Uncovering Coordinated Communities on Twitter During the 2020 U.S. Election

Renan S. Linhares, José M. Rosa, Carlos H. G. Ferreira
Universidade Federal de Ouro Preto
{renan.linhares, jose.rosa}@aluno.ufop.edu.br
chgferreira@ufop.edu.br

Fabricio Murai
Worcester Polytechnic Institute
fmurai@wpi.edu

Gabriel Nobre, Jussara Almeida
Universidade Federal de Minas Gerais
{gabrielnobre, jussara}@dcc.ufmg.br

***Abstract*—A large volume of content related to claims of election fraud, often associated with hate speech and extremism, was reported on Twitter during the 2020 US election, with evidence that coordinated efforts took place to promote such content on the platform. In response, Twitter announced the suspension of thousands of user accounts allegedly involved in such actions. Motivated by these events, we here propose a novel network-based approach to uncover evidence of coordination in a set of user interactions. Our approach is designed to address the challenges incurred by the often sheer volume of noisy edges in the network (i.e., edges that are unrelated to coordination) and the effects of data sampling. To that end, it exploits the joint use of two network backbone extraction techniques, namely Disparity Filter and Neighborhood Overlap, to reveal strongly tied groups of users (here referred to as communities) exhibiting repeatedly common behavior, consistent with coordination. We employ our strategy to a large dataset of tweets related to the aforementioned fraud claims, in which users were labeled as *suspended*, *deleted* or *active*, according to their accounts status after the election. Our findings reveal well-structured communities, with strong evidence of coordination to promote (i.e., retweet) the aforementioned fraud claims. Moreover, many of those communities are formed not only by suspended and deleted users, but also by users who, despite exhibiting very similar sharing patterns, remained active in the platform. This observation suggests that a significant number of users who were potentially involved in the coordination efforts went unnoticed by the platform, and possibly remained actively spreading this content on the system.**

***Index Terms*—Coordinated Behavior, Twitter, Backbone Extraction, Community Detection**

I. Introduction

Current Online Social Networks (OSNs) have enjoyed a mostly steady increase in the number of active users bolstered by ever-changing ways of communication[1]. On the flip side, social media platforms have become powerful political tools frequently abused to disseminate information and manipulate public opinion [1]. One of the most marked examples is the set of events related to the 2020 U.S. election, which ultimately resulted in an invasion of the Capitol[2]. Such events were widely debated on Twitter and, in response, the platform reported the suspensionof thousands of user accounts associated with incitation of violence and deliberate sharing of misinformation, notably extremist (QAnon) content[3,4].

[1]https://www.statista.com/statistics/number-of-worldwide-social-network-users/

[2]https://www.theguardian.com/us-news/us-capitol-breach

[3]https://blog.twitter.com/en_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington--

[4]https://www.theguardian.com/us-news/2021/jan/08/donald-trump-twitter-ban-suspended

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

Indeed, the literature displays strong evidence that many events during the U.S. election were influenced by coordinated user actions on Twitter [2, 3, 4, 5]. However, many questions remain to be answered. For example, how effective were the aforementioned account suspensions towards restraining the dissemination of the harmful content, and what is their impact on preventing further coordinated actions on the platform that could potentially harm the electoral process? Thus, a number of relevant prior efforts emerged to study how this phenomenon played out on Twitter while relying on many distinct network models [6, 7, 8, 9]. Such efforts often the links established between users (i.e., nodes in the network) and investigate the emergence of tightly connected groups of nodes in the network, referred to as communities, as evidence of groups of users purposely acting with a common goal, such as promoting the same pieces of content.

Unfortunately, such prior studies neglected two important aspects that can fundamentally undermine a network driven analysis of coordinated actions. The first one is the (potentially large) presence of *noisy edges* in the network, that is, marginal edges that are not related to coordination and, as such, do not contribute to (and may in fact obfuscate) the study of such phenomenon. Consider, for instance, an undirected network model where weighted edges connect users based on the number of common messages retweeted by both. In this case, noisy edges emerge from sporadic and weak interactions among users or even from independent behavior (e.g., a very popular tweet retweeted by many users). As such, these edges offer little (if any) evidence of coordinated activity. Indeed, recent work has advocated for the importance of removing such noisy edges, focusing on the remaining (salient) edges, or the so-called *network backbone* [10, 11], for the study of various phenomena, including coordination to disseminate content [11, 12, 13, 14, 15]. Amongst other features, it is expected that the backbone provides a clearer view of the phenomenon under study [10, 11].

Another neglected aspect is that often only a *partial view of the social network*, consisting of a sample of all user interactions, is available for analysis. This is certainly the case of data collected using Twitter's API. This partial view may indeed bias the understanding of how some users interact with each other in the network [16]. Moreover, as we show here, data sampling may indeed obfuscate the identification of tightly connected communities in the network, hurting interpretability of the results.

In this context, we are here driven by the general goal of investigating the formation of communities of Twitter users potentially involved in coordinated efforts to spread content related to claims of fraud during the 2020 U.S. election. However, unlike prior work,

we pay particular attention to the two aforementioned aspects: the presence of noisy and unrelated edges in the network and the potential effects of data sampling on community identification. To this end, we propose a novel approach to identify edges connecting groups of users who retweeted the same content with strong evidence of consistent and coordinated behavior. Our strategy explores the tandem combination of two backbone extraction methods, namely Disparity Filter [17] and a threshold-based method applied to the neighborhood overlap [18] of the two participating nodes. We refer to the latter simply as Neighborhood Overlap. Though both techniques have already been employed (in isolation) to study various phenomena [11, 19, 20], including coordinated actions [12, 21], to our knowledge, we are the first to use them together as a strategy to remove noisy edges while also minimizing the effects of data sampling on community identification.

The first method, Disparity Filter, removes weak and sporadic edges by identifying (and retaining only) those that are significantly stronger than the other links connected to the respective adjacent nodes, and as such offer evidence of consistent and repeated behavior. As a second step, Neighborhood Overlap removes peripheral and bridge connections, retaining only edges between pairs of users with a common neighborhood of nodes, which, in turn, also have similar sharing patterns. This second step is key to mitigating the negative effects of sampling on community extraction, as it narrows our focus to the strongly connected parts of the network.

We apply our proposed approach to the VoterFraud dataset [8], consisting of tweets related to fraud claims gathered during the 2020 U.S. election. VoterFraud also includes user labels categorized into those whose accounts remained *active*, were *suspended* or were *deleted*, after the election period. Our main findings are:

- Our two-step backbone extraction strategy is able to reveal well-structured communities of users involved in the dissemination of election fraud claims that cannot be easily identified neither in the original network nor in the backbone extracted by employing only Disparity Filter.
- Our results also reveal that, surprisingly, the identified communities consist of not only *suspended* users but also by *deleted* and *active* user accounts, which have gone unnoticed by Twitter despite sharing the same content. We hypothesize that the deleted accounts may have been created for temporary use only, being deleted by the users themselves afterwards. The active accounts, in turn, which correspond to the majority of the members of some communities, remain spreading (possibly fake) information on the platform.
- Finally, we find that *active* users are the ones more likely to remain in their communities over time, offering further evidence that these users may have actively participated in the coordinated efforts to spread fraud-related content. Nevertheless, they managed to escape Twitter's banning actions, which raises concerns as to whether such actions were indeed effective in containing the spread of that content.

The remainder of this paper is organized as follows. Related work is discussed in Section II. Section III presents our proposed strategy, while Section IV briefly describes the VoterFraud dataset. Our main results and findings are then discussed in Section V. Conclusions and future work are offered in Section VI.

II. Related Work

Prior studies related to our present effort fall into two bodies of work, namely: (i) network-oriented analyses of coordinated behavior on social media platforms, and (ii) analyses of Twitter data during the 2020 U.S. elections. We briefly review relevant studies in each category next.

A. Coordinated Behavior on Social Media Platforms

A number of prior studies have employed network models to detect and analyze coordinated actions to disseminate content on social media platforms, notably Twitter. For example, Pacheco et al. [22] investigated disinformation campaigns against the Syrian Civil Defense on Twitter. By exploring a network model connecting users who posted similar tweets, they uncovered groups of users who repeatedly shared the same content, offering strong evidence of coordinated behavior. Vargas et al. [23], in turn, explored a number of features derived from coordination network analysis to develop a supervised predictor of disinformation campaigns on Twitter. Similarly, Weber and Neumann [24] proposed a temporal window approach to uncover latent networks of cooperating Twitter users by exploring account interactions and metadata alone. Keller et al. [25], in turn, explored the network of retweets to study ad coordination during the presidential election in South Korea, whereas Nizzoli et al. [26], proposed a network-based framework to detect various degrees of coordinated social media efforts during the 2019 UK general election.

All the aforementioned studies explored the full set of network edges in their analyses. Yet, recent work has highlighted the importance of removing the so-called noisy edges and revealing the *backbone* to study a number of phenomena, including coordinated actions [2, 12, 13, 14, 19]. For example, Pacheco et al. [14] proposed to identify coordinated actions in social media, notably Twitter, by employing a simple backbone extraction method that removes edges whose weights fall below a given threshold. Others have employed more sophisticated backbone extraction methods. For example, the Disparity Filter method [17] has been used to uncover groups of users who spread similar content during the 2018 Brazilian elections on WhatsApp [12, 19], as well as to analyze conversations about climate change in Qatar on Twitter [21].

We here also advocate for the removal of noisy edges (i.e., backbone extraction) prior to uncovering evidence of coordinated behavior. However, unlike prior studies, we here propose a new strategy that employs two methods in tandem, so as to also minimize the effects of Twitter data sampling on community identification. Specifically, we first employ Disparity Filter to identify edges connecting pairs of users whose weights significantly deviate from the interaction patterns each user has with their other peers. As a second step, we explore the neighborhood overlap metric [18] jointly with a threshold-based method to further remove peripheral and bridge edges and only retain those building up well structured communities. To our knowledge, the only prior study that jointly used multiple backbone extraction methods focused on a quite different phenomenon, namely the formation of polarized ideological groups in networks of Congress members [20]. However, those

authors used two threshold-based backbone extraction methods, one applied on edge weights and one applied on the neighborhood overlap of the connected nodes. We are the first to jointly use the latter with Disparity Filter to study coordinated actions.

B. Studies on 2020 U.S. election

A number of studies examined the events of the 2020 U.S. election focusing on textual analysis of Twitter content. These studies varied in purpose, including hate speech detection [3], sentiment analysis before and after election [5], the reasons for account suspension [4] and the prevalence of content supplied by other platforms (notably YouTube and BitChute) [27].

Others have employed network-based models. For instance, Sharma et al. [9] employed cascade models to analyze narratives related to misinformation and conspiracy theories. Their findings suggest that Twitter's efforts to contain the QAnon conspiracy may not be effective. Ferrara et al. [2] examined election manipulation on Twitter from the perspective of the presence of bots and fact distortion. Similarly, Tran [6] explored a network of users sharing the same hashtags to investigate the presence of bots spreading candidate-specific hashtags. Others have used the network of retweets to analyze the formation of communities with similar ideologies [7] as well as to identify communities and characterize the presence of *suspended* users in the set of members [8].

Out of the aforementioned studies, the work by Abilov et al. [8] is the closest one in purpose to ours. Indeed we use the same dataset, kindly provided by the authors. Yet, there are some key differences between the two works. In [8], the authors modeled a directed retweet network and extracted communities from the complete network, without considering the presence of noise in the graph. In contrast, we here adopt a co-retweet network, by connecting users who retweeted the same original tweet. Thus, our study provides an orthogonal view: rather than capturing the communities which captures the flow of information we here aim to identify groups of users retweeting the same content, possibly from specific sources (i.e., accounts who posted the original tweets). To that end, we propose a novel backbone extraction strategy that combines two methods to remove noisy, peripheral and bridge edges, thus revealing better structured user communities with strong evidence of coordination.

III. Detection and Characterization of Coordinated User Communities

In this section, we describe our novel approach to detect and characterize communities of users acting in coordinated fashion to promote the sharing of specific pieces of content. We start by describing the network model adopted (Section III-A), and then introduce our proposed strategy to extract its backbone using a two-step approach to reduce noise while also minimizing the effects of data sampling (Section III-B). Finally, we discuss how we detect groups of users potentially involved in coordinated actions by extracting communities from the backbone and characterizing their content sharing patterns (Section III-C).

A. Network Model

Considering that user coordinated actions are naturally dynamic and may exhibit different patterns over time, we start by first discretizing the period of interest into fixed-size non-overlapping time windows. We then define a sequence of snapshots to be analyzed $T = (1, 2, ..., t)$ corresponding to those time windows. For each time window $\Delta_t \in T$, we build a network model as an undirected and weighted graph $G_{\Delta_t} = (V_{\Delta_t}, E_{\Delta_t})$. V_{Δ_t} is the set of nodes, representing users who retweeted during Δ_t, and E_{Δ_t} is the set of weighted edges connecting pairs of nodes v_i and v_j by the number w of retweets in common they made during Δ_t.

We adopt a co-retweet network to explicitly represent connections among users who aimed at promoting certain pieces of content (original tweets) by retweeting them. While the retweet network model used by Abilov et al. [8] (directed and weighted graph) captures the information dissemination by a chain of retweets, our co-retweet network focuses on the presence of groups of users who often retweet the same content, and, as such, have greater chance of being coordinating to promote that content. However, as we argue next, this network may contain a number of spurious and random edges that are *not* related to coordination[5]. As such, it is important to extract and work on the network *backbone*.

B. Backbone Extraction: A two-step approach

We here advocate that the network model described above may contain a large number of spurious and weak edges, which are not relevant for identifying coordinated efforts. In our context, such edges may emerge as side effect from the natural heterogeneity of user activity and content popularity that exists in social media. For instance, a particularly very popular tweet may be retweeted by multiple users acting completely independently (i.e., with no coordination). Similarly, a very active user may retweet the same content as many other users by pure chance, as side effect of their frequent retweets. These factors may indeed generate a large number of weak edges, which may obfuscate those that indeed are related to strong coordinated efforts. Thus, we need to identify (and focus on) co-retweeting patterns that are *more likely* to reflect coordination. The set of edges capturing such patterns, referred to as *salient* edges, is called the network backbone.

Moreover, when building the network from sampled data, as is the case of data gathered by the Twitter API, one must be aware that its structure (as that of the extracted backbone) may be affected by the partial view of user behavior [16]. In particular, the extraction of communities from the backbone may be disturbed by the presence of edges that, despite representing strong co-retweeting patterns (thus in the backbone), are only peripheral and, if retained, may blur the boundaries of the identifiable communities. For example, consider that users a, b, and c form a clique in the backbone. A fourth user d is also in the backbone but with a single edge connecting it to a. Although all four users exhibit some unusual co-retweeting patterns (and, as such, are kept in the backbone), peripheral nodes and edges (such as d and its edge to a) do not contribute to build a strong community structure and may indeed confound the identification of such structures. The same holds for nodes representing bridges and tree structures. While such structures may occur regardless of whether the data

[5]Indeed, the large presence of noisy edges may occur in different network models, not only the co-retweet network.

is sampled or not, we expect them to occur more often when only a (possibly very small) portion of the user interactions are represented in the network (and thus in the backbone).

To address the aforementioned challenges, we propose to combine, in tandem fashion, two backbone extraction methods, in order to extract a robust subset of nodes and edges that can reveal tightly connected user communities most likely involved in coordination. Our approach relies on the assumption that the removal of some (salient) edges is acceptable if such removal facilitates the identification of well structured communities.

Specifically, for each network G_{Δ_t} we employ the following two steps. First, we apply Disparity Filter [17] to extract an initial backbone B_{Δ_t}, retaining only heavy edges for which we can identify strong evidence of coordination. Disparity Filter relies on a reference model built on the assumption that the weights of all edges incident to a node should be uniformly distributed. Only edges that significantly deviate from such assumption (based on a predefined significance level α) are considered salient and kept in B_{Δ_t}. Since our networks are undirected, each edge is tested twice, and only those that significantly exceed the expected value (according to the uniform assumption) for both incident nodes are kept in the backbone. The intuition is to retain only edges connecting users with a number of retweets in common much larger than expected given their typical behavior, captured by the other incident edges connecting them to their neighbors. Such deviation from typical behavior is seen as strong evidence that some coordination between the two specific users is taking place. Indeed, this approach has been shown to be quite effective in investigating coordinated actions in various domains [12, 19, 21].

As a second step, aiming at dealing with the potential effects of data sampling, we further remove edges from B_{Δ_t} so as to retain only those belonging to strongly connected communities. That is, we focus on edges connecting users with a sufficiently large number of neighbors in common, acting similarly. We refer to this final set of edges, which build up the backbone to be analyzed, as C_{Δ_t} Note that $C_{\Delta_t} \subseteq B_{\Delta_t} \subseteq G_{\Delta_t}$. Note also that by removing peripheral and bridge edges, we may be losing some strong edges with respect to co-retweets. This is a conservative strategy, aimed at facilitating the discovery of tightly connected non-trivial (i.e., larger) communities, which is our primary goal. There might be a concern as to whether this approach may cause larger communities to be broken into multiple smaller communities. As we discuss in Section V, this seems not to be the case. Yet, the edges and nodes removed from B_{Δ_t} may still be of interest and we leave an analysis of their role in the coordinated actions for future work.

To build C_{Δ_t}, we make use of a network metric called neighborhood overlap [18]. Let (a, b) be an edge in the backbone and let $N(a)$ and $N(b)$ be the sets of neighbors of nodes a and b, respectively. The neighborhood overlap between a and b, $N_O(a, b)$ captures the strength of the relationship between nodes a and b by the Jaccard similarly of sets $N(a)$ and $N(b)$, that is, $N_O(a, b) = |N(a) \cap N(b)|/|N(a) \cup N(b)|$. We further remove peripheral and bridge edges from B_{Δ_t} by first computing the distribution of neighborhood overlap values for all its edges and then applying a threshold to retain only the top k% edges with larger values, where k is an input parameter. We refer to this second backbone extraction method as simply Neighborhood Overlap.

C. Community Detection and Characterization

The final step of our approach consists of identifying and characterizing tightly connected communities in backbone C_{Δ_t}. Given our proposed two-step backbone extraction approach, each such *community* refers to a group of users who retweeted the same original content (tweet) with disproportionately large frequency (compared to their respective typical patterns) and thus, offer strong evidence of coordination for massive content promotion.

Towards uncovering such communities, we employ the widely used Louvain's community detection algorithm, which considers communities as non-overlapping groups of nodes in a graph [28]. In summary, it is a greedy algorithm that aims at optimizing a metric called *modularity*. Modularity quantifies the relative density of edges within communities in comparison to a network with the same degree sequence but randomly connected nodes. Its value ranges from -1/2 to +1; while values equal to 0.4 or higher are considered as reliable indicatives of well-formed communities [29].

We characterize each identified community in terms of number of members of each class (*active*, *deleted* or *suspended* account) and content shared (e.g., use of hashtags associated with extremist groups, popular tweets and hashtags). We also analyze the communities' temporal dynamics by means of two metrics: *persistence* and *Normalized Mutual Information*. The former is the fraction of users in the backbone in window Δ_t who remained in the backbone in the next window Δ_{t+1} [13]. The latter is used to assess whether users who appear in the backbone across successive windows tend to remain in the same community or not [30]. Given two sets of partitions X_{Δ_t} and $X_{\Delta_{t+1}}$ defining community assignments to nodes in windows Δ_t and Δ_{t+1}, respectively, the mutual information of X_{Δ_t} and $X_{\Delta_{t+1}}$ represents the informational overlap between them. NMI values range from 0 (all users changed their communities) to 1 (all users remained in the same community).

IV. Case Study

Our case study is based on the VoterFraud[6] dataset provided by the authors of [8]. This is a collection of tweets gathered using the Twitter's API between October 23^{th} and December 16^{th} 2020[7]. The crawling was performed by searching for a set of predefined keywords and hashtags related to fraud allegations in the 2020 U.S. election. Examples of keywords include *#voterfraud*, *#discardedballots*, *#stopthesteal*, and *#stopcheating*.

Following the end of the data collection and the invasion on the Capitol on January 6^{th} 2021, Twitter suspended 70,000 accounts associated with extremist content[8] and radical groups (e.g., QAnon)[9]. A few days after the suspensions (on January 10^{th}) the authors of the VoterFraud dataset labeled all tweets/retweets based on the status of their users as either *suspended, deleted* or *active*, depending on whether the account had been suspended by Twitter, deleted by the user or still active on the platform, respectively.

[6]https://voterfraud2020.io/

[7]The week of the election started on November 3^{rd} 2020.

[8]https://www.bbc.com/news/technology-55638558

[9]https://blog.twitter.com/en_us/topics/company/2020/suspension

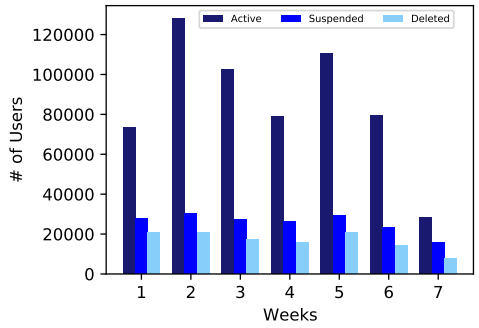


Fig. 1: Categories of users who shared content on each week.

We here use these labels as user categories and characterize the identified communities with respect to their members' categories.

We divided the dataset into 9 weekly intervals. Due to considerably lower amount of data, we chose to discard the first two weeks, thus analyzing 7 consecutive weeks. We also concentrated our analysis on the spread of content posted by the most popular accounts. To that end, we considered only tweets by accounts who received at least 5,000 retweets per week in total (i.e., across all their tweets). In total, over a period of 7 weeks, we selected 186 accounts who posted a total of 4,328 tweets and received 4,545,021 retweets from 323,912 unique user accounts (which, in turn, were used to build the co-retweet network).

Figure 1 shows the numbers of *active*, *suspended* and *deleted* accounts that shared content (tweet or retweet) on each week. *Active* users are the majority, as expected, followed by *suspended* and *deleted* users. Though smaller, the number of *suspended* users, who were actively spreading content during the election period, is quite large, around 30k on most weeks. Similarly, the number of accounts that were eventually deleted by the users is also noticeable, remaining above 15k in most weeks.

V. Results

In this section we discuss the main results of applying our proposed approach on the VoterFraud dataset. We start by analyzing backbone extraction and community identification (Section V-A) and then present our community characterization (Section V-B).

A. Backbone and Community Extraction

We first characterize the backbones built using our approach by highlighting the topological differences between them and the original network. We also include the backbone obtained only by using the Disparity Filter, which is a intermediary step, to compare with our complete approach. For the disparity filter, we assume $\alpha = 0.05$, and for the filter based on the neighborhood overlap metric, we assume the threshold given by the 95_{th} percentile of the neighborhood overlap distribution for each week.

Table I shows the properties of the three network models for the fist time window here analyzed. Looking at the properties of the original network, we find that there is a large number of edges, high average degree (Avg. Degree) and high density. Nevertheless, one can notice a small number of communities (# Comm.) identified by the Louvain algorithm that are poorly structured according to the computed modularity (Mod.). When only the disparity filter is applied, we observe that a large number of nodes and edges are discarded. Moreover, the few identified communities are more weakly structured than the original network. This suggests that the nodes remaining in the backbone do not form well-structured communities. As mentioned earlier, we assume that this is a side effect of Twitter sampling, as evidenced mainly by the fact that the number of connected components remains equal to one, indicating possible tree-like structures in the network.

TABLE I: Characteristics of the original network and the backbone extracted in Δ_1.

Netwok Model	# Nodes	# Edges	Avg. Degree	Density	# C. C.	# Comm.	Mod.
Original	121992	651835172	19686.5	0.0876	1	12	0.24
DF	28310	1709735	120.8	0.0043	2	8	0.22
DF and NB	13525	314142	46.4	0.0030	26	89	0.51

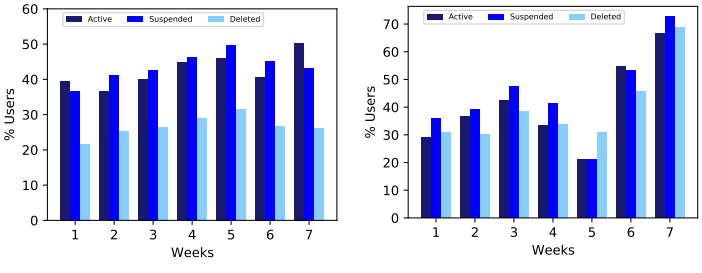
We then present the results of our approach that combines the Disparity Filter and Neighborhood Overlap (NB and DF). In week Δ_1, we found a threshold for neighborhood overlap of 0.39. This moderate value suggests that when the disparity filter is applied, many users have at least one edge that is significantly different from others in the original network, but most of them do not have many neighbors with the same characteristic. In other words, many neighborhood edges do not exist in the backbone of the Disparity Filter, indicating the vulnerability of this unique filter to the sampling problem. By tackling it, such backbone kept only 20% of the nodes whereas it shows much more structured communities that allow us to investigate the presence of potentially coordinated user groups.

After illustrating how our methodology works, we calculate and characterize the extracted backbones for each week. Table II summarizes the results obtained for the weeks we analyzed here. In general, our methodology captures a reasonable number of users that form the backbone. On average, these users are well-tied according to the clustering coefficient (Avg. clustering), and are divided into very well-structured communities, as shown by modularity (Mod.). The large number of communities (# Comm.) is to be expected as the number of components in the network increases through the application of backbone methods. By construction, these communities consist of users who co-retweet more frequently than their respective neighbors in the original network. Moreover, these users have a significant number of common neighbors in the backbone who do the same. To understand the similarity in content between the communities, we calculated the pairwise Jaccard index of the tweets they retweeted and found that the median and third quartile were below 0.18 and 0.3, respectively, for the seven weeks analyzed. This suggests that some communities have some similarity in content, but the larger portion is specific to each community.

Moreover, we examined the percentage of each class (*active*, *suspended*, and *deleted*) at each step of backbone extraction in order to understand how the users classes are related to the users permanence in the backbone. Figures 2a and 2b show the relative percentage of each of the classes for the extraction that uses only the Disparity Filter and for our approach that combines two filters, respectively. Focusing at Figure 2a, it shows that using

TABLE II: Breakdown of the backbones extracted by our methodology over the analyzed period.

Week	# Nodes	# Edges	Avg. Degree	Density	Avg. Clustering	# C. C.	# Comm.	Mod.
1	13525	314142	46.4	0.003	0.35	26	89	0.51
2	23589	1540198	130.5	0.005	0.44	51	135	0.48
3	24662	1855157	150.4	0.006	0.40	45	74	0.44
4	18392	2241889	243.7	0.133	0.51	40	99	0.53
5	15978	2612256	326.9	0.021	0.55	35	106	0.44
6	25010	1887246	150.9	0.0060	0.39	38	116	0.38
7	15944	317498	39.8	0.002	0.37	13	43	0.48



(a) DF (b) DF and NB

Fig. 2: Percentage of users remaining in the backbones according to the backbone extraction step.

only the Disparity Filter leaves a higher percentage of *active* and *suspended* than *deleted* users in the backbone. In turn, Figure 2b shows that when the second stage of the backbone is applied, the percentage for all classes is quite similar compared to the first filter. Therefore, a large fraction of the *active* and *deleted* users show similar connectivity and activity patterns as the *suspended* users.

Our backbones capture the idea of users retweeting a large number of common tweets while still being connected to common neighbors in the network. We also found that such backbones are formed by users of the three classes considered. To quantify the extent of which they spread the same content, we use the Gini index to measure the contribution of each user class to the dissemination of each tweet [31].

Figure 3 shows the distribution of the Gini index of tweets retweeted by users in the backbone for the entire period analyzed. We also considered the top 10% and top 50% most retweeted tweets. The values of the Gini index range from 0.23 to 0.67, showing that user participation by class in the distribution of tweets is unequal overall. To explain an idea of what these values mean, a Gini index of 0.237 means that retweets are distributed in three classes with the following percentages: 48.5%, 38.6%, and 12.8%. On the other hand, a tweet whose Gini index is 0.66 indicates the upper bound where only one class would be responsible for the distribution of all retweets. In general, Figure 3 shows that the most popular tweets (top 10%) have a much lower inequality than the others (top 50% and all tweets). Although the most widely shared tweets are assumed to be shared by more users, this suggests that users of different classes are involved in spreading the same tweets.

To summarize our analysis of the similarity of the behavior of the remaining users in the backbones with respect to their class, we apply principal component analysis (PCA) to the vectors representing the behavior of the users. PCA is used for dimensionality

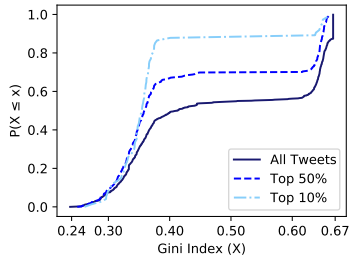
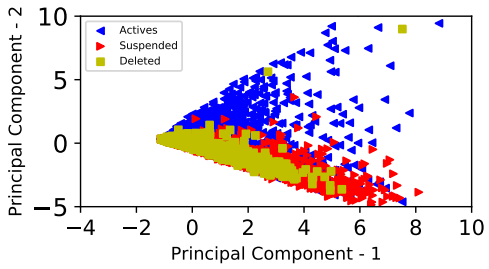
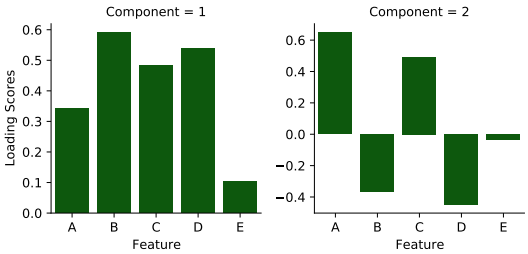


Fig. 3: Distribution of the Gini index among the tweets of the users who stayed on the backbone, considering the whole analyzed period.



(a) PCA



(b) Description of the 2 first principal components

Fig. 4: (a) 2-D representation of users using PCA. (b) Explainability of the two principal components. The bar represents the loading scores for the components (positive or negative).

reduction in multivariate analysis [32]. Our idea is to project users according to a set of characteristics about their behavior onto two principal components (PCs), i.e., along the axes that capture most of the variance in the data. To this end, we calculate the following metrics for each user: (A) # tweets created, (B) # retweets, (C) # words associated with extremist groups, (D) degree centrality, and (E) # weeks (out of 7) that the user stays on the backbone. Almost all of these features were extracted from our database and network properties. In particular, we use the vocabulary about extremist groups during the 2020 U.S. elections identified in [9].

Figure 4a shows the 2-D representation obtained for the users using the two principal components. The colors correspond to the classes and indicate that users of all classes have some overlap, especially the *suspended* and *deleted* ones. This means that the behavior of some users in the different classes is quite similar with respect to the given features. To understand which metrics best distinguish the users in Figure 4b, we show the *loading scores* for the two principal components. The Loading Score quantifies the contribution of each metric to a component. The larger the value (either positive or negative), the more the metric contributes

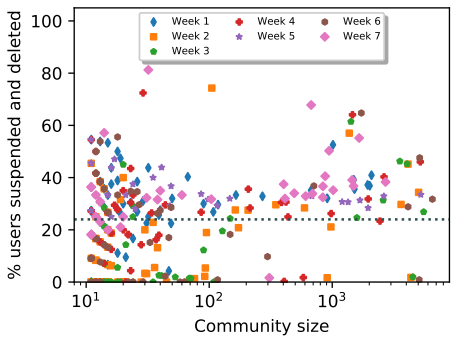


Fig. 5: Percentage of *suspended* and *deleted* users present in communities per week.

to the component (positive or negative). In Figure 4b, the bars represent the magnitude of the Loading Scores for each metric for Principal Component 1 (left) and Principal Component 2 (right). PC 1 (left) can be mainly associated with (B) # retweets, (C) # words associated with extremist groups and (D) degree centrality. Figure 4a shows that high values to these features for PC 1 are more common among users who belong to the *suspended* and *deleted* classes than the *active* ones.

For PC 2, the large positive values are associated with (A) # tweets created and (C) # words associated with extremist groups. Also, the large negative values are observed for (B) # retweets and (D) degree centrality (see Figure 4b, right plot). In Figure 4a, *suspended* and *deleted* users are concentrated in the $y \in [-5, 1]$ region and overlap with a substantial fraction of *active* users whose points are distributed along the y axis.

We conclude that *deleted* and *suspended* users are more homogeneous in terms of their behavior than *active* ones. Moreover, there is a fraction of users in the three classes that exhibit very similar behavior. In particular, they create less tweets, retweet more, spread content containing more words associated with extremist groups, and they perform the previous actions while surrounded by a larger number of users according to the degree centrality.

B. Characterization of Potential Coordinated Communities

We begin our analysis by examining the percentage of users by class in each community. For this purpose, we consider the classes *suspended* and also *deleted* together. The latter are potentially representative of temporary accounts that were used to disseminate information during the election process and deleted afterwards. Moreover, according to our analysis in Figure 4a, users of these classes exhibit similar behavior. Figure 5 shows the percentage of users whose accounts were *suspended* or *deleted* in each community, separated according to analyzed week[10]. In addition, the dashed line shows the expected value of *deleted* and *suspended* users in a community, given the original distribution of users belonging to these classes in all backbones. It is possible to observe a large number of communities formed by *deleted* and *suspended* users, with some of them exceeding the expected value. In some

[10]For visualization purposes, communities with fewer than 10 members were filtered out

TABLE III: Top 5 most retweeted tweets from two communities with the largest number of *suspended* and *deleted* users.

Most retweeted tweets by a community in week 4
"Breaking: There is a report that Biden is getting pressure by intelligence communities to concede due to the mounting of election fraud law suits."
"Can you feel it? There is a great shifting going on for the mounting pressure on @JoeBiden to concede, not Trump due to massive election fraud exposure?"
"Do you know why Kamala is still holding and not giving up her senate seat? Because she knows better that they wouldn't make it due to massive election fraud exposure that was executed by @JoeBiden. She knows they are doomed."
"That @SidneyPowell1 's website is deemed too toxic by Twitter, it explains to everyone there was massive voter fraud"
Most retweeted tweets by a community in week 6
"Judge asks Powell if Trump wins Georgia, can he win the election. Powell answers "Yes, he can." Fraud cannot be allowed to stand any where it occurs and there is fraud worse than Georgia's in other states."
"Why did the judge ask Sidney Powell if President Trump won Georgia could he win the election? Why was he looking past a request to investigate fraud further to the possible outcome of that investigation? I see red flags."
"There are currently only 8 of 29 Senators that have signed the Georgia petition for a special session. Obviously, they don't care about fixing the massive election fraud problem especially with the run-off coming up in January. People of Georgia must get on the phones NOW! #fixit"
"Re-read this. Then look up Insurrection Act. https://t.co/Tcc0gEpfdl"
"There's a turning of the narrative towards C h i n a: - their purchasing of U.S. politicians - their infiltrating of U.S. colleges and universities - their interference in the U.S. election"

cases, communities with one thousand to three thousand users are found, formed by 40-50% of the users of these classes.

We then looked at some communities to analyze some characteristics of the content spread by them. To do this, we selected two communities with more than one thousand members that contained the most *suspended* and *deleted* users. Following the Figure 5, we selected the two communities located on the rightmost and in the uppermost part of the figure indicated by the red (week 4) and brown (week 6) dots. Each of them has about 1500 users, 65% of whom have been suspended and deleted.

Table III summarizes the most retweeted tweets by community members. In general, we can see that these communities are strongly associated with tweets about election fraud allegations. We also examined the top 5 hashtags used by users of these communities and found for the two examples: *#stopthesteal*, *#voterfraud*, *#electionfraud*, *#ripjournalism*, *#fightbackforamerica*, *#fightfortrump*, and *#maga*. Among the keywords and hashtags associated with extremist groups, we find the most mentions to *WWG1WGA*, *kag*, *maga*, *thestorm*, and *cabal*. In short, such communities consist of many users, some of whom have not been suspended or deleted from Twitter, and who have also been massively involved in spreading election fraud messages.

Finally, we analyze the dynamics of the communities during the 7 weeks of observation. By computing the metrics *persistence* and *NMI* (see Section III), we examine the extent to which users persist in the network backbone and whether they stay in their communities over weeks. We also want to investigate whether these temporal aspects vary depending on their class membership. In other words, we want to test whether there is a class in which users are more likely to remain than in another, and to what extent this is the case. To this end, we perform a separate analysis for all users and for each class for consecutive weeks Δ_t and Δ_{t+1}.

TABLE IV: Temporal evolution of community membership

All users						
Sequential Weeks	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**
Persistence	31.07%	36.10%	31.72%	10.26%	15.26%	31.06%
NMI	0.075	0.365	0.272	0.061	0.042	0.143
Activeusers						
Sequential Weeks	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**
Persistence	26.62%	30.10%	24.39%	9.93%	15.10%	23.39%
NMI	0.092	0.408	0.294	0.082	0.052	0.171
Suspended users						
Sequential Weeks	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**
Persistence	44.08%	56.48%	53.20%	10.85%	19.92%	53.10%
NMI	0.078	0.214	0.224	0.076	0.078	0.097
Deleted users						
Sequential Weeks	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**
Persistence	23.46%	38.46%	36.30%	10.86%	9.12%	36.94%
NMI	0.123	0.203	0.199	0.173	0.119	0.097

Table IV reports a *Persistence* and *NMI* for consecutive weeks. Note that the metrics are calculated for all users and for each class. Considering all users, the *Persistence* shows that about 30% of the users stay in the backbone in the initial weeks closer to the election week and in the last two weeks. However, the NMI shows that they are quite dynamic in terms of community membership. Focusing on the classes, *suspended* and *deleted* users stay in the backbone more, but the *active* users retain communities more over the period.

VI. Conclusions and Future Work

In this paper, we examined the spread of fraud allegations during the 2020 U.S. election on Twitter. For this, we used a strategy to model how interactions between these users on Twitter occurred. It takes into account previously overlooked aspects of studying Twitter data and focuses on uncovering communities of users that coordinate in the dissemination of content.

By applying this to a large dataset covering a seven-week period, we observed better-structured communities formed by users who massively disseminated information related to those rumours. We found that such communities consisted of (i) *suspended*, (ii) *deleted*, and (iii) still *active* users that exhibit behavior patterns similar to group (i). We also observed that the users who remained active during the observed period were found in backbones and even grouped in the same communities, thus exhibiting quite similar behavior. Taken together, our results suggest that the account ban imposed by Twitter after the 2020 election may not have effectively stopped the spread of information and conspiracy theories and should have captured a larger number of users.

As future work, we intend to perform a deeper analysis of the content shared in these communities, such as the discussion topics during the analyzed period, and to investigate the presence of bots. Our methodology can also be applied to other phenomena related to information dissemination on Twitter, such as other political contexts and public health.

References

[1] S. Bradshaw and P. N. Howard, "The global organization of social media disinformation campaigns," *Journal of International Affairs*, 2018.

[2] E. Ferrara, H. Chang, E. Chen, G. Muric, and J. Patel, "Characterizing social media manipulation in the 2020 us presidential election," *First Monday*, 2020.

[3] L. Grimminger and R. Klinger, "Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection," in *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2021.

[4] F. Chowdhury, D. Saha, K. Hasan, Saha, and A. Mueen, "Examining factors associated with twitter account suspension following the 2020 us presidential election," in *Advances in Social Network Analysis and Mining*, 2021.

[5] H. N. Chaudhry, Y. Javed, F. Kulsoom, Z. Mehmood, Z. I. Khan, U. Shoaib, and S. H. Janjua, "Sentiment analysis of before and after elections: Twitter data of us election 2020," *Electronics*, 2021.

[6] H. Tran, "Studying the community of trump supporters on twitter during the 2020 us presidential election via hashtags# maga and# trump2020," *Journalism and Media*, 2021.

[7] Y. Dai, "Using 2020 u.s. presidential election to study patterns of user influence, community formation and behaviors on twitter," Ph.D. dissertation, The Pennsylvania State University, 2021.

[8] A. Abilov, Y. Hua, H. Matatov, O. Amir, and M. Naaman, "Voterfraud2020: a multi-modal dataset of election fraud claims on twitter," in *International Conference on Web and Social Media*, 2021.

[9] K. Sharma, E. Ferrara, and Y. Liu, "Characterizing online engagement with disinformation and conspiracies in the 2020 u.s. presidential election," *International AAAI Conference on Web and Social Media*, 2022.

[10] M. Coscia and F. M. Neffke, "Network backboning with noisy data," in *International Conference on Data Engineering*, 2017.

[11] C. H. Gomes Ferreira, F. Murai, A. P. Silva, M. Trevisan, L. Vassio, I. Drago, M. Mellia, and J. M. Almeida, "On network backbone extraction for modeling online collective behavior," *Plos one*, vol. 17, no. 9, p. e0274218, 2022.

[12] G. P. Nobre, C. H. Ferreira, and J. M. Almeida, "A hierarchical network-oriented analysis of user participation in misinformation spread on whatsapp," *Information Processing & Management*, 2022.

[13] C. Ferreira, F. Murai, A. Silva, J. M. Almeida, M. Trevisan, L. Vassio, I. Drago, and M. Mellia, "Unveiling community dynamics on instagram political network," in *ACM Conference on Web Science*, 2020.

[14] 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 *International Conference on Web and Social Media*, 2021.

[15] C. Ferreira, F. Murai, A. Silva, J. Almeida, M. Trevisan, L. Vassio, M. Mellia, and I. Drago, "On the dynamics of political discussions on instagram: A network perspective," *Online Social Networks and Media*, 2021.

[16] A. Campan, T. Atnafu, T. M. Truta, and J. Nolan, "Is data collection through twitter streaming api useful for academic research?" in *IEEE Big Data*, 2018.

[17] M. Á. Serrano, M. Boguná, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *National Acad. of Sciences*, 2009.

[18] A.-L. Barabási *et al.*, *Network Science*. Cambridge Press, 2016.

[19] G. P. Nobre, C. H. G. Ferreira, and J. M. Almeida, "Beyond groups: Uncovering dynamic communities on the whatsapp network of information dissemination," in *International Conference on Social Informatics*, 2020.

[20] C. H. Ferreira, F. Murai, B. Matos, and J. M. Almeida, "Modeling dynamic ideological behavior in political networks," *Web Science Journal*, vol. 7, 2019.

[21] S. Abbar, T. Zanouda, L. Berti-Equille, and J. Borge-Holthoefer, "Using twitter to understand public interest in climate change: The case of qatar," in *International AAAI Conference on Web and Social Media*, 2016.

[22] D. Pacheco, A. Flammini, and F. Menczer, "Unveiling coordinated groups behind white helmets disinformation," in *The Web Conference*, 2020.

[23] L. Vargas, P. Emami, and P. Traynor, "On the detection of disinformation campaign activity with network analysis," in *ACM Cloud Computing Security Workshop*, 2020, pp. 133–146.

[24] D. Weber and F. Neumann, "Amplifying influence of coordinated behaviour in social networks," *Social Network Analysis and Mining*, 2021.

[25] F. Keller, D. Schoch, S. Stier, and J. Yang, "Political astroturfing on twitter: How to coordinate a disinformation campaign," *Political Comm.*, 2020.

[26] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and M. Tesconi, "Coordinated behavior on social media in 2019 uk general election," in *ICWSM*, 2021.

[27] M. C. Childs, C. Buntain, M. Z. Trujillo, and B. D. Horne, "Characterizing youtube and bitchute content and mobilizers during us election fraud discussions on twitter," in *Web Science*, 2022.

[28] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[29] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, 2004.

[30] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, vol. 11, 2010.

[31] S. Yitzhaki, "Relative deprivation and the gini doefficient," *The Quarterly Journal of Economics*, 1979.

[32] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B*, 1999.