Detecting the Hidden Dynamics of Networked Actors Using Temporal Correlations

Citation: Erhardt, Keeley and Albassam, Dina. "Detecting the Hidden Dynamics of Networked Actors Using Temporal Correlations."

As Published: https://doi.org/10.1145/3543873.3587672

Publisher: ACM|Companion Proceedings of the ACM Web Conference 2023

Persistent URL: https://hdl.handle.net/1721.1/150610

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Detecting the Hidden Dynamics of Networked Actors Using Temporal Correlations

Keeley Erhardt
keeley@mit.edu
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Dina Albassam
dalbassam@kacst.edu.sa
King Abdulaziz City for Science and Technology
Riyadh, Saudi Arabia

ABSTRACT

Influence campaigns pose a threat to fact-based reasoning, erode trust in institutions, and tear at the fabric of our society. In the 21st century, influence campaigns have rapidly evolved, taking on new online identities. Many of these propaganda campaigns are persistent and well-resourced, making their identification and removal both hard and expensive. Social media companies have predominantly aimed to counter the threat of online propaganda by prioritizing the moderation of "coordinated inauthentic behavior". This strategy focuses on identifying orchestrated campaigns explicitly intended to deceive, rather than individual social media accounts or posts. In this paper, we study the Twitter footprint of a multi-year influence campaign linked to the Russian government. Drawing from the influence model, a generative model that describes the interactions between networked Markov chains, we demonstrate how temporal correlations in the sequential decision processes of individual social media accounts can reveal coordinated inauthentic activity.

CCS CONCEPTS

• **Networks** → **Network dynamics**; *Topology analysis and generation*; **Network dynamics**; **Topology analysis and generation**; • **Applied computing** → Sociology; Sociology.

KEYWORDS

Markov model, temporality, dynamics, propaganda, social media, hidden influence

ACM Reference Format:
Keeley Erhardt and Dina Albassam. 2023. Detecting the Hidden Dynamics of Networked Actors Using Temporal Correlations. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543873.3587672

1 INTRODUCTION

Social media platforms have become a place where people connect with one another, engage in conversation, and consume news. They are now also being used as a tool for bad actors to amplify the reach of propaganda. Propaganda aims to shape public opinion and influence how people perceive events. Though typically deceptive in nature, propaganda does not always contain explicitly false information. It can be more subtle – presenting images out of context, agenda-setting, or flooding the information environment with irrelevant content to confuse and distract [10, 19]. Given the multitude of ways that propaganda can manifest, it follows that examining content alone may be insufficient for identifying problematic content and combating its spread.

In this paper, we instead study the dynamics of a state-backed influence campaign, showing how temporal correlations in the posting activity of networked accounts can provide insight into which accounts are involved in the coordinated campaign. We show that these coordinating accounts would not be easily detectable through examining each account's posting activity in isolation (i.e. they do not all exhibit simple bot-like behavior), and demonstrate the value of quantifying their causal relationships. Further, we compare the implicit, behavior-driven connections between the coordinated accounts to their explicit, user-mention network, illustrating why implicit relationships are key to understanding the campaign.

The remainder of this paper is organized as follows. Section 2 briefly summarizes related work on detecting and describing information operations and online propaganda. Sections 3 and 4 describe the Twitter dataset that we leverage and the methodology we employ to examine temporal correlations in the network. Section 5 presents our results, and we conclude with a discussion on the implications of this work, limitations, and future directions.

2 RELATED WORK

Prior work on detecting and countering influence campaigns can be broadly categorized as descriptive or detection-focused. Descriptive studies may focus on understanding the goals of individual campaigns and the strategies their coordinators employ. For example, [6] investigated influence campaigns linked to Iran. The authors found that the campaigns primarily targeted the Arab world, promoting third-party websites aligned with Iran's foreign policy objectives. Other studies use insights gleaned from accounts that have been previously identified as engaging in state-linked information operations to find additional, complicit accounts. For example, [14] compared the social networks and interactions of Twitter accounts linked to Turkey's ruling party that were taken down by Twitter's moderation team to the social networks and behaviors of still-active accounts collected over the same time period. A number of accounts that the researchers flagged as suspicious based on their social interactions were later suspended by Twitter. Still others aim to describe campaigns across social media platforms [9, 20].

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587672

The detection-focused papers generally propose some new method for 1) identifying false rumors, low-credibility content, or propagandistic rhetoric in individual posts, 2) classifying social media accounts as authentic or inauthentic based on their profile information and/or behaviors, or 3) flagging coordinated activity. Most focus on either content [1, 13, 16, 18], network structure [15, 17], or temporal activity [3, 7, 12]. A number of studies have tried to detect disinformation campaigns in online social networks (OSNs) by creating user similarity networks and running community detection algorithms to identify clusters of similar users. In [12], the authors built a user similarity network by computing the Jaccard similarity between the hashtags posted by each user and the accounts that they follow and/or retweet. They then weighted users' similarity and ran a community detection algorithm to select the most coordinated sub-networks. [17] similarly used clustering to uncover coordinating groups of accounts. The authors constructed coordination networks based on behavioral traces common between accounts such as shared images, hashtag sequences, and retweets.

Temporal approaches range from constructing multi-view coordination networks based on tweet behaviors (hashtags and URLs) [12], to iteratively building clusters of similar users and performing offline analysis to distinguish between organic and inorganic community development [3], to studying the influence between coordinating accounts using temporal correlations [7]. Our study builds on the work of [7] in three ways. First, by demonstrating the applicability of the approach to a novel dataset, showing that it successfully discriminates coordinated accounts associated with influence campaigns originating from two different countries (Russia and China). Second, by comparing the results to those obtained when considering the explicit user-mention network, which we demonstrate is insufficient for revealing coordination. Third, by accounting for a broader set of tweet behaviors than previously considered – the sharing of similar images, in addition to sharing common hashtags or URLs, or mentioning the same users.

3 DATA

We study a state-backed influence campaign linked to Russia, originally identified and removed by Twitter, and subsequently published as part of their transparency reports.[1] In particular, we use the accompanying dataset released by Guo and Vosoughi that provides background data (negative samples) to accompany the state-linked accounts (positive samples) flagged by Twitter [8]. The negative samples were collected from the Twitter Stream Grab of Internet Archive [2], which draws from Twitter's 1% sample stream of real-time tweets. Guo and Vosoughi then filtered the positive and negative samples to ensure that tweets from both classes were focused on the same topics, measured by the use of shared hashtags. More details on the data collection and filtering methodology are available in their paper. We further filtered out accounts that posted less than ten times in the dataset to reduce the likelihood of seeing coordination based on random chance alone, which resulted in 122,815 tweets from 269 accounts – 174 in the positive class and 95 in the negative class. Summary information on the dataset is shown in Table 1.

4 METHODS

We hypothesize that influence, a measure of causality, is a strong indicator of coordination given the behaviors of coordinating accounts could be expected to be predictive of one another. To uncover these hidden dynamics, we leverage the *influence model*. The influence model describes the relationships between networked Markov chains, and defines a set of evolution equations for how each chain evolves according to its status and the status of its neighbors. We choose the model because it can surface temporal correlations that may not be discernible from a simple time series analysis, revealing hidden coordination between accounts participating in an orchestrated campaign. We use the Python implementation of the influence model presented in [7], and available on PyPI [3]. The influence model was first proposed in [2] and extended in [4], [5], and [11].

For the state-linked campaign, we have a system of A Twitter accounts. We represent all accounts as nodes in a network graph and their labels (positive or negative) are hidden. Each account a can be in one of two possible states at time t, denoted by $a[t] \in \{0, 1\}$. An account's state is dictated by whether the account engaged in a particular tweet behavior b at time t. We define four tweet behaviors: a tweeted hashtag, a tweeted URL, a tweeted image, or a user mention. As an example, if account a tweets a given hashtag at time t, then $a_b[t] = 1$, otherwise 0.

The influence model is a generative model and its parameters can be learned from observations. For each account, we generate an observation vector representing all one hour time blocks from the campaign's start in March 2015 to its end in December 2019. The value at index t is a binary indicator representing the account's state at that time. We are primarily interested in reconstructing the state-transition matrices, which describe how an account's activity in the next time step is influenced by the current activity of the other accounts in the network. We do this using a maximum-likelihood estimate, similar to the approach in [4]. To speed up the learning process, we learn the parameters for each behavior of interest (hashtag, URL, image hash, or user mention) individually.

We can then obtain a scalar influence measure for each pair of accounts by computing the Frobenius inner product of the pair's state-transition matrix and the identity matrix, where zero represents maximum positive coordination (copycat behavior). This influence measure represents the degree to which each account is influenced by each other account in the network. We hypothesize that accounts that are highly influential on each other's online behavior are coordinating offline.

5 RESULTS

We find that across the tweet behaviors studied, the state-linked accounts exhibit more coordination than background accounts.

5.1 Coordination network

To create a coordination network, we represent the Twitter accounts from the positive and negative classes as nodes, where a directed edge exists from User A to User B if User A influences User B (defined as an influence score below 0.5). Each edge is weighted by the number of behaviors that show influence. Figure 1 presents a graph visualization of this coordination network. Node sizes are

[1] https://transparency.twitter.com/

[2] https://archive.org/details/twitterstream

[3] https://pypi.org/ project/influence-model

Table 1: Tweet characteristics in the Russia (May 2020) dataset

Class	Includes a Hashtag	Includes a URL	Mentions a User	Includes an Image
Positive	1.0	.88	.12	.23
Negative	1.0	.77	.02	.06

scaled according to their outdegree, i.e. the number of other nodes that they influence. For clarity, only nodes that exert influence or are themselves influenced are shown. Two insights emerge – the vast majority of accounts in the influence network are from the positive class, and there appear to be two distinct communities of state-linked accounts. Figure 2 shows the adjacency matrix with all accounts from the positive and negative classes. The block structure of the state-linked accounts (positive class) is apparent.

5.2 User-mention network

We can compare the coordination network to the static user-mention network in which a directed edge exists from User A to User B if User A mentions User B. However, as shown in Table 1, users in the dataset rarely mention other users and the users they do mention are outside the network. Figure 3 shows the user-mention network. Similar to the coordination network, it reveals more state-linked accounts than background accounts. However, Figure 4 emphasizes that the network is very sparse, providing limited insight.

6 DISCUSSION

This work demonstrates how an intuitive Markov model can shed light on the hidden dynamics of online activity associated with a state-backed information operation. We show how an approach based on the *influence model* reveals a higher degree of coordination between state-linked accounts compared to background data using temporal tweet behaviors alone. This finding is particularly interesting given that the two classes were chosen to have a high degree of similarity in their behaviors. All tweets contain at least one hashtag, and the tweets that make up each class were selected based on containing a common set of hashtags. Moreover, as shown in Table 1, the ratio of tweets with a URL is similar between the two classes and though more state-linked tweets mention a user or contain an image, the percentage is relatively low for both classes.

We do note a limitation of this approach. The influence model uses a fixed time window, so the results are dependent on the time window selected – in our case, one hour. If User A always shares the same image as User B five minutes after User B first posts the image, we only capture the influence if it takes place in the final 5 minutes of the fixed time window. It further follows that under a uniform distribution of behaviors, the influence model will only find 50% of connections between accounts. Despite this limitation, it appears that even the subset of influence detected is sufficient to highlight coordinating accounts.

We conclude by highlighting two advantages of this approach. First, it requires no contextual understanding of the post content or associated entities. We did not aim to identify which hashtags were associated with state-linked tweets, nor which URLs were being promoted by state-linked accounts. Instead, our approach only uses hashtags, URLs, user mentions, and image hashes as fingerprints

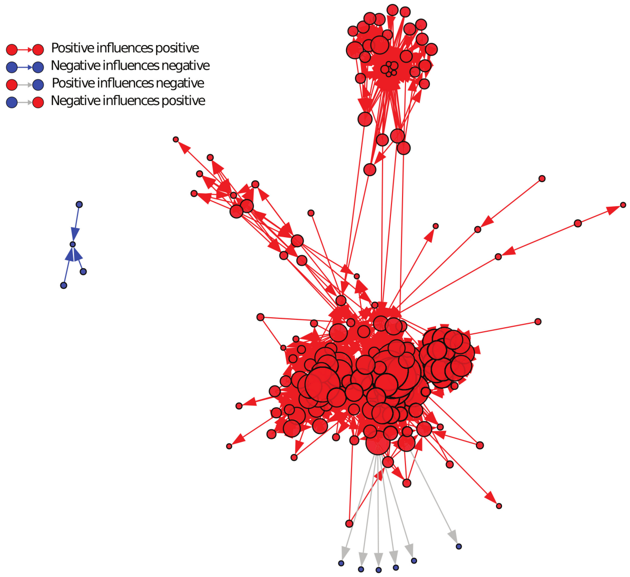
Positive influences positive
Negative influences negative
Positive influences negative
Negative influences positive

Figure 1: A graph view of coordinating accounts in the Russia (May 2020) dataset. Nodes are scaled by their outdegree and displayed using the Fruchterman-Reingold layout, which places adjacent nodes spatially close to one another.

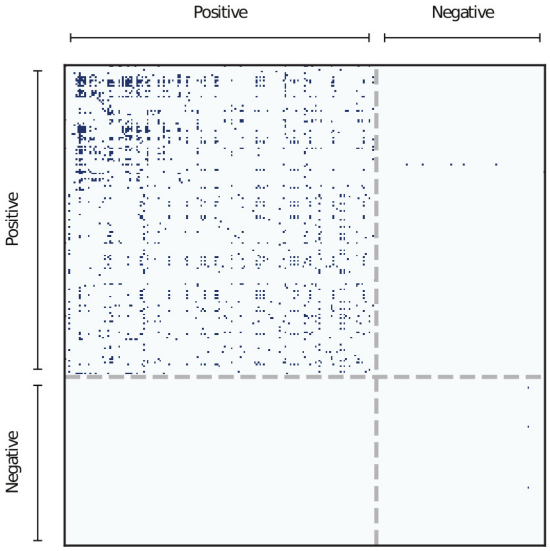

Figure 2: An adjacency matrix m representing state-linked and background accounts from the Russia (May 2020) dataset. $m_{i,j}$ is colored if account a_i exerted influence on account a_j for one or more tweet behaviors.

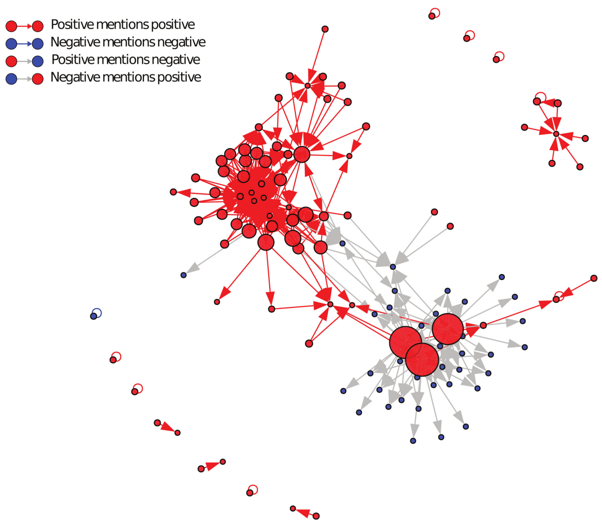
Positive mentions positive
Negative mentions negative
Positive mentions negative
Negative mentions positive

Figure 3: A graph view of the user-mention network in the Russia (May 2020) dataset. Nodes are scaled by their outdegree and displayed using the Fruchterman-Reingold layout, which places adjacent nodes spatially close to one another.

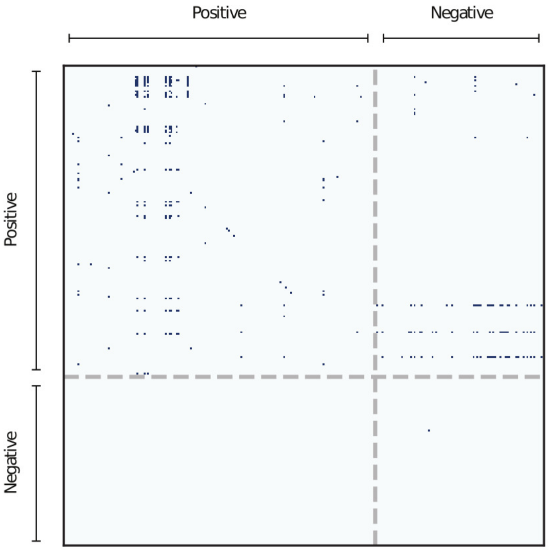

Figure 4: An adjacency matrix m representing state-linked and background accounts from the Russia (May 2020) dataset. $m_{i,j}$ is colored if account a_i mentions account a_j.

to identify the behavior an account is engaging in. Second, the approach works in the absence of an explicit network structure. In this case, we did not have access to the follower network or retweet information. Further, as shown in Figure 4, the user-mention network is too sparse to provide useful insight into the coordinating accounts. Despite these data limitations, we were able to construct a dynamic, implicit network based on behaviors. This network quantifies the causal relationships between users' behaviors, revealing hidden influence in the network and showing promise as a technique to use for identifying influence campaigns in OSNs.

REFERENCES

[1] Meysam Alizadeh, Jacob N Shapiro, Cody Buntain, and Joshua A Tucker. 2020. Content-based features predict social media influence operations. *Science advances* 6, 30 (2020).

[2] Chalee Asavathiratham. 2001. *The influence model: A tractable representation for the dynamics of networked markov chains.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[3] Dennis Assenmacher, Lena Clever, Janina Susanne Pohl, Heike Trautmann, and Christian Grimme. 2020. A two-phase framework for detecting manipulation campaigns in social media. In *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22.* Springer, 201–214.

[4] Sumit Basu, Tanzeem Choudhury, Brian Clarkson, Alex Pentland, et al. 2001. Learning human interactions with the influence model. NIPS.

[5] Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Sandy Pentland, Fabio Pianesi, and Massimo Zancanaro. 2007. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on Multimodal interfaces.* 271–278.

[6] Mona Elswah and Mahsa Alimardani. 2021. Propaganda Chimera: Unpacking the Iranian Perception Information Operations in the Arab World. *Open Information Science* 5, 1 (2021), 163–174.

[7] Keeley Erhardt and Alex Pentland. 2022. Detection of Coordination Between State-Linked Actors. In *Social, Cultural, and Behavioral Modeling: 15th International Conference, SBP-BRiMS 2022, Pittsburgh, PA, USA, September 20–23, 2022, Proceedings.* Springer, 144–154.

[8] Xiaobo Guo and Soroush Vosoughi. 2022. A Large-scale Longitudinal Multimodal Dataset of State-backed Information Operations on Twitter. https://doi.org/10.7910/DVN/NO3I34

[9] NG Kin Wai, Sameera Horawalavithana, and Adriana Iamnitchi. 2021. Multi-platform information operations: Twitter, facebook and youtube against the white helmets. In *Proceedings of The Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM)(SocialSens' 21). Atlanta, USA.*

[10] Gary King, Jennifer Pan, and Margaret E Roberts. 2017. How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American political science review* 111, 3 (2017), 484–501.

[11] Yan Leng, Tara Sowrirajan, and Alex Pentland. 2020. Interpretable Stochastic Block Influence Model: measuring social influence among homophilous communities. *CoRR* (2020). arXiv:2006.01028 https://arxiv.org/abs/2006.01028

[12] Thomas Magelinski and Kathleen M Carley. 2020. Detecting coordinated behavior in the Twitter campaign to Reopen America. In *Center for Informed Democracy and Social-cybersecurity annual conference, IDeaS.*

[13] G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696* (2020).

[14] Maya Merhi, Sarah Rajtmajer, and Dongwon Lee. 2021. Information Operations in Turkey: Manufacturing Resilience with Free Twitter Accounts. *arXiv preprint arXiv:2110.08976* (2021).

[15] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. 2021. Coordinated behavior on social media in 2019 UK general election. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 443–454.

[16] Denisa A Olteanu Roberts. 2021. Multilingual Evidence Retrieval and Fact Verification to Combat Global Disinformation: The Power of Polyglotism. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43.* Springer, 359–367.

[17] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2021. Uncovering coordinated networks on social media: methods and case studies. In *Proceedings of the international AAAI conference on web and social media*, Vol. 15. 455–466.

[18] Steven T Smith, Edward K Kao, Erika D Mackin, Danelle C Shah, Olga Simek, and Donald B Rubin. 2021. Automatic detection of influential actors in disinformation networks. *Proceedings of the National Academy of Sciences* 118, 4 (2021).

[19] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[20] Tom Wilson. 2021. *Understanding the Structure and Dynamics of Multi-platform Information Operations.* Ph. D. Dissertation. University of Washington Libraries.