The Grey Spot of Twitter Bot Moderation: Studying the traits of accounts who survived amongst the batch of suspended accounts.

Pujan Paudel[1] , Andrew H. Sung[2]

School of Computing Sciences and Computer Engineering,
The University of Southern Mississippi, Hattiesburg MS 39406, USA
pujan.paudel@usm.edu,andrew.sung@usm.edu

Abstract. Automated accounts on Twitter have been a major problem since the earlier days of the existence of social media and they have been evolving as key players in Twitter with the maturity and wide adoption of the social network platform. The Twitter platform has been successfully able to detect various natures of Twitter bots and constantly suspend the accounts who seemingly violate the policies of the platform. Despite its efforts, a significant portion of social bots that are part of bigger campaigns is still active on Twitter. The active bots possess a threat of being re-purposed across different political or advertisement campaigns. In this work, we perform a comparison from different dimensions of suspended and active accounts belonging to the same bot campaign for two different previously identified bot campaigns. We compare and contrast the network-based and activity-based characteristics of the suspended and persistent bots and discuss what would have made these bots go undetected from Twitter's platform moderation. We conclude our paper by discussing the avenues of opportunities and challenges for detecting these active accounts who survived the batch moderation of the bot network and how we can propose new features moving ahead to possibly identify many such active accounts in a grander scale.

Keywords: social bots · twitter · social network · detection

1 Introduction

Automated accounts or bots in Twitter pose an ever-increasing threat to disrupt the flow of conversation on the social networks and pollute the platform with unwanted traffic of text and interactions. The platform design of Twitter allows the usage of automated accounts in the social network to promote large scale business interactions and introduce the ever-availability of identity through powerful ways of scripted interaction in the platform. Twitter also has its own rules regarding the do's and don't of the automated accounts and their specific set of instructions on how they would not want automated accounts to behave in the platform.

The primary inspiration behind this work comes from our observation of Twitter suspending only a certain portion of bots from a bigger bot network in the two social bot campaigns identified by [1]. We wanted to investigate why Twitter might have not suspended the bots which are still active out there by analyzing if they pose any statistical advantage in their activity patterns or network position which might have made them go undetected against the detection system of Twitter. Even though our preliminary analysis revealed that the bots which are active have not demonstrated any activity dating back to 2015, the observation in the work of [4] where social bots were re-purposed across different political disinformation operations inspired this work to understand the existing accounts and the threats they possess. Similarly, in [3] the authors discuss the differential impact of the different types of potentially suspend-able users. This also motivated us to assess how the differing traits of these suspend-able accounts from the suspended accounts could help us in building better systems to detect more robust active social bots. A parallel study to our work in a different setting, conducted by [5] analyzes the "Gamergaters" who were left unsuspended by Twitter despite their active abusive behavior in the incident.

The rest of the paper is organized as follows. Section 2 details the data collection process. Section 3 discusses the set of methodologies and setup for our analysis. Section 4 discusses the different behavior-based and network-based experiments to investigate the characteristics of the two types of accounts. We conclude our paper by discussing the opportunities and challenges and future works in Section 5.

2 Data Collection

Most of the previous studies and the datasets made available had the limitation of sharing only the Tweet Ids and User Ids as a part of the dataset. This limitation did not allow us to collect the user metadata and tweets of users who were already deleted from the platform. Because of these limitations, we were restricted to the dataset used in the work of [1], which coincidentally had the collection of the tweet metadata object and user metadata object as their full dataset. There were multiple reasons which makes this dataset a suitable fit for our study. Primarily, the dataset had already been manually annotated by human annotators for identification of the bot accounts, and the author had separated the bot accounts belonging to two different use cases. The major research design of our study is to study the suspended accounts and the active accounts which are part of the same campaign, not just two unrelated automated accounts that are deployed for entirely different purposes. Pertaining to our research design, this dataset allowed us to directly use the accounts and tweets in our study as we did not need any extra set of pipelines to extract the related bot accounts belonging to the same operation. We polled the user accounts against the Twitter API and checked the error code of the request to determine if the user accounts were suspended or they are still active in the platform. The overview of the resulting dataset collected is displayed in Table 1.

Table 1: Overview of the dataset used in the study

Dataset	**# Initial accounts**	**# Current accounts**	**% active**
Amazon Spammers [1]	461	380	84 %
Italian Political Bots [1]	991	356	35 %

We believe the nature of the datasets used in our study, one of them being a product affiliate campaign and the other being a political botnet provides a certain extent of diversity to the nature of the bot accounts we are studying. One other observation to note is that the two datasets have contrastingly different proportions of active accounts and suspended accounts. The majority of the amazon spammers are still active, while the majority of the Italian political bots are suspended. The nature of this class imbalance would represent two different scenarios of under moderation and over moderation by the Twitter platform. For the rest of the paper, we will be referring to Amazon Spammers collected from [1] as Amazon Bots, and the Italian Political Bots from [1] as Political Bots.

3 Methodology

3.1 Separation of Campaign related and non-campaign related tweets.

In sections that follow in our work, we make a distinction of tweets, and activities of user as : campaign-based and non-campaign based. The idea behind separating the campaign based tweets develops from the observations that the new waves of social bots use deceptive patterns of mixing campaign-related tweets with genuine-looking discussions [6]. In case of the Amazon Bots, the campaign-related tweets were marked by the presence of URLs that redirected to any form of Amazon product campaigns. Since the majority of the URLs were shortened URLs, we expanded the shortened URLs to obtain the final location to which they redirected to. Similarly, in the case of Political Bots, we studied two different types of campaign characteristics. The first type of campaign characteristics studies the behavior of retweeting the tweets of verified political accounts affiliated with the political party the bots were believed to be deployed by. The other type of campaign characteristics investigates the activity of using campaign-related hashtags on tweets authored by the accounts.

3.2 Communication Networks

We construct multiple networks of communication signals, such as Retweet Network, URL Network, Mention Network and Hashtag Network. We could only work with the URL network of the Amazon Bots as their other forms of communication network was extremely sparse and did not contribute any significant results to our studies.

Retweet (RT) Network: is an undirected network where there is an edge from Bot A to Bot B if both of them have retweeted 5 or a higher number of similar Twitter users.

The other networks : Mention Network , Hashtag (HT) Network and URL Network were created in similar way as the RT Network.

3.3 Signal Based TimeSeries Analysis

We perform a detailed temporal analysis of the users by considering two different types of activity sequences. The first type of analysis, similar to the traditional time series analysis considers the raw volume of the tweets posted by a user as magnitude in the Y-axis across equal time buckets of 2 hours along the X-axis. We select an appropriate time frame of 2 months for the starting point and ending point of the analysis as the activity duration where all of the accounts were actively involved.

The other type of time series analysis, which we call Signal Analysis is inspired by the observation of the new wave of social bots using a deceptive strategy of inter-mixing their campaign-related tweets with normal-looking tweets only in fixed intervals of posting duration. We were interested to investigate if there is a certain underlying pattern that could explain how they embed their campaign-related tweets along their posting timeline. We sort all of the tweets of a user in ascending order by their timestamp and divide them into 100 different time buckets. Instead of taking the raw magnitude of tweets posted during the time bucket, we take the percentage of campaign-related tweets during the bucket. We use the same mechanism of separating campaign-related tweets as we discussed before.

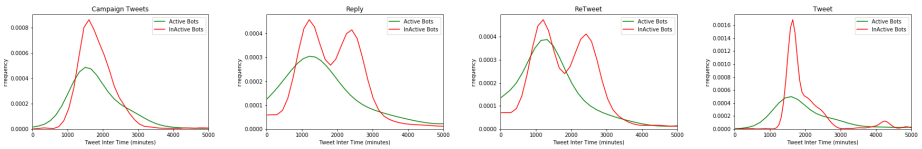
4 Comparative Analysis of Active Users and Suspended Users

4.1 Activity Distance Analysis

Social bots are controlled by scripts running in the background with the component of business logic added triggering to interaction they perform in the Twitter platform. Because of their scripted nature and the social signals that could possibly trigger the activity of the social bots, mining the distance of tweeting patterns demonstrated by the bots is an important dimension of analysis. The main purpose of this activity pattern analysis is to ask the question if the temporal patterns of the suspended users are more closer to each other than the accounts who are still active. After formulating the two different types of time series as discussed in the previous section: signal based and activity based, we compared how the time series of an individual user compared with the rest of the users. For this comparison, we needed an appropriate time series distance measure. We chose Dynamic Time Warping (DTW) to measure dissimilarity between temporal sequences of user activity.

We also wanted to quantify how semantically similar the tweets authored by the user accounts were. It has been observed that the automated accounts use Markov chain generators, and reuse existing text from a large corpus of tweets from human users to mix between their campaign related tweets. We used the distance metric of Word Movers Distance (WMD) from available semantic distance metrics to analyze the similarity of tweets within the tweet corpus posted by a user. Before computing the distance metrics, we apply basic preprocessing to all of the tweets being studied, such as removing emoticons and special characters, removing all hashtags, URLs and HTML tags. Since the tweets were initially in the Italian language and our calculation of WMD used the *word2vec-google-news-300* model, we had to convert the tweets into the English language using Google Translation API. We converted our entire vocabulary to lower case, lemmatized them and expanded common English contraction words.

We observe in Figure 1 that the active users and the suspended users displayed similar inter tweet arrival distance when we consider all of their tweets. But, they were separated when we observe the inter campaign arrival distance with the active bots displaying relatively larger delay time between successive tweets with their campaigns. In terms of the temporal distance metrics displayed in Figure 2, the active bots had lower activity-based DTW distance and higher signal-based DTW distance while the WMD distance distribution between the two types of bots were almost identical. This suggests that the suspended bots were more suspicious in their signal based properties (inter-campaign duration) and lower signal distance, which could have led to their moderation.

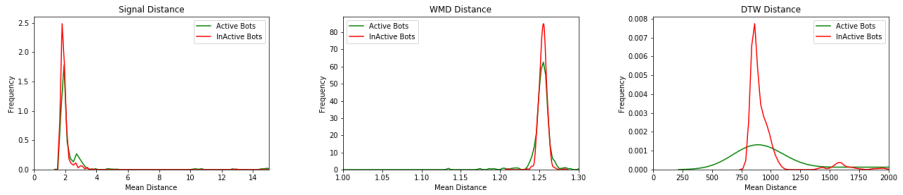


(a) (b) (c) (d)

Fig. 1: Inter Tweet Distances (in Hours) a) Campaign Tweets, b) Reply, c) ReTweet & d) Tweet

4.2 Distribution of Coreness of Users

For the complex network analysis-based studies that follow, we work with different types of communication networks the bots make as discussed in Section 3.2. The k-core of a graph is a maximal subgraph in which each vertex has at least degree k. The coreness of a node is k if it belongs to the k-core but not to the (k+1)-core. We use the distribution of coreness of nodes in the communication network to quantify how embedded were the nodes in their respective communication networks. If the networks were shallowly embedded, they would

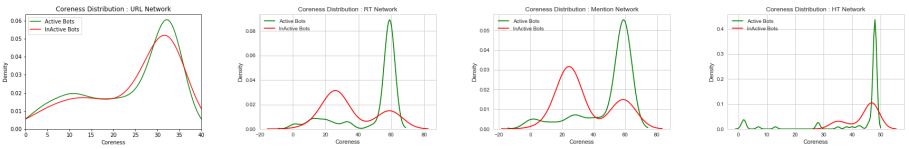


(a) (b) (c)

Fig. 2: Distance a) Signal and b) WMD c) DTW

have relatively smaller values of coreness than the deeply embedded nodes. The spread of the coreness of nodes in a communication network would also give us an idea of the structural embedding of the active users and the suspended users.

The distribution of coreness density is plotted in Figure 3. It can be observed that the active accounts were more deeply embedded within their communication networks with a higher value of coreness. The observation of a relatively larger measure of coreness on the part of active users in comparison to the suspended users suggests that the active users are embedded in more deeper k-core of the communication networks. It also calls for network-based detection measures to dig more deeper into the communication graphs to search for the possibility of suspend-able bot accounts.



(a) (b) (c) (d)

Fig. 3: Network Coreness Distribution : a) URL , b) ReTweet, c) Mention & d) HashTag

4.3 Inter-community and Intra-community interaction

Social bots, while primary deployed to amplify a central set of accounts, or retweet/mention other users from their respective botnets, interact with other users outside of their own community to avoid detection and depict a more social image of themselves. We study this behavior of social bots from two different perspectives.

The first perspective we want to get about the organization of the suspended and active bots is how diverse are their inter-community and intra-community posting behaviors. Do the bots restrict themselves in interacting within their

primary community, or they have more intra-community interactions spreading across diverse communities? We use the Louvain community detection algorithm [2] in the different communication networks discussed in Section 3.2 and separate the communities. We then calculate the inter-community and intra-community edges. Our secondary dimension of the study was to study the raw incoming and outgoing interactions (retweets, mentions) within the bots and the human users they try to influence or garner attention from. This would give us a measure of the reciprocity of interactions, and eventually the effectiveness of their presence, and the percentage of unsolicited interactions which could have possibly resulted in their moderation from the platform.

Table 2: Community-based Interactions of Political Bots, RT,Mention, Hashtag and URL Network

Network	**Active Bots**	**Suspended Bots**
Retweet Network (RT)	Inter: 53 % Intra : 47 %	Inter: 60 % Intra : 39 %
Mention Network	Inter: 47 % Intra : 52 %	Inter: 58 % Intra : 41 %
HashTag Network (HT)	Inter: 42 % Intra : 57 %	Inter: 44 % Intra : 55 %
URL Network (URL)	Inter: 59 % Intra : 40 %	Inter: 69 % Intra : 30 %

It can be seen in Table 2 that in the Mention and Hashtag networks, the active bots spread out their interactions across communities, with more intra-community interactions than the interaction between the modular community of their own belongings. This might be one of the reasons which make it challenging to detect them. Whereas, in the Retweet network and URL Network, they have more inter-community interactions. It would be an interesting future work to study the detection of undiscovered bots through these channels of communication across these graphs.

In terms of raw retweet interaction, for both the active bots as well as the suspended bots, the traffic was heavily directed to outside users. But, active bots had retweeted from slightly more outside users (97 %) compared to suspended bots (93 %). It was also be observed that the active bots retweeted inactive bots slightly less than the proportion in which the suspended bots retweeted the active ones. Similar to retweet interaction, for both the active bots as well as the suspended bots, the mention traffic was heavily directed to outside users. But, small differences can be observed. The active bots mention the remaining active bots comparatively more than the inactive bots, who direct more of their remaining interactions towards the active users. We can further investigate the mention channel to see if we can further detect the accounts which could be suspended.

4.4 Sub-campaign Analysis

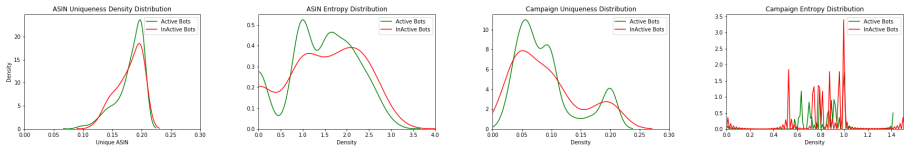
Spam campaigns such as Amazon product campaigns, giveaway campaigns use the concept of tracking the affiliates using campaign id and affiliate id. This metadata can be utilized to infer the campaign and sub-campaign level organization of the spam accounts. A common example of an Amazon affiliate URL campaign is :

`https://www.amazon.com/gp/product/B002RSBTVM?ie=UTF8&camp=213733&creative=393185&creativeASIN=B002RSBTVM&linkCode=shr&tag=activepubs`

In this example, the url parameter *camp* gives us the campaign affiliate identifier they are campaigning for, and the parameter *creativeASIN* refers to the amazon product. In case of the political bots, we study the candidate level and hashtag level organization of sub-campaigns , which are analogous to the *Asin* and *Camp* organization of the Amazon bots.

We calculate two different types of measures: compositional and temporal to get an idea of the organization of the bot accounts in terms of their sub-campaign strategies. The first of our measures, uniqueness gives us an estimate of the spread of sub-campaign organization by the bots. A higher uniqueness would represent that the bots tweeted about a larger variety of amazon products, or retweeted from more political candidates, or employed wider political hashtags. The second measure, entropy gives us an idea of the randomness in the timeline of their sub-campaigns . A higher entropy would imply that the bot accounts were deceptively mixing their sub-campaign along the timeline of their tweets, making their posting behavior look more complex.

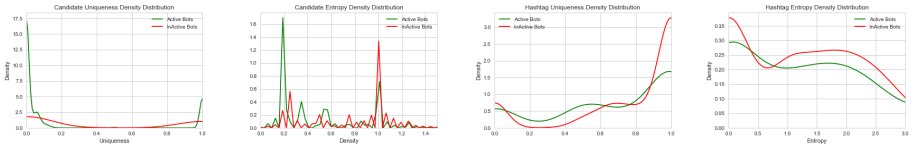
For calculating the entropy-based measures, we calculate the metrics on user's tweets sorted by the ascending order of their timestamps. By doing so, we hope to capture the variance and randomness in the posting behavior of a user across the timeline. We use Shannon entropy for calculating the entropy measure.



(a) (b) (c) (d)

Fig. 4: Amazon Sub-Campaign: a) Asin Uniqueness , b) Asin Entropy, c) Camp Uniqueness & d) Camp Entropy

We get interesting results from the sub-campaign analysis of the Amazon bots in Figure 4 and Figure 5. The active bots campaigned for more unique products and campaigns than the suspended bots, but the entropy of their affiliate posting strategy was less than the suspended accounts. This leads us to the observation that the suspended accounts were moderated possibly because they used less



(a) (b) (c) (d)

Fig. 5: Political Bots Sub-Campaign: a) Candidate Uniqueness , b) Candidate Entropy, c) HT Uniqueness & d) HT Entropy

diversified products and campaigns in their tweets. As a result, one possible strategy for detecting these suspend-able accounts could be studying the entropy of the patterns of products and campaigns they post about, as we found out that the accounts that are still out there have relatively lower entropy in their campaign patterns. In the case of the political bots, the suspended bots had higher values of entropy and uniqueness for both the candidate campaigns and hashtags campaign than the suspended accounts. Surprisingly, the uniqueness of the candidate campaigns was almost identical for the active bots as well as the suspended bots. Similar to the observation of Amazon bots, lower entropy of the patterns of the candidate campaigns and hashtags campaigns can be a good line of feature for detecting them.

5 Discussion and Conclusion

Our analysis of the activity signals and network positions of the active accounts in comparison to the suspended accounts reveals some new avenues for detecting those suspend-able accounts, and also the challenges that come with it. Our formulation of signal based temporal analysis allowed us to better distinguish the active users. But, practically the major caveat lies in determining the set of communication markers that can help us build these signal based properties. In our case, we had the prior knowledge of the campaigns the bots were involved in, so we could extract the signal based properties. It will be particularly challenging to extract the common theme or campaign signals the bots are driven by; when we are given with the challenge of detecting new suspicious accounts in the wild.

Our network-based analysis was also extremely helpful in separating the active bots from the suspended bots. We also discovered the active bots rank much higher in their betweenness centrality measures in the communication networks. So, we could detect them through their centrality metrics. We can use this observation to create the initial set of suspicious users on the platform and investigate them in depth if needed. The sub-campaign level entropy of active bots was determined to be very less than the suspended accounts. It presents us with opportunities to create better entropy level features that can detect these accounts.

Our analysis also suggests that how challenging it can get to detect them. The active bots were more deeply embedded within their communication networks,

and it can get especially complicated to detect them in the inner core of the communication networks. The active bots were also discovered to spread out their interactions across different communities of users, which makes them particularly interesting as well as difficult to detect. They detected mature sub-campaign level characteristics by taking part in more unique affiliate campaigns. It will be particularly difficult to detect them if they spread out their sub-campaign level strategies. Apart from the challenge of detecting these suspend-able accounts, we could also have the problem of possibly confusing them with normal users.

In this work, we studied how the Twitter bots which were not suspended by Twitter differ from the bots that were suspended and were part of the same campaign by studying two different datasets. Our findings on the temporal, community-based and network-based properties of the still existing bots suggest to us what might have led to these accounts still being active on the platform. We plan to extend this work further by converting our analytical results to statistical features which could be used to emulate a platform moderation engine and detect the suspend-able accounts through machine learning techniques. Detecting the suspend-able accounts using those features and existing meta-data of already suspended users through unsupervised clustering is also a possible research direction. We also believe it is extremely important to study the neighboring graph of those suspend-able bots and perform a susceptibility analysis of the human users to study if there are chances of the bot accounts manipulating them further down the road.

References

1. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. and Tesconi, M., 2017, Apr. 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). International World Wide Web Conferences Steering Committee.
2. De Meo, P., Ferrara, E., Fiumara, G. and Provetti, A., 2011, Nov. Generalized louvain method for community detection in large networks. In 2011 11th International Conference on Intelligent Systems Design and Applications (pp. 88-93). IEEE.
3. Wei, W., Joseph, K., Liu, H. and Carley, K.M., 2015, Aug. The fragility of Twitter social networks against suspended users. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 9-16). IEEE.
4. Ferrara, E., 2017, Aug. Disinformation and social bot operations in the run up to the 2017 French presidential election. First Monday, 22(8).
5. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G. and Vakali, A., 2017, Jul. Hate is not binary: Studying abusive behavior of # gamergate on twitter. In Proceedings of the 28th ACM conference on hypertext and social media (pp. 65-74). ACM.
6. Paudel, P., Nguyen, T.T., Hatua, A. and Sung, A.H., 2019. How the Tables Have Turned: Studying the New Wave of Social Bots on Twitter Using Complex Network Analysis Techniques. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 501-508). ACM.