitter Search API v2, selecting for tweets posted between April 20, 2019 and April 5, 2021 with at least one of the keywords or hashtags: "xinjiang", "uighur", "uighurs", "uyghur", "uyghurs", "uygur", "uygurs", "uigur", or "uigurs". This search query returned a total of 14,728,582 tweets from 2,665,001 unique accounts.

To ensure a reasonable number of observations (tweets) for each account, we only consider tweets from accounts in the top one percent of accounts by total number of tweets. This means that an account must tweet at least 60 times over the two-year period to be included in the analysis. After downselecting tweets to only those posted by the most prolific accounts, we are left with 10,889 tweets from 103 state-linked accounts and 6,231,955 tweets from 27,003 unaffiliated accounts. From these unaffiliated accounts, we randomly select 103 accounts (corresponding to the number of state-linked accounts) and their associated tweets to analyze. Our final dataset then includes 31,521 tweets from 206 accounts (50% state-linked and 50% unaffiliated).

4 Methodology

Each account in our dataset is represented as a site in a network graph. The two classes of accounts (state-linked and unaffiliated), as well as the true network structure (the follower-following relationships), are not known a priori. Our goal is to quantify the "influence" that determines the status of each site in the network using observed behaviors.

4.1 Constructing Observations

Sites interact by posting messages (tweets), the observed behavior. If a site posts a message at discrete-time instant k, we consider the site "active" at time k. At any given time, a site can be in one of two states, *Active* or *Inactive*. We choose to discretize tweets into 1-h time blocks to ensure enough granurity to differentiate explicitly coordinated behavior from topics that begin to trend, while still ensuring a reasonable number of accounts are likely to be *Active* at any given time. The sequence of observations for each account represents the account's status over time.

Algorithm 1: Constructs a sequence of observations for each site

```
function GetObservations ($posts, accounts, start, end$)
Input  : All posts, accounts, and the time range of interest
Output: Mapping from accounts to observations
delta $\leftarrow$ 1 hour
foreach $account \in accounts$ do
    while $start < end$ do
        k $\leftarrow$ time range from start to delta
        if $account\ posted\ at\ time\ k$ then
            status $\leftarrow$ 1
        else
            status $\leftarrow$ 0
        AddToObservations (account, status)
        start $\leftarrow$ start + delta
```

Given we expect coordinated actors to collectively promote similar narratives, we are less interested in overall post activity and more interested in post activity by topic. We choose a simple definition for “topic": any entity is a topic. Each message contains zero or more entities, defined as hashtags, URLs, or user mentions. We first extract all entities from posts and then construct observation sequences for each entity individually, across all sites. For example, for the entity #hashtag, we only consider an account *Active* if the account posts a message that includes #hashtag. We exclude any entities that were used as search terms in collecting accounts from the Twitter API. And, we normalize URLs by stripping the protocol, subdomain(s), and any query parameters.

4.2 Learning the State-Transition Matrices

In the influence model, the status of each site varies over time based on the “influence" of the other sites in the network. This influence is represented in part by the state-transition matrices covered previously. Given sequences of observations for each site, we can reconstruct the state-transition matrices using a maximum-likelihood estimate, similar to the approach in [4]. Each state-transition matrix is 2×2 representing the two possible statuses, *Active* and *Inactive*. If $site_j$ perfectly follows the behavior of $site_i$ (positive coordination), then A_{ij} is the identity matrix. To obtain a scalar coordination measure for each state-transition matrix, we compute the Frobenius inner product of A_{ij} and the identity matrix. The coordination measure can range $[0, 2]$. Zero represents maximum positive coordination, $site_i[k-1] = site_j[k] \, \forall \, k$, and two represents maximum negative coordination, $site_i[k-1] \neq site_j[k] \, \forall \, k$. By averaging these coordination measures across all entities, we can determine the master state-transition matrix for each pair of sites.

5 Results

We find that the accounts engaged in the most coordinated activity are overwhelmingly the accounts controlled by state-linked actors. Additionally, we discover that the accounts at the center of networks of coordination would not have been identifiable through analysis of the more traditional follower-following relationship network (even if it were available), as these accounts predominantly had few to no followers.

5.1 Account Clusters

To assess clusters of accounts with high-levels of coordinated activity, we construct a coordination network from the pairwise coordination measures. A directed edge (i, j) in the coordination network represents that $site_i$ exhibits apparent influence on $site_j$ with an edge weight equal to one minus the coordination measure. We are primarily interested in positive coordination—when an account mimics the behavior of another account—so only create an edge if the coordination measure is less than one (recall that zero corresponds to maximum positive coordination). This filtering means that not all accounts are represented in the coordination network. If an account does not positively “influence" another account and is not itself “influenced", it will be absent. We find that the clusters of accounts with high degrees of coordination are primarily controlled by state-linked actors, and that each cluster is typically made up of all state-linked or all unaffiliated accounts. This corresponds to our intuition that accounts will exhibit differences in the accounts that they coordinate with based on class membership.

We observe differences in how coordination is expressed when we examine the three entity types individually. In all cases, state-linked accounts make up the majority of the accounts engaged in coordinated activity and almost exclusively coordinate with other state-linked accounts. The unaffiliated accounts are most represented in the network through URL shares, potentially due to the rapid rate at which emerging news stories can diffuse through an OSN.

5.2 Coordinated URL Sharing

For the state-linked accounts, an English-language article from Xinhua News Agency, the official state press agency of the PRC, revealed the most coordinated activity. The story condemned sanctions imposed by the United States (US) for alleged human rights violations in Xinjiang. For the unaffiliated accounts, a Chinese-language Facebook post from the Photographic Society Of Hong Kong Media Limited (PSHK Media) describing the “sinicization" of the Uyghur population in Xinjiang by CCP officials revealed the most coordination. The post accused CCP officials of coercing the ethnic, Muslim minority into celebrating a traditional Chinese holiday and consuming pork. Interestingly, Facebook blocks redirects to PSHK Media’s official site from its platform and, as of the writing of this paper, the site appears to have been suspended by its hosting provider.

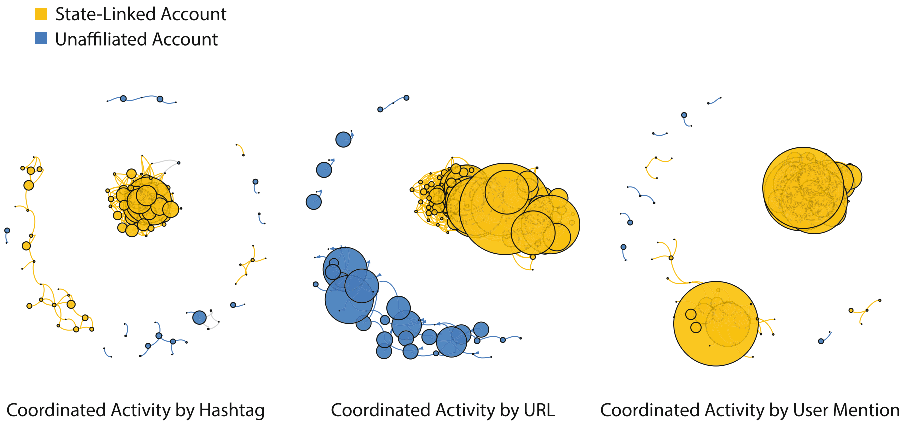

🟨 State-Linked Account
🟦 Unaffiliated Account
Coordinated Activity by Hashtag
Coordinated Activity by URL
Coordinated Activity by User Mention

Fig. 1. The coordination between accounts. An edge is colored yellow if it represents coordinated activity between a state-linked account and another state-linked account, blue if the coordination is from an unaffiliated account to an unaffiliated account, and gray if the edge connects accounts from different classes. The size of a node is scaled by the total "influence" the account exerts. (Color figure online)

5.3 Top Influencer

Averaging across the hashtag, URL, and user mention coordination networks produces a new network consisting of 81 accounts, 75 state-linked accounts and six unaffiliated. In this network, we find that one account exhibits a much higher degree of coordination than any other account. This "top influencer" is state-linked, and exclusively coordinates with other state-linked accounts. Interestingly, this account did not follow any other users and *had no followers.*

The account posted 87 times during the two-years of the PRC information operation. 59 tweets included a hashtag, the most popular being "xinjiang", "xinjiangonline", and "stopxinjiangrumors". 28 included URLs, referencing stories from eight news or informational sites owned by the Chinese government in addition to the People's Daily, a newspaper of record for the CCP. 71 of the user's tweets contained user mentions. The tweets range from argumentative, countering allegations of state-mandated sterilizations and forced labor in Xinjiang, to upbeat, describing the happy, peaceful, and productive lives of people in the region.

5.4 Spike in State-Linked Tweets

On January 19, 2021, Mike Pomepo's last day as US secretary of state, he released a press statement accusing China of “ongoing" genocide perpetuated against the Uyghur population in Xinjiang [12]. The statement appears to have triggered a dramatic increase in tweet activity from state-linked accounts (Fig. 2).

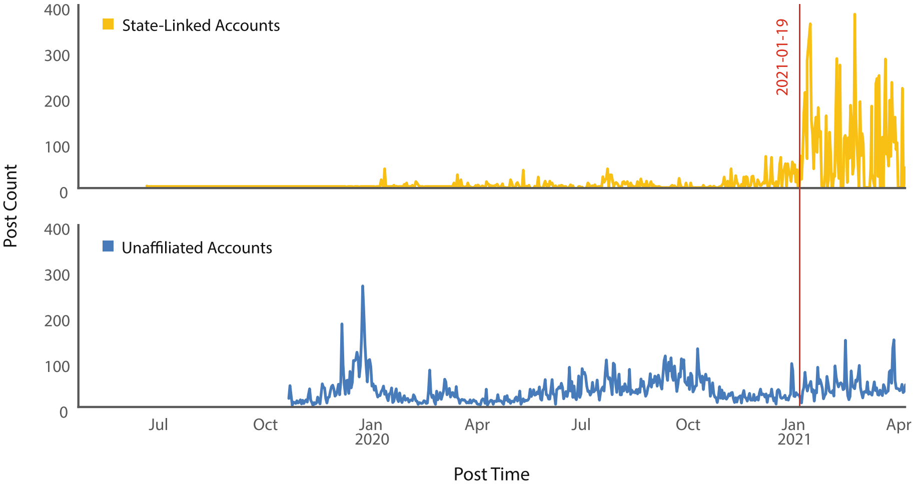
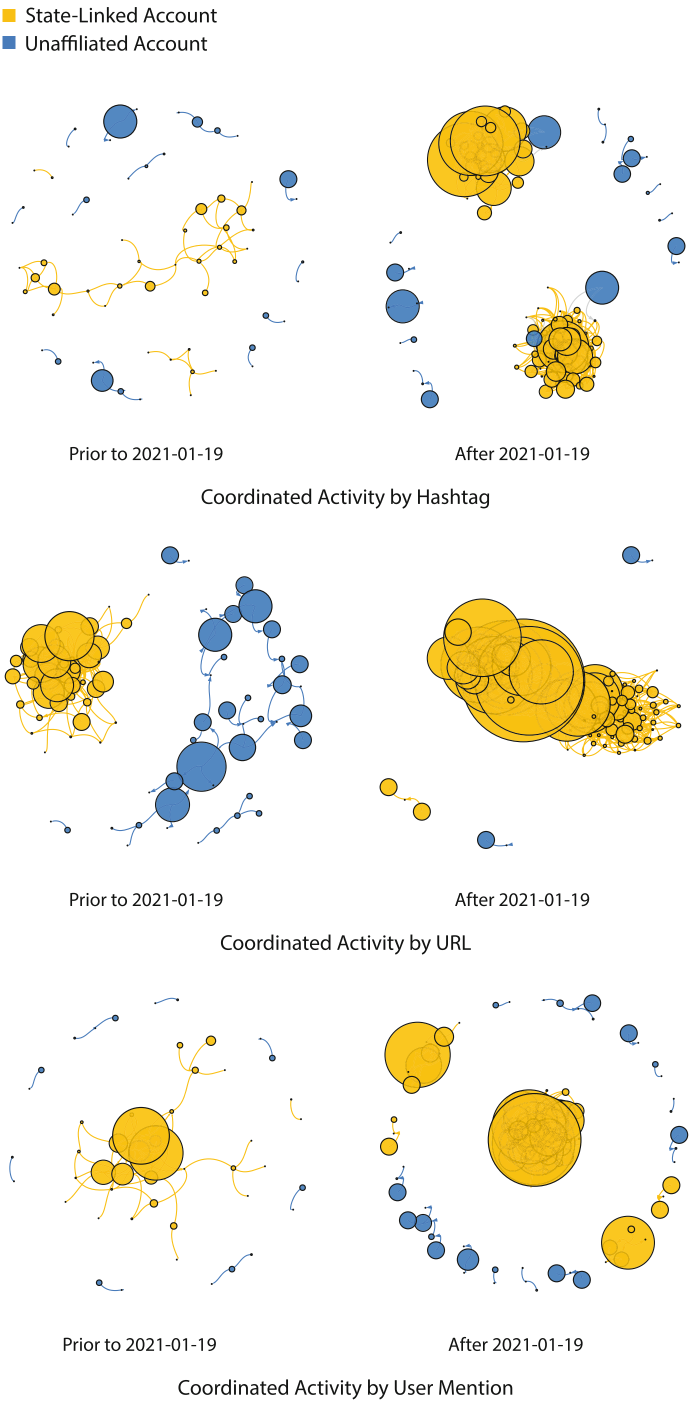


Fig. 2. The tweet count per day across state-linked and unaffiliated accounts.

Given the high volume of state-linked tweets between January and April 2021, we were curious how the coordination network compared between low-activity and high-activity periods. We computed the same networks as in Fig. 1, this time subdividing tweets into two groups: tweets posted before January 19, 2021, and tweets posted after. We find similar results as when we considered the entire two-year period, though the detected coordination between the state-linked accounts is more prevalent following Pompeo's public statement when state-linked tweet volume is highest (Fig. 3).



Prior to 2021-01-19

After 2021-01-19

Coordinated Activity by Hashtag

Prior to 2021-01-19

After 2021-01-19

Coordinated Activity by URL

Prior to 2021-01-19

After 2021-01-19

Coordinated Activity by User Mention

Fig. 3. Coordinated activity between state-linked and unaffiliated accounts before and after Mike Pompeo publicly accuses the PRC of genocide.

6 Discussion and Future Work

We believe that this work represents a unique approach to detecting coordinated information operations, rooted in a well-studied model with broad utility. As an immediate next step, we would like to re-run the analysis on the entirety of the unaffiliated accounts and tweets that we collected, rather than a sample. This will require exploring approaches to minimize the number of sites for which to compute pairwise coordination measures. N sites will always have $N!$ ordered pairs, resulting in a high runtime for large networks. As another area of research, it would be interesting to consider behaviors beyond post activity. And finally, we are interested in exploring how our method performs on additional information operations. Twitter has released dozens of datasets containing accounts and tweets from over 40 state-linked information operations. We would like to see how our model performs on this wide-range of campaigns.

References

1. Alizadeh, M., Shapiro, J.N., Buntain, C., Tucker, J.A.: Content-based features predict social media influence operations. Sci. Adv. **6**(30), eabb5824 (2020)
2. Asavathiratham, C.: The Influence Model: A Tractable Representation for the Dynamics of Networked Markov Chains. Ph.D. thesis, Massachusetts Institute of Technology (2001)
3. Asavathiratham, C., Roy, S., Lesieutre, B., Verghese, G.: The influence model. IEEE Control Syst. Mag. **21**(6), 52–64 (2001)
4. Basu, S., Choudhury, T., Clarkson, B., Pentland, A.: Learning human interactions with the influence model. NIPS (2001)
5. Dong, W., Lepri, B., Cappelletti, A., Pentland, A., Pianesi, F., Zancanaro, M.: Using the influence model to recognize functional roles in meetings. In: Proceedings of the 9th International Conference on Multimodal Interfaces, pp. 271–278 (2007)
6. Erhardt, K., Pentland, A.: Disambiguating disinformation: Extending beyond the veracity of online content. Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (2021)
7. Facebook: Threat report the state of influence operations 2017–2020. https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf. Accessed 24 Jun 2022
8. King, G., Pan, J., Roberts, M.E.: How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. Am. Polit. Sci. Rev. **111**(3), 484–501 (2017)
9. Luceri, L., Giordano, S., Ferrara, E.: Detecting troll behavior via inverse reinforcement learning: a case study of Russian trolls in the 2016 us election. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 417–427 (2020)
10. Magelinski, T., Carley, K.M.: Detecting coordinated behavior in the twitter campaign to reopen America. In: Center for Informed Democracy & Social-Cybersecurity Annual Conference, IDeaS (2020)
11. Pan, W., Dong, W., Cebrian, M., Kim, T., Fowler, J.H., Pentland, A.: Modeling dynamical influence in human interaction: Using data to make better inferences about influence within social systems. IEEE Signal Process. Mag. **29**(2), 77–86 (2012)

12. Pompeo, M.: Determination of the secretary of state on atrocities in xinjiang. https://2017-2021.state.gov/determination-of-the-secretary-of-state-on-atrocities-in-xinjiang/index.html. Accessed 24 Jun 2022
13. Rheault, L., Musulan, A.: Efficient detection of online communities and social bot activity during electoral campaigns. J. Inf. Technol. Politics **18**(3), 324–337 (2021)
14. Starbird, K., Arif, A., Wilson, T.: Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. In: Proceedings of the ACM on Human-Computer Interaction 3(CSCW), pp. 1–26 (2019)
15. Twitter: Transparency report: Information operations. https://transparency.twitter.com/en/reports/information-operations.html. Last accessed 24 Jun 2022
16. Vargas, L., Emami, P., Traynor, P.: On the detection of disinformation campaign activity with network analysis. In: Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop, pp. 133–146 (2020)