Int. J. Inf. Secur.
DOI 10.1007/s10207-016-0321-5

REGULAR CONTRIBUTION

If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts

Abdullah Almaatouq[1] · Erez Shmueli[1] · Mariam Nouh[2] · Ahmad Alabdulkareem[1] · Vivek K. Singh[1] · Mansour Alsaleh[3] · Abdulrahman Alarifi[3] · Anas Alfaris[1] · Alex 'Sandy' Pentland[1]

Abstract Spam in online social networks (OSNs) is a systemic problem that imposes a threat to these services in terms of undermining their value to advertisers and potential investors, as well as negatively affecting users' engagement. As spammers continuously keep creating newer accounts and evasive techniques upon being caught, a deeper understanding of their spamming strategies is vital to the design of future social media defense mechanisms. In this work, we present a unique analysis of spam accounts in OSNs viewed through the lens of their behavioral characteristics. Our analysis includes over 100 million messages collected from Twitter over the course of 1 month. We show that there exist two behaviorally distinct categories of spammers and that they employ different spamming strategies. Then, we illustrate how users in these two categories demonstrate different individual properties as well as social interaction patterns. Finally, we analyze the detectability of spam accounts with respect to three categories of features, namely content attributes, social interactions, and profile properties.

Keywords Online social networks · Microblogging · Account abuse · Spam detection · Spam analysis

Abdullah Almaatouq and Erez Shmueli have contributed equally to this work.

✉ Abdullah Almaatouq
amaatouq@mit.edu

Erez Shmueli
shmueli@mit.edu

Mariam Nouh
mariam.nouh@cs.ox.ac.uk

Ahmad Alabdulkareem
kareem@mit.edu

Vivek K. Singh
singhv@mit.edu

Mansour Alsaleh
maalsaleh@kacst.edu.sa

Abdulrahman Alarifi
aarifi@kacst.edu.sa

Anas Alfaris
anas@mit.edu

Alex 'Sandy' Pentland
pentland@mit.edu

[1] Massachusetts Institute of Technology, Cambridge, MA, USA

[2] University of Oxford, Oxford, UK

[3] King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

1 Introduction

Spam exists across many types of electronic communication platforms, including e-mail, Web discussion forums, text messages (SMS), and social media. Today, as social media continues to grow in popularity, spammers are increasingly abusing such media for spamming purposes. According to a recent study [33], there was a 355% growth in social spam during the first half of 2013. Twitter company's initial public offering (IPO) filing indicates spam as a major threat in terms of undermining their value to advertisers and potential investors, as well as negatively affecting users' engagement [50].

While there is a growing literature on social media in terms of developing tools for spam detection [30,42,53] and analyzing spam trends [45,58,59], spammers continue to evolve and change their penetration techniques. Therefore, there is a continuous need for understanding the evolving and diverse properties of malicious accounts in order to combat them properly [33,50].

Published online: 20 February 2016

In this paper, we present an empirical analysis of suspended spam accounts on Twitter, in terms of profile properties and social interactions. To perform the study, we collected over 100 million tweets over the course of 1 month (from March 5, 2013, to April 2, 2013) generated by approximately 30 million distinct user accounts (see Sect. 3). In total, over 7 % of our dataset accounts are suspended or removed accounts due in part to abusive behaviors and other violations.

Our preliminary analysis for comparing the behavioral properties of normal and malicious users shows a tendency for a bimodal distribution in the case of spam accounts (see Sect. 4). Bimodal distributions commonly arise as a mixture of uni-modal distributions corresponding to a mixture of populations. Accordingly, we separate the subpopulations within spammers, using Gaussian mixture models (GMMs), resulting in two distinct subpopulations (categories) of spammers.

We then investigate the individual properties as well as the social interaction patterns of the two categories of spammers (see Sect. 5). We observe that the two categories exhibit different spamming patterns and employ distinct strategies for reaching their victims. More specifically, by analyzing the spam accounts profile attributes, we identify a cluster of malicious accounts that seems to be originally created and customized by legitimate users, whereas the other cluster deviates from the baseline significantly. Also, through network analysis of multiple social interactions, we reveal a set of diverse strategies employed by spammers for reaching audiences. We focus on the *mention* function as it is one of the most common ways in which spammers engage with users, bypassing any requirement of sharing a social connection (i.e., follow/following relationship) with a victim.

We analyze the detectability of spam accounts with respect to three categories of features, namely content attributes, social interactions, and profile properties (see Sect. 6). The goal is to highlight the importance of behavioral characteristics (i.e., profile and social interactions) as an enabling methodology for the detection of malicious users in OSNs. The conclusion and future work of our study are discussed in Sect. 8. In summary, we frame our contributions as follows:

- We categorize spam accounts based on their behavioral properties and find that Twitter spammers belong to two broad categories.
- We analyze the different properties of spam accounts in terms of their profile attributes and use the attributes of legitimate accounts as a baseline.
- Through network analysis of multiple social interactions, we reveal a set of diverse strategies employed by spammers for reaching audiences.
- By examining the detectability of spam accounts with respect to multiple categories of features, we highlight the importance of behavioral characteristic as an enabling methodology for OSNs spam detection.

Finally, we note that a portion of this paper has appeared previously as a conference publication [2]. Our main contributions for the journal version include highlighting the importance of behavioral characteristic as an enabling methodology for OSNs spam detection, adding more discussion, references, as well as in-depth analysis.

2 Background

Twitter is a microblogging platform and an online social network (OSN), where users are able to send *tweets* (i.e., short text messages limited to 140 characters). According to a recent study, Twitter is the fastest growing social platform in the world [23]. In 2013, Twitter estimated the number of active users at over 200 million, generating 500 million tweets per day [50].

Twitter spam is a systemic problem [45]. While traditional e-mail spam usually consists of spreading bulks of unsolicited messages to numerous recipients, spam on Twitter does not necessarily comply to the volume constraint, as a single spam message on Twitter is capable of propagating through social interaction functions and reaches a wide audience. In addition, previous studies showed that the largest suspended Twitter accounts campaigns directed users via affiliate links to some reputable Web sites that generate income on a purchase, such as Amazon [45]. Such findings blur the line about what constitutes as OSN spam. According to the "Twitter Rules," what constitutes *spamming* will evolve as a response to new tactics employed by spammers [49]. Some of the suspicious activities that Twitter considers as indications for spam [49] include: (1) aggressive friending; (2) creating false or misleading content; (3) spreading malicious links; and (4) trading followers.

Spam content can reach legitimate users through the following functions: (i) *home timeline*: a stream showing all tweets from those being followed by the user or posts that contain @*mention* requiring no prior follow relationship; (ii) *search timeline*: a stream of messages that matches a search query; (iii) *hashtags*: tags used to mark tweets with keywords or topics by incorporating the symbol # prior to the relevant phrase (very popular hashtags are called *trending topics*; (iv) *profile bio*: spam accounts generate large amounts of relationships and favorite random tweets from legitimate users with the hope that victims would view the spammer account profile which often contains a URL embedded in its bio or description; and (v) *direct messages*: private tweets that are sent between two users.

Accounts distributing spam are usually in the form of: (i) *fraudulent accounts* that are created solely for the purpose of

sending spam; (ii) *compromised accounts* created by legitimate users whose credentials have been stolen by spammers; and (iii) legitimate users posting spam content. While multiple previous studies focused on fraudulent accounts [45,46], the compromised accounts are more valuable to spammers as they are relatively harder to detect due to their associated history and network relationships. On the other hand, fraudulent accounts exhibit a higher anomalous behavior at the account level, and hence are easier for detection [18].

3 Datasets

Our Twitter dataset consists of 113,609,247 tweets, generated by 30,391,083 distinct users, collected during a 1-month period from March 5, 2013, to April 2, 2013, using the Twitter public stream APIs [48]. For each tweet, we retrieve its associated attributes (e.g., tweet text, creation date, client used) as well as information tied to the account who posted the tweet (e.g., the account's number of following, followers, date created). On average, we receive over 4 million tweets per day. We lack data for some days due to network outages, updates to Twitter's API, and instability of the collection infrastructure (using Amazon EC2 instances). A summary of tweets collected each day and outage periods is shown in Fig. 1.

In order to label spammer accounts in our dataset, we rely on Twitter's account suspension algorithm described in [45]. Given that the implementation of the suspension algorithm is not publicly available, we verify whether an account has been flagged as spam by checking the user's profile page. In case an account has been suspended or removed, the crawler request will be redirected to a page describing the user statues (i.e., suspended or does not exist). While all of the removed/suspended user's information is no longer available through the Twitter's API, we were able to reconstruct their information based on the collected sample. In total, over 7 % of our dataset are suspended/removed accounts. As we rely on Twitter suspension mechanism, this dataset contains caught spam accounts on Twitter by the suspension mechanism, where uncaught accounts are treated as legitimate users. Also, Twitter's policy page states that other activities such as publishing malicious links, selling usernames, and using obscene or pornographic images may also result in suspension or deletion [49]. Also, removed accounts may include users that deactivated their accounts during the data collection period, which will cause them to be treated as spam accounts in our analyses. Previous study [45] validated that the vast majority (i.e., 93 % true-positive rate) of suspensions are rooted in spamming behaviors and that Twitter's suspension algorithm has false-negative rate bound of ±3.3 % at 95 % confidence intervals.

4 Identifying subpopulations

The results of the initial analysis to compare the collective tweeting patterns and social behavior of normal and malicious users showed tendency for bimodality in the case of spam accounts. This was less evident in the case of legitimate users (see Fig. 2). This pattern occurs across multiple attributes (i.e., tweets count, favorites count, followers count). The bimodal distributions commonly arise as a mixture of uni-modal distributions corresponding to mixture of populations. Accordingly, we separated the subpopulations within spammers, using Gaussian mixture models (GMMs), in order to reveal distinct spamming strategies and behaviors.

In order to identify subsets of malicious accounts, we use Gaussian mixture models (GMMs). GMM is a probabilistic model that assumes that data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. To determine the number of components (i.e., subpopulations or clusters), we fit multiple GMMs with different numbers of Gaussians and then calculate the Bayesian information criteria (BIC) score for each fit. The use of BIC penalizes models in terms of the number of parameters or complexity. Hence, complex models (i.e., high number of free parameters) will have to compensate with how well they describe the data. This can be denoted as follows:

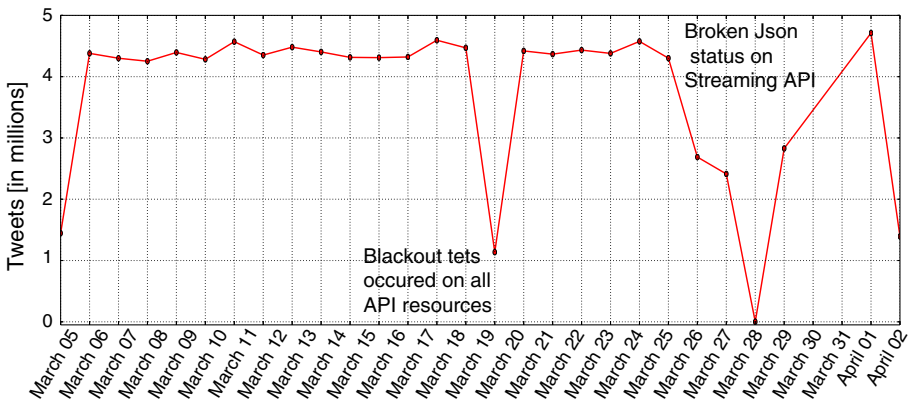

Fig. 1 Tweets received per day. On average, we receive 4 million tweets per day. We lack complete data for some days due to network outages, updates to Twitter's API, and instability of the collection infrastructure

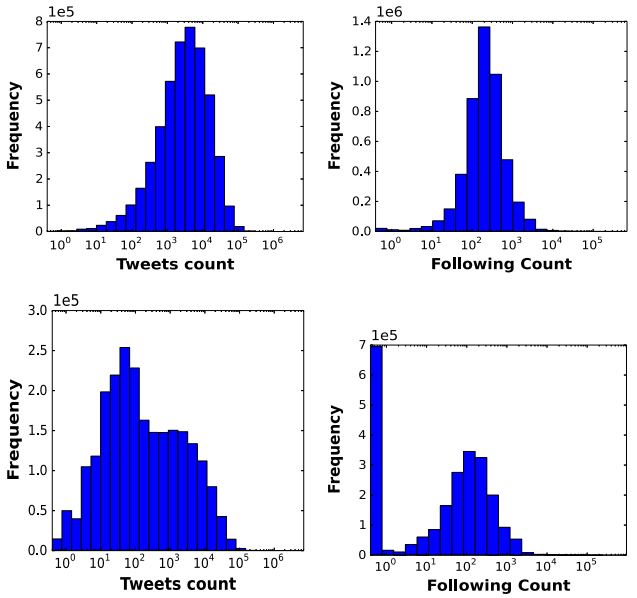

Fig. 2 An illustration of different tweeting patterns and following behaviors for normal and spam accounts. The first row (*top figures*) represents the tweets and following count frequencies for normal users. The second row (*bottom figures*) represents the tweets and following count frequencies for spam accounts

$$BIC(M_c) = -2 \cdot \ln P(x|M_c) + \ln N \cdot k \tag{1}$$

where x is the observed data, N is the number of observations, k is the number of free parameters to be estimated, and $P(x|M_c)$ is the marginal likelihood of the observed data given the model M with c number of components.

A GMM with two components and spherical covariance gives the lowest BIC score (see Fig. 3). The results of the clustering exhibit two classes of spam accounts $C_1 \subset C$ and $C_2 \subset C$, where C is the set of all accounts. We refer to the normal class (i.e., legitimate accounts) as C_{normal}. The results of the separation in one dimension (i.e., tweets count) are shown in Fig. 3.

Based on the separation, we can further investigate the properties and activity patterns of the different identified classes. This separation aids in developing taxonomies and exploiting meaningful structures within the spam accounts communities.

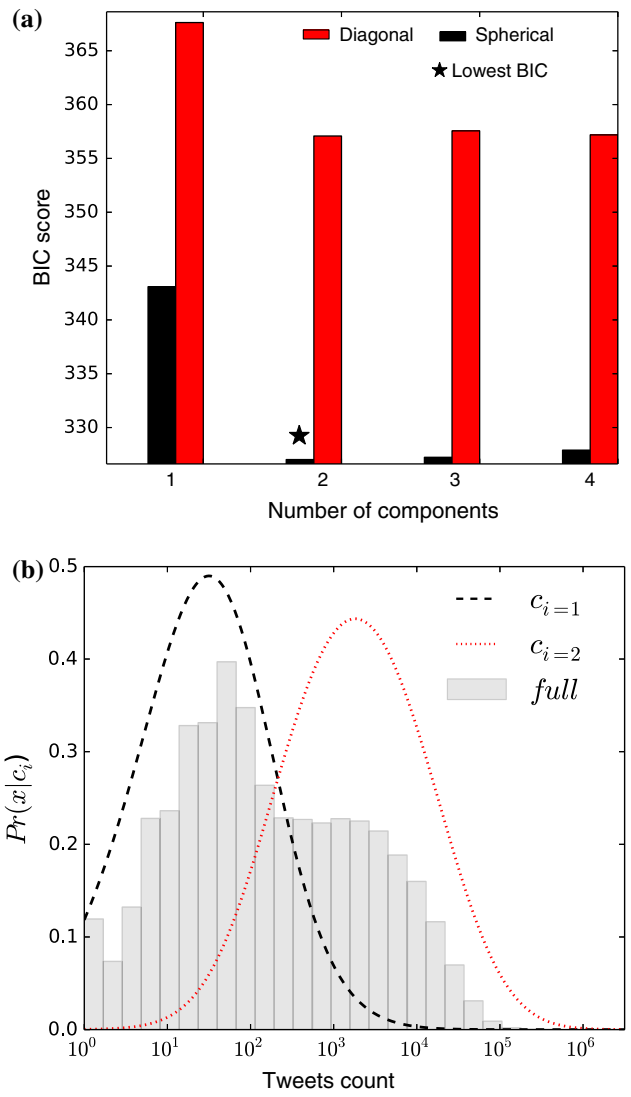

Fig. 3 A GMM with two components and spherical covariance gives the lowest BIC score. The results of the separation in one dimension (i.e., tweets count) are shown in **b**

5 Behavioral analysis

5.1 Profile properties

In order to further investigate the different identified classes, we examine the empirical cumulative distribution functions (ECDFs) of different attributes for each class (see Fig. 4). We find that 50 % of the accounts in C_1 have less than 29 tweets; however, for C_{normal} and C_2, 50 % of the accounts have tweeted around 2000 times. Furthermore, we find that almost 90 % of the accounts in C_1 have no favorites (i.e., tweets added to their favorites list), whereas C_2 and C_{normal} show closely matching patterns, with 50 % of the accounts having less than 50 favorite tweets.

We continue to observe similar patterns across multiple attributes, where C_2 and C_{normal} have similar distributions and C_1 deviates from the baseline. We explain this observation through the hypothesis that C_2 mainly consists of *compromised accounts*, while C_1 consists of *fraudulent accounts* as defined in Sect. 2.

The similarity between C_{normal} and C_2 in the basic profile attributes, such as the percentage of accounts with default profile settings, default profile images, profile descriptions, and profile URLs (see Table 1), might indicate that C_2

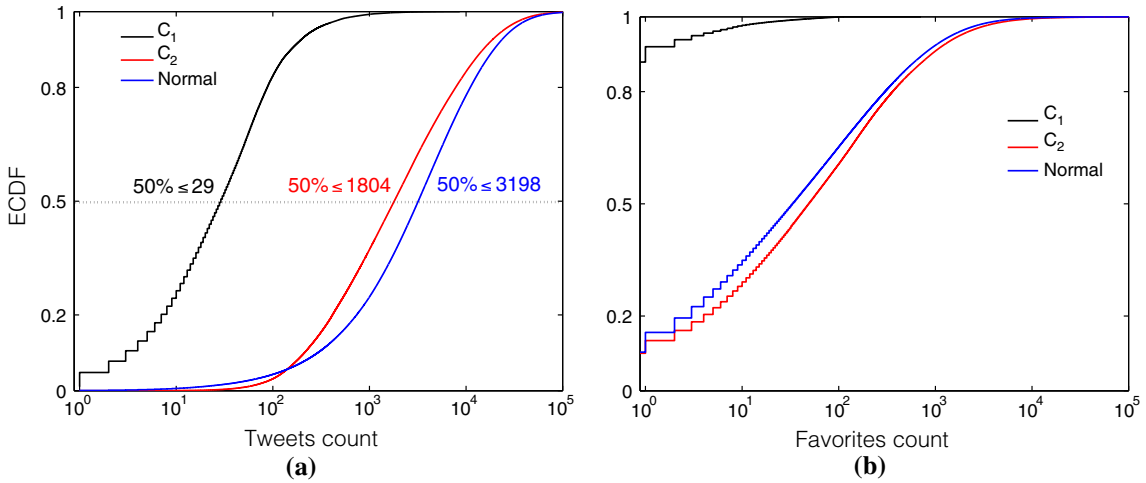

Fig. 4 Comparison between the three classes C_1, C_2, and C_{normal} in terms of tweeting and following behaviors after the GMM clustering

Table 1 Summary of basic profile attributes

	Default profile (%)	Default image (%)	URL (%)	Biography (%)
C_{normal}	22	1.3	29	83.6
C_1	76	14	4	60
C_2	36	1.5	20	84.7

We notice that C_{normal} and C_2 have relatively similar patterns

accounts were originally created and customized by *legitimate users*. For example, we notice that only 22 % of C_{normal} and 36 % of C_2 accounts kept their default profile settings unchanged, in comparison with 76 % in the case of C_1.

5.2 Social interactions

In this section, we analyze users behavior in terms of the follow relationship and mention functions, from the topological point of view. We approach this by incorporating multiple measures that are known to signify network characteristics (differences and similarity). Through this analysis, we reveal sets of behavioral properties and diverse strategies employed by spammers for engaging with victims and reaching audiences.

5.2.1 Preliminaries

Let $G = (V, E)$ be the graph that represents the topological structure of a given function (i.e., follow or mention), where V is the set of nodes and E is the set of edges. An edge in the graph is denoted by $e = (v, u) \in E$, where $v, u \in V$. Note that in the follow and mention networks, a node v corresponds to a Twitter user and an edge corresponds to an interaction between a pair of users. If two nodes have an edge between them, they are adjacent and we refer to them as neighbors.

We define the neighborhood of node v as the subgraph $H = (V', E') \mid V' \subset V$ and $E' \subset E$ that consists of all the nodes adjacent to v (alters) excluding v (we refer to v as ego) and all the edges connecting two such nodes. The 1.5 egocentric network $E_{1.5}(v)$ of node v is defined as the neighborhood subgraph including v itself. Therefore, the neighborhood can be denoted as $N(v) := \{u \mid (u, v) \in E$ or $(v, u) \in E\}$ and the 1.5 ego network as $E_{1.5}(v) := \{N(v) \cup \{v\}\}$.

Focusing on the egocentric networks around the nodes allows for studying the local graphical structure of a given user and signifies the types of interactions that develop within their social partners [4]. Figure 5 shows an illustration of different levels of egocentric networks. From this, we can define node properties and measure the relative importance of a node within its egocentric network such as node degree $d(v)$, node out-degree $d_{out}(v)$, in-degree $d_{in}(v)$, and reciprocal relationship $d_{bi}(v)$.

$$\begin{aligned} d_{out}(v) &= |\{u \mid (v, u) \in E_{1.5}(v)\}| \\ d_{in}(v) &= |\{u \mid (u, v) \in E_{1.5}(v)\}| \\ d(v) &= d_{in} + d_{out} \\ d_{bi}(v) &= |\{u \mid (u, v) \in E_{1.5}(v) \wedge (v, u) \in E_{1.5}(v)\}| \end{aligned} \tag{2}$$

From the properties defined in Eq. 2, we can derive the in-degree density $density_{in}(v)$, out-degree density $density_{out}(v)$, and the density of reciprocal relationships $density_{bi}(v)$.

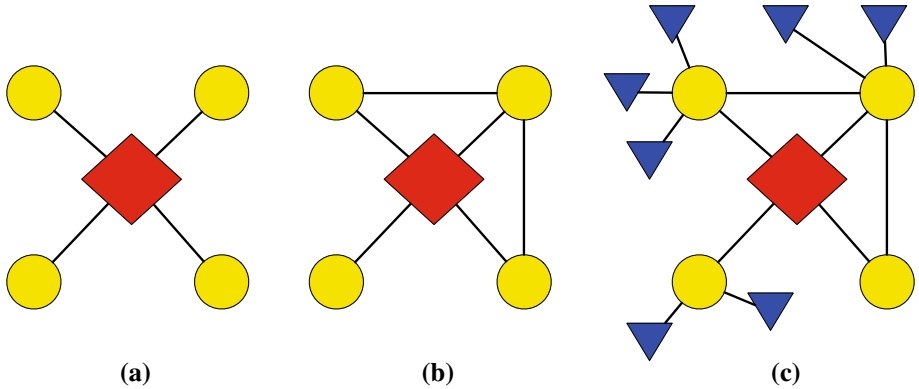

Fig. 5 An illustration of the **a** 1.0 egocentric network; **b** the 1.5 egocentric network; and **c** the 2.0 egocentric network. The ego node is marked in *red* (*diamond*), and its connections (*alters*) are marked in *yellow* (*circles*), and the alters' connections are marked in *blue* (*triangles*) (color figure online)

$$\begin{aligned} density_{in}(v) &= \frac{d_{in}(v)}{d(v)} \\ density_{out}(v) &= \frac{d_{out}(v)}{d(v)} \\ density_{bi}(v) &= \frac{d_{bi}(v)}{d(v)} \end{aligned} \tag{3}$$

In addition, we calculate the betweenness centrality for each ego node in order to quantify the control of such node on the communication between other nodes in the social network [21]. The measure computes the fraction of the shortest paths that pass through the node in a question v within its egocentric network $E_{1.5}(v)$. Therefore, the betweenness centrality $C_B(v)$ can be computed as [11]:

$$C_B(v) = \sum_{u \neq w \in N(v)} \frac{\sigma_{uw}(v)}{\sigma_{uw}} \tag{4}$$

where σ_{uw} is the total number of shortest paths from node u to node w and $\sigma_{uw}(v)$ is the number of those paths that pass through the node v. Therefore, $C_B(v) = 0$ in the case where all the alters are directly connected to each other and $C_B(v) = 1$ when the alters are only connected to each other through the ego node.

We also compute the closeness centrality $C_C(v)$ which measures the inverse of the sum of the shortest path distances between a node v and all other nodes $u_0, u_1, .., u_n \in N(v)$ normalized by the sum of minimum possible distances. This can be formulated as follows:

$$C_C(v) = \frac{n-1}{\sum_{u \in N(v)} \sigma(v,u)} \tag{5}$$

where $\sigma(u, v)$ is the shortest path distance between v and u, and n is the number of nodes in the egocentric graph.

A network is strongly connected if there is a path between every node to every other node in a directed graph. We define the number of strongly connected components in the egocentric networks $E_{1.5}(v)$ and open neighborhood $N(v)$ to be $SCC_{E_{1.5}}(v)$ and $SCC_N(v)$, respectively. By replacing all of the directed edges with undirected edges, we compute the number of weakly connected components for the egocentric network and open neighborhood as $WCC_{E_{1.5}}(v)$ and $WCC_N(v)$, respectively. The *SCC* and *WCC* are used to measure the connectivity of a graph.

5.2.2 Relationship graph

Twitter follow relationship is modeled as a directed graph, where an edge between two nodes $e = (v, u) \in E$ means that v is following u. For the follow relationship, we only have the number of followers and following for each account, and not the actual relationship list. Therefore, in order to compare relationships formed by both C_1 and C_2, we aggregate following and follower data from both classes.

Figure 6 shows the number of followers and following represented by the in-degree d_{in} (follower) and out-degree d_{out} (following) for each class. We find that spam accounts that belong to C_1 are heavily skewed toward following rather than followers, which could indicate a difficulty in forming reciprocal relationships. Furthermore, we observe a low $density_{in}$ for C_1 with an average of 0.16 and high $density_{out}$ with an average of 0.4. On the other hand, C_2 has more balanced densities with approximately 0.5 for both.

While Twitter does not constrain the number of followers a user could have, the number of following (i.e., d_{out}) is limited [47]. Every user is allowed to follow 2000 accounts in total; once an account reaches this limit, they require more followers in order to follow more users [47]. This limit is based on the followers to following ratio.

Furthermore, as shown in Fig. 6c, almost 50 % of C_1 accounts have no followers (i.e., they did not embed themselves within the social graph) and almost 75 % of these accounts have less than ten followers. We find that C_2 accounts are more connected in terms of social relationships, which makes them harder to detect and hence contribute

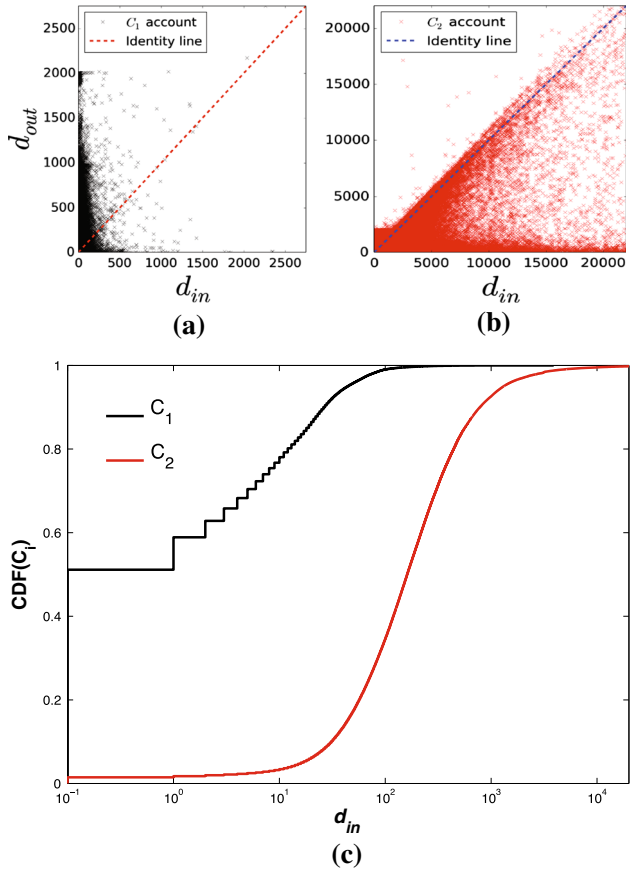

Fig. 6 Illustration of the different relationship behaviors for C_1 in **a** and C_2 in **b**. We find that spam accounts that belong to C_1 are heavily skewed toward following rather than followers or the identity line. The effect of the number of following limit (i.e., 2000 d_{out}) is apparent/observed in both classes

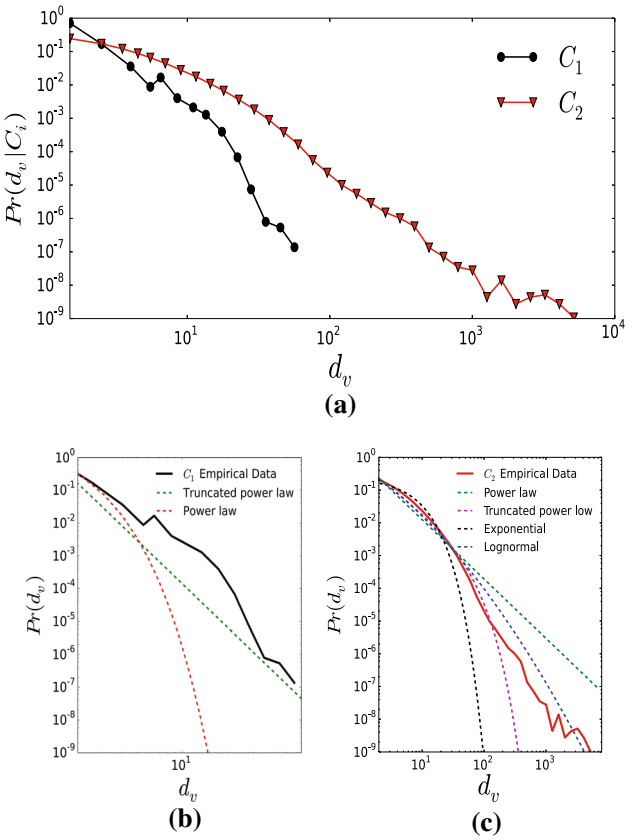

Fig. 7 The *top figure* shows the distribution of the frequency of mentions $d(v)$ for C_1 (*black circles*) and C_2 (*red triangles*). The *bottom figures* compare the empirical distribution obtained with best fits of other heavy-tailed distributions (color figure online)

more content. These findings adhere to a known phenomenon observed in multiple security contexts. For example, [3] showed that in many cases (especially in social networks), optimal attack strategies (i.e., causing greater damage or spreading more spam content) exhibit slow spreading patterns rather than spreading aggressively.

The compromised account population that exists within C_2 can utilize the associated history and network relationships of the original account owner to aid them in increasing the visibility of their spam content.

5.2.3 Mention graph

The mention function is one of the most common ways in which spammers engage with users; unlike the *direct messages (DM)*, it bypasses any requirement of prior social connection with a victim.

The mention network is constructed as a simple, weighted, and directed graph, such that an edge between two nodes $e = (v, u) \in E$ means that user v mentioned user u during our collection period. We extract the 1.5 egocentric network $E_{1.5}(v)$, where v are the accounts in C_1 and C_2.

Figure 7 shows the degree distribution of the mention network. Although multiple studies observed that the degree for the mention network follows heavy-tailed distributions (e.g., [27], in order to understand the topological structure, we further investigate the concrete goodness of fit [2]. The scale-free nature of the mention network (i.e., degree distribution that follows a power law) implies a very high heterogeneity level in user behavior, which is expected for human activity phenomena [9,32]. In addition, the figure shows a clear difference between the lengths of the tail of the distributions between the two classes C_1 and C_2.

Table 2 shows the comparison between two centrality measures for the mention network, namely the betweenness C_B and closeness C_C centralities. We observe that the average betweenness centrality for C_2 is significantly higher than C_1, which indicates that C_1 accounts target users that mention each other (i.e., communities and clusters of users). This is somewhat a surprising outcome, as we expect C_2 accounts to utilize the associated relationships of the original account owner, where the nodes in the neighborhood are real friends

Table 2 Comparing different centrality measures for the mention network for C_1 and C_2 accounts

Class	Betweenness (C_B)		Closeness (C_C)	
	μ	σ	μ	σ
C_1	0.014	0.08	0.97	0.12
C_2	0.096	0.14	0.77	0.25

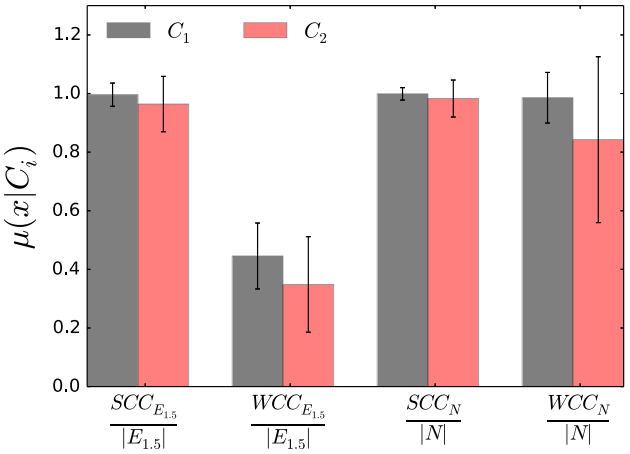

Fig. 8 The density of connected components in the mention network for C_1 and C_2

and are more likely to mention one another. The relatively low betweenness in C_1 can be explained by at least three possibilities:

- *Conversations hijacking*. We observe that 51.5 % of the tweets captured by C_1 contain mentions and 43.3 % of these mentions are replies. In addition, only 1.2 % of their mentions were reciprocated ($density_{bi} = 0.0127$), which arouses suspicion that C_1 accounts intrude on ongoing conversations between legitimate users and thus have resulted in a low betweenness centrality.
- *Targeting hubs*. Due to the scale-free nature (i.e., degree distribution that follows a power law) of the mention network, mentioning or replying to hubs (nodes that are highly connected to other nodes in the network) increases the chance that the alters will be connected, and hence the low betweenness score.
- *Crawling profiles*. It is also possible that C_1 accounts target communities and connected users in the mention graph by crawling profiles (i.e., visiting the followers and following lists or users' *timeline* of the seed targeted profile).

Figure 8 shows high average densities of strongly connected components for both the egocentric network and the neighborhood network in classes C_1 and C_2 (i.e., $\frac{SCC_N}{|N|}$ and $\frac{SCC_{E_{1.5}}}{|E_{1.5}|}$). This observation indicates a difficulty in forming reciprocal mention relationships as discussed earlier. Also, a higher score in the densities of weakly connected components ($\frac{WCC_N}{|N|}$ and $\frac{WCC_{E_{1.5}}}{|E_{1.5}|}$) for C_1 explains the lower betweenness centrality score observed in Table 2.

The discrepancy in network measures (i.e., degree distribution, centralities, and connectivity) between C_1 and C_2 indicates the existence of different strategies for reaching audiences employed by each class accounts.

6 Detection analysis

In this section, we analyze the detectability of spam accounts with respect to three categories of features, namely content attributes, social interactions, and profile properties (see Sect. 6.1). Our goal here is to highlight the importance of behavioral characteristics (i.e., profile and social interactions) as an enabling methodology for the detection of malicious users in OSNs. As Twitter spammers are constantly evolving to evade existing detection features, content-based features (e.g., tweet similarity and duplicate tweet count) will easily be evaded. In our work, we investigate new and robust features to detect Twitter spammers. Therefore, unlike previous studies (e.g., [42], we focus on comparing the different categories of features in terms of their relative classification performance (see Sect. 6.2) and information gain (see Sect. 6.3), rather than on achieving a high absolute classification performance. Moreover, although our algorithm relies on a "labelled dataset" that was extracted from Twitter, it does not mean that these labels were generated by an automatic spam detection algorithm. It could have been the case that a large amount of the suspended accounts (that we consider as spam accounts) were suspended manually (e.g. if legitimate users reported these accounts as spam accounts).

Our analyses included four different classification tasks: (1) distinguishing spam accounts ($C_1 \cup C_2$) from normal accounts (C_{normal}), (2) distinguishing C_1 spam accounts from normal accounts (C_{normal}), (3) distinguishing C_2 spam accounts from normal accounts (C_{normal}), and (4) distinguishing C_1 spam accounts from C_2 spam accounts.

In order to reduce computation time, all of the experiments reported in this section were conducted on a stratified sample of Twitter accounts, which was obtained by sampling 2.5 % of the accounts in each of the three subpopulations in our dataset.

6.1 Features extraction

As mentioned above, we experimented with three categories of features:

Table 3 Content features summary

Feature	Description	Type
Mean tweets similarity	The average pairwise tweets similarity based on the term frequency inverse document frequency	Float
Sampled tweets count	The number of sampled tweets appearing in our dataset for a specific user. This feature is an indication of the account activity level during the data collection period	Integer
Tweets with mentions	The number of tweets containing mentions to other users (i.e., if one tweet contains more than one mentioned user, it still counts as one)	Integer
Tweets with hashtags	The number of tweets containing hashtags (i.e., if one tweet contains more than one hashtag, it still counts as one)	Integer
Hashtags density	The number of hashtags (i.e., one tweet can include more than one hashtag) normalized by the number of tweets	Float
Tweets with links	The number of tweets containing URLs (i.e., if one tweet has more than one URL, it still counts as one)	Integer
Links density	The number of URLs normalized by the number of tweets	Float

Table 4 Summary of profile features

Feature	Description	Type
Total number of tweets	The total number of tweets posted by the user	Integer
Favorite count	The total number of tweets that the user has marked as favorite	Integer
Verification status	Whether the user account is verified by twitter. Verification is currently used to establish authenticity of identities of key individuals and brands on Twitter	Integer
Default profile image	Whether the user is using Twitter's default avatar image	Boolean
Listed count	The number of Twitter lists on which the user appears	Integer
Geo enabled	Whether the geographical location of the user account is activated	Boolean
Account-age	The number of days between the time of creation of the account until the date of the last tweets captured in our dataset	Float

Content features capture linguistic cues and specific properties of the tweet text posted by a user. Given that our dataset contains multiple tweets for each user, we extract the densities, averages, or frequencies of content attributes. A summary of the features used and their description is given in Table 3 Features are inspired by [6,20,30,39,57].

Profile features are based on Twitter meta-data related to an account, including language, geographic locations, and account creation time (see Table 4). Similar features were used in [6,26].

Social interaction features capture various dimensions of information diffusion patterns. We build networks based on mentions, replies, and follow relationships, and extract their statistical features. Examples include degree distribution and centrality measures (see Table 5). Several of these features have been used previously in the literature [6,20,26,37,39,53].

A note on categorical features: While categorical features can easily be coded as integers, where each integer value represents a different category, such integer values may be

Table 5 Summary of social interaction features

Feature	Network	Description	Type
Out-degree	Follow	The number of accounts a user is following (i.e., following count)	Integer
Out-degree density	Follow	The density of followings	Float
In-degree	Follow	The number of accounts following the user (i.e., followers count)	Integer
In-degree density	Follow	The density of followers	Float
In-degree	Mention	The number of accounts mentioning the user of interest	Integer
Weighted in-degree	Mention	The number of time the user of interest was mentioned	Integer
Weighted in-degree density	Mention	The number of time the user of interest was mentioned normalized by the number of accounts mentioning the user	Float
Out-degree	Mention	The number of accounts mentioned by the user of interest	Integer
Weighted out-degree	Mention	The number of time the user of interest mentioned other users	Integer
Weighted out-degree density	Mention	The number of time the user of interest mentioned other users normalized by the number of accounts that mentioned the user	Float
Bidegree	Mention	The number of reciprocal mention relationships	Integer
Weighted bidegree	Mention	The weighted reciprocal relationship or conversations length	Integer
Closeness centrality	Mention	The closeness centrality of the node with respect to the 1.5 ego network	Float
Betweenness centrality	Mention	The betweenness centrality of the user with respect to the 1.5 mention ego network	Float
Relative edges density	Mention	The total degree of the user normalized by the total number of edges in the 1.5 ego network	Float
Open strongly connected components	Mention	The number of strongly connected components in the neighborhood of the user (i.e., excluding the user of interest)	Integer
Open weakly connected components	Mention	The number of weakly connected components in the neighborhood of the user(i.e., excluding the user of interest)	Integer
Ego strongly connected components	Mention	The number of strongly connected components in the 1.5 ego network of the user (i.e., including the user of interest)	Integer
Ego weakly connected components	Mention	The number of weakly connected components in the 1.5 ego network of the user (i.e., including the user of interest)	Integer

misinterpreted as being ordered, which may result in undesired behaviors. Therefore, in our experiments, we used the 1-of-K encoding [1,34] technique to convert a categorical feature with k possible values to a set of k binary features.

6.2 Classification performance

For each one of the four binary classification tasks and each one of the three categories of features (i.e., content, profile, and social features), we trained and tested seven different machine learning algorithms (i.e., ZeroR, Bayesian network, naive Bayes, logistic regression, decision trees, and random forest) in a fivefold cross-validation manner to compute the average area under the ROC curve (AUROC) and the standard deviation.

In our first experiment, we attempted to distinguish spam accounts from legitimate users. Focusing on the best performing algorithm (decision tree) in Fig. 9, we observe that the social interaction features outperform profile and content features and hence seem to be a better indicator for clas-

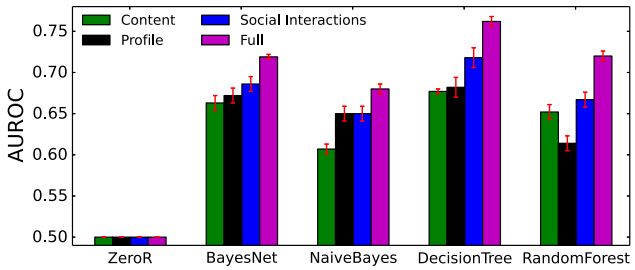

Fig. 9 The results of experiment #1 where we try to distinguish C_{normal} from $C_1 \cup C_2$

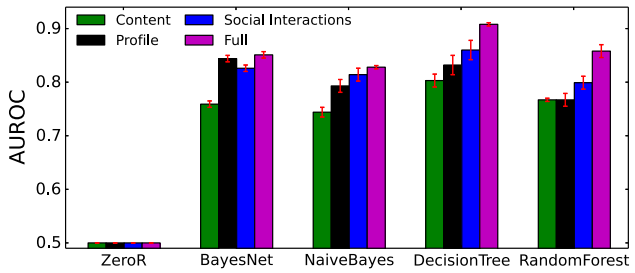

Fig. 10 The results of experiment #2 where we try to distinguish C_1 types of spam accounts

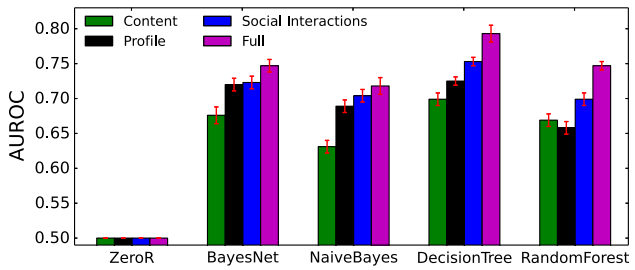

Fig. 11 The results of experiment #3 where we try to distinguish C_2 types of spam accounts from C_{normal}

sifying spam accounts. We also notice that profile features outperform content features in this case.

The second experiment focused on separating C_1 spam accounts from normal accounts (C_{normal}). As shown in Fig. 10, we observe again a similar pattern where the social interaction features achieve the highest detection score (with the only exception of the Bayes network classifier).

It is also important to notice the scale in this experiment; the detection AUC score is relatively higher than the scores obtained in the previous experiment. This result is quite expected from our previous analyses due to the fact that C_1 spam accounts deviate significantly from C_{normal} accounts across different attributes.

In the third experiment (see Fig. 11), we study the detectability of C_2 spam accounts from normal accounts (C_{normal}). We find that social interaction features provide a better indication in comparison with other types of features. However, in this experiment the reported AUC scores

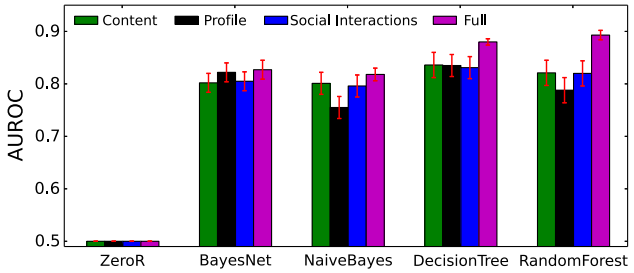

Fig. 12 The results of experiment #4 where we try to distinguish the different types of spam accounts

are lower than the ones that were reported in the second experiment (i.e., C_1 vs. C_{normal}). Again, this result is quite expected due to our earlier observations that C_2 and C_{normal} manifested similar patterns across multiple attributes.

In our fourth experiment, we focused on spam accounts only (see Fig. 12). Surprisingly, although we used some of the profile features to infer the separation between the two classes in Sect. 4, the content features (generally) provided a better detection signal than the profile features and a comparable signal to the social interaction features. The discrepancy between the results obtained in the first three experiments above and this experiment might be explained as follows. Both C_1 and C_2 users engage with their environment in an anomalous manner compared to C_{normal} users, and hence both types can be distinguished relatively easily from C_{normal} using such features. However, comparing C_1 users to C_2 users becomes difficult since they both exhibit anomalous social interaction patterns, and therefore, content features become more important.

6.3 Information gain results

Finally, it is worth mentioning that in all four experiments, and for all seven machine learning algorithms, the composite model (involving all features presented in this work) performed significantly better than single-category models (e.g., content or profile based).

While the four experiments (presented in Sect. 6.2 focused on evaluating the different categories of features, in this section we evaluate individual features in terms of their information gain. Note, however, that as opposed to the previous approach, this approach does not capture the dependencies between the different features.

The information gain IG for an attribute α in each experiment's training examples T evaluates the worth of α by measuring the IG with respect to the class C. This concept can be formulated as follows:

$$IG(T, a) = H(T) - H(T|\alpha)$$

where H is the is the information entropy (i.e., the average amount of information contained in each attribute).

The results of the analysis in this subsection conform with our findings in the previous subsection and with our findings in Sect. 5 as we proceed to explain. Table 6 shows the top ten attributes, ranked in terms of their information gain scores, for each one of the four classification tasks. As shown in the table, the social interaction features account for 90, 60, 60, and 40 % of the top ten features for the four classification tasks, respectively.

More specifically, we can see that social interaction features outperform other type of features such as profile and content. Moreover, in the case of the fourth classification task, the granularity of content becomes relatively more discriminative.

7 Related work

We discuss prior related work on OSNs' spam and network analysis. Although we focus on spam accounts analysis, our first-in-its-kind approach of spam behavioral categorization (i.e., identifying subpopulations), analyzing the different classes of spam accounts, and analyzing the mention interactions, provides a unique view in looking at spam trends in OSNs.

7.1 Spam in social networks

With the rapid growth of OSNs popularity, we are witnessing an increased usage of these services to discuss issues of public interest and hence shape public opinions [16]. This model of users as an information contributors has provided researchers, news organizations, and governments with a tool to measure (to some degree) representative samples of populations in real time [24,29,43]. However, [28] identified *propagandists* Twitter accounts that exhibit opinions or ideologies to either sway public opinion, disseminate false information, or disrupt the conversations of legitimate users. The study focused on accounts connected to two political events: (i) the 2010 Nevada senate race and (ii) the 2011 debt-ceiling debate. A similar campaign has been analyzed by [44], in which spam accounts flood out political messages following the announcement of Russia's parliamentary election results. In addition, classical forms of abuse such as spam and criminal monetization exist in Twitter including phishing scams [15], spreading malware [36], and redirecting victims to reputable Web sites via affiliate links [45] to generate income.

Table 6 Summary of the information gain evaluation of individual features for the four experiments

#	Feature	Type	IG
Experiment #1			
1	Mention out-degree density	Social interaction	0.044
2	Mention in-degree density	Social interaction	0.043
3	Follow in-degree	Social interaction	0.040
4	Follow out-degree	Social interaction	0.039
5	Total number of tweets	Profile	0.028
6	Mention out-degree	Social interaction	0.028
7	Mention weighted out-degree density	Social interaction	0.023
8	Follow in-degree density	Social interaction	0.019
9	Follow out-degree density	Social interaction	0.019
10	Mention closeness centrality	Social interaction	0.015
Experiment #2			
1	Mention out-degree density	Social interaction	0.075
2	Mention in-degree density	Social interaction	0.074
3	Follow in-degree	Social interaction	0.066
4	Follow out-degree	Social interaction	0.059
5	Total number of tweets	Profile	0.049
6	Favorites count	Profile	0.030
7	Links density	Content	0.020
8	Tweets with links	Content	0.019
9	Follow out-degree density	Social interaction	0.019
10	Follow in-degree density	Social interaction	0.019
Experiment #3			
1	Mention out-degree density	Social interaction	0.056
2	Mention in-degree density	Social interaction	0.055
3	Follow in-degree	Social interaction	0.050
4	Follow out-degree	Social interaction	0.047
5	Total number of tweets	Profile	0.035
6	Favorites count	Profile	0.019
7	Links density	Content	0.014
8	Tweets with links	Content	0.014
9	Follow out-degree	Social interaction	0.014
10	Follow in-degree density	Social interaction	0.013
Experiment #4			
1	Mention out-degree density	Social interaction	0.316
2	Mention in-degree density	Social interaction	0.315
3	Follow in-degree	Social interaction	0.254
4	Follow out-degree	Social interaction	0.223
5	Total number of tweets	Profile	0.218
6	Favorites count	Profile	0.184
7	Links density	Content	0.140
8	Tweets with links	Content	0.139
9	Replies density	Content	0.118
10	Mean tweets similarity	Content	0.110

7.2 Social network spam analysis

Due to the popularity of social media services, several studies measured and analyzed spam in OSNs. [57] provided an analysis of some of the evasive techniques utilized by spammers and discussed several detection features. In addition, [58] performed an empirical analysis of the social relationship in Twitter (i.e., following relationship) in the spam community. The study showed that spam accounts follow each other and form small-world networks. [41] examined *Twitter account markets* and investigated their association with abusive behaviors and compromised profiles. [46] performed a study in collaboration with Twitter to investigate the *fraudulent accounts* marketplace. The study discussed prices, availability, and fraud perpetrated by 27 merchants generating 127 to 459K US dollars for their efforts over the course of 10 months. In another study [45], the authors examined tools, techniques, and support infrastructure spam accounts rely upon to sustain their campaigns. Surprisingly, the study showed that three of the largest spam campaigns in Twitter direct users to legitimate products appearing on reputable Web sites via affiliate links that generate income on a purchase (e.g., *Amazon.com*. However, the authors considered only tweets that contained URLs, and thus overlook malicious accounts that employ other spamming strategies, such as: i) embedding *non-hyperlink URL* by encoding the ASCII code for the dot; ii) *follow spam accounts* that generate large amounts of relationships for the hope the victim account would reciprocate the relationship or at least view the criminal's account profile which often has a URL embedded in its bio [22] investigated the spammers' mechanism of forming social relationship (link framing) in Twitter and found that vast majority of spam accounts are followed by legitimate users who reciprocate relationships automatically (social capitalists). The dataset used in this study contained 41,352 suspended Twitter accounts that posted a blacklisted URL. However, [25] discussed the ineffectiveness of blacklisting at detecting social network spam in a timely fashion and also suggested the existence of subpopulations of spam accounts.

Moreover, Boshmaf et al. [10] evaluated how OSNs are vulnerable to large-scale infiltration campaign caused by social bots by building and coordinating a group of programmable social bots on Facebook for 8 weeks then evaluated the collected data and studied the effects for the spamming campaigns and users behavior. Influenced by Boshmaf et al. [10] work, Elyashar et al. [19] studied infiltration targeting specific organizations' employees using Facebook. They have created social bots which were able to get connected with 50–70 % of organizations' employees and get access to their personal information.

7.3 Social network spam detection

A number of detection and combating techniques proposed in the literature rely on machine learning. [7] manually labeled 8,207 Twitter accounts and developed a classifier to detect spammers based on the URL and hashtag densities, followers to following ratio, account-age, and other profile-based features. The account-age and number of URLs sent were the most discriminating features. Stringhini et al. [42] created a diverse set of "honey-profiles" and monitored activities across three different social networks (Facebook, Twitter, and MySpace) for approximately 1 year. They also built a tool to detect spammers on Twitter and successfully detected and deleted 15,857 spam accounts in collaboration with Twitter.

Another approach is presented by [56], where they designed and implemented a system that recognizes legitimate users early in OSNs. They utilized an implicit vouching process, where legitimate users help in identifying other legitimate users. Additionally, [55] investigated the feasibility of utilizing crowdsourcing as the enabling methodology for the detection of *fraudulent accounts*. This study analyzed the detection accuracy by both "experts" and "turkers" (i.e., workers from Amazon Mechanical Turk under a variety of conditions). Moreover, [30] used traditional classifiers to detect spam users in Twitter. They defined a collection of content-based and user-based features. Similarly, [53] proposed content-based and graph-based features to facilitate spam detection using different classification algorithms. His results show that the Bayesian classifier generates best overall performance. [52] proposed a new system that predicts whether a user will interact with the social bots in Twitter using a set of selected features and six classifiers (5-nearest neighbor, logistic regression, multilayer perceptron, naive Bayes, and random forest). Wang et al. [54] presented a new sybil detection system using server-side clickstream models for Renren which is a large Chinese social network. The clickstream models are created by clustering clickstream into behavioral clusters.

However, most of the work in the literature did not consider the behavioral features. This is highly important as spammers continue to adopt different techniques and workarounds to overcome the standard detection methods. One of the recent works that incorporated behavioral features into the detection mechanism is the work by [20]. They designed a framework for detecting Twitter social bots, where they identified several classes of features ranging from users and content-based features, to behavioral network-based features, to distinguish between bot and human behavior.

Moreover, [5] surveyed sybil defenses approaches that leverage the structural properties of social networks for accurate identification of sybil accounts. The authors also

provided an analysis of these approaches and highlighted their strengths and weaknesses. Beeutel et al. [8] focused on Page Likes generated by spammer on Facebook and proposed a new approach based on social graphs that capture Page Likes, users who created these likes, and the times at which the likes are created in order to identify detection patterns of spammers using iterative and approximate-based algorithms. Another work by Cao et al. [12,13] proposed a social graph-based tool, called SybilRank, to detect sybil accounts in Tuenti which is the largest OSN in Spain. Their tool was deployed and tested in Tuenti operation center which helps the Tuenti system to detect 18 times more sybil accounts than before. SybilRank is based on the observation that short random walks from non-Sybil accounts on the social network tend to stay within the non-Sybil region of the network, and another tool, called SynchroTrap, was also proposed to detect malicious accounts in online social networks, and it relies on the observation that malicious accounts tend to perform loosely synchronized actions relative to benign accounts. SynchroTrap was implemented and deployed at Facebook and Instagram and resulted in detecting more than two million malicious accounts.

Beyond detection, Wagner et al. used a set of network, behavioral, and linguistic features to build a predictive model to estimate users' level of *susceptibility* for Twitter using data from the Social Bot Challenge 2011 [51]. Stein et al. [40] built Facebook immune system which checks and classifies every action in real time and provides explicit and implicit user feedbacks and protects its users from malicious activities including spamming. The classification is built using various machine learning-based classifiers such as random forests, SVM, and logistic regression.

7.4 Social bots for the greater good

Although social bots are typically referred to as an evil entity conducting malicious behavior, several social bots actually perform benign useful functions in online social networks. Therefore, not all of the identified bots should be suspended as many of them actually serve useful functions. For instance, social bots that aggregate content are being used for delivering news feeds, hot topics, and breaking news occurring in a user's social network. One example is Fuego [35], a Twitter bot designed to deliver the future of journalism by monitoring a user's universe of people and returning the links and stories they are sharing. Another example for a useful social bot that reports about hazardous events is Earthquake Robot [38]. It gathers information from the U.S. Geological Survey (USGS) and updates users about earthquakes as they happen. Other benign social bots are used by companies to provide customer care and gather their feedback. Some marketers use social bots that detect specific keywords and send automated replies/follow requests to customers. The main challenge here is to be able to distinguish between benign and harmful social bots.

Although some social bots may be designed with good intentions, the fact that they are fully automated may sometimes make them dangerous by spreading rumors and causing social panic. A recent study demonstrates that Twitter followers perceive Twitter bots as credible attractive sources [17]. Thus, false information spread by automated accounts is regarded as credible and may lead to false accusations as happened in the Boston marathon bombing [14].

8 Conclusion and future work

This paper presents a unique look at spam accounts in OSNs through the lens of the behavioral characteristics and spammers' techniques for reaching victims. We find that there exist two main classes of spam accounts that exhibit different spamming patterns and employ distinct strategies for spreading spam content and reaching victims. We find that C_2 (i.e., category 2 of spammers) and C_{normal} (i.e., legitimate users) manifest similar patterns across multiple attributes. We attempt to explain this observation through the hypothesis that C_2 mainly consists of compromised accounts, while the accounts in C_1 (i.e., category 1 of spammers) are fraudulent accounts, as we find support for the hypothesis throughout our analysis of profile properties. It is also possible that fraudulent and compromised accounts can gain more followers by purchasing them from online services [2] to evade detection [57,58]. In terms of the relationship graph, we find that spam accounts that belong to C_1 are heavily skewed toward following rather than followers, which indicates difficulty in forming reciprocal relationships. Furthermore, we observe a low in-degree density for C_1, while C_2 has a more balanced in-/out-degree densities. We show that the betweenness centrality for C_1 in the mention graph is significantly lower than C_2, which might be a result of hijacking conversations, targeting hubs, or crawling profiles.

Following the behavioral analysis, we also investigated the detectability of spam accounts with respect to three categories of features, namely content attributes, social interactions, and profile properties, focusing on two types of analysis: (1) relative classification performance and (2) information gain. The results of these analyses highlighted the importance of social interaction features when distinguishing between legitimate users and spammers. However, once we attempt to distinguish the two types of spammers, the very obvious features (i.e., social interaction and profile) diminish and the details (i.e., content) become more relevant. Generally, in all classification tasks, using the union of all feature types provided the highest classification performance. The sociobehavioral features demonstrated to work with relatively few examples in the learning phase, before

automatically detecting spamming accounts with minimal processing time. Thus, there is a good chance that the proposed sociobehavioral features are more robust (i.e., harder to evade by spammers) and will allow for the detection of such accounts much faster than Twitter's current approach. We cannot currently test this hypothesis since we do not know the exact time in which the accounts were suspended.

We acknowledge that our analysis may contain some bias. We have a partial view of the activities occurring during the data collection period due to the at most 1 % sampling limit imposed by Twitter. However, the work of [31] showed that the implications of using the Twitter Streaming API depend on the coverage and type of analysis. Generally, the streaming API can be sufficient to provide representative samples that get better with higher coverage, for certain types of analysis (i.e., top hashtags, topics, retweet network measures). Furthermore, we lack the absolute ground truth labels for the accounts presented in the dataset and primarily rely on Twitter's suspension algorithm. This might impose a lower bound on the number of spam accounts in our dataset (i.e., uncaught spam accounts are treated as legitimate users). In addition, there might be a fraction of legitimate users who deactivated their accounts during the collection period and hence would be labeled as removed. We also lack the appropriate resolution for important attributes used in the analysis; for example, we only have the number of followers and following for each user, and not the actual relationships list. Finally, our sample suffers from other technical limitations, such as a number of service outages that affected the collection during some days throughout the accounted month. Despite such limitations, we believe that our first-in-its-kind analysis of twitter functions and spam behavioral categorization describes well the current trends and phenomenon of OSN's spam and can be leveraged in designing OSN spam detectors and resilient architectures.

In our future work, we will design and test alternative labeling and validation mechanisms for the analyzed accounts. In particular, given that the compromised accounts are very different from the fraudulent accounts, sudden changes in the behavior of compromised accounts could be detected, which would indicate the time at which the account got compromised. This will require collecting and analyzing a new dataset with more frequent checking for suspension in order to provide accurate time stamp of when the suspension occurred. In addition, we plan to further investigate the differences between the spam accounts utilizing other interactions functions (e.g., hashtag, retweet, and favorite). We also intend to quantify the success of spam campaigns and explore the tools, techniques, and spam underground markets utilized by spam accounts to spread their content and evade many of the known detection mechanisms.

References

1. Almaatouq, A., Alabdulkareem, A., Nouh, M., Alsaleh, M., Alarifi, A., Sanchez, A., Alfaris, A., Williams, J.: A malicious activity detection system utilizing predictive modeling in complex environments. In: 2014 IEEE 11th Consumer Communications and Networking Conference (CCNC), pp. 371–379 (2014). doi:10.1109/CCNC.2014.6866597
2. Almaatouq, A., Alabdulkareem, A., Nouh, M., Shmueli, E., Alsaleh, M., Singh, V.K., Alarifi, A., Alfaris, A., Pentland, A.S.: Twitter: Who gets caught? Observed trends in social microblogging spam. In: Proceedings of the 2014 ACM Conference on Web Science, WebSci '14, pp. 33–41. ACM, New York, NY, USA (2014). doi:10.1145/2615569.2615688
3. Altshuler, Y., Aharony, N., Pentland, A., Elovici, Y., Cebrian, M.: Stealing reality: when criminals become data scientists (or vice versa). IEEE Intell. Syst. **26**(6), 22–30 (2011). doi:10.1109/MIS.2011.78
4. Altshuler, Y., Fire, M., Shmueli, E., Elovici, Y., Bruckstein, A., Pentland, A., Lazer, D.: The social amplifier reaction of human communities to emergencies. J. Stat. Phys. **152**(3), 399–418 (2013). doi:10.1007/s10955-013-0759-z
5. Alvisi, L., Clement, A., Epasto, A., Lattanzi, S., Panconesi, A.: Sok: The evolution of sybil defense via social networks. In: 2013 IEEE Symposium on Security and Privacy (SP), pp. 382–396. IEEE (2013)
6. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on twitter. In: Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS), vol. 6, p. 12 (2010)
7. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on Twitter. In: Proceedings of the Seventh Annual Collaboration, Electronic messaging, Anti-abuse and Spam Conference (CEAS) (2010)
8. Beutel, A., Xu, W., Guruswami, V., Palow, C., Faloutsos, C.: Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 119–130. International World Wide Web Conferences Steering Committee (2013)
9. Borondo, J., Morales, A.J., Losada, J.C., Benito, R.M.: Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish Presidential election as a case study. Chaos Interdiscip. J. Nonlinear Sci. **22**(2), 023138 (2012)
10. Boshmaf, Y., Muslukhov, I., Beznosov, K., Ripeanu, M.: Design and analysis of a social botnet. Comput. Netw. **57**(2), 556–578 (2013)
11. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**, 163–177 (2001)
12. Cao, Q., Sirivianos, M., Yang, X., Pregueiro, T.: Aiding the detection of fake accounts in large scale social online services. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, pp. 15–15. USENIX Association (2012)
13. Cao, Q., Yang, X., Yu, J., Palow, C.: Uncovering large groups of active malicious accounts in online social networks. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 477–488. ACM (2014)
14. Cassa, C.A., Chunara, R., Mandl, K., Brownstein, J.S.: Twitter as a sentinel in emergency situations: lessons from the Boston marathon explosions. PLoS Curr **5** (2013). http://currents.plos.org/disasters/article/twitter-as-a-sentinel-in-emergencysituations-lessons-from-the-boston-marathon-explosions/
15. Chhabra, S., Aggarwal, A., Benevenuto, F., Kumaraguru, P.: Phi.sh/$ocial: The phishing landscape through short urls. In: Proceedings of the 8th Annual Collaboration, Electronic Messaging,

Anti-Abuse and Spam Conference, CEAS '11, pp. 92–101. ACM, New York, NY, USA (2011). doi:10.1145/2030376.2030387
16. Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., Menczer, F.: Political polarization on twitter. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM) (2011)
17. Edwards, C., Edwards, A., Spence, P.R., Shelton, A.K.: Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. Comput. Hum. Behav. **33**, 372–376 (2014). doi:10.1016/j.chb.2013.08.013
18. Egele, M., Stringhini, G., Kruegel, C., Vigna, G.: COMPA: Detecting compromised accounts on social networks. In: ISOC Network and Distributed System Security Symposium (NDSS) (2013)
19. Elyashar, A., Fire, M., Kagan, D., Elovici, Y.: Homing socialbots: intrusion on a specific organization's employee using socialbots. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1358–1365. ACM (2013)
20. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. arXiv preprint arXiv:1407.5225 (2014)
21. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1977)
22. Ghosh, S., Viswanath, B., Kooti, F., Sharma, N.K., Korlam, G., Benevenuto, F., Ganguly, N., Gummadi, K.P.: Understanding and combating link farming in the twitter social network. In: Proceedings of the 21st International Conference on World Wide Web, WWW '12, pp. 61–70. ACM, New York, NY, USA (2012). doi:10.1145/2187836.2187846
23. GlobalWebIndex: Global Web Index: Q4 2012 (2013). http://www.thesocialclinic.com/the-state-of-social-media-in-saudi-arabia-2012-2
24. González-Bailón, S., Borge-Holthoefer, J., Rivero, A., Moreno, Y.: The dynamics of protest recruitment through an online network. Sci. Rep. (2011). doi:10.1038/srep00197
25. Grier, C., Thomas, K., Paxson, V., Zhang, M.: @spam: The underground on 140 characters or less. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS '10, pp. 27–37. ACM, New York, NY, USA (2010). doi:10.1145/1866307.1866311
26. Hua, W., Zhang, Y.: Threshold and associative based classification for social spam profile detection on twitter. In: 2013 Ninth International Conference on Semantics, Knowledge and Grids (SKG), pp. 113–120. IEEE (2013)
27. Kato, S., Koide, A., Fushimi, T., Saito, K., Motoda, H.: Network analysis of three twitter functions: favorite, follow and mention. In: Richards, D., Kang, B. (eds.) Knowledge Management and Acquisition for Intelligent Systems, Lecture Notes in Computer Science, vol. 7457, pp. 298–312. Springer, Berlin (2012)
28. Lumezanu, C., Feamster, N., Klein, H.: bias: Measuring the tweeting behavior of propagandists. In: ICWSM (2012)
29. Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., Miller, R.C.: Twitinfo: aggregating and visualizing microblogs for event exploration. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, pp. 227–236. ACM, New York, NY, USA (2011). doi:10.1145/1978942.1978975
30. McCord, M., Chuah, M.: Spam detection on twitter using traditional classifiers. In: Proceedings of the 8th International Conference on Autonomic and Trusted Computing, ATC'11, pp. 175–186. Springer, Berlin (2011). http://dl.acm.org/citation.cfm?id=2035700.2035717
31. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. In: Proceedings of ICWSM (2013). http://www.public.asu.edu/~fmorstat/paperpdfs/icwsm2013
32. Newman, M.E.J.: Power laws, pareto distributions and Zipf's law. Contemp. Phys. **46**, 323–351 (2005)
33. Nguyen, H.: 2013 State of Social Media Spam. Technical report, Nexgate (2013)
34. Passerini, A., Pontil, M., Frasconi, P.: New results on error correcting output codes of kernel machines. IEEE Trans. Neural Netw. **15**(1), 45–54 (2004). doi:10.1109/TNN.2003.820841
35. Phelps, A.: OpenFuego: Nieman Journalism Lab. http://niemanlab.github.io/openfuego/ (2013). Accessed 16 Sept 2014 (online)
36. Sanzgiri, A., Hughes, A., Upadhyaya, S.: Analysis of malware propagation in twitter. In: 2013 IEEE 32nd International Symposium on Reliable Distributed Systems (SRDS), pp. 195–204. IEEE (2013)
37. Sharma, P., Biswas, S.: Identifying spam in twitter trending topics. In: American Association for Artificial Intelligence (2011)
38. Snitzer, B.: EarthQuakes Bot. http://eqbot.com/ (2009). Accessed 16 Sept 2014 (online)
39. Song, J., Lee, S., Kim, J.: Spam filtering in twitter using sender-receiver relationship. In: Recent Advances in Intrusion Detection, pp. 301–317. Springer (2011)
40. Stein, T., Chen, E., Mangla, K.: Facebook immune system. In: Proceedings of the 4th Workshop on Social Network Systems, p. 8. ACM (2011)
41. Stringhini, G., Egele, M., Kruegel, C., Vigna, G.: Poultry markets: on the underground economy of twitter followers. In: Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks, WOSN '12, pp. 1–6. ACM, New York, NY, USA (2012). doi:10.1145/2342549.2342551
42. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, p. 1. ACM Press, New York, New York, USA (2010). doi:10.1145/1920261.1920263. http://portal.acm.org/citation.cfm?doid=1920261.1920263
43. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment in Twitter events. J. Am. Soc. Inf. Sci. Technol. **62**(2), 406–418 (2011). doi:10.1002/asi.21462
44. Thomas, K., Grier, C., Paxson, V.: Adapting social spam infrastructure for political censorship. In: Proceedings of the 5th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET) (2012). https://www.usenix.org/conference/leet12/adapting-social-spam-infrastructure-political-censorship
45. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect: an analysis of twitter spam. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, IMC '11, pp. 243–258. ACM, New York, NY, USA (2011). doi:10.1145/2068816.2068840
46. Thomas, K., McCoy, D., Grier, C., Kolcz, A., Paxson, V.: Trafficking fraudulent accounts: the role of the underground market in twitter spam and abuse. In: Proceedings of the 22nd Usenix Security Symposium (2013)
47. Twitter: Following Rules and Best Practices. https://support.twitter.com/articles/68916-following-rules-and-best-practices (2012). Accessed 22 Oct 2013 (online)
48. Twitter: Public Stream. https://dev.twitter.com/docs/streaming-apis/ (2012). Accessed 1 Oct 2013 (online)
49. Twitter: Rules. https://support.twitter.com/articles/18311-the-twitter-rules (2012) Accessed 1 Oct 2013 (online)
50. Twitter: Initial Public Offering of Shares of Common Stock of Twitter, Inc (2013). Accessed 5 Oct 2013 (online)
51. Wagner, C., Mitter, S., Körner, C., Strohmaier, M.: When social bots attack: modeling susceptibility of users in online social networks. Making Sense of Microposts (# MSM2012) p. 2 (2012)
52. Wald, R., Khoshgoftaar, T.M., Napolitano, A., Sumner, C.: Predicting susceptibility to social bots on twitter. In: 2013 IEEE 14th International Conference on Information Reuse and Integration (IRI), pp. 6–13. IEEE (2013)

53. Wang, A.H.: Don't follow me: spam detection in Twitter. In: Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT), pp. 1–10 (2010)
54. Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., Zhao, B.Y.: You are how you click: clickstream analysis for sybil detection. In: USENIX Security, pp. 241–256 (2013)
55. Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M.J., Zheng, H., Zhao, B.Y.: Social turing tests: crowdsourcing sybil detection. In: NDSS. The Internet Society (2013)
56. Xie, Y., Yu, F., Ke, Q., Abadi, M., Gillum, E., Vitaldevaria, K., Walter, J., Huang, J., Mao, Z.M.: Innocent by association: early recognition of legitimate users. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS '12, pp. 353–364. ACM, New York, NY, USA (2012). doi:10.1145/2382196.2382235
57. Yang, C., Harkreader, R., Gu, G.: Die free or live hard? Empirical evaluation and new design for fighting evolving twitter spammers. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) Recent Advances in Intrusion Detection, Lecture Notes in Computer Science, vol. 6961, pp. 318–337. Springer, Berlin (2011)
58. Yang, C., Harkreader, R., Zhang, J., Shin, S., Gu, G.: Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on Twitter. In: Proceedings of the 21st International Conference on World Wide Web, WWW '12, pp. 71–80. ACM, New York, NY, USA (2012). doi:10.1145/2187836.2187847
59. Zhang, C.M., Paxson, V.: Detecting and analyzing automated activity on Twitter. In: Proceedings of the 12th International Conference on Passive and Active Measurement, PAM'11, pp. 102–111. Springer, Berlin (2011). http://dl.acm.org/citation.cfm?id=1987510.1987521