This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier XX.XXXX/ACCESS.2023.XXXXXXX

Coordinated Behavior in Information Operations on Twitter

LORENZO CIMA[1,2], LORENZO MANNOCCI[1,2], MARCO AVVENUTI[1], MAURIZIO TESCONI[2], STEFANO CRESCI[2]

[1]Department of Information Engineering, University of Pisa, Italy
[2]Institute of Informatics and Telematics (IIT), National Research Council (CNR), Italy

Corresponding author: Lorenzo Cima (e-mail: lorenzo.cima@phd.unipi.it).

This work was partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU – NGEU; by the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n. 871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics"; by the PNRR-M4C2 (PE00000013) "FAIR-Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded under the NextGeneration EU program; and by the Italian Ministry of Education and Research (MUR) in the framework of the FoReLab project (Departments of Excellence).

ABSTRACT Online information operations (IOs) refer to organized attempts to tamper with the regular flow of information and to influence public opinion. Coordinated online behavior is a tactic frequently used by IO perpetrators to boost the spread and outreach of their messages. However, the exploitation of coordinated behavior within large-scale IOs is still largely unexplored. Here, we build a novel dataset comprising around 624K users and 4M tweets to study how online coordination was used in two recent IOs carried out on Twitter. We investigate the interplay between coordinated behavior and IOs with state-of-the-art network science and coordination detection methods, providing evidence that the perpetrators of both IOs were indeed strongly coordinated. Furthermore, we propose quantitative indicators and analyses to study the different patterns of coordination, uncovering a malicious group of users that managed to hold a central position in the discussion network, and others who remained at the periphery of the network, with limited interactions with genuine users. The nuanced results enabled by our analysis provide insights into the strategies, development, and effectiveness of the IOs. Overall, our results demonstrate that the analysis of coordinated behavior in IOs can contribute to safeguarding the integrity of online platforms.

INDEX TERMS Coordinated behavior, information operations, disinformation, Twitter

I. INTRODUCTION

Online social networks (OSNs) are increasingly central to the dissemination of information in our society, allowing users to express their opinions in unprecedented ways. However, most of the information that spreads through OSNs comes from unverified sources. As a consequence, false information is pervasive across OSNs and many attempts are constantly made to manipulate the online information landscape [1]. OSNs like Twitter/X and Facebook refer to *information operations* (IOs) when describing organized communicative activities that attempt to circulate problematically inaccurate or deceptive information [2]. Some of these are planned and carried out by governmental entities, especially in the run up to major political events. The state-backed IOs carried out by the United Arab Emirates, Honduras, China, and Iran are but some recent and notable examples of this kind [3]–[5]. The most infamous case is however related to the activities of the Russian Internet Research Agency (IRA), who set up troll farms to tamper with the 2016 US Presidential election [6]. During many IOs, the organizers purposely entangle orchestrated manipulations with organic grassroots activities, up to the point that genuine audiences may become "willing but unwitting" collaborators that contribute to achieving the IO's goals [2]. Despite the growing relevance of state-sponsored disinformation and IOs, the activity of the different types of agents linked to such efforts has not been thoroughly studied. Indeed, most of the studies focused on automated accounts (i.e., bots) and paid trolls [7], [8]. However, while many IOs rely on these two types of accounts, others also rely on the support of unaware genuine users or on a combination of multiple strategies and agents [2].

Independently of the strategies, tools, and agents used to carry out an IO, a certain degree of coordination among the perpetrators is needed for the IO to spread and obtain

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

a significant outreach, and ultimately to be effective [9]. Because of the relevance of *coordinated behavior* (CB) in large-scale manipulations, scholarly interest on the detection and investigation of CB has arisen. As an example, some methods were recently proposed for detecting coordinated groups of users [10], [11] and for measuring the extent of coordination among them [9]. Nonetheless, a large share of the existing literature on the detection of online manipulation is still based on the analysis and characterization of individual accounts, as done in the well-known tasks of bot and troll detection [6], [12]–[14]. These traditional approaches are however limited in their capacity to contrast complex IOs, given the multitude of different accounts involved in the manipulations. Thus, instead of attempting to accurately classify the nature of each account, which is a notoriously error-prone task [15], [16], it is more favorable to detect and investigate suspicious patterns of coordination among them. An emerging stream of research focused specifically on this task, providing promising results [17]–[19]. Despite these findings however, the study of online coordination is still relatively new and the role played by CB and coordinated groups of accounts in the spread of IOs is still unclear. More in detail, IOs originate from a core of coordinated users and quickly spread through social networks. Effective IOs eventually spread beyond the core perpetrators by reaching and influencing other unaware users. Indeed, multiple groups of coordinated accounts might be involved in the spread of a single IO, each with their own motivations and dynamics of coordination [2]. Thus, another important facet of IOs that is poorly understood is the interplay between maliciously coordinated groups of accounts (e.g., the original promoters of the IO) and the unaware genuine users that are affected by the IO [20]. A common cause for the scarcity of results on the interplay between CB and IOs is the lack of reference datasets that encompass both malicious and legitimate forms of online coordination.

A. CONTRIBUTIONS

To advance research on the adoption of CB in IOs, we built two datasets related to two important state-sponsored IOs detected on Twitter/X: one from Honduras and the other from the United Arab Emirates. Both datasets contain a ground-truth of malicious users who perpetrated the IOs, as well as a large number of unaware users who discussed the main topics of the IOs while those were unfolding (i.e., at the same time of the IOs). Next, we analyzed the two datasets with a state-of-the-art method for detecting coordinated groups of users, thus investigating the interplay between CB and the two IOs. By construction, our datasets encompass both malicious and genuine users, enabling us to compare inauthentic and harmful patterns of coordination with organic ones.

Our results describe the existence of different patterns of coordination. In particular, we uncovered a malicious community of coordinated users that managed to hold a central position in the network and that was strongly connected to genuine users. At the same time, we also discovered small groups of malicious coordinated users that remained at the periphery of the network, with limited interactions with genuine users. Other than shedding light on the relationship between CB and IOs, our results are also useful for understanding the strategies adopted by the malicious users involved in the two IOs, and their effectiveness at influencing unaware users. In addition, our datasets and results might foster future research on the automatic detection of harmful and harmless coordination. Overall, our contributions can be summarized as follows:

- We built two datasets including both malicious and genuine users involved in two large IOs on Twitter/X. Collectively, our datasets contain around 624K users and 4M tweets. These datasets are publicly available for research purposes, as thoroughly explained in Section III.
- We investigated the adoption of CB in two large, yet little studied, IOs. Our results reveal that the perpetrators of both IOs made use of CB, albeit with important differences.
- We studied coordination networks and the resulting patterns of coordination, uncovering commonalities and differences between the two IOs. We discussed these results in terms of the characteristics, strategies, and effectiveness of the IOs.
- We proposed measures to quantify the degree of separation between malicious and other users in a coordination network, and we discussed the implications towards the automatic detection of IOs.

B. SIGNIFICANCE

This work builds upon and improves the ongoing studies on the interplay between coordinated behavior and information operations [3]. To this end, we provide a novel tool to analyze the strategies and the effectiveness of large-scale information manipulation campaigns. Our work also makes important contributions for the characterization of different types of coordinated behavior. As such, it can inform future methods for distinguishing between harmful and harmless coordination, which still represents a largely open problem [20].

II. RELATED WORK

This section surveys previous work in the areas of information operations and coordinated behavior, addressing each in a separate subsection.

A. INFORMATION OPERATIONS

In recent years many works investigated major information operations (IOs), which were studied in different contexts and from different perspectives. For what concerns our two IOs analyzed in the present work, [20] carried out an analysis of the Honduras IO, generating features for distinguish the activity of a disinformation campaign from legitimate Twitter activity, while [3]–[5] studied the IO in the United Arab Emirates. In particular, network science is used to find on each IO coordination patterns [3] and drivers [5]. Further, the authors in [4] proposed a framework to identify bots and coordinated

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

inauthentic behaviors. However, the largest and most studied IO was carried out by the Internet Research Agency (IRA), a Russian company operating in the information sector and involved in multiple information manipulation campaigns [21], [22]. In 2016, the IRA exploited thousands of fake human-operated accounts to influence political events in the US [6] and other countries [23]. Leveraging longitudinal Twitter/X data, [24] qualitatively studied the evolution of the activities and behavior of the IRA accounts, by means of temporal user-hashtag graphs [24]. Pavliuc also applied her methodology to some other IOs detected by Twitter/X through the years.[1] Similarly, the authors of [25] investigated the news shared by IRA accounts and compared them with the ones reported by trusted sources. These analyses aimed at quantifying the amount of disinformation shared by the IRA and their agenda-setting capabilities.

1) Tactics and actors involved in IOs

Although Twitter/X represents by far the most frequent source of data on IOs, these typically unfold and spread across multiple platforms [26]. For this reason, the study of IOs on alternative platforms, or even in multi-platform settings, is particularly valuable. [26] studied the contested online debate about the While Helmets [27], uncovering a network of alternative social media platforms where IO content is produced before being integrated into mainstream platforms [26]. The analysis also revealed the use of the alternative platforms as a way to circumvent possible ''censorship'' actions by the strictly-moderated mainstream platforms [28], [29].

Another thriving area of research on IOs concerns investigating the accounts that take part in, or that are affected by, the manipulations. The authors of [2] analyzed three IOs that employed different tactics to influence public opinion, ultimately distinguishing between fully orchestrated IOs, explicitly coordinated ones, and the emergent, organic behaviors of online crowds [2]. They also highlighted the need to move beyond studies that solely consider bots or trolls [6], [13] as the perpetrators of IOs, by also considering the role played by unaware genuine users. The interplay between malicious accounts and unaware users also has important implications for the detection of IOs. In the case of an IO that tampered with the #BlackLivesMatter protests, it was shown that the malicious perpetrators imitated genuine users to systematically micro-target different audiences, enhancing divisions and undermining trust in authoritative information [30].

2) Automatic detection of IOs

The above descriptive works aimed to ''dissect'' known IOs in order to make sense of their features, tactics, and effects. Instead, others leveraged results on the characterization of past IOs to develop methods for automatically detecting new ones. For example, [31], [32] used standard classification algorithms to analyze the textual content of shared messages, demonstrating its informativeness towards detecting content and users that are part of an IO. Other detection techniques analyze network structures with the goal of identifying groups of users that are involved in an IO [3], [10], [19], [20], and to estimate their influence on the rest of the network [33]. These techniques are typically based on community detection or focal structure analysis methods. Other related literature is about the detection of online campaigns. To this end, [34] tackled the early detection of promoted campaigns with supervised machine learning, by leveraging a combination of features derived from diffusion patterns, content, timing, and user information. Similarly, the authors of [35] released a framework to identify users engaged in spreading conspiracy theories. Others circumvented the scarcity of high-quality ground-truth datasets by adopting unsupervised approaches based on text stream clustering [36].

Independently of the approach, detection results in the majority of previous works are mixed, showing good detection performance for already known IOs, but rather poor generalization capabilities.

3) Our contributions to the IO literature

In spite of the growing body of work investigating recent IOs, we still have a partial understanding of the interplay between the different types of accounts involved therein, and their strategies of coordination. Our present study builds upon recent work that jointly considered multiple coordinated actors involved in IOs [3], but covers new ground by carrying out a full-fledged analysis of coordinated behavior, identifying coordinated communities, and investigating their patterns of coordination. In turn, better knowledge about the coordinated groups of accounts involved in an IO might inform future strategies for promptly detecting unfolding IOs, which still represents an open problem [20]. Moreover, out of all the IOs detected and shared by Twitter/X, the vast majority of works analyzed the one related to the activities of the IRA [2], [6], [23]–[25], [37]–[39]. Other IOs that received some scholarly attention are those perpetrated by Egypt [20], [39], China [31], [39], and Iran [40]. Instead, our present study complements and extends the existing literature by analyzing the IOs carried out by Honduras and the United Arab Emirates, which are still largely unexplored despite their global scale and relevance. In fact, both involved a large number of users and mostly spread English language messages.

B. COORDINATED BEHAVIOR

Since its introduction by Facebook,[2] the analysis of coordinated online behavior (CB) has become frequent among studies on online manipulation. This is because CB underpins the very fabric of information disorder in digital spaces, including the spread of disinformation, online propaganda, and infor-

[1] https://medium.com/swlh/watch-six-decade-long-disinformation-operations-unfold-in-six-minutes-5f69a7e75fb3 (accessed: 11/30/2023)

[2] https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/ (accessed: 11/30/2023)

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

mation operations. For example, CB emerges as one of the features of online propaganda – a form of communication that attempts to achieve a response that furthers the desired intent of the propagandist [41] – as a consequence of the reliance on computational tools such as automation and algorithms to disseminate and amplify discourses for ideological control and manipulation [17], [42]. Similarly, CB is instrumental for the success of large-scale IOs, since it enables the involvement of online communities and the effective spread of the IO narratives [3], [43]. CB can thus serve as a linchpin for comprehending large-scale online information manipulations. Its systematic and thorough analysis is useful for discerning the motives, tactics, and impact of orchestrated efforts, fostering the development of effective countermeasures to safeguard the integrity of online discourse and information ecosystems.

1) Automatic detection of CB

The majority of existing approaches for detecting CB are based on network science. In these works, coordination is often defined as an unexpected or exceptional similarity between the actions of two or more users. User similarity networks are thus built based on common activities between users, and studied, for example by means of community detection algorithms [9]–[11], [44], [45]. The typical output of these methods is a network where coordinated groups of users are highlighted. Such rich networks lend themselves to many subsequent investigations, such as those aimed at distinguishing between genuine and malicious forms of coordination [20], [43]. Some network-based methods do not only detect coordination as in a binary classification task but also quantify the extent of coordination between users, thus providing more nuanced results [9]. Other than with user networks, online coordination was also studied by means of temporal point processes that model user activities on an OSN as the realization of a stochastic process [46], [47] or was studied using a text stream clustering [36]. Another important area of analysis is the coordination among astroturfing or automated agents. Here, state-of-the-art works move beyond the traditional classification of single accounts, by adopting sophisticated pattern recognition approaches aimed at identifying the inorganic coordination behind fake grassroots movements [19], [48], [49] and that distinctive of groups of automated agents (e.g., social bots) [12], [16]. Finally, other works adopted traditional feature engineering approaches to find similarities between users [50], [51], or exploited language processing to find common content alphabets [4], or focused on some specific action such as URL sharing [52].

2) Patterns of CB

Rather than proposing methods for detecting CB, some scholars focused on analyzing the patterns of coordination and the behavior of the coordinated groups of users. As an example, [45] investigated coordinated authentic and inauthentic groups based on their URL-sharing behavior. They traced back to the websites of the posted URLs, assessed their reliability, and finally drew insights into the harmfulness or harmlessness of the CB [45]. The study described in [43] had the similar aim of differentiating harmful versus harmless patterns of coordination. To reach their goal, [43] adopted machine learning methods for identifying the use of propaganda techniques in tweets, and cross-checked their use by coordinated groups of users, labeling as harmful those coordinated communities featuring extensive use of propaganda [43]. Finally, a challenging – and thus frequently overlooked – research question is the one about the effectiveness of coordinated behaviors. Authors in [53] leveraged reconstructed information cascades on Twitter/X to analyze the extent to which non-coordinated users that participated in a cascade were influenced by the coordinated ones, finding that the latter had a significant influence on the former. One of the emerging challenges is the lack of ground-truth information on CB [54], which explains why several works tried to differentiate harmful and harmless coordinated communities after their detection [10], [20], [43], [45], [52], [55].

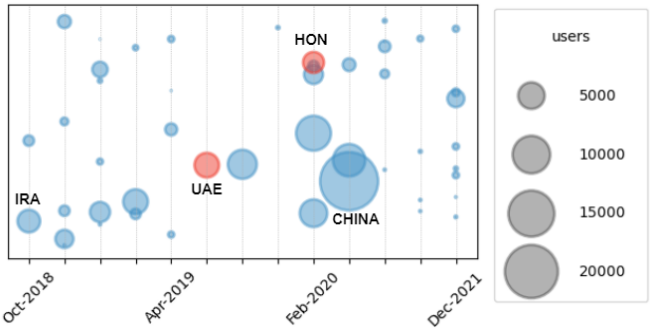


FIGURE 1: Univariate bubble chart of the IOs detected and shared by Twitter/X at the time of writing. Bubbles are chronologically ordered along the *x* axis. Vertical space along the *y* axis results from the application of some jitter to mitigate overplotting. Bubble size is proportional to the number of malicious accounts in the IO. Red bubbles denote the HON and UAE IOs studied in this work.

3) Our contributions to the CB literature

We extend the extant literature on CB by applying a state-of-the-art coordination detection method [11] to two novel datasets, thus yielding new results on the activity of two overlooked threat actors and, more in general, on the exploitation of CB for online information manipulation. Furthermore, our design choices allow for building datasets where malicious coordinated users are known beforehand. We therefore contribute to increasing the limited availability of such datasets [54], thus providing an orthogonal contribution with respect to the majority of the existing works. Finally, our analyses of the interactions between the malicious and the other users involved in an IO, allow us to draw insights into the strategies and effectiveness of the coordinated groups of users, an area for which few results were achieved so far [2].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

III. DATA

A. MALICIOUS USERS

We built our datasets for this work starting from official data provided by Twitter/X's Moderation Research Consortium (TMRC).[3] Since 2018, Twitter/X boosted its transparency efforts in moderation by providing tweets and account information of banned users involved in state-sponsored IOs. Through the years, many datasets were published by TMRC, ranging from tens to thousands of removed users and covering different languages and regions of the world, as sketched in Figure 1. Because of this, Twitter/X's official datasets are often considered as an authoritative ground-truth of malicious users involved in IOs [31], [56], [57]. Out of all the IOs detected and removed by Twitter/X, we focused on one promoted by the government of Honduras between 2019 and 2020 (HON) and on another one promoted by the United Arab Emirates in 2019 (UAE). Both are red-colored in Figure 1. This choice allows us to focus and analyze two recent and large – yet essentially unstudied – IOs that involved thousands of malicious users. Notably, while both HON and UAE encompass a large number of English tweets, which eases content analyses, the two IOs employed different strategies. According to Twitter/X, HON includes 3,104 inauthentic users that showed fake grassroots support for the Honduras President Juan Orlando Hernández, in office from January 2014 to January 2022, by artificially boosting the popularity and engagement of his tweets. They did so by mass-retweeting the President's account (*@JuanOrlandoH*) from a single IP range in Honduras. Instead, UAE involves 4,248 users operating uniquely from the United Arab Emirates. Their activity was mainly directed against Qatar and Yemen by employing false personae tweeting about controversial and divisive regional issues, such as the Yemeni Civil War [58] and the Houthi Movement [59]. For each removed IO, Twitter/X releases anonymized versions of all tweets published by all banned users since their creation (i.e., their full timelines). While analyzing full timelines may enable interesting longitudinal analyses [24], we constrained our study to the activity that the users carried out in the last 4 months prior to their ban, which represents a good coverage of their involvement in the IOs. For HON, this results in analyzing all tweets produced between 11 September, 2019 and 8 January, 2020. For UAE instead we analyzed all tweets produced between 27 January and 26 May, 2019. Accounts that did not tweet in the 4 months time span of their respective IO were not included in our datasets.

B. GENUINE USERS

In spite of the usefulness and authoritativeness of Twitter/X's datasets of IOs, such datasets are seldom used in computational works on the study and detection of online manipulations since they only include data about *malicious users*. For the datasets to be really useful, comparable data about *genuine users* should also be collected [3]. To reach this goal, we built a new dataset leveraging Twitter/X APIs with elevated Academic access,[4] which contains genuine users that discussed the same topics of the two IOs, while the IOs were unfolding. This allowed to complement Twitter/X provided data about malicious users. Similar sampling strategies have been profitably adopted in some recent works on disinformation [31], astroturfing [3], and on the study of online behaviors [4]. In detail, for each IO, we selected the top hashtags used by the malicious users. Then, we collected all tweets that included at least one of such hashtags that were published during our 4 months observation window by non-banned users. We chose to rely on hashtags, which are the most widely used method of performing content-based Twitter data collection and filtering [4], [60]. Based on the ranked lists of each IO's top hashtags, our data collection step stops upon collecting data about a few hundreds of thousands of genuine users [9], [43]. We imposed this constraint to make our analyses computationally feasible, given that studying CB entails constructing and analyzing massive user interaction networks [11]. We remark that similar choices and limitations are frequent in computational works on CB [9], [10], [27], [43]. In summary, regarding HON we collected tweets about the following top-9 hashtags: #AlivioDeDeuda, #ParqueVidaMejor, #NavidadCatracha, #HondurasEnLaONU, #FiestasPatrias2019, #VivaHonduras, #VidaMejor, #EEUU, #FeriadoMorazanico. Instead for UAE we leveraged the following top-3 hashtags: #UnitedArabEmirates, #Yemen, #Video. Finally, for each IO we merged Twitter/X data about malicious users with our data about genuine users, obtaining the final HON and UAE datasets reported in Table 1.

IO	class	users		tweets		time span
		number	*%*	*number*	*%*	
HON	malicious	1,867	0.83	137,252	10.87	11/09/19 – 08/01/20
	genuine	222,819	99.17	1,125,578	89.13	
	total	224,686	100.00	1,262,830	100.00	
UAE	malicious	1,991	0.50	387,338	13.59	27/01/19 – 26/05/19
	genuine	397,025	99.50	2,462,130	86.41	
	total	399,016	100.00	2,849,468	100.00	

TABLE 1: Statistics about malicious and genuine users included in the HON and UAE datasets, together with the tweets they produced during the considered time span.

C. REMARKS

As shown, despite using only the top-3 hashtags for UAE, we ended up with significantly more genuine accounts than HON. Table 1 also reports the percentages of malicious and genuine users, and of their tweets, with respect to the total users and tweets in our datasets. This highlights the large imbalance between genuine and malicious users, which is representative of the reality of OSNs [61], [62]. Since the malicious users are removed from Twitter/X, the collected

[3]https://transparency.Twitter/X.com/en/reports/information-operations.html (accessed: 11/30/2023)

[4]Twitter/X Academic APIs have been unavailable since February 2023. However, results reproduction can be achieved via Twitter/X's paid API access levels: https://developer.twitter.com/en/docs/twitter-api.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

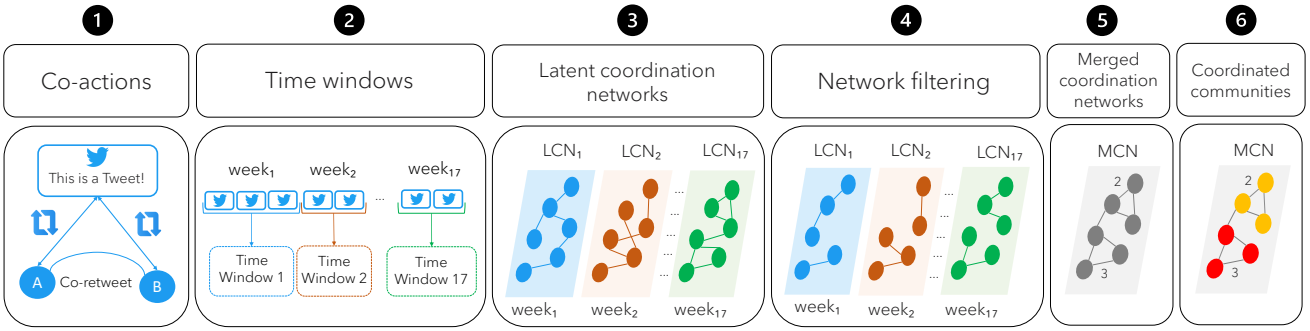


FIGURE 2: Overview of our coordination detection method.

datasets do not include direct interactions (e.g., retweets) between genuine and malicious users. However, this does not represent a limitation of our study given that we are interested in detecting and studying latent coordination between users, rather than direct interactions [11]. On the contrary, surfacing meaningful patterns of coordination without considering direct interactions increases the relevance of our results. For these reasons, studying coordination networks based on co-retweets is an established and profitable way to investigate coordinated online behaviors, including those carried out as part of malicious campaigns [9], [11], [27].

D. DATA AVAILABILITY

Our datasets are publicly available for research purposes.[5] In detail, we are releasing for scientific purposes the user IDs of the genuine users in our datasets, while information about the malicious users – that are now suspended – are available from Twitter/X upon motivated request.

IV. METHOD

For our quantitative analysis of CB, we adapted the state-of-the-art, network-based, coordination detection method proposed by [11]. We specifically selected this framework over other comparable state-of-the-art methods such as those proposed by [9], [10] because the authors of the former provided a publicly available Python implementation.[6] Our adapted method is composed of the following 6 analytical steps, which are also summarized in Figure 2:

1) **Co-actions.** As anticipated in Section II, many different co-actions might be indicative of CB [10]. Therefore, each coordination detection method starts with the selection of the co-action to use for modeling coordination among users. On Twitter/X, common examples of co-actions are co-retweets, co-mentions, and co-hashtags. Here, we used co-retweets (i.e., two users who retweet the same tweet) to model coordination, as done in the majority of existing works [9], [10], [48], [63].
2) **Time windows.** For a co-retweet to be indicative of CB, the two users must have retweeted the same tweet at around the same time. In other words, co-actions must occur within a given time window [11].In literature, short time windows (e.g., some seconds or minutes) are preferred by works that specifically aim to detect malicious behaviors, as this choice contributes to highlighting inorganic activity [10], [44]. On the contrary, works that explicitly consider all instances of coordination, including spontaneous coordination by independent users, tend to use long time windows (e.g., some days or weeks) [9], [43]. Here, we set the time window length = 7 days, which allows evaluating both medium- and long-term interactions [64], as well as to model the activity of both malicious and genuine users in our networks. A second methodological choice regards the adoption of overlapping or non-overlapping time windows. Here we resort to non-overlapping time windows since these are preferred when relatively long windows are used [20], [63], such as in our work, and also as done in [11] from which our method is derived. Therefore, given that both our datasets cover 4 months in time, each is split into 17 non-overlapping time windows. Then, we sort tweets (and retweets) in chronological order and we assign them to the corresponding time window. Only co-retweets that occur within the same time window are used for computing coordination.
3) **Latent coordination networks.** We build a latent coordination network (LCN) for each time window. An LCN is a weighted undirected user similarity network $G(V, E, W)$, where V is the set of nodes (i.e., the users) and E is the set of edges between them [11]. An edge between two users exists if they co-retweeted at least one tweet within the time window. Edges are weighted proportionally to the number of co-retweets so that users who perform many co-actions are strongly tied in the network G, which makes strong ties a good proxy for coordination [9]. W is the set of edge weights in G.
4) **Network filtering.** The LCNs resulting from the analysis of real-world IOs are typically too big to be analyzed quantitatively, and often even to be visualized [9]. For this reason, all coordination detection methods proposed to date carry out some sort of filtering [9]–[11]. To make

[5] https://zenodo.org/doi/10.5281/zenodo.10619747
[6] https://github.com/weberdc/find_hccs (accessed: 11/30/2023)

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

the analysis of our coordination networks feasible and meaningful, we first prune each LCN by discarding all edges whose weight $= 1$, which is a common choice in literature [3], [19]. The resulting disconnected nodes are discarded as well. While this quick filtering operation allows discarding all trivial interactions, it is however insufficient to adequately reduce the size of our LCNs. For this reason we also perform a more fine-grained filtering, as suggested by [9]. Specifically, we apply FSA_V to aggregate adjacent nodes that form influential communities [11], [65]. FSA_V is an agglomerative clustering algorithm with a stopping criterion, which aggregates nodes connected by strong ties. The stopping criterion makes it so that only nodes connected by edges with a significant weight are aggregated. For this reason, FSA_V can be used both as a community detection and as a network filtering algorithm [11]. Here we use it for the latter goal, to further prune our LCNs.

5) **Merged coordination network.** The previous filtering step allows obtaining LCNs with a tractable size. Since we are interested in studying the overall patterns of coordination among users of our datasets, we can now merge the LCNs [11]. This process produces a merged coordination network (MCN), where nodes and edges are obtained as the union of the nodes and edges of the filtered LCNs. Then, each edge weight in the MCN is recomputed as the sum of the weights that the same edge had, if present, in the filtered LCNs.
6) **Coordinated communities.** The last analytical step in coordination detection methods involves detecting coordinated communities [11]. We perform community detection on the MCN with the well-known greedy modularity algorithm [66], [67]. Since our MCN contains nodes and edges that encode significant coordination, performing community detection on MCN allows identifying coordinated groups of users [9], [10].

We applied the above steps to both our HON and UAE datasets, obtaining one coordination network per IO. The networks include all users that are highly coordinated independently of their class (i.e., malicious or genuine), and are complemented with information about the coordinated communities that took part in the online debate. In the next section we provide results of the analysis of the coordination networks.

V. RESULTS

We initially present and analyze the HON and UAE coordination networks. Subsequently, we investigate patterns of coordination and we leverage the ground-truth in our dataset to introduce measures for quantifying the extent to which a coordination network separates malicious users from the rest. Finally, we evaluate the robustness of our results.

A. HONDURAS

Some first insights into the presence of CB in the HON IO can be obtained by evaluating the results of the filtering step of our method. Table 2 reports the number and fraction of users in the HON dataset before and after the application of the coordination detection method. Overall, the filtering step discarded around 89% of the initial users due to low coordination. Interestingly, the fraction of discarded users varies markedly based on the user class. While around 90% of all genuine users exhibited low coordination, only 35% of the malicious users were discarded. This initial result already highlights the presence of CB in the HON IO. Specifically, it shows that the malicious users who organized the IO did so in a largely coordinated fashion.

	HON			UAE		
	all users	*coord. users*	*%*	*all users*	*coord. users*	*%*
malicious	1,867	1,217	65.18	1,991	745	37.42
genuine	222,819	23,564	10.58	397,025	28,228	7.11
total	224,686	24,781	11.03	399,016	28,973	7.26

TABLE 2: Number and fraction of highly coordinated users in HON and UAE. For both IOs, the filtering step discarded the majority of genuine users due to low coordination. Conversely, a much larger fraction of malicious users were retained as highly coordinated.

Next, we qualitatively visualize and quantitatively analyze the HON coordination network. Across the 4 months of our observation time, users in the HON dataset produced 980,699 retweets that were split across the 17 time windows. Each LCN in this IO, which corresponds to one time window, includes on average 9K nodes and 668K edges. Instead, the MCN obtained by merging all the LCNs includes 561,565 edges and 24,781 nodes, as reported in Table 2. The coordinated communities found in the HON MCN are shown in Figure 3. To visualize our coordination networks we adopted the Yifan Hu proportional layout algorithm, which combines a high efficiency for large networks and the high-quality of force-directed drawing algorithms [68]. As shown in Figure 3, the HON coordination network is characterized by 3 large and dense communities (i.e., yellow-, blue-, and orange-colored in figure), and by a multitude of smaller communities that hold peripheral positions in the network. Figure 4 shows the same network where the nodes are colored according to their class (i.e., either malicious or genuine). By comparing the communities highlighted in Figure 3 with the positions of the red-colored malicious users shown in Figure 4, we uncover that almost all malicious users involved in the HON IO belong to the blue-colored community. Table 3 reports quantitative results about the main coordinated communities in HON and about the presence of malicious users in each community. For any given community i we report the number of nodes (n_i) and edges (e_i), the number of malicious nodes (m_i), and the percentages of malicious users with respect to all users in that community (m_i/n_i) and with respect to all malicious users in the network (m_i/m_{tot}). Table 3 highlights that the blue-colored community of Figure 3 is composed of 1,816 users, out of which 1,126 (62%) are malicious. Furthermore, the malicious

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

FIGURE 3: Coordination network for the HON dataset, obtained at the end of the last step of the coordination detection method described in Figure 2. Each node represents a user and is colored according to the coordinated community to which it belongs. The network is characterized by 3 large and dense communities (yellow-, blue-, and orange-colored), and by several other smaller and peripheral ones.

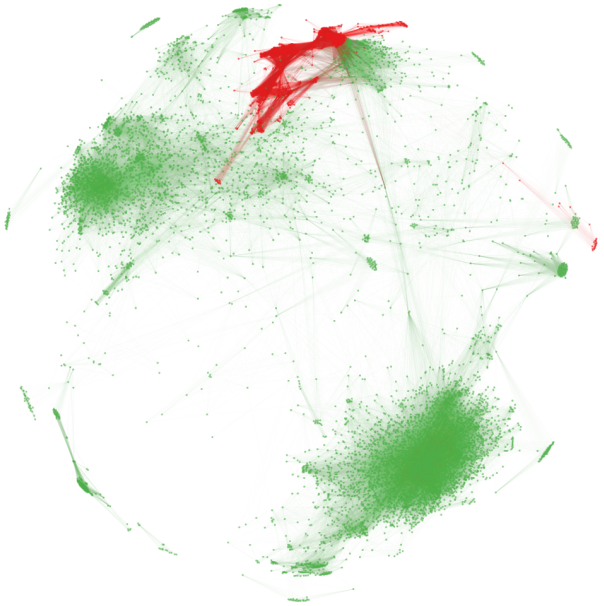


FIGURE 4: Coordination network for the HON dataset, where nodes are colored based on their class (malicious or genuine). Red nodes correspond to malicious users and green nodes to genuine ones. The vast majority of malicious users are clustered in a portion of the network that corresponds to the blue-colored coordinated community of Figure 3.

community	nodes (n_i)	edges (e_i)	malicious (m_i)	m_i/n_i	m_i/m_{tot}
	10,022	269,102	0	0%	0%
	8,591	98,853	0	0%	0%
	1,816	156,264	1,126	62.00%	92.52%
	830	8,402	0	0%	0%
	830	6,412	0	0%	0%
	786	7,013	0	0%	0%
	211	6,345	0	0%	0%
others	1,695	4,236	91	5.37%	7.48%
total	24,781	561,565	1,217	4.91%	100.00%

TABLE 3: Statistics about the coordinated communities in the HON dataset. Communities are color-coded as in Figure 3. The highlighted table row shows that 92.52% of all malicious users in the network are clustered in the blue-colored community.

users in this community account for 92.52% of all malicious users in HON.

These results provide interesting insights into the behavior of the perpetrators of the HON IO. Results in Table 2 and Table 3 support the finding that the malicious users involved in the IO were indeed coordinated. This is evident from both the large fraction (65.18%) of malicious users that are part of the coordination network, as well as from their position in the network. Indeed, the vast majority (92.52%) of malicious coordinated users in HON are tightly clustered in a single community (blue-colored in Figure 3). At the same time however, that community also contains many genuine users, which account for a minority yet significant share (38%) of all nodes in the community. These results can have multiple implications about the strategies used by the perpetrators of the IO, their success at influencing unaware genuine users, and the possibility of automatically detecting IO perpetrators. These points are discussed in Section VI.

B. UNITED ARAB EMIRATES

Similarly to our analysis of the HON IO, we begin our study of UAE by looking for evidence of CB. Table 2 reveals that, again, the vast majority (93%) of all users in the UAE dataset were filtered out due to low coordination. This time however, contrarily to HON, also the majority (62.58%) of malicious users were discarded. Nonetheless, malicious users still display overall higher coordination than genuine ones (37.42% versus 7.11%). These results surface that the perpetrators of the UAE IO made use of CB to spread their messages, but to a lower extent than in HON.

Next, we focus on the UAE coordination network. Users involved in this IO produced 2,183,767 retweets during the 4 months of observation. Each LCN in this IO includes, on average, 15K nodes and 711K edges per weekly time window. Finally, after the merging step, the MCN is composed of 288,247 edges and 28,973 nodes, as reported in Table 2. Figures 5 and 6 show the UAE coordination network. We observe

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

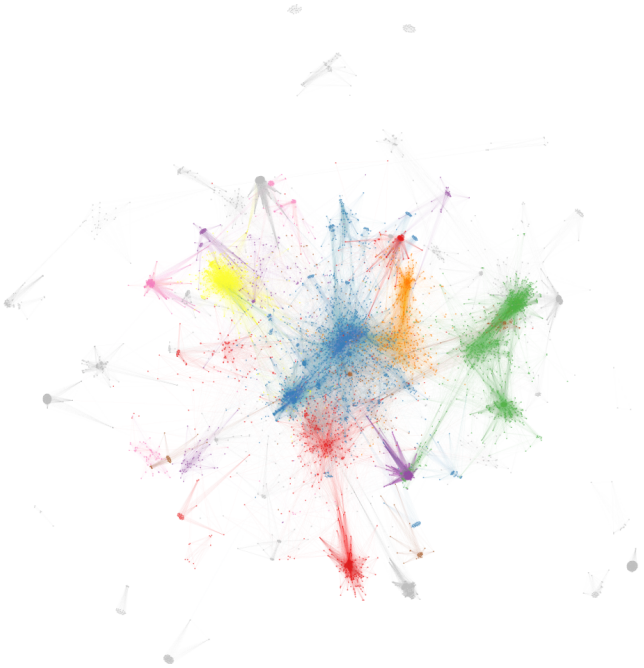


FIGURE 5: Coordination network for the UAE dataset, obtained at the end of the pipeline described in Figure 2. Each node represents a user and is colored according to the coordinated community to which it belongs. Contrary to Figure 3, this network does not feature any large and central community. Instead, the network is sparse and characterized by a relatively large number of small communities.

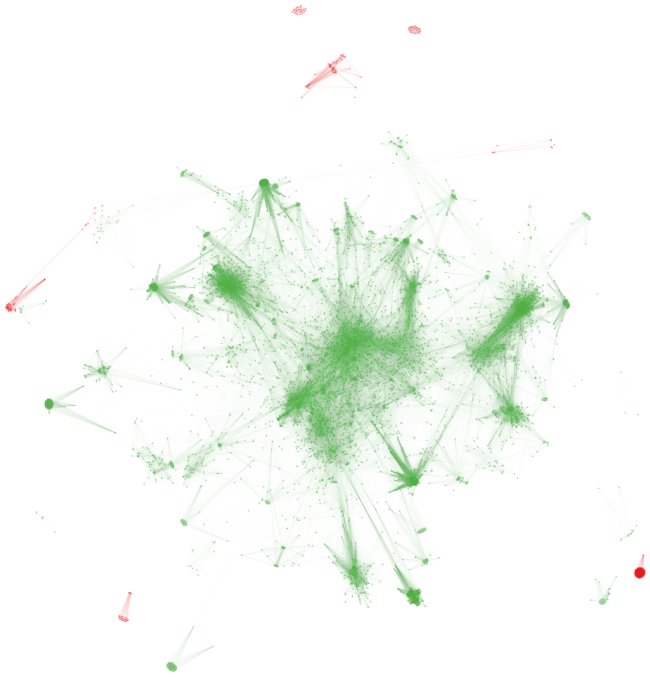


FIGURE 6: Coordination network for the UAE dataset, where nodes are colored based on their class (malicious or genuine). Red nodes correspond to malicious users and green nodes to genuine ones. Malicious users in this network are grouped in several small and peripheral communities.

that this network is more fragmented than that of HON. This is reflected by the larger number of small communities. In particular, the many communities that are displayed as grey-colored in Figure 5 are composed of up to a few hundreds of nodes and lay in the periphery of the network. Figure 6 shows the same network where nodes are colored according to their class (i.e., either malicious or genuine). Interestingly, many of the grey-colored small and peripheral communities shown in Figure 5 correspond to groups of malicious coordinated users (red-colored). Conversely, the main communities of the UAE coordination network, both according to their size and central position in the network, are completely composed of genuine users. Table 4 provides analytical results for each of the UAE coordinated communities, confirming the previous observations. In particular, the six rows highlighted in table correspond to small and peripheral communities that are exclusively composed of malicious users. Together, they account for 69% of all malicious users in the network. The remaining malicious users belong to other, even smaller, communities.

Overall, our results about the use of CB in UAE are slightly different from those of HON. In fact, on the one hand both IOs made use of coordination to amplify their messages. On the other hand however, a larger fraction of malicious users in HON appeared to be coordinated. Such malicious users are also more central in the network and strongly intertwined with genuine users. Conversely, fewer malicious users in UAE were coordinated. Moreover, they are scattered throughout the periphery of the network. These latter results about UAE hold also when modifying the initial selection of hashtags used

community	nodes (n_i)	edges (e_i)	malicious (m_i)	m_i/n_i	m_i/m_{tot}
●	5,803	64,052	0	0%	0%
●	5,319	28,697	0	0%	0%
●	2,095	41,459	0	0%	0%
●	1,696	23,321	0	0%	0%
●	1,669	15,750	0	0%	0%
●	1,574	25,880	0	0%	0%
●	1,269	6,453	0	0%	0%
●	912	8,340	0	0%	0%
●	223	2,133	223	100%	29.93%
●	102	4,949	102	100%	13.69%
●	58	80	58	100%	7.79%
●	52	154	52	100%	6.98%
●	49	103	49	100%	6.58%
●	29	84	29	100%	3.89%
● *others*	8,123	56,174	232	2.86%	31.14%
total	28,973	288,247	745	2.57%	100.00%

TABLE 4: Statistics about the coordinated communities in the UAE dataset. Communities are color-coded as in Figure 5. The highlighted table rows show that the malicious users in this IO are grouped in several small, homogeneous, and peripheral communities. Contrarily, the main communities (topmost rows) are fully genuine.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

to collect the genuine users, as discussed in Section V-D4. The implications of these findings are instead discussed in Section VI.

C. INVESTIGATING PATTERNS OF COORDINATION

In the previous sections we analyzed the overall structure of the HON and UAE coordination networks. Here we delve deeper by investigating the patterns of coordination of the different communities within such networks, with a particular focus on the communities of malicious users. Given that, in general, different IOs make use of different strategies, tools, and agents to carry out the manipulations [2], this analysis allows for a more nuanced understanding of the characteristics of the two IOs and how they unfolded.

1) Quantifying the extent of coordination

We investigate the patterns of coordination by following the approach introduced in [9]. In detail, we characterize each coordinated community with some network-based measures that provide information about their structure and organization. Additionally, we repeat this characterization at multiple degrees of coordination, so as to uncover the characteristics of the most coordinated users and to compare them with those of the weakly coordinated ones. Likewise, this approach also allows comparing the characteristics of the communities of malicious users with those of genuine users, drawing insights into the tactics of the former. In practice, let $G(V, E, W)$ be a coordination network and $W_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,N}\} \subseteq W$ be the set of weights associated to the edges of node $i \in V$. We define the coordination score c_i of node i as the maximum weight of its edges: $c_i = \max(w_{i,j}) \ \forall j \in V$. To this end, we recall that the weight of an edge between two nodes in a coordination network is a proxy for the extent of coordination between the two linked nodes [9], [11]. Finally, for the sake of clarity we normalize coordination scores in the $[0, 1]$ range, so that $\max(c_i) = 1$. All other coordination scores are rescaled proportionally.

We are now able to analyze the different communities in light of the coordination scores of their members. Specifically, we compute the size, density, and assortativity [69] of each community by only considering nodes whose coordination score exceeds a certain threshold $c_i \geq \varphi \ \forall i \in V$. To explore the full spectrum of coordinated behavior in our data, we iteratively repeat this analysis by progressively increasing the threshold, starting from $\varphi = 0$ up to $\varphi = 1$. Increasing the threshold allows finding, at each iteration k, a set of nodes that is more coordinated than those at iteration $k-1$. This effectively avoids fixing an arbitrary coordination threshold, as typically done in those works that exclusively focus on malicious behaviors [10], [52]. Conversely, this analysis allows characterizing the patterns of coordination of each community in terms of standard network measures, as a function of the extent of coordination between users. Figure 7 shows the results of this analysis, which we discuss in the following.

2) Community size, density, and assortativity

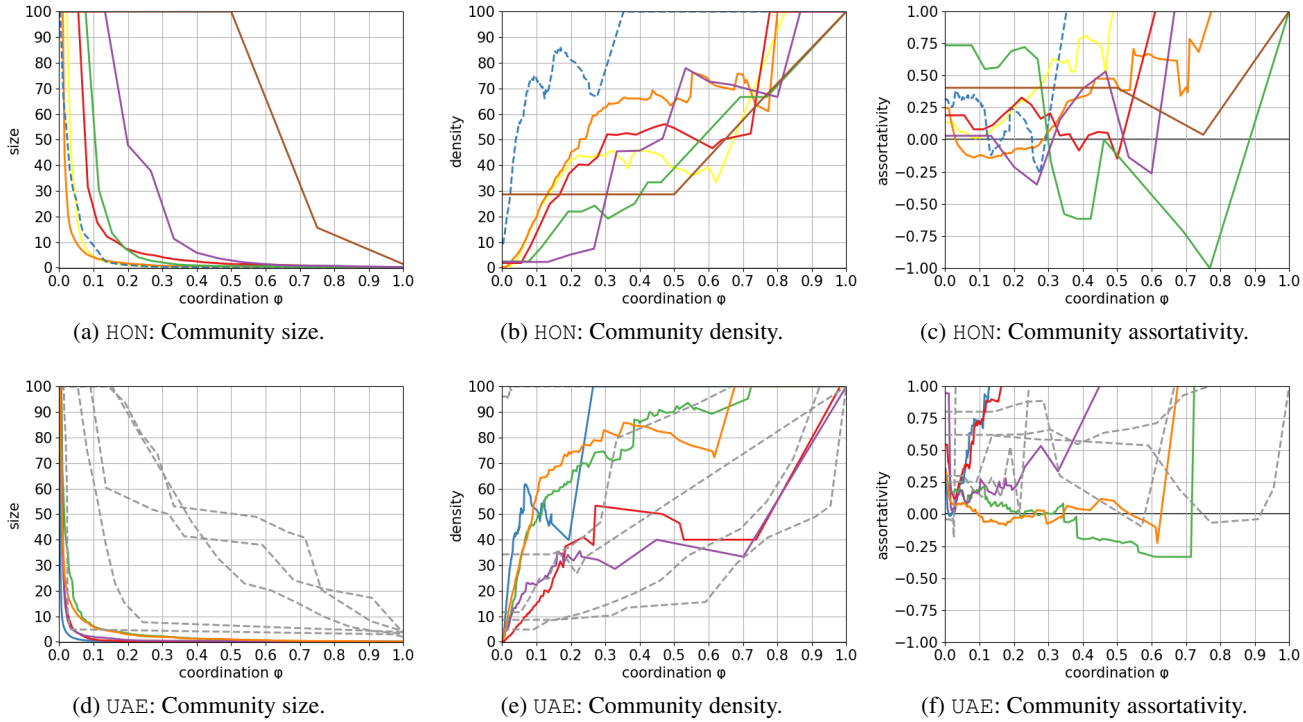
Figure 7a shows how the size of the different communities in the HON network change when considering increasingly coordinated users. In figures, for each community, size is expressed as the fraction of users whose coordination score $c \geq \varphi$, with respect to the total number of users in that community. In other words, the figure shows whether the members of the different HON communities are strongly or weakly coordinated. Figure 7a reveals that the majority of communities are composed of weakly coordinated users, as shown by the trends in community size rapidly plummeting when $\varphi \geq 0.2$. Two communities, brown- and purple-colored, stand out as significantly more coordinated than the others. By cross-checking with Figure 3 and Table 3, we note that such communities are small, very peripheral in the network, and exclusively composed of genuine users. A particularly relevant community in the HON network is the blue-colored one, which contains $\sim$93% of all perpetrators of the IO. However, regarding its size, Figure 7a does not show significant differences with respect to the communities of genuine users. Figure 7b, however, reveals that the blue-colored community of malicious users features the highest density out of the whole HON network. Density is a measure of the degree of interconnectedness in a network or community [69]. Therefore this result indicates that the malicious users responsible for the HON information operation are very well connected between one another. Furthermore, the density trend in Figure 7b steeply increases, up to the point that, for $\varphi \geq 0.35$, the malicious users form a clique. This result reinforces the idea that the perpetrators of this IO are well organized. Finally, Figure 7c shows trends in assortativity, a measure of the extent to which nodes with a high degree are connected to other nodes with a high degree, and vice versa. In the context of coordinated communities, this property is interesting as it indicates whether influential users in a community are connected to other influential users, which again provides information on the structure and organization of the community [9]. Figure 7c shows that the blue-colored community in the HON network is largely non-assortative,[7] which reflects the lack of correlation between the degree of malicious nodes and that of their neighbors. Contrarily, some genuine communities exhibit opposite behaviors. For example, the yellow-, red-, and orange-colored communities are strongly assortative when considering strongly coordinated users ($\varphi \geq 0.5$). Instead, the green-colored community appears as largely disassortative. Overall, these results highlight the informativeness of investigating the patterns of coordination as a way to make sense of the internal structure and organization of coordinated communities.

Figure 7d shows interesting results about the size and coordination of the communities in the UAE network. Indeed, all but one of the grey-colored communities are composed

[7]When interpreting results of Figures 7c and 7f, we recall that cliques are perfectly assortative, which explains why all assortativity trends eventually reach the maximum value of 1, and particularly so for large values of coordination.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482



(a) HON: Community size. (b) HON: Community density. (c) HON: Community assortativity.

(d) UAE: Community size. (e) UAE: Community density. (f) UAE: Community assortativity.

FIGURE 7: Network measures computed for some coordinated communities in each IO, as a function of the extent of coordination among the members of such communities. Dashed lines denote communities characterized by malicious users, while solid lines denote genuine communities. HON communities (top row, subfigures *a*, *b*, *c*) are colored as in Figure 3 and Table 3. UAE communities (bottom row, subfigures *d*, *e*, *f*) are colored as in Figure 5 and Table 4. Diverging trends in community size (*a*, *d*), density (*b*, *e*), and assortativity (*c*, *f*) are shown for some communities, revealing marked differences in the patterns of coordination exhibited by those communities.

of users whose coordination scores are markedly larger than those of genuine users. This result is particularly relevant considering that the grey-colored communities in the UAE network are solely composed of malicious users and collectively account for ∼69% of all perpetrators of the IO, as reported in Table 4. Results for community density and assortativity are respectively presented in Figures 7e and 7f. Among the notable findings in Figure 7e is that a grey-colored community of malicious users is fully connected independently of the coordination threshold (i.e., even for $\varphi \sim 0$). All other malicious communities have lower densities. Finally, Figure 7f shows that all malicious UAE communities are moderately or even strongly assortative. Three genuine communities also feature strong assortativity – namely, the blue- and red-colored communities, and to a lower extent, the purple-colored one – while the two remaining ones are slightly disassortative. Taken together, results about the UAE coordination network reveal that, although holding a peripheral position in the network, the malicious users involved in the IO were strongly coordinated and organized. In Section VI we further compare results about the patterns of coordination found in the HON and UAE information operations with those reported in previous works [9], [43].

D. FORMALIZATION AND SENSITIVITY ANALYSIS

The previous analyses provided nuanced results about the different use of CB in the HON versus the UAE IO. Despite the differences however, the analysis of CB provided valuable results for both IOs. Among the interesting findings is that the malicious communities showed clear signs of internal organization and featured peculiar and distinctive coordinated behaviors with respect to the genuine communities, in both IOs. For example, malicious HON users were almost entirely grouped in a single coordinated community, albeit mixed with some genuine users. Conversely, malicious UAE users spread across six small and peripheral communities, which however did not contain a single genuine user. In other words, malicious users in both HON and UAE featured some degree of *separation* with respect to the genuine users. These results indicate that the analysis of CB in IOs can provide valuable information for identifying the organized groups of users behind IOs. In the following, we provide measures to formalize the previous intuitions and to quantify the separation between malicious and genuine users in a coordination network. Furthermore, we investigate whether the previous results are robust to small variations of the parameters in our datasets and methods.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

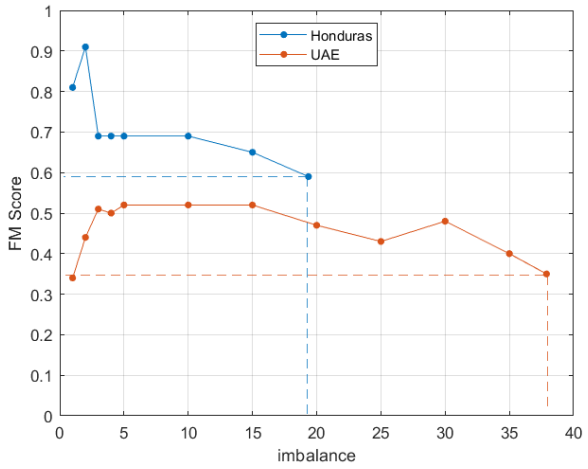


FIGURE 8: Relationship between *FM-score* (i.e., separation between malicious and genuine users) and class imbalance.

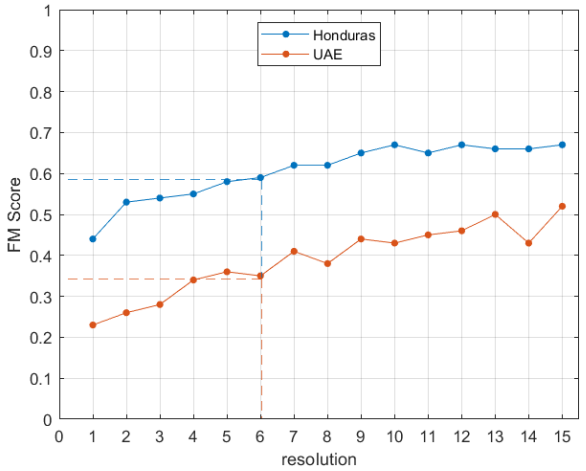


FIGURE 9: Relationship between *FM-score* (i.e., separation between malicious and genuine users) and the resolution γ of the community detection algorithm.

1) Quantifying differences in CB between malicious and genuine users

The results reported in Tables 3 and 4 highlight two desirable properties of coordination networks: *(i)* the capacity to group together all perpetrators of an IO, and *(ii)* the capacity to differentiate the perpetrators from the remaining users in the network. We formalize and measure the extent to which coordination networks possess the two aforementioned properties by leveraging the *Fowlkes-Mallows score* (*FM-score*) [70]. The *FM-score* is a well-known external evaluation metric that can be used to compare the results of a clustering with some ground-truth labels. Here we employ it to measure the extent to which the perpetrators of the HON and UAE IOs are separated from the other unaware users in the respective coordination network. Let *true positives* (*TP*) be the number of pairs of users that belong to the same cluster in both the ground-truth clustering and in our clustering, *false negatives* (*FN*) the number of pairs of users that belong to the same cluster in the ground-truth clustering but not in our clustering, and *false positives* (*FP*) the number of pairs of users that belong in the same cluster in our clustering but not in the ground-truth clustering. Then, the *FM-score* is defined in the $[0, 1]$ range as:

$$FM\text{-}score = \frac{TP}{\sqrt{(TP + FP) \cdot (TP + FN)}} \tag{1}$$

$FM\text{-}score = 1$ occurs in case a clustering yields only two communities: one exclusively composed of malicious users and the other exclusively composed of genuine users, which implies perfect separation. Instead, $FM\text{-}score \simeq 0$ occurs when a clustering results in many communities that feature a balanced mix of both genuine and malicious users, a much less informative scenario. We can now leverage the *FM-score* to quantitatively evaluate the characteristics of the HON and UAE coordination networks. In Section V-A we showed that the blue-colored community in HON includes almost all of the malicious users in the network, although mixed with some genuine users. This is correctly reflected by $FM\text{-}score = 0.59$, denoting a considerable separation between malicious and genuine users. Differently, in Section V-B we highlighted that all the UAE communities are completely homogeneous (i.e., exclusively composed of either genuine or malicious users), which contributes to increasing the *FM-score* because of $FN = 0$. On the other hand however, the UAE coordination network features almost twice the number of communities of the HON network, which lowers the *FM-score* due to a large number of *FP*. In fact, the resulting *FM-score* for UAE $= 0.35$, versus $FM\text{-}score = 0.59$ for HON. In spite of these differences, these results demonstrate that the coordination networks obtained from the application of our method provide valuable information for telling apart the genuine and malicious users that take part in IOs.

2) Sensitivity analysis: Class imbalance

In this paragraph we delve deeper into the properties of the HON and UAE coordination networks. In particular, we evaluate their capacity to separate malicious and genuine users, quantified by *FM-score*, in relation to the imbalance between the two classes of users. As typically done in classification tasks, we measure imbalance as the proportion between the majority and the minority classes [71]. Thus in our context, imbalance can be measured as the proportion between the number g_{tot} of genuine users with respect to the number m_{tot} of malicious ones: $\frac{g_{tot}}{m_{tot}}$. To assess the sensitivity of our results to class imbalance, we started from the maximum imbalance in our dataset, highlighted with dashed lines in Figure 8 and corresponding to the statistics in Table 1, and we progressively lowered it by discarding genuine users based on their activity, so that users who tweeted less recently were discarded first. Figure 8 shows that both datasets are extremely imbalanced, which is typical of studies on online harms [61], [62]. In addition, *FM-score* is always higher in

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

HON than in UAE, which is mainly due to the higher number of communities in the UAE network, which leads to a higher value of *FP*. Interestingly, we also observe that *FM-score* (i.e., separation) increases between malicious and *active* genuine users. This is evident in both low and high imbalance areas of the plot. In the former, *FM-score* initially increases for both HON and UAE as we increase the imbalance, which corresponds to adding active genuine users to the malicious ones. In the high imbalance conditions, *FM-score* drops as we add increasingly less active users. This result suggests that malicious users in HON and UAE were not particularly active, in spite of their coordination.

3) Sensitivity analysis: Resolution

The previous results showed that the different *FM-score* measured for HON and UAE largely depend on the number and size of the communities in the two networks. Thus, here we investigate the relationship between *FM-score* and the *resolution* (γ) of our community detection algorithm. Specifically, larger γ yield fewer but larger communities, while decreasing it results in more communities of small size.[8] Figure 9 shows the relationship between *FM-score* and γ. Dashed lines denote *FM-score* at $\gamma = 6$, which is the setting used for the results presented in Sections V-A and V-B. Generally, we measured a slightly increasing trend in *FM-score* when (γ) is increased, mainly due to the smaller number of detected communities, which is a fundamental influencing factor for *FM-score*, because the higher the number of communities, the higher the value of *FP*.

4) Sensitivity analysis: Hashtags selection

In Section III-B we explained the criterion that we used to collect a comparable set of genuine users with which to enrich Twitter/X's datasets of malicious users involved in IOs. Specifically, for each IO we collected genuine users that used in their tweets the most frequent hashtags also used by the perpetrators of that IO, at around the same time. We applied this criterion equally for both the HON and UAE IOs, so as to avoid biasing our data collection. However, differences in the content tweeted by the malicious users involved in the two IOs resulted in selecting many IO-specific hashtags for the HON dataset, contrarily to UAE for which many of the most tweeted hashtags were rather generic. Considering that the choice of hashtags directly influences the genuine users that are part of the coordination networks, we are interested in evaluating whether selecting a different set of hashtags for one of the IOs yields qualitatively different results. For this reason, we repeated the data collection step for the UAE dataset and all subsequent analyses. This time we only used IO-specific hashtags for UAE, similarly to what we already did for HON. In particular, we selected genuine users with which to contrast the malicious UAE users based on the following two highly-specific hashtags: #Yemen and #Houthi.[9] The resulting new UAE dataset contains 132K users and 1.3M tweets, respectively corresponding to 33% of the users and 46% of the tweets of the original UAE dataset described in Table 1.

Figure 10 shows the coordination network obtained with the application of our method to the new dataset, which we compare to the original network of the UAE dataset displayed in Figure 5. As shown, the network obtained by using only IO-related hashtags is smaller as a consequence of the reduced set of genuine users. Nonetheless, its overall topology and structure are similar to that of Figure 5: a few sizeable communities in the center of the network, surrounded by many smaller and peripheral ones. More importantly, also the positioning and clustering of the malicious users in the networks are similar, as shown in Figures 11 and 6. Indeed, in both networks the malicious users (red-colored) lay at the periphery and belong to a few small and homogeneous communities. In summary, results in Figures 10 and 11 show that considering a different selection of hashtags did not significantly alter our results for the UAE dataset, which were robust to this choice.

VI. DISCUSSION

Our analyses provide multiple interesting results about the adoption of CB in two so far unexplored IOs, as summarized in the following:

- The perpetrators of the HON and UAE IOs were both coordinated and organized.
- The perpetrators of the HON IO held a central position in the network and were well-connected with many genuine users.
- The perpetrators of the UAE IO were split across multiple small coordinated groups and held a marginal position in the network, with few connections with genuine users.

We verified that the above findings are robust to a number of variations in our data selection criteria and in the parameters of our method. Furthermore, in addition to being relevant on their own, our results also open up the possibility to explore the possible strategies, organization, and influence exerted by the two IOs. We discuss these and other points in the remainder of this section.

A. STRATEGIES

In addition to surfacing coordinated behaviors, our analyses also uncovered marked differences between the two IOs. While the majority of malicious HON users ended up tightly clustered in a single community, only a subset of malicious UAE users were considered to be significantly coordinated. Moreover, such coordinated users ended up scattered across multiple small communities, rather than grouped together. A

[8]This is the Gephi 0.9.7 implementation of greedy modularity and its γ parameter. Other implementations use γ differently: https://tinyurl.com/networkx-greedy-modularity (accessed: 11/30/2023).

[9]Twitter/X Moderation Research Consortium described the IO as in the following: "We suspended a separate group of 4,248 accounts operating uniquely from the UAE, mainly directed at Qatar and Yemen. These accounts were often employing false personae and tweeting about regional issues, such as the Yemeni Civil War and the Houthi Movement."

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

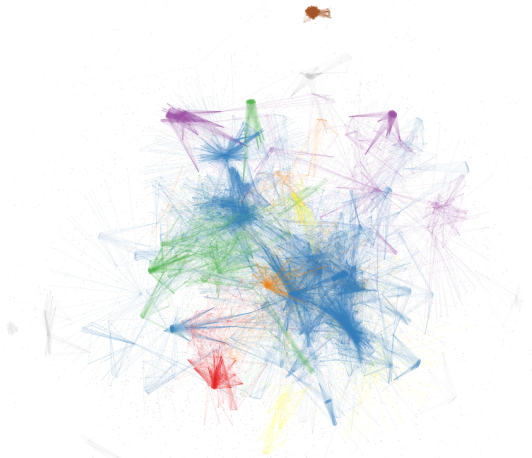


FIGURE 10: Coordination network for the UAE dataset built using only two IO-specific hashtags. Similarly to Figure 5, the network is relatively sparse and characterized by many small communities.

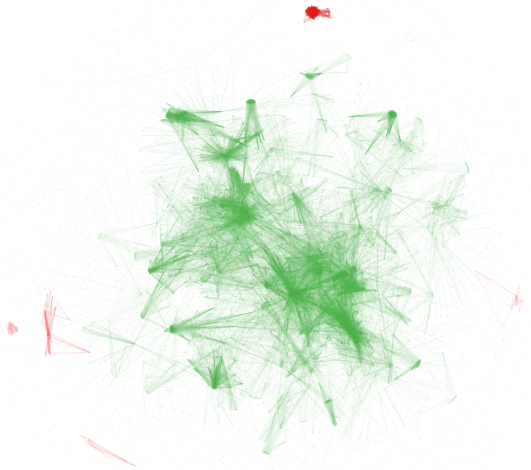


FIGURE 11: Coordination network for the UAE dataset built using only two IO-specific hashtags. Nodes are colored based on their class. Similarly to Figure 6, the malicious nodes (red-colored) are split across a few small communities in the network periphery.

possible explanation for this result is related to the strategies exploited in the two IOs. As described in Section III, malicious HON users massively retweeted the former Honduras President. As such, our choice of using co-retweets to model user similarities allowed capturing the full extent of coordination exhibited by HON users. Conversely, malicious UAE users mainly tampered with hashtags, in an effort to sow discord in Qatar and Yemen. In this case, we likely captured only part of their coordinated behaviors, which could explain the lower coordination found for the UAE IO. Therefore, on the one hand, we showed the extent to which the analysis of CB is capable of surfacing important differences between IOs in terms of their strategies and tactics [2]. On the other hand however, our own results could be influenced by the different strategies adopted by the perpetrators of the two IOs. We expand on this latter point in Section VI-E, while in the following we discuss the implications of this observation for future work in this area.

Indeed, we broaden the above observation by noting that, in general, thorough investigations of CB require the simultaneous modeling of multiple user activities (e.g., co-retweets, co-mentions, co-hashtags, and more) [9], [10], which however are seldom considered conjointly. For example, this can be achieved by studying coordination with multi-layer networks, where each layer models one activity type [44]. Therefore, a promising direction for future work involves experimenting with rich and multidimensional user representations, such as multi-layer networks [44], [64], [72] or multivariate time series [73], to jointly model the multifaceted behavior of online users, thus surpassing one of the limitations of existing coordination detection methods that almost exclusively rely on unidimensional networks.

B. ORGANIZATION

Investigating the patterns of coordination can provide valuable information towards understanding the structure and organization of the coordinated communities. To this end, in Section V-C2 we computed standard network measures, such as density and assortativity, for individual communities rather than for the whole network, as typically done. Furthermore, we examined trends in such measures as a function of the extent of coordination. For example, computing the density of a coordinated community with this approach allows assessing the interconnectedness of the most coordinated users, which is a proxy for the degree of orchestration and organization of that community [9]. In this regard, our findings confirm those of previous work, showing that malicious communities indeed exhibit high density when considering their strongly coordinated users. Then, analyzing the assortativity of a coordinated community allows to draw insights into the possible inorganic nature of such coordination. Indeed, recent works showed that disassortative network structures are associated with the inorganic behaviors of malicious botnets [74], [75]. Conversely, assortative network structures were recently linked to the behavior of grassroots movements [9]. With respect to this body of literature, our results provide new find-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

ings. In fact, we measured moderate to strong assortativity for all of the malicious communities involved in the HON and UAE IOs. Considering that IOs are typically carried out by trolls rather than bots, this result remarks the differences between these two types of agents that are often involved in online manipulations [76]. In turn, this can inform future strategies for detecting these threats [6]. Moreover, our results also mandate care in interpreting assortativity as a sign of grassroots behavior [9].

Despite these compelling results, however, still little is known about the ways in which complex IOs unfold. In this regard, future research and experimentation efforts should aim at investigating a broader array of measures. This effort, possibly in combination with the analysis of ground-truth datasets where the activity of the malicious actors is known *a priori*, would allow gathering deeper understanding of the ways in which the analysis of CB can contribute to contrast online information manipulations.

C. INFLUENCE

In studies on online information manipulation, the position of the actors in a network is used to estimate the influence exerted by such actors [77], [78]. Therefore, our results about the centrality of the malicious HON and UAE users in their respective networks also provide insights into the possible influence obtained by the two IOs. Specifically, malicious HON users held a central position in their network. Additionally, such users were also tightly interconnected with one another, with the genuine users in their own community, as well as with those of neighboring communities. This network layout is akin to that of highly influential users, both genuine [65], [79] and malicious [77]. Contrarily, malicious UAE users were dispersed in the periphery of their network and featured limited connections with genuine users. As such, it is unlikely that they managed to exert a strong external influence. Notably, this scenario resembles the one recently measured for some botnets involved in political manipulation [78]. In any case, the examples provided by the analysis of the HON and UAE IOs demonstrate the usefulness of coordination networks for gaining insights into the effectiveness of online manipulations – an important yet largely unexplored area of research [53]. To this regard, our present work complements and extends the existing literature on IOs, which rarely investigated the position of the involved actors in the discussion networks [3], [4], [20], as well as their interactions, including possible coordination.

D. DETECTION

Our results could also have future implications for improving the automatic detection of IOs and their perpetrators (e.g., state-sponsored trolls) [6]. Current state-of-the-art works in this area either leverage the content shared by the malicious users [31], [32], or interaction and coordination networks [3], [10], [20] such as those studied in our work. However, different from our work, the common assumption when studying coordination networks is that the most coordinated groups are those responsible for the manipulations [3], [10]. As a consequence, all strongly coordinated users are typically flagged as malicious. However, this assumption was proven wrong both in recent literature [43] and in our present work, where we uncovered multiple genuine – yet strongly coordinated – communities. Specifically, here we go beyond the existing literature, by analyzing coordination networks that contain both strongly coordinated genuine and malicious users. Nonetheless, we demonstrate that carefully-built coordination networks still embed a certain degree of separation between genuine and malicious users, as shown by our results about the HON and UAE IOs. In Section V-D we quantified this characteristic by leveraging the *FM-score* [70]. In the future, we could leverage the analysis of coordination networks to compute even other informative machine learning features to be used for this challenging task, thus possibly improving current detection performances. The *FM-score* with which we experimented in the present work already proved informative when used in isolation. In the future, it could therefore be even more valuable when used in conjunction with other additional features.

E. LIMITATIONS

1) Data selection and generalizability

Our data selection for this study is based on two large-scale IOs that used the English language. In addition, we also aimed at studying campaigns carried out in relatively understudies contexts, because that contributes to increasing the diversity in an area that is dominated by studies on manipulations targeted at the US [20], [80] or the main European countries [9], [43], or carried out by well-known threat actors such as Russia [22], [24], Iran [40], and China [31], [42]. Nonetheless, our choice of studying some little-explored datasets negatively affects our work in terms of the possibility of comparing our results to those of other studies. In addition, having tested our approach only against two datasets provides relatively limited evidence of the usefulness of our proposed approach. Unfortunately however, carrying out network analyses of massive online social datasets typically require multiple weeks of processing. This represents a further drawback that hinders the possibility to analyze multiple large datasets. In fact, this is one of the reasons why the majority of existing studies on coordinated online behavior are based on the analysis of either one single dataset [43], [53], [63], [81], or two [11], [24]. In this respect, our present work is on par with the current state-of-the-art in the field. For future work however, it would be important to carry over our methodology to one or more additional reference datasets, in order to assess the generalizability of our findings.

Furthermore, an additional limitation arises from the significant imbalance between the volume of malicious and genuine content within our dataset. Specifically, malicious tweets constitute merely 11% of the entire HON dataset and 14% of the UAE dataset. Moreover, the segments of our datasets pertaining to the activities of malicious users encompass all tweets authored by such users within the designated time

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

frame. Conversely, as delineated in Section III, our datasets concerning the activities of genuine users solely encompass tweets containing at least one of the most prevalent hashtags circulated within the same time frame. Introducing either the inclusion of all tweets from genuine users or the exclusion of certain tweets from malicious users would exacerbate this imbalance. It is noteworthy that this limitation is widely acknowledged in the literature on online harms and even in the realm of security in general, resulting in various adverse implications for downstream tasks [61], [62].

2) Data collection and transparency

Another limitation arises from the use of historical Twitter/X APIs for data collection, in relation to possible data deletions and suspensions. Specifically, spontaneous user deletions are relatively infrequent and, as a consequence, they typically have a minor impact on social media analyses. On the contrary, suspensions can involve the mass removal of a large number of accounts and tweets. As such, failing to account for Twitter/X suspensions can severely affect an analysis. However, suspensions are related to the behavior of malicious users, for which we obtained data directly from Twitter/X via its repository of IOs, rather than from the APIs. As such, our approach gives access to all malicious users (as detected by Twitter/X), almost all genuine users, and all interactions of malicious users with genuine content. What it *does not* give access to are possible interactions between genuine users and malicious content, for which there currently exists no viable solution. Moreover, data about the malicious users involved in the two considered IOs was provided directly by Twitter/X. However, Twitter/X's methodology for detecting such accounts is currently unknown. As such, there is no guarantee that Twitter/X identified all malicious accounts involved in the IOs, nor that all accounts flagged as malicious were actually so. Based on these limitations, it is thus impossible to assess the impact that Twitter/X's detection methodology could have had on our results.

On the one hand, the above considerations highlight the importance of transparency in content moderation, since that can enable research and validation that would otherwise be impossible [82]. At the same time however, even more transparency is needed in order to gain a thorough understanding of current content moderation processes and of their impact on the integrity of the online environment [83], [84].

3) Methodological choices and validation

As explained in Section IV, some of our methodological choices can have significant repercussions on the results of our analyses. For example, the adoption of overlapping *vs.* non-overlapping time windows, and the length of such time windows, are two important parameters of our methodology [85]. Similarly, as anticipated in Section VI-A, also the choice of the action with which to compute user similarities (e.g., co-retweets, as in our case) can deeply influence the shape and structure of coordination networks [3]. Here, we based the selection of the parameters of our method on the current best practices and on the latest results in the field [11], [64], [85]. In addition, in Section V-D we carried out extensive sensitivity analyses to assess the robustness of our results to small variations of the parameters in our datasets and method. Nonetheless, a thorough validation of our method would require extensive experiments on an authoritative reference or ground-truth dataset. However, building a ground-truth dataset of coordinated behaviors currently represents a crucial open challenge [19]. For this reason, reference or ground-truth datasets of coordinated behavior are still few and far between, and the majority of recent works resorted to partial or reconstructed ground-truths as done in the present work [3], [4], [20].

VII. CONCLUSIONS

We investigated the presence and patterns of coordinated behavior (CB) in two large information operations (IOs) on Twitter/X. By leveraging two novel datasets and a state-of-the-art coordination detection method, we found that perpetrators of both IOs were markedly coordinated. Additionally, our nuanced results revealed that while the perpetrators of the first IO held a central position in their network and established strong connections with other users, the perpetrators of the second IO remained in the periphery of their network, with limited interconnections with genuine users. We discussed these results in terms of the strategies, organization, and influence of these IOs. Finally, we proposed measures to quantify the extent to which the analysis of CB can contribute to distinguishing between malicious and genuine users. This latter contribution goes in the direction of improving the automatic detection of IOs and their perpetrators, which we will tackle in future works. For the future, we also aim to extend current coordination detection methods by leveraging multilayer networks to conjointly analyze multiple dimensions of online user behavior. Finally, a better characterization of the detected communities, for example in terms of the discussed topics, would allow a deeper understanding of the content they produced and, in turn, of their aims and intent.

REFERENCES

[1] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policymaking," Council of Europe, Tech. Rep., 2017.

[2] K. Starbird, A. Arif, and T. Wilson, "Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations," in *The 22th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'19)*, 2019, pp. 1–26.

[3] D. Schoch, F. B. Keller, S. Stier, and J. Yang, "Coordination patterns reveal online political astroturfing across the world," *Scientific Reports*, vol. 12, no. 1, pp. 1–10, 2022.

[4] A. C. Nwala, A. Flammini, and F. Menczer, "A language framework for modeling social media account behavior," *EPJ Data Science*, vol. 12, no. 1, p. 33, 2023.

[5] L. Luceri, V. Pantè, K. Burghardt, and E. Ferrara, "Unmasking the web of deceit: Uncovering coordinated activity to expose information operations on Twitter," in *The 33rd ACM Web Conference (WWW'24)*, 2024.

[6] F. Ezzeddine, O. Ayoub, S. Giordano, G. Nogara, I. Sbeity, E. Ferrara, and L. Luceri, "Exposing influence campaigns in the age of LLMs: A behavioral-based AI approach to detecting state-sponsored trolls," *EPJ Data Science*, vol. 12, no. 1, pp. 1–21, 2023.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

[7] D. L. Linvill and P. L. Warren, "Troll factories: The Internet Research Agency and state-sponsored agenda building," Resource Centre on Media Freedom in Europe, Tech. Rep., 2018.

[8] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Emergent properties, models and laws of behavioral similarities within groups of Twitter users," *Computer Communications*, vol. 150, pp. 47–61, 2020.

[9] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and M. Tesconi, "Coordinated behavior on social media in 2019 UK General Election," in *The 15th International AAAI Conference on Web and Social Media (ICWSM'21)*, 2021, pp. 443–454.

[10] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering coordinated networks on social media: Methods and case studies," in *The 15th International AAAI Conference on Web and Social Media (ICWSM'21)*, 2021, pp. 455–466.

[11] D. Weber and F. Neumann, "Amplifying influence through coordinated behaviour in social networks," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 111, 2021.

[12] S. Cresci, M. Petrocchi, A. Spognardi, and S. Tognazzi, "On the capability of evolved spambots to evade detection via genetic engineering," *Online Social Networks and Media*, vol. 9, pp. 1–16, 2019.

[13] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn, "Who let the trolls out? Towards understanding state-sponsored trolls," in *The 11th ACM Conference on Web Science (WebSci'19)*, 2019, pp. 353–362.

[14] J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, and E. Gilbert, "Still out there: Modeling and identifying Russian troll accounts on Twitter," in *The 12th ACM Conference on Web Science (WebSci'20)*, 2020, pp. 1–10.

[15] A. Rauchfleisch and J. Kaiser, "The false positive problem of automatic bot detection in social science research," *PLoS One*, vol. 15, no. 10, 2020.

[16] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 561–576, 2017.

[17] I. Alieva, L. H. X. Ng, and K. M. Carley, "Investigating the spread of russian disinformation about biolabs in ukraine on twitter using social network analysis," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 1770–1775.

[18] I. Alieva, J. Moffitt, and K. M. Carley, "How disinformation operations against russian opposition leader alexei navalny influence the international audience on twitter," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 80, 2022.

[19] F. B. Keller, D. Schoch, S. Stier, and J. Yang, "Political astroturfing on Twitter: How to coordinate a disinformation campaign," *Political Communication*, vol. 37, no. 2, pp. 256–280, 2020.

[20] L. Vargas, P. Emami, and P. Traynor, "On the detection of disinformation campaign activity with network analysis," in *The 11th Cloud Computing Security Workshop (CCSW'20)*, 2020, pp. 133–146.

[21] M. Bastos and J. Farkas, ""Donald Trump is my President!": The Internet Research Agency propaganda machine," *Social Media + Society*, vol. 5, no. 3, 2019.

[22] D. L. Linvill, B. C. Boatwright, W. J. Grant, and P. L. Warren, ""THE RUSSIANS ARE HACKING MY BRAIN!" Investigating Russia's Internet Research Agency Twitter tactics during the 2016 United States presidential campaign," *Computers in Human Behavior*, vol. 99, pp. 292–300, 2019.

[23] C. Llewellyn, L. Cram, A. Favero, and R. L. Hill, "Russian troll hunting in a Brexit Twitter archive," in *The 18th ACM/IEEE Joint Conference on Digital Libraries (JCDL'18)*, 2018, pp. 361–362.

[24] C. Kriel and A. Pavliuc, "Reverse engineering Russian Internet Research Agency tactics through network analysis," *Defence Strategic Communication*, vol. 6, 2019.

[25] C. Ehrett, D. L. Linvill, H. Smith, P. L. Warren, L. Bellamy, M. Moawad, O. Moran, and M. Moody, "Inauthentic newsfeeds and agenda setting in a coordinated inauthentic information operation," *Social Science Computer Review*, 2021.

[26] T. Wilson and K. Starbird, "Cross-platform information operations: Mobilizing narratives & building resilience through both 'Big' & 'Alt' tech," in *The 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'21)*, 2021, pp. 1–32.

[27] D. Pacheco, A. Flammini, and F. Menczer, "Unveiling coordinated groups behind White Helmets disinformation," in *The 29th Web Conference Companion (WWW'20 Companion)*, 2020, pp. 611–616.

[28] S. Jhaver, C. Boylston, D. Yang, and A. Bruckman, "Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter," in *The 24th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW'21)*. ACM, 2021, pp. 1–30.

[29] A. Trujillo and S. Cresci, "Make Reddit Great Again: Assessing community effects of moderation interventions on r/The_Donald," in *The 25th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW'22)*. ACM, 2022, pp. 1–28.

[30] A. Arif, L. G. Stewart, and K. Starbird, "Acting the part: Examining information operations within #blacklivesmatter discourse," in *The 21th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'18)*, 2018, pp. 1–27.

[31] M. Alizadeh, J. N. Shapiro, C. Buntain, and J. A. Tucker, "Content-based features predict social media influence operations," *Science Advances*, vol. 6, no. 30, 2020.

[32] B. Ghanem, D. Buscaldi, and P. Rosso, "TexTrolls: Identifying trolls on Twitter with textual and affective features," in *The 2020 Workshop on Online Misinformation- and Harm-Aware Recommender Systems (OHARS'20)*, 2020.

[33] M. Alassad, B. Spann, and N. Agarwal, "Combining advanced computational social science and graph theoretic techniques to reveal adversarial information operations," *Information Processing & Management*, vol. 58, no. 1, 2021.

[34] O. Varol, E. Ferrara, F. Menczer, and A. Flammini, "Early detection of promoted campaigns on social media," *EPJ data science*, vol. 6, pp. 1–19, 2017.

[35] M. Gambini, S. Tardelli, and M. Tesconi, "The anatomy of conspiracy theorists: Unveiling traits using a comprehensive twitter dataset," *Computer Communications*, 2024.

[36] D. Assenmacher, L. Adam, H. Trautmann, and C. Grimme, "Towards real-time and unsupervised campaign detection in social media," in *The 33rd International FLAIRS Conference (FLAIRS'20)*. AAAI, 2020.

[37] A. Toney, A. Pandey, W. Guo, D. Broniatowski, and A. Caliskan, "Automatically characterizing targeted information operations through biases present in discourse on Twitter," in *The 15th IEEE International Conference on Semantic Computing (ICSC'21)*, 2021, pp. 82–83.

[38] S. Bradshaw and A. Henle, "The gender dimensions of foreign influence operations," *International Journal of Communication*, vol. 15, p. 23, 2021.

[39] B. De Clerck, F. Van Utterbeeck, J. Petit, B. Lauwens, W. Mees, and L. E. Rocha, "Maximum entropy networks applied on Twitter disinformation datasets," in *The 10th International Conference on Complex Networks and Their Applications (CNA'22)*, 2022, pp. 132–143.

[40] A. Burns and B. Eltham, "Twitter free Iran: An evaluation of Twitter's role in public diplomacy and information operations in Iran's 2009 election crisis." Communications Policy and Research Forum, 2009.

[41] G. S. Jowett and V. O'donnell, *Propaganda & persuasion*. Sage publications, 2018.

[42] S. C. Woolley and P. N. Howard, *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press, 2018.

[43] K. Hristakieva, S. Cresci, G. Da San Martino, M. Conti, and P. Nakov, "The spread of propaganda by coordinated communities on social media," in *The 14th International ACM Web Science Conference (WebSci'22)*, 2022, pp. 191–201.

[44] T. Magelinski, L. Ng, and K. Carley, "A synchronized action framework for detection of coordination on social media," *Journal of Online Trust and Safety*, vol. 1, no. 2, 2022.

[45] C. Cao, J. Caverlee, K. Lee, H. Ge, and J. Chung, "Organic or organized? Exploring URL sharing behavior," in *The 24th ACM International Conference on Information and Knowledge Management (CIKM'15)*, 2015, pp. 513–522.

[46] K. Sharma, Y. Zhang, E. Ferrara, and Y. Liu, "Identifying coordinated accounts on social media through hidden influence and group behaviours," in *The 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD'21)*, 2021, pp. 1441–1451.

[47] Y. Zhang, K. Sharma, and Y. Liu, "VigDet: Knowledge informed neural temporal point process for coordination detection on social media," in *The 35th Annual Conference on Neural Information Processing Systems (NeurIPS'21)*, 2021, pp. 3218–3231.

[48] F. Keller, D. Schoch, S. Stier, and J. Yang, "How to manipulate social media: Analyzing political astroturfing using ground truth data from south korea," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 564–567.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

[49] V. Chomel, M. Panahi, and D. Chavalarias, "Manipulation during the french presidential campaign: coordinated inauthentic behaviors and astroturfing analysis on text and images," in *International Conference on Complex Networks and Their Applications*. Springer, 2022, pp. 121–134.
[50] C. Francois, V. Barash, and J. Kelly, "Measuring coordinated versus spontaneous activity in online social movements," *New Media & Society*, 2021.
[51] H. S. Heidenreich, M. I. Mujib, and J. R. Williams, "Investigating coordinated 'social' targeting of high-profile Twitter accounts," *arXiv:2008.02874*, 2020.
[52] F. Giglietto, N. Righetti, L. Rossi, and G. Marino, "It takes a village to manipulate the media: Coordinated link sharing behavior during 2018 and 2019 Italian elections," *Information, Communication & Society*, vol. 23, no. 6, 2020.
[53] M. Cinelli, S. Cresci, W. Quattrociocchi, M. Tesconi, and P. Zola, "Coordinated inauthentic behavior and information spreading on Twitter," *Decision Support Systems*, vol. 160, pp. 1–12, 2022.
[54] J. S. Pohl, D. Assenmacher, M. V. Seiler, H. Trautmann, and C. Grimme, "Artificial social media campaign creation for benchmarking and challenging detection approaches," in *The 2022 Workshop on Novel Evaluation Approaches for Text Classification Systems on Social Media (NEATCLasS'22)*. AAAI, 2022.
[55] A. Gruzd, P. Mai, and F. B. Soares, "How coordinated link sharing behavior and partisans' narrative framing fan the spread of covid-19 misinformation and conspiracy theories," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 118, 2022.
[56] X. Wang, J. Li, E. Srivatsavaya, and S. Rajtmajer, "Evidence of inter-state coordination amongst state-backed information operations," *Scientific Reports*, vol. 13, no. 1, 2023.
[57] Q. Kong, P. Calderon, R. Ram, O. Boichak, and M.-A. Rizoiu, "Interval-censored Transformer Hawkes: Detecting information operations using the reaction of social systems," in *The ACM Web Conference*, 2023, pp. 1813–1821.
[58] J. M. Sharp and I. A. Brudnick, "Yemen: Civil war and regional intervention," Congressional Research Service, Tech. Rep., 2019.
[59] J. Palik, "Dancing on the heads of snakes": The emergence of the Houthi movement and the role of securitizing subjectivity in Yemen's civil war," *Corvinus Journal of International Affairs*, vol. 2, no. 2-3, pp. 42–56, 2017.
[60] X. Guo and S. Vosoughi, "A large-scale longitudinal multimodal dataset of state-backed information operations on twitter," in *The 16th International AAAI Conference on Web and Social Media (ICWSM'22)*, vol. 16, 2022, pp. 1245–1250.
[61] R. Robertson, "Uncommon yet consequential online harms," *Journal of Online Trust and Safety*, vol. 1, no. 3, 2022.
[62] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Information Sciences*, vol. 513, pp. 429–441, 2020.
[63] R. S. Linhares, J. M. Rosa, C. H. Ferreira, F. Murai, G. Nobre, and J. Almeida, "Uncovering coordinated communities on Twitter during the 2020 US election," in *The IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'22)*, 2022, pp. 80–87.
[64] S. Tardelli, L. Nizzoli, M. Tesconi, M. Conti, P. Nakov, G. Da San Martino, and S. Cresci, "Temporal dynamics of coordinated online behavior: Stability, archetypes, and influence," *arXiv:2301.06774*, 2023.
[65] F. Şen, R. Wigand, N. Agarwal, S. Tokdemir, and R. Kasprzyk, "Focal structures analysis: identifying influential sets of individuals in a social network," *Social Network Analysis and Mining*, vol. 6, pp. 1–22, 2016.
[66] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, 2004.
[67] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On modularity clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 172–188, 2007.
[68] Y. Hu, "Efficient, high-quality force-directed graph drawing," *Mathematica Journal*, vol. 10, no. 1, pp. 37–71, 2005.
[69] A.-L. Barabási, "Network science," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1987, 2013.
[70] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, 1983.
[71] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019.
[72] M. Magnani, O. Hanteer, R. Interdonato, L. Rossi, and A. Tagarelli, "Community detection in multiplex networks," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–35, 2021.
[73] L. Mannocci, S. Cresci, A. Monreale, A. Vakali, and M. Tesconi, "MulBot: Unsupervised bot detection based on multivariate time series," in *The 10th IEEE International Conference on Big Data (BigData'22)*, 2022, pp. 1485–1494.
[74] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," *Social Networks*, vol. 39, pp. 62–70, 2014.
[75] G. Yan, "Peri-Watchdog: Hunting for hidden botnets in the periphery of online social networks," *Computer Networks*, vol. 57, no. 2, pp. 540–555, 2013.
[76] K. Starbird, "Disinformation's spread: bots, trolls and all of us," *Nature*, vol. 571, no. 7766, 2019.
[77] M. Mendoza, M. Tesconi, and S. Cresci, "Bots in social and interaction networks: Detection and impact estimation," *ACM Transactions on Information Systems*, vol. 39, no. 1, pp. 1–32, 2020.
[78] S. González-Bailón and M. De Domenico, "Bots are less central than verified accounts during contentious political events," *Proceedings of the National Academy of Sciences*, vol. 118, no. 11, 2021.
[79] D. Bucur, "Top influencers can be identified universally by combining classical centralities," *Scientific Reports*, vol. 10, no. 1, pp. 1–14, 2020.
[80] L. Luceri, S. Giordano, and E. Ferrara, "Detecting troll behavior via inverse reinforcement learning: A case study of Russian trolls in the 2016 US election," in *The 14th International AAAI Conference on Web and Social Media (ICWSM'20)*, 2020, pp. 417–427.
[81] K. Burghardt, A. Rao, S. Guo, Z. He, G. Chochlakis, B. Sabyasachee, A. Rojecki, S. Narayanan, and K. Lerman, "Socio-linguistic characteristics of coordinated inauthentic accounts," *arXiv preprint arXiv:2305.11867*, 2023.
[82] M. Kubli, E. Hoes, and N. Umansky, "The blackbox of social media content moderation: A first look into a novel Twitter dataset," *SocArXiv Preprint*, 2023.
[83] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, vol. 7, no. 1, 2020.
[84] N. P. Suzor, S. M. West, A. Quodling, and J. York, "What do we mean when we talk about transparency? toward meaningful transparency in commercial content moderation," *International Journal of Communication*, vol. 13, p. 18, 2019.
[85] D. Weber and L. Falzon, "Temporal nuances of coordination network semantics," *arXiv preprint:2107.02588*, 2021.

LORENZO CIMA received the master's degree in computer engineering from the University of Pisa. He is a Ph.D. student in information engineering at the University of Pisa. He is also associated with the Institute of Informatics and Telematics (IIT) of CNR. His research interests include social media analysis with a focus on coordinated online behaviour and content moderation.

LORENZO MANNOCCI received the master's degree in Data Science and Business Informatics from the University of Pisa. He is a Ph.D. student in the Italian National Ph.D. in Artificial Intelligence for Society at the University of Pisa and Cyber Intelligence Lab, at the Institute of Informatics and Telematics (IIT) of CNR. His research interests include social media, with a focus on coordinated inauthentic behavior and disinformation's spread.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3393482

MARCO AVVENUTI is a Full Professor of computer systems with the Department of Information Engineering, University of Pisa. He received a Ph.D. degree in information engineering from the University of Padua. He is chair of the master's degree in Artificial Intelligence and Data Engineering. His research interests include human-centric sensing and social network analysis.

MAURIZIO TESCONI received the Ph.D. degree in information engineering from the University of Pisa. He is a Researcher in computer science and leads the Cyber Intelligence research unit at the Institute of Informatics and Telematics (IIT) of CNR. His research interests include big data, web mining, social network analysis, and visual analytics within the context of open source intelligence. He is a member of the Permanent Team of the European Laboratory on Big Data Analytics and Social Mining, performing advanced research and analyses on the emerging challenges posed by big data.

STEFANO CRESCI received the Ph.D. degree in information engineering from the University of Pisa. He is a Researcher at IIT-CNR, Italy. His interests lay at the intersection of web science and data science, with a focus on content moderation and coordinated online behavior. For his achievements he received multiple awards, including an ERC grant on data-driven and user-centered content moderation (DEDUCE), the ERCIM Cor Baayen Young Researcher Award, and the IEEE Next-Generation Data Scientist Award.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/