Zouzou and Varol *EPJ Data Science* (2024) 13:62
https://doi.org/10.1140/epjds/s13688-024-00499-6

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Unsupervised detection of coordinated fake-follower campaigns on social media

Yasser Zouzou[1] and Onur Varol[1,2*]

*Correspondence:
onur.varol@sabanciuniv.edu
[1]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey
[2]Center of Excellence in Data Analytics, Sabanci University, Istanbul, Turkey

Abstract

Automated social media accounts, known as bots, are increasingly recognized as key tools for manipulative online activities. These activities can stem from coordination among several accounts and these automated campaigns can manipulate social network structure by following other accounts, amplifying their content, and posting messages to spam online discourse. In this study, we present a novel unsupervised detection method designed to target a specific category of malicious accounts designed to manipulate user metrics such as online popularity. Our framework identifies anomalous following patterns among all the followers of a social media account. Through the analysis of a large number of accounts on the Twitter platform (rebranded as X after the acquisition of Elon Musk), we demonstrated that irregular following patterns are prevalent and are indicative of automated fake accounts. Notably, we found that these detected groups of anomalous followers exhibited consistent behavior across multiple accounts. This observation, combined with the computational efficiency of our proposed approach, makes it a valuable tool for investigating large-scale coordinated manipulation campaigns on social media platforms.

Keywords: Computational social science; Fake-followers; Bots; Online coordinated activities; Misinformation

1 Introduction

Twitter was originally established as a personal social networking platform, where users can follow each other and share messages with their followers. In recent years, Twitter has been used by leading politicians and large organizations worldwide for sharing information and news. This, along with the ease with which information cascades can form on social media, has rendered Twitter a vulnerable space for the spread of misinformation [18, 47]. Misinformation is false or misleading information that is deliberately or inadvertently propagated [49]. Disinformation is the subset of misinformation that is deliberately propagated [19]. To efficiently spread disinformation, automated accounts, known as *bots*, have been widely used on Twitter [10, 17]. Bots are run through APIs or other mechanisms to circumvent detection mechanisms and are considered legitimate as long as they openly state on the platform that they are bots [1, 50]. However, bots are also maliciously used on social media platforms to spread disinformation and manipulate user popularity and engagement metrics [6, 39, 48]. Identifying malicious bots is crucial in suspending such

Springer

accounts and conducting research to understand the role of bots in manipulation campaigns on social media [21, 46, 50]. Recently, the acquisition of Twitter by Elon Musk also raised concerns about bot prevalence on the platform, since the amount of bot accounts reported in SEC documents and academic research differs and the definitions of social bots or spam accounts are not easy to be agreed on [43]. Musk claims reduced activity of automated activities and hate speech; however, researchers reach opposite conclusion [2, 20].

In this study, we propose a novel unsupervised approach to detect groups of coordinated anomalous followers. Specifically, our detection approach targets accounts that are created on similar dates and follow Twitter users in coordination. A recent study on coordinated online influence campaigns indicated that accounts created in bursts in short periods exhibited similar behavior amongst themselves and were more likely to be bots [3]. Another study showed that 47% of all Twitter accounts following US Senators before the 2018 elections were created within six months in 2017; 41% percent of these accounts never posted any tweets and had an average of 1.7 followers and 99 friends [41]. Furthermore, a New York Times investigation that tracked fake accounts sold in bulk as fake followers showed that these accounts tend to have similar creation dates and follow the target user successively [9]. A subsequent study on journalists on Twitter identified similar patterns in fake followers which were used to increase online popularity and manipulate the online perception of journalist accounts [48]. While coordination in account activity does not necessarily coincide with automation [36, 37], accounts created on similar dates and following users successively are more likely to be automated [3, 48]. Indeed, this is intuitive because humans may engage in similar social media activities, such as posting about hot topics. However, there is no reason for accounts created on similar dates to follow the same users simultaneously. Throughout this study, we refer to the users that we aim to detect as anomalous followers or fake followers. This study is divided into two parts: (i) developing a method to detect groups of anomalous followers (ii) utilizing the developed detection method to investigate anomalous followers in Turkish political Twitter.

In the first part of this study, we tested four existing anomaly detection algorithms with hand-crafted features and introduced a method for this detection task that doesn't rely on engineered features. Due to the absence of a labeled dataset of anomalous followers as we define them, we first tested the existing algorithms and our suggested method on a synthetic dataset to choose the most suitable approach for this detection task before applying it on a real dataset. It is important to note that we only used unsupervised methods to avoid learning patterns that are specific to the synthetic data. The synthetic dataset consisted of simulated anomalous followers inserted into the followers of users on the Dribbble platform (dribbble.com), which is a social networking platform for digital designers. We chose the Dribbble platform for two reasons: (i) As a professional network for designers, it is less susceptible to having fake followers (ii) Follower data collected from Dribbble includes following times, which allowed us to validate the follow-time estimation method [31] that we used in our analyses in the second part of the study.

In the second part of the study, we applied the best-performing detection method on the followers of 1318 Twitter accounts of Turkish politicians and media outlets taken from the #Secim2023 dataset [34]. Since our approach addresses all the followers of a user, we explored its ability to identify users with irregular followers out of a pool of users, i.e., given a large set of Twitter accounts, which are the accounts that have anomalous followers

amongst their followers? Then, we explored the detected anomalous follower accounts and qualitatively verified that they were indeed fake accounts. Finally, we conducted an analysis to explore the coordinated activities of the detected fake accounts.

2 Related works

2.1 Bot detection

Bot detection methods aim to differentiate human-operated social media accounts from automated accounts. Cresci thoroughly reviewed bot detection approaches for the last decade [10]. Algorithmically, bot detection methods can be categorized into supervised, semi-supervised, and unsupervised methods. Supervised methods rely on training classifiers on labeled datasests of bots and human accounts, using information such as profile metadata, user network, shared content, and temporal activity. [45]. This information can be incorporated as either manually engineered statistical features [11, 32, 50] or features that are learnt by deep neural networks [15, 28]. The reliance of supervised methods on labeled datasets limits their ability to generalize to unseen types of bots [16]. This limitation has been addressed by training ensembles of specialized classifiers to detect different types of bots [28, 38]. Furthermore, to address the constant evolution of bots on social media, a new path of study has focused on adverserially creating synthetic bots that simulate the evolution of bots to evade detection methods [14, 24]. Besides fully supervised approaches, semi-supervised approaches that rely on a small set of labeled bot accounts have also been utilized [23, 33]. Semi-supervised detection methods rely on a network representation of user relationships and interactions, in which accounts that are most similar to the labeled bot accounts are considered suspicious. Unsupervised approaches [7, 13, 29, 30] rely on clustering users based on their similarities and identifying clusters that have suspicious properties or behaviors. Unsupervised detection methods do not require labeled datasets and are therefore not susceptible to the limitations of supervised methods that originate from biases in the available labeled datasets. However, unsupervised detection still include an intrinsic bias that originates from the definition of anomalous activity used to identify bots. For instance, RTBust detects bots based on their temporal retweeting behavior [30]; DeBot uses temporal correlations between user activities to identify bots [7]; MulBot leverages multivariate time series of daily user activities to identify groups with similar behavior [29]; and Social Fingerprinting utilizes sequences of user actions on Twitter to identify similar users with suspicious behavior [13].

2.2 Coordination detection

Coordination detection methods aim to identify groups of users that exhibit exceptional or suspicious similarity in their behavior on social media [8]. These methods assume that user activities on social media are mostly independent and a significant interdependence in their activities indicates coordination. The most common similarity measures used for coordination detection are shared content similarity and temporal activity similarity. Nizzoli et al. used the similarity in retweeted tweets to uncover coordinated behaviors in a network of Twitter users by iteratively filtering the similarity network between these users [36]. Pacheco et al. utilized the similarity in hashtag sequences and shared images among other similarities to detect coordinated Twitter accounts [37]. Coordination detection methods that use temporal similarity model user activities on social media as ordered sequences of events and detect groups with highly concerted activity patterns [13, 37, 40].

While most studies on coordination examine static snapshots of user networks, Tardelli et al. conducted a dynamic analysis of coordinated activities and showed that users may move between different coordinated communities throughout time [42]. Furthermore, since coordination can occur organically, as is the case with grassroots activism, efforts have been made to combine propaganda measures with coordination detection to differentiate between harmful and harmless coordinated behavior [22].

Our suggested approach is an unsupervised method that leverages the interplay between account creation dates and following behaviors to identify anomalous patterns. Specifically, we detect coordinated followers exhibiting unexpected similarities in their creation dates, which have been shown to indicate suspicious behavior [3, 41].

3 Methods

3.1 Datasets

Synthetic dataset: We obtained the profile data of the followers of 2834 users on the Dribbble platform, a social networking platform for digital designers. The 2834 collected users have follower counts ranging between 1000 and 110,000. The distribution of the follower counts and data preparation details are given in SI-Sect. 1 (Additional file 1). We produce two types of synthetic anomalous follower patterns that we derive from the previously observed anomalous follow patterns [48]. Type 1 followers represent a batch of follower accounts that were created in a limited range of time in the past and followed the user consecutively. Type 2 followers represent a batch of follower accounts that followed the user consecutively and almost immediately after being created. By varying the count of synthetic followers, the spread of these followers, and combinations of the two synthetic types, we generated 55 permutations of the original dataset, resulting in a total of 55×2834 synthetic datasets, each having distinct synthetic anomalous followers. Further description of the parameters used to generate these synthetic followers is presented in SI-Sect. 2 (Additional file 1).

Twitter dataset: We used the Twitter accounts of 1318 Turkish politicians and media outlets from the #Secim2023 dataset [34] to experiment with our anomalous follower detection approach. This dataset collected various information available on Twitter API, including list of followers and profile metadata of the follower accounts. Although this is an unlabeled dataset, it allowed us to explore the types of anomalous followers that our approach is capable of capturing in real Twitter data. For additional validation, we also conduct an annotation study to measure performance of the anomalous follower detection model. The number of followers of these accounts ranges between 1000 and 20 million. The distribution of the follower counts and data preparation details are given in SI-Sect. 1 (Additional file 1).

3.2 Follower map

To better visualize the groups of anomalous followers observed in previous studies and that we aim to detect, we introduce an instrument that we call *follower map*. A follower map is a graph that plots all the followers of a certain account based on their follow rank (x-axis) and their account creation dates (y-axis). It is important to note that the x-axis corresponds to the order of following and not the exact time of following. Figure 1a is a segment of the follower map of a user in our Twitter dataset showing the first 15,000 followers of this account. The rising upper bound (blue line) at each rank represents the

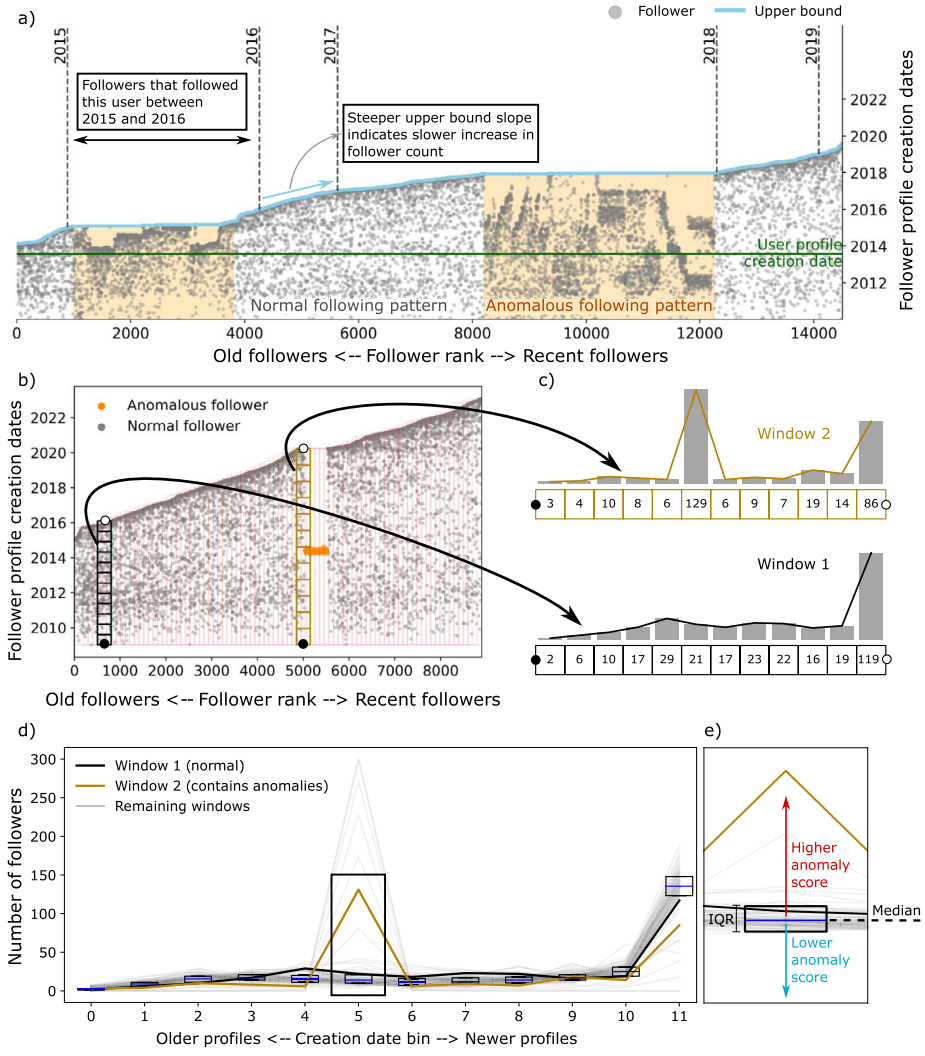

Figure 1 *Follower map and Sliding Histogram illustration* (a) A follower map with zones of anomalous following patterns highlighted in orange. The vertical dashed lines mark the beginning of each year based on the estimated following times. (b) A follower map with inserted synthetic irregular followers (orange) showing all sliding windows (light gray) with two of them highlighted in black and orange. (c) The histograms corresponding to the two highlighted windows in the follower map. Window 1 only includes normal followers and Window two includes anomalous followers. The numbers are the count of followers that fall within each bin. (d) All histograms plotted together as line plots, with the black and orange lines corresponding to the black and orange windows above. (e) A zoom-in on bin No. 5 showing the median and interquartile range (IQR) of all histograms at this bin

most recent profile creation date up to the current follower rank. Given the fact that the time at which a follower starts following is certainly after the creation date of all previous followers, the upper bound can be used as a proxy to estimate follow times. The beginning of each year based on the estimated following times are shown in the figure. The exact details of the follow time estimation algorithm presented in a work by Meeder *et al.* [31]. In areas where the follow pattern is normal, followers are evenly distributed along the y-axis, with a slightly denser zone near the upper bound. The dense zone beside the upper bound is attributed to accounts following this user just after being created. In the anomalous zones of the map (highlighted in orange), we see batches of follower accounts created on

similar dates and following the user consecutively. In these anomalous zones, the upper boundary remains horizontal and then returns to its original slope after the anomalous follower batch, suggesting that this group followed the user within a short time. While these anomalous patterns can be visually identified, we aim in this study to provide a detection method that can automatically identify similar patterns among the followers of a social media user.

3.3 Anomalous follower detection

We experimented with two approaches to detect the anomalous followers in a follower map: (i) employing existing unsupervised anomaly detection methods with manually generated features from the follower map (ii) using a novel unsupervised approach specifically designed for this task without relying on engineered features.

3.3.1 Anomaly detection methods

To apply the unsupervised anomaly detection algorithms, we generated features from the follower map that can help isolate the anomalous followers in the feature space. The anomalous following patterns we are interested in detecting consist of followers that have similar profile creation dates and follow an account consecutively. This translates to atypically dense zones in the follower map. Therefore, we used features that describe the local density around followers in the follower map. We also used features to describe the position of a follower in the follower map to prevent mislabeling dense zones that are typically dense in normal follower maps, such as those near earlier ranks or close to the upper bound. The features are described in detail in SI-Sect. 3 (Additional file 1). We evaluated 4 unsupervised anomaly detection algorithms using the engineered features: (1) Isolation Forest [27] (2) Local Outlier Factor (LOF) [5] (3) Empirical-Cumulative-distribution-based Outlier Detection (ECOD) [26] (4) Gen2Out [25].

Isolation Forest: In the isolation forest algorithm [27], a forest of decision trees with random splits is grown, and higher anomaly scores are given to points that have a shorter average path in the forest. The path length is the number of splits from the root node required to isolate a data point in a leaf node. This definition of anomaly score is based on the fact that anomalies, by definition, are "few and different". Therefore, by randomly splitting nodes in a decision tree, we expect anomalies to be isolated earlier than normal points since they reside in sparser areas of the feature space. Isolation forest trees are created using sub-samples of the dataset to avoid two common problems in anomaly detection: swamping and masking. The isolation forest algorithm requires two main hyperparameters: number of trees in the forest and sub-sample size. In our experiment, we use 200 trees and a sub-sample size of 256, which is the size recommended by the authors

Local Outlier Factor: The Local Outlier Factor algorithm (LOF) [5], is designed to detect local outliers, i.e., points that lie in areas with less density than that of the nearest cluster of points. A point is assigned a high anomaly score if the average distance between this point and its nearest neighbors is greater than the average distance between its nearest neighbors and their nearest neighbors. The main hyperparameter in this algorithm is the number of nearest neighbors to be considered (*MinPts*). Since we are dealing with groups of anomalous followers, we expect them to be clustered together in the feature space. Thus, *MinPts* should be set to a value greater than the number of anomalies in a group of anomalous followers. Otherwise, this cluster of anomalies would be assigned low

anomaly scores since all the nearest neighbors would be inside the same cluster. However, the fact that we do not have prior information about the number of anomalies that we expect to see in one group makes it hard to choose the value of *MinPts*. In our experiment, we set *MinPts* to 3% of the total number of followers of each user. Although users may have an anomaly ratio greater than 3% in their followers, larger values of *MinPts* result in prohibitive run times and memory usage for users with a large number of followers.

ECOD: The Empirical-Cumulative-distribution-based Outlier Detection (ECOD) method assigns high outlier scores to data points that have a low tail probability under the joint cumulative distribution function (CDF) of the data [26]. The joint CDF is estimated by assuming that the dimensions (features) of the data are independent. Thus, the product of the univariate empirical CDFs (ECDF) of all dimensions is used as an estimate of the joint CDF. Data points that have extreme feature values, based on the distribution of the corresponding feature, receive high outlier scores. This method does not require any hyperparameter tuning and is computationally efficient. However, due to the independence assumption, the interactions between features are not considered in this method.

Gen2Out: The Gen2Out method relies on the same concept of the Isolation Forest method, i.e., an anomalous point tends to have a shorter average path from the root node to its leaf node in a forest of random decision trees, referred to as *AtomTrees* in this study [25]. However, instead of growing full trees on subsets of the dataset, trees are grown to a predefined maximum depth using all of the data points. The path length of each data point (q) to its leaf node is then estimated using Eq. (1), where h_0 is the path length up to the final node that the data point q falls in, l_{busy} is the number of points in that node, and $H(l_{busy})$ the estimated depth of an *AtomTree* grown using l_{busy} points.

$$h(q) = h_0 + H(l_{busy}) \tag{1}$$

The authors demonstrate that a linear relationship exists between the depth of the *AtomTree* and the logarithm of the count of data points used to construct the tree, regardless of the distribution of the data. Based on this observation, a number of *AtomTrees* are grown using several subsets of the data set to fit a linear function H that maps the logarithm of the count of points to the depth of a fully grown *AtomTree*. The anomaly score assigned to a point q is then computed using Eq. (2), where n is the number of points in the considered data set and $E[h(q)]$ is the average path length of point q in the forest.

$$s(q, n) = 2^{-\frac{E[h(q)]}{H(n)}} \tag{2}$$

3.3.2 Task specific approach - sliding histogram

Our proposed approach specifically addresses anomalous groups defined in this study, i.e., dense groups of followers created in a tight time range that follow a user consecutively. This is achieved by finding groups of followers that have a local distribution in the follower map that is significantly different from the overall distribution of the followers of the same user. The steps of this method are described as follows:

- A window with a predefined width (b) is slid along the rank axis of the follower map. The window stretches on the timestamp axis between the lower and upper bounds of the follower timestamps at that position (Fig. 1b).

- At each position, the window is divided into a predefined number of bins (N_{bins}) in the creation date axis and the number of followers in each bin is computed (Fig. 1c). These histograms are shown as line plots in Fig. 1d.
- At each creation date bin, the median and inter-quartile range (IQR) of all windows (histograms) are computed.
- An anomaly score is assigned to each bin in each window using Eq. (3). Thus, each bin of each window is assigned a score that is the number of IQRs between the follower count in that window's bin and the median of follower counts in all windows at the same creation date bin (Fig. 1e).

$$A_{ij} = \frac{H_{ij} - M_j + 1}{IQR_j + 1} \tag{3}$$

Where H_{ij} is the count of followers in the bin j of the window i, and M_j and IQR_j are the median and IQR of follower counts in the bin j across all windows, respectively.

- Since we are using a sliding window, each follower appears in more than one window. Thus, an anomaly score can be assigned to each individual follower f using a weighted average of all bin scores A_{ij} that include the follower f. The weight λ_{fi} (Eq. (4)) takes its maximum value when the follower f is in the center of the bin and its minimum value when the follower f is at the edge of the bin. The anomaly score is then computed using Eq. (5)

$$\lambda_{fi} = 1_{f \in W_i} \left(\frac{\frac{b}{2} - |R_f - C_i| + 1}{\sum_j \frac{b}{2} - |R_f - C_j| + 1} \right) \tag{4}$$

Where b is the width of the sliding window, R_f is the rank of the follower f, and C is the center of the sliding window.

$$score_f = \sum_{j}^{N_{bins}} \sum_{i}^{N_{windows}} \lambda_{fi} A_{ij} 1_{f \in W_i} 1_{f \in B_{ij}} \tag{5}$$

4 Results

4.1 Evaluation on synthetic dataset

Table 1 shows the performance metrics of the anomaly detection methods using three different window sizes, averaged across all the synthetic Dribbble datasets. We evaluated the methods using the area under ROC curve (AUC), average precision (AP), and precision when the top 50 scores are considered (precision@50). Empirical-Cumulative-distribution-based Outlier Detection (ECOD) performs best among the feature-based models. However, our suggested method clearly outperforms the feature-based methods in this task, especially when looking at the precision measures.

Although we know that Turkish politicians on Twitter have anomalous followers, we still conducted an experiment by inserting synthetic data on this dataset. Our approach estimated anomaly scores, and we used them to assess how well we could detect synthetic data. The false positives we encountered in these experiments are probably real-world anomalous followers. Despite these limitations, our model achieved an average AUC score of 0.91 and an average precision of 0.61. These results are comparable to the experimental results obtained from Dribbble dataset.

Table 1 *Results on synthetic dataset.* Area under ROC curve, average precision, and precision at 50 mean (std) values for all methods using different window sizes

Window	Method	AUC	AP	P@50
W51	ECOD	0.71 (0.22)	0.31 (0.13)	0.26 (0.21)
	Gen2Out	0.62 (0.31)	0.26 (0.10)	0.15 (0.18)
	IsolationForest	0.61 (0.30)	0.24 (0.11)	0.09 (0.17)
	LocalOutlierFactor	0.54 (0.20)	0.28 (0.18)	0.49 (0.31)
	SlidingHistogram	**0.86 (0.15)**	**0.69 (0.23)**	**0.72 (0.39)**
W101	ECOD	0.70 (0.21)	0.29 (0.13)	0.21 (0.19)
	Gen2Out	0.63 (0.30)	0.25 (0.11)	0.12 (0.18)
	IsolationForest	0.62 (0.28)	0.23 (0.11)	0.07 (0.16)
	LocalOutlierFactor	0.51 (0.19)	0.25 (0.18)	0.46 (0.37)
	SlidingHistogram	**0.87 (0.15)**	**0.71 (0.23)**	**0.72 (0.39)**
W201	ECOD	0.66 (0.21)	0.26 (0.13)	0.16 (0.17)
	Gen2Out	0.59 (0.27)	0.22 (0.12)	0.05 (0.11)
	IsolationForest	0.58 (0.26)	0.20 (0.12)	0.02 (0.09)
	LocalOutlierFactor	0.45 (0.17)	0.21 (0.16)	0.37 (0.38)
	SlidingHistogram	**0.87 (0.15)**	**0.69 (0.24)**	**0.70 (0.40)**

4.2 Applications on real-world Twitter dataset

We applied the SH method on 1318 accounts comprising Turkish politicians and media outlets from the #Secim2023 dataset [34] to explore the anomalous follow patterns that this method can uncover. We used a window size of 200 since on the Twitter dataset. On Twitter dataset, politicians have a significantly larger number of followers than the Dribbble users; however, our approach focuses on users in each sliding window together. We show that the results are robust towards different windows sizes (see SI-Fig. 5 (Additional file 1)).

Validating anomalous accounts in real-world datasets is a challenging task. In supervised learning algorithms, testing on a held-out dataset is a standard way to validate an algorithm. Unsupervised approaches can also use manually annotated datasets to measure precision and recall metrics on a sorted list of items. To validate our approach, we conduct a manual annotation study and additional explanatory data analysis to investigate anomalous accounts. Our approach assigns anomaly scores for each follower of an account. We adopt the strategy to evaluate our anomaly detection framework following a previous work where the authors developed a botnet detection framework by exploiting temporal patterns [30].

In our manual annotation study, two annotators had an accuracy of 0.93 and 0.98 on classifying anomalous accounts. Their inter-annotator agreement scores were 87% for percent agreement and 0.735 for Cohen's Kappa score. Furthermore, we examined the behavioral similarities between accounts during annotation as described in the literature [12] and noticed significant similarities in the content-sharing behavior of anomalous accounts (see SI-Fig. 18 (Additional file 1)). We present the details of the annotation study and how we selected anomalous and normal accounts described in SI-Sect. 8 (Additional file 1). Results of our manual annotation experiments indicate that our approach accurately identifies anomalous accounts.

Using the anomaly detection framework that we present, we conduct in-depth analysis on Turkish politicians. We divide our analysis of the Twitter dataset into three parts: (i) Retrieving user accounts that have anomalous followers (ii) Identifying individual anoma-

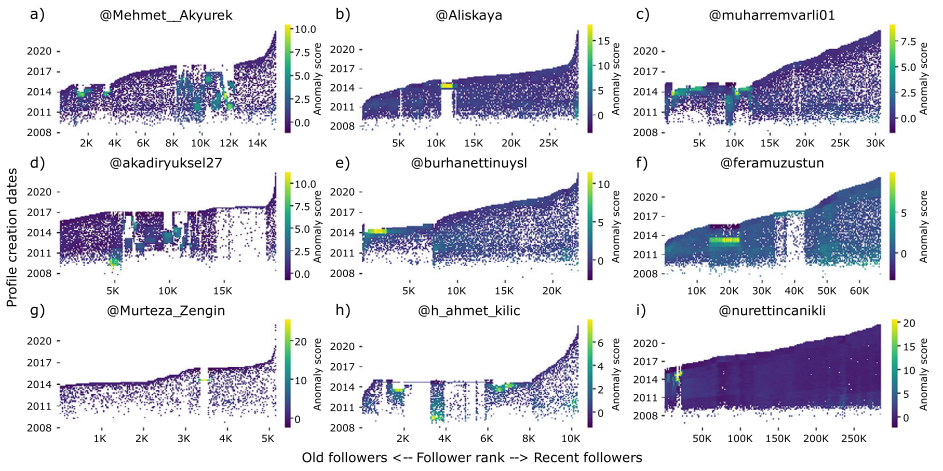

Figure 2 *Retrieving users with anomalous followers* Follower maps of the 9 Twitter accounts with the highest average anomaly score across all of their followers. The colors represent the average anomaly scores of all followers that fall in each bin (cell) of the heat map

lous follower accounts (iii) Exploring the coordinated behavior of the detected anomalous followers.

4.2.1 Retrieving users with anomalous followers

In order to detect the users that have anomalous follow patterns among their followers, we first looked at the 9 Twitter accounts with the highest average anomaly score across all their followers (Fig. 2). The follower maps are shown here as heat maps instead of scatter plots since these users have large numbers of followers. Irregular follow patterns can be observed in all of the follower maps of these users. Since the average anomaly score across all followers is generally lower for popular accounts, we can alternatively look at the average anomaly score of the highest N anomaly scores of a user's followers. Figure 3 shows the deviant followers of two popular Twitter accounts from our dataset. More examples can be seen in SI-Figs. 9-11 (Additional file 1).

4.2.2 Identifying individual anomalous accounts

Next, we looked at the individual accounts that constitute the groups of anomalous followers. First, we looked at these accounts' bot scores as computed by BotometerLite [51]. The BotometerLite only uses features that can be extracted from the account information, making it applicable to our dataset. We refer to the scores computed by the BotometerLite as *bot scores*. Figure 3 shows two cases, (A) anomalous followers having high bot scores (B) anomalous followers having low bot scores. To validate that the anomalous followers in the second case are indeed suspicious accounts, we manually observe a sample of these accounts. SI-Figs. 18-21 (Additional file 1) show samples of Twitter profiles of irregular followers of three accounts from our datasets, including the two accounts shown in Fig. 3. Snapshots of these anomalous profile webpages can be accessed through the Wayback Machine (Internet Archive) links provided in SI-Table 3 (Additional file 1). We observed that many of these accounts share the same tweets and share many of their friends. Additionally, the usernames of these accounts are in many cases meaningless combinations of letters. Figure 3d and Fig. 3h show the distribution of the friend, follower, and status

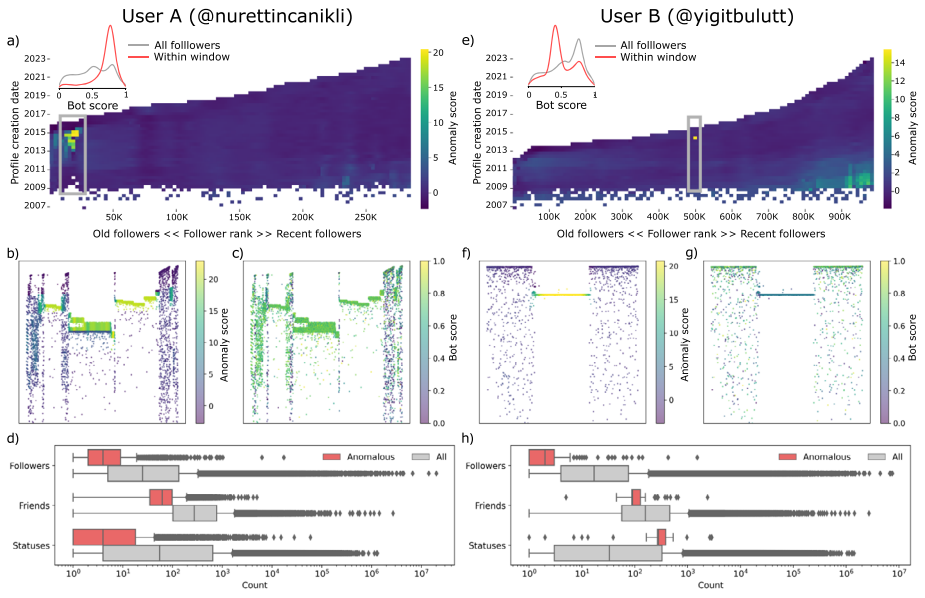

Figure 3 *Detailed analysis of anomalous followers* User A: Anomalous followers have high bot scores. User B: Anomalous followers have low bot scores. Anomalous regions are zoomed in for User A (b, c) and User B (f, g). Profile statistics for regular and all followers are also compared for these users in subplots (d) and (e)

counts of the anomalous followers compared to that of all the followers of the same account. In both cases A and B, the anomalous accounts tend to have a lower number of followers. In case A, the anomalous followers have a low number of shared posts, indicating that they are mainly aimed at increasing the follower counts. On the other hand, the anomalous followers in case B share a lot of posts, indicating that they are used to spread information. These results show that our approach can capture bots that act in coordination, even though their bot scores as computed by other methods may not necessarily be high.

4.2.3 Exploring anomalous follower group behavior

We explored the following patterns of the detected groups of anomalous followers and studied when they follow other users in our Twitter dataset. Are they always showing suspicious following patterns for other politicians, or is it specific to the particular user that we made the observation for? Firstly, we looked for accounts in our dataset that are followed by at least 30% of the suspicious followers of users A and B (Fig. 3). We found 0 accounts followed by the anomalous followers of user A and 12 accounts followed by the anomalous followers of user B. Since the anomalous accounts following user A do not follow any other users from our dataset, we resumed our analysis for user B only. We estimated the dates that the anomalous followers followed each of the 13 Twitter accounts using the method suggested in [31]. SI-Sect. 5 (Additional file 1) presents results for evaluation of the follow time estimation method on the Dribbble dataset, which provides ground truth values for follow times. Figure 4(a) and Fig. 4(b) show the following times and anomaly scores, respectively, of the anomalous followers (red) and the followers shared across the 13 users (gray) for comparison. The anomalous followers follow each user almost simultaneously, which demonstrates that they are automated accounts that

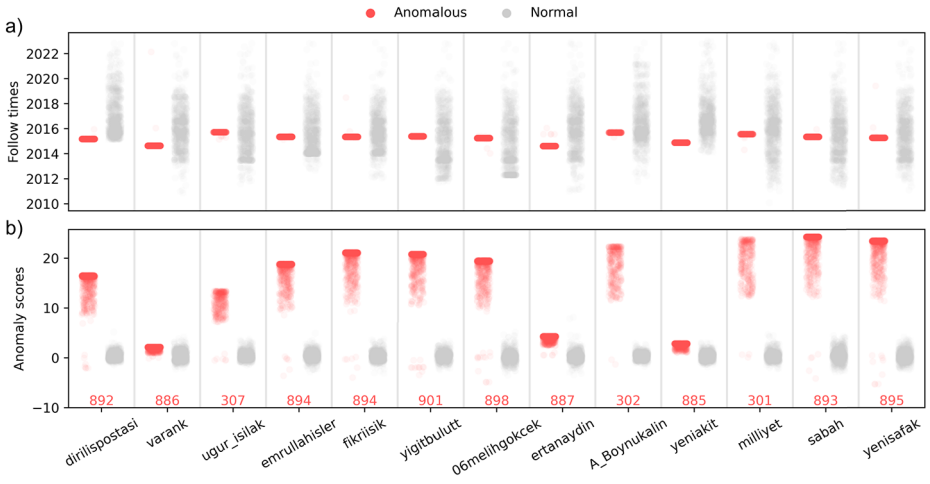

Figure 4 *Coordinated behavior of anomalous followers* Follow times (top) and anomaly scores (bottom) of the shared anomalous followers (red) and the shared non-anomalous followers (gray) across 13 users that are followed by the same batch of anomalous followers shown in Fig. 3(f)

work in coordination. Furthermore, the anomalous followers followed all of the 13 users between the years 2014 and 2016. Finally, our approach correctly assigned significantly higher anomaly scores to the anomalous followers as compared to the normal followers in 10 out of the 13 cases shown in Fig. 4.

We expanded the analysis of the group behavior of anomalous followers to uncover other groups of accounts that share the same suspicious followers. For this purpose, we created a similarity network based on the shared anomalous followers. The similarity between each pair of accounts is the cosine similarity between the two anomaly score vectors of the followers shared across the pair of users. Since our method assigns anomaly scores based on the follower map, a follower that follows accounts U1 and U2 will have two different anomaly scores computed for U1 and U2. Thus, a pair of accounts that share followers who were assigned high anomaly scores in both follower maps will have a high similarity. The Louvain community detection algorithm was then used to detect the communities in the network [4]. Figure 5 shows the two communities with the highest pairwise average anomaly scores across all edges in the community. For each community, we show the follower maps of a user pair corresponding to one of the edges in the community. The follower maps are colored by the ratio of shared followers between the pair of users in each bin. This allows us to capture concentrations of shared followers in both users' maps, which appear as reddish regions in the follower map. We observe that the concentrated regions of shared followers exhibit anomalous following patterns in both follower maps. This finding supports our hypothesis that anomalous followers work in coordination. More details about this network analysis and other samples of anomalous follower groups appearing in different users' follower maps are presented in SI-Sect. 7 (Additional file 1).

5 Discussion

Our analysis of the followers of 1318 Twitter accounts supports the earlier findings regarding the existence of coordinated accounts that exhibit anomalous following patterns on Twitter [9, 48]. Exogenous events like political campaigns and elections can change the

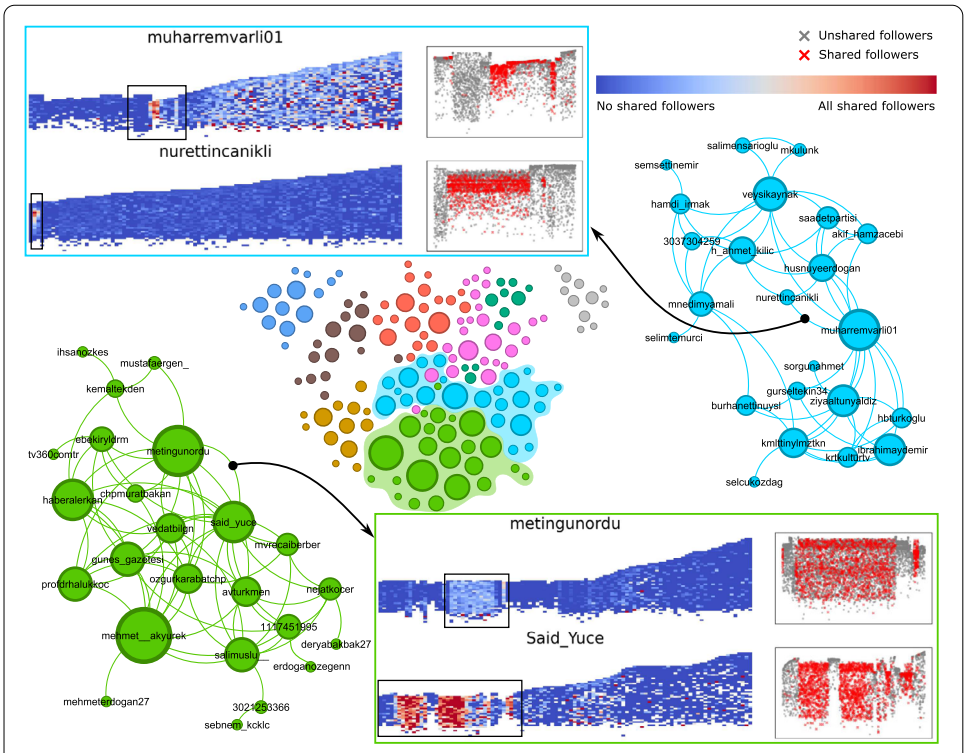

Figure 5 *Similarity network based on the shared anomalous followers.* The full network is shown in the middle of the figure, where nodes are colored based on communities and sized based on their degrees. The two communities with the highest pairwise average anomaly scores are highlighted and shown in detail along with the follower maps of one edge in each community

composition of followers, and Twitter accounts gain new followers around these events [44]. Our manual observation of samples of the accounts that engage in anomalous following patterns indicates that these accounts are likely to be fake. Moreover, we showed that batches of anomalous followers tend to follow Twitter accounts almost simultaneously, suggesting that they are automated accounts managed from one place. The fact that many of the detected anomalous followers in this study are still active accounts indicates that Twitter has not yet identified these accounts as malicious. Although these are clearly fake and automated accounts, we cannot make any conclusions about their intentions. We hypothesize three possible scenarios: (i) The user purchased these followers to gain popularity (ii) The anomalous followers followed the user to gain credibility or fit in a specific persona (iii) The anomalous followers targeted the user to serve propaganda for the user's opponents.

The recent changes in the API policy have rendered Twitter data less accessible. However, our method applies to any other social media platform that provides an ordered list of followers and their creation dates. Other than Twitter, platforms like Mastodon provides similar information about accounts and our proposed approach can be applied to these platforms. Furthermore, detecting this type of behavior allows for understanding coordinated activities and disinformation-spreading campaigns that may have happened earlier on Twitter.

Our model has advantage of being less data intensive method. It requires two simple information: account creation time and follow rank. We also tested our approach by sam-

pling followers to simulate different scenarios of imperfect data collections and our model can still identify groups of anomalies as we presented in robustness analysis section on SI-Sect. 4 (Additional file 1). Since the required data for creating follower maps exists on other platforms, our methodology can be applied for platforms other than Twitter/X. As a demonstration, we analyzed a Mastodon dataset and present results for a sample account (see SI-Fig. 26 (Additional file 1)).

We also demonstrated that our methodology can be applied to detect anomalous amplification activities by taking into account content creation times and reshare timestamps. Using this approach, we were able to detect cases of a single user retweeting large numbers of tweets of the same user consecutively, including a user that retweeted more than 300 tweets of a another user in less than 12 minutes. These observations conform with the previous observations made by Cresci et al., where they observed several anomalous retweeting behaviors, including one in which a user retweets a large number of another user's tweets in a short time [30]. While RTBust looked at individual retweeters, our method finds the same patterns by looking at the retweeted user, therefore uncovering all the anomalous retweeters of that user. Additionally, we found another interesting behavior where one user would retweet the same tweet multiple times (up to 41 times in one case). We shared results for these accounts and other examples in SI-Sect. 10 (Additional file 1).

The main limitation of our approach is that it tends to compute anomaly scores that lead to higher false negatives when the ratio of anomalous followers is high. SH assigns high anomaly scores to followers that deviate from the main follower distribution, as defined by the set of histograms. Therefore, the method will not assign the right anomaly scores when the account's followers are dominated by anomalous followers. SI-Fig. 10 (Additional file 1) shows some cases with high ratios of anomalous followers, where the scores are wrongly assigned.

Detecting coordinated disinformation campaigns on social media platforms has become crucial in recent years [52]. In this paper, we presented a method to detect a previously unaddressed type of anomalous followers on social media platforms. We demonstrated that the detected anomalous followers act in coordination and, in many cases, exhibit similar anomalous behavior across more than one account. Using this approach, further analyses can be applied to uncover coordinated disinformation activities on social media platforms.

Abbreviations

AP, average precision; AUC, area under ROC curve; ECOD, Empirical-Cumulative-distribution-based Outlier Detection; IQR, interquartile range; LOF, Local Outlier Factor; SH, Sliding Histogram.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-024-00499-6.

Additional file 1. Supplementary information (PDF 18.1 MB)

Acknowledgements

We thank Rossano Schifanella for providing access to Dribbble dataset. We also thank Baris Temel for his earlier work on the topic.

Author contributions

OV and YZ conceived and designed the research. OV and YZ analyzed the data. YZ conducted research and prepared figures. OV and YZ prepared manuscript. All authors read and approved the final manuscript.

Funding

This work is partly supported by TUBITAK projects 121C220 and 222N311.

Data availability
The Twitter data analyzed in this study was taken from the #Secim2023 dataset [34] which is publicly available in Harvard Dataverse [35]. The Dribbble data was provided by Rossano Schifanell and can be acquired upon reasonable request. The code is available at github.com/ViralLab/FollowerAnalyzer.

Declarations

Competing interests
The authors declare no competing interests.

Received: 6 February 2024 Accepted: 26 September 2024 Published online: 07 October 2024

References

1. Alkulaib L, Zhang L, Sun Y, et al (2022) Twitter bot identification: an anomaly detection approach. In: 2022 IEEE international conference on big data (big data). IEEE, pp 3577–3585
2. Auten T, Matta J (2023) Retweeting Twitter hate speech after musk acquisition. In: International conference on complex networks and their applications. Springer, Berlin, pp 265–276
3. Bellutta D, Carley KM (2023) Investigating coordinated account creation using burst detection and network analysis. J Big Data 10(1):1–17
4. Blondel VD, Guillaume JL, Lambiotte R, et al (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
5. Breunig MM, Kriegel HP, Ng RT, et al (2000) LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data, pp 93–104
6. Bruno M, Lambiotte R, Saracco F (2022) Brexit and bots: characterizing the behaviour of automated accounts on Twitter during the uk election. EPJ Data Sci 11(1):17
7. Chavoshi N, Hamooni H, Mueen A (2016) Debot: Twitter bot detection via warped correlation. In: Icdm, pp 28–65
8. Cinelli M, Cresci S, Quattrociocchi W, et al (2022) Coordinated inauthentic behavior and information spreading on Twitter. Decis Support Syst 160:113819
9. Confessore N, Dance GJ, Harris R, et al (2018) The follower factory. https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html [Online; accessed 04-December-2018]
10. Cresci S (2020) A decade of social bot detection. Commun ACM 63(10):72–83
11. Cresci S, Di Pietro R, Petrocchi M, et al (2015) Fame for sale: efficient detection of fake Twitter followers. Decis Support Syst 80:56–71
12. Cresci S, Di Pietro R, Petrocchi M, et al (2017) The paradigm-shift of social spambots: evidence, theories, and tools for the arms race. In: Proceedings of the 26th international conference on world wide web companion, pp 963–972
13. Cresci S, Di Pietro R, Petrocchi M, et al (2017) Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling. IEEE Trans Dependable Secure Comput 15(4):561–576
14. Cresci S, Petrocchi M, Spognardi A, et al (2019) Better safe than sorry: an adversarial approach to improve social bot detection. In: Proceedings of the 10th ACM conference on web science, pp 47–56
15. Ding J, Chen Z (2023) How to find social robots exactly? In: Proceedings of the 2023 6th international conference on software engineering and information management, pp 12–18
16. Echeverrìa J, De Cristofaro E, Kourtellis N (2018) LOBO: evaluation of generalization deficiencies in Twitter bot classifiers. In: Proceedings of the 34th annual computer security applications conference, pp 137–146
17. Ferrara E, Varol O, Davis C, et al (2016) The rise of social bots. Commun ACM 59(7):96–104
18. Goel S, Anderson A, Hofman J, et al (2016) The structural virality of online diffusion. Manag Sci 62(1):180–196
19. Guess AM, Lyons BA (2020) Misinformation, disinformation, and online propaganda. In: Social media and democracy: The state of the field, prospects for reform, vol 10
20. Hickey D, Schmitz M, Fessler D, et al (2023) Auditing elon musk's impact on hate speech and bots. In: Proceedings of the international AAAI conference on web and social media, pp 1133–1137
21. Himelein-Wachowiak M, Giorgi S, Devoto A, et al (2021) Bots and misinformation spread on social media: implications for COVID-19. J Med Internet Res 23(5):e26933
22. Hristakieva K, Cresci S, Da San Martino G, et al (2022) The spread of propaganda by coordinated communities on social media. In: Proceedings of the 14th ACM web science conference 2022, pp 191–201
23. Jia J, Wang B, Gong NZ (2017) Random walk based fake account detection in online social networks. In: 2017 47th annual IEEE/IFIP international conference on dependable systems and networks (DSN). IEEE, pp 273–284
24. Le T, Tran-Thanh L, Lee D (2022) Socialbots on fire: modeling adversarial behaviors of socialbots via multi-agent hierarchical reinforcement learning. In: Proceedings of the ACM web conference, pp 545–554
25. Lee MC, Shekhar S, Faloutsos C, et al (2021) Gen 2 out: detecting and ranking generalized anomalies. In: 2021 IEEE international conference on big data (big data). IEEE, pp 801–811
26. Li Z, Zhao Y, Hu X, et al (2023) ECOD: unsupervised outlier detection using empirical cumulative distribution functions. IEEE Trans Knowl Data Eng 35:12181–12193
27. Liu FT, Ting KM, Zhou ZH (2008) Isolation forest. In: 2008 eighth ieee international conference on data mining. IEEE, pp 413–422
28. Liu Y, Tan Z, Wang H, et al (2023) BotMoE: Twitter bot detection with community-aware mixtures of modal-specific experts. arXiv preprint. arXiv:2304.06280
29. Mannocci L, Cresci S, Monreale A, et al (2022) Mulbot: unsupervised bot detection based on multivariate time series. In: 2022 IEEE international conference on big data (big data). IEEE, pp 1485–1494
30. Mazza M, Cresci S, Avvenuti M, et al (2019) Rtbust: exploiting temporal patterns for botnet detection on Twitter. In: Proceedings of the 10th ACM conference on web science, pp 183–192
31. Meeder B, Karrer B, Sayedi A, et al (2011) We know who you followed last summer: inferring social link creation times in Twitter. In: Proceedings of the 20th international conference on world wide web, pp 517–526

32. Mendoza M, Providel E, Santos M, et al (2024) Detection and impact estimation of social bots in the Chilean Twitter network. Sci Rep 14(1):6525
33. Mendoza M, Tesconi M, Cresci S (2020) Bots in social and interaction networks: detection and impact estimation. ACM Trans Inf Syst 39(1):1–32
34. Najafi A, Mugurtay N, Demirci E, et al (2022) # secim2023: first public dataset for studying Turkish general election. arXiv preprint. arXiv:2211.13121
35. Najafi A, Mugurtay N, Zouzou Y, et al (2022) # Secim2023: first public dataset for studying Turkish general election. https://doi.org/10.7910/DVN/QJA1ZW
36. Nizzoli L, Tardelli S, Avvenuti M, et al (2021) Coordinated behavior on social media in 2019 UK general election. In: Proceedings of the international AAAI conference on web and social media, pp 443–454
37. Pacheco D, Hui PM, Torres-Lugo C, et al (2021) Uncovering coordinated networks on social media: methods and case studies. In: Proceedings of the international AAAI conference on web and social media, pp 455–466
38. Sayyadiharikandeh M, Varol O, Yang KC, et al (2020) Detection of novel social bots by ensembles of specialized classifiers. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 2725–2732
39. Shao C, Ciampaglia GL, Varol O, et al (2018) The spread of low-credibility content by social bots. Nat Commun 9(1):1–9
40. Sharma K, Zhang Y, Ferrara E, et al (2021) Identifying coordinated accounts on social media through hidden influence and group behaviours. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp 1441–1451
41. Takacs R, McCulloh I (2019) Dormant bots in social media: Twitter and the 2018 us senate election. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, pp 796–800
42. Tardelli S, Nizzoli L, Tesconi M, et al (2024) Temporal dynamics of coordinated online behavior: stability, archetypes, and influence. Proc Natl Acad Sci 121(20):e2307038121
43. Varol O (2023) Should we agree to disagree about Twitter's bot problem? Online Soc Netw Media 37:100263
44. Varol O (2023) Who follows Turkish presidential candidates in 2023 elections? In: 2023 31st signal processing and communications applications conference (SIU). IEEE, pp 1–4
45. Varol O, Davis CA, Menczer F, et al (2018) Feature engineering for social bot detection. In: Feature engineering for machine learning and data analytics, vol 311
46. Varol O, Ferrara E, Davis C, et al (2017) Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the international AAAI conference on web and social media, pp 280–289
47. Varol O, Uluturk I (2019) Deception strategies and threats for online discussions. arXiv preprint. arXiv:1906.11371
48. Varol O, Uluturk I (2020) Journalists on Twitter: self-branding, audiences, and involvement of bots. J Comput Soc Sci 3(1):83–101
49. Wu L, Morstatter F, Carley KM, et al (2019) Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD Explor Newsl 21(2):80–90
50. Yang KC, Varol O, Davis CA, et al (2019) Arming the public with artificial intelligence to counter social bots. Hum Behav Emerg Technol 1(1):48–61
51. Yang KC, Varol O, Hui PM, et al (2020) Scalable and generalizable social bot detection through data selection. In: Proceedings of the AAAI conference on artificial intelligence, pp 1096–1103
52. Zhang Y, Sharma K, Liu Y (2023) Capturing cross-platform interaction for identifying coordinated accounts of misinformation campaigns. In: European conference on information retrieval. Springer, Berlin, pp 694–702

Publisher's Note