10-2016

On profiling bots in social media

Richard Jayadi OENTARYO

Arinto MURDOPO

Philips Kokoh PRASETYO

Ee-peng LIM
Singapore Management University, eplim@smu.edu.sg

On Profiling Bots in Social Media

Richard J. Oentaryo(✉), Arinto Murdopo, Philips K. Prasetyo, and Ee-Peng Lim(✉)

Living Analytics Research Centre, Singapore Management University, Singapore, Singapore
{roentaryo,arintom,pprasetyo,eplim}@smu.edu.sg

Abstract. The popularity of social media platforms such as Twitter has led to the proliferation of automated bots, creating both opportunities and challenges in information dissemination, user engagements, and quality of services. Past works on profiling bots had been focused largely on malicious bots, with the assumption that these bots should be removed. In this work, however, we find many bots that are benign, and propose a new, broader categorization of bots based on their behaviors. This includes *broadcast*, *consumption*, and *spam* bots. To facilitate comprehensive analyses of bots and how they compare to human accounts, we develop a systematic profiling framework that includes a rich set of features and classifier bank. We conduct extensive experiments to evaluate the performances of different classifiers under varying time windows, identify the key features of bots, and infer about bots in a larger Twitter population. Our analysis encompasses more than 159K bot and human (non-bot) accounts in Twitter. The results provide interesting insights on the behavioral traits of both benign and malicious bots.

Keywords: Bot profiling · Classification · Feature extraction · Social media

1 Introduction

In recent years, we have seen a dramatic growth of people's activities taking place in social media. Twitter, for example, has evolved from a personal microblogging site to a news and information dissemination platform. The openness of the Twitter platform, however, has made it easy for a user to set up an automated social program called *bot*, to post tweets on his/her behalf.

The proliferation of bots has both good and bad consequences [4,8]. On the one hand, bots can generate benign, informative tweets (e.g., news and blog updates), which enhance information dissemination. Bots can also be helpful for the account owners, e.g., bots that aggregate contents from various sources based on the owners' interests. On the other hand, spammers may exploit bots to attract regular accounts as their followers, enabling them to hijack search engine results or trending topics, disseminate unsolicited messages, and entice users to visit malicious sites [8,10,11]. In addition to deteriorating user experience and

E. Spiro and Y.-Y. Ahn (Eds.): SocInfo 2016, Part I, LNCS 10046, pp. 92–109, 2016.
DOI: 10.1007/978-3-319-47880-7_6

trust, malicious bots may cause more severe impacts, e.g., creating panic during emergencies, biasing political views, or damaging corporate reputation [8,21].

It is thus important to characterize different types of bots and understand how they compare with human users. Recent studies have shown the importance of profiling bots in social media [1,2,4,8,10–13,17,18,20,21], but these works have focused mainly on malicious (e.g., spam) bots, failing to account for other types of benign bots. With the rise of new services and intelligent apps in Twitter, benign bots are increasingly becoming prominent as well.

Comprehensive profiling of both malicious and benign bots would offer several major benefits. In information dissemination and retrieval, knowing the activity traits of both bot types and the nature of their tweet contents can improve search and recommendation services by separating tweets of bots from those of humans, returning more relevant, personalized search results, and promoting certain products/services more effectively. For social science research, a more accurate understanding of human interactions and information diffusion patterns [8,9] can also be obtained by filtering out activity biases generated by bots. In turn, these would benefit the overall user community as well.

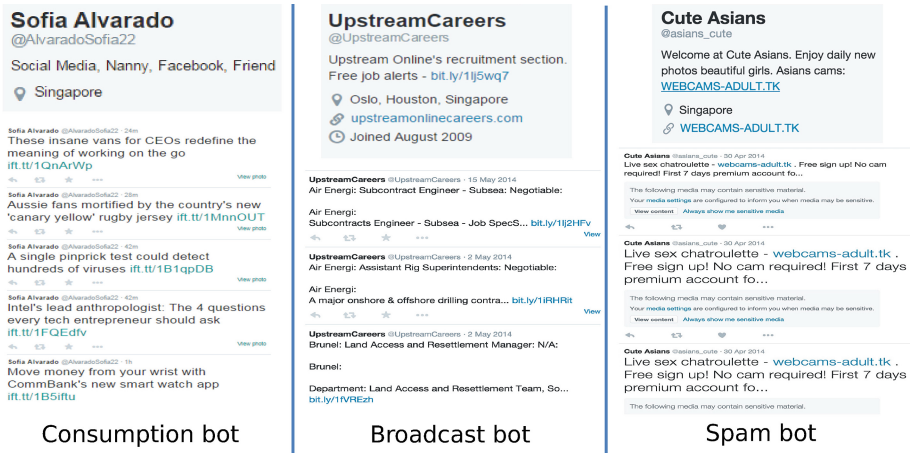


Fig. 1. Examples of broadcast, consumption and spam bots in Twitter

To illustrate the usefulness of profiling bots, consider the examples in Fig. 1, of different types of benign and malicious bots (which we further describe in Sect. 3). The first example is a user who utilizes the IFTTT service[1] to gather contents from diverse sources for her own consumption. Knowing that she uses a consumption bot, Twitter can provide a new service to organize the unstructured contents, or recommend new contents that match her interest. The second example involves a broadcast bot managed by a job agency to advertise job openings.

[1] https://ifttt.com.

Twitter recently introduced a new feature called *promoted tweets*[2] and, knowing it is a (benign) broadcast bot, Twitter can recommend the feature to help the agency reach a wider audience. The last example shows a malicious, spam bot that lures users to visit adult websites, posssibly containing harmful malware. For such a bot, Twitter may develop a strategy to demote—or even filter out—its posts, so that the followers do not see them on their tweet streams.

Contributions. In this paper, we present a new categorization of bots based on long-term observations on the behaviors of various automated accounts in Twitter. To our best knowledge, this work is the first extensive study on both *benign* and *malicious* Twitter bots, with detailed analyses on both their static and dynamic patterns of activity. In recent years, Twitter bots have evolved rapidly, and so our work also provides a more timely study that offers updated insights on the bot characteristics. Our findings should also benefit social science and network mining researches. We summarize our key contributions below:

- We propose a new categorization of Twitter bots based on their behavioral traits. In contrast to past studies that focus largely on malicious bots, our study encompasses more detailed examinations of both malicious and benign bots, as well as how they compare to human accounts. For this, we have studied a large dataset of more than 159K Twitter accounts, out of which we have manually labeled 1.6K bot and human accounts.
- To facilitate comprehensive analyses on bots, we develop a systematic profiling framework that includes a rich set of numeric, categorical, and series features. This enables us to examine both the static and dynamic patterns of bots, which span various user profile, tweet, and follow network entities. Our framework also features a classifier bank that includes prominent classification algorithms, thus allowing us to comprehensively evaluate various algorithms so as to identify the best approach for bot profiling.
- We carry out extensive empirical studies to evaluate the performance of our classifiers under different time windows and to identify the most relevant, discriminating features that characterize both benign and malicious bots. We also conduct a novel study to assess the generalization ability of our method on unseen, unlabeled Twitter accounts, based on which we infer the behavioral traits of bots in a larger Twitter population.

2 Background and Related Work

A number of studies have been conducted to identify and profile bots in social media. To detect spam bots, Wang [21] utilized content- and graph-based features, derived from the tweet posts and follow network connectivity respectively. Chu *et al.* [4] investigated whether a Twitter account is a human, bot, or cyborg. Here a bot was defined as an aggresive or spammy automated account, while cyborg refers to a bot-assisted human or human-assisted bot. Different from our

[2] https://business.twitter.com/solutions/promoted-tweets.

work, the bots defined in [4] are more of malicious nature, and the study did not provide further categorization/analysis of benign and malicious bots in Twitter.

To investigate on spam bots, Stringhini *et al.* [17] created honey profiles on Facebook, Twitter and MySpace. By analyzing the collected data, they identified anomalous accounts who contacted the honey profiles and devised features for detecting spam bots. Going further, Lee *et al.* [13] conducted a 7-month study on Twitter by creating 60 social honeypots that try to lure "content polluters" (a.k.a. spam bots). Users who follow or message two or more honeypot accounts are automatically assumed to be content polluters. There are also related works on spam bot detection based on social proximity [10] or both social and content proximities [11]. Tavares and Faisal [19] distinguished between personal, managed, and bot accounts in Twitter, according to their tweet time intervals.

Ferrara *et al.* [8] built a web application to test if a Twitter account behaves like a bot or human. They used the list of bots and human accounts identified by [13], and collected their tweets and follow network information. This study, however, covers only malicious bots. Dickerson *et al.* [5] used network, linguistic, and application-oriented features to distinguish between bots and humans in the 2014 Indian election. Abokhodair *et al.* [1] studied on a network of bots that collectively tweet about the 2012 Syrian civil war. This study covers both malicious (e.g., phishing) and benign (e.g., testimonial) bots. In contrast to our work, however, their findings are tailored to a specific event (i.e., the civil war) and may not be applicable to other bot types in a larger Twitter population.

There are also studies aiming to quantify the susceptibility of social media users to the influence of bots [2,12,20]. By embedding their bots into the Facebook network, Boshmaf *et al.* [2] demonstrated that users are vulnerable to phishing (e.g., exposing their phone number or address). The susceptibility of users is also evident in Twitter [12,20]. Freitas *et al.* [9] tried to reverse-engineer the infiltration strategies of malicious Twitter bots in order to understand their functioning. Most recently, Subrahmanian *et al.* [18] reported the winning solutions of the DARPA Twitter Bot Detection Challenge. Again, however, all these studies deal mainly with malicious bots and ignore benign bots.

3 New Categorization of Bots

We define a bot as a Twitter account that generates contents and interacts with other users automatically—at least according to human judgment. Our definition thus includes *both* benign and malicious bots. Based on long-term observations on Twitter data, we propose to categorize Twitter bots into three main types:

- **Broadcast bot**. This bot aims at disseminating information to general audience by providing, e.g., benign links to news, blogs or sites. Such bot is often managed by an organization or a group of people (e.g., bloggers).
- **Consumption bot**. The main purpose of this bot is to aggregate contents from various sources and/or provide update services (e.g., horoscope reading, weather update) for personal consumption or use.

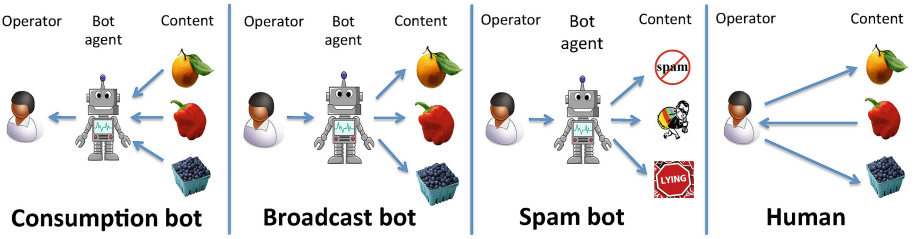


Fig. 2. Bot and human accounts in Twitter

– **Spam bot**. This type of bots posts malicious contents (e.g., to trick people by hijacking certain account or redirecting them to malicious sites), or promotes harmless but invalid/irrelevant contents aggressively.

Figure 2 illustrates the three bot types, where the arrow direction represents the flow of information. It is worth noting that our proposed categorization is more general than the taxonomy put forward in [15], which covers mainly malicious bots. Our categorization is also general enough to cater for new, emerging types of bot (e.g., chatbots can be viewed as a special type of broadcast bots).

4 Dataset

Data collection. Our study involves a Twitter dataset generated by users in Singapore and collected from 1 January to 30 April 2014 via the Twitter REST and streaming APIs[3]. Starting from popular seed users (i.e., users having many followers), we crawled their follow, retweet, and user mention links. We then added those followers/followees, retweet sources, and mentioned users who state Singapore in their profile location. With this, we have a total of 159,724 accounts.

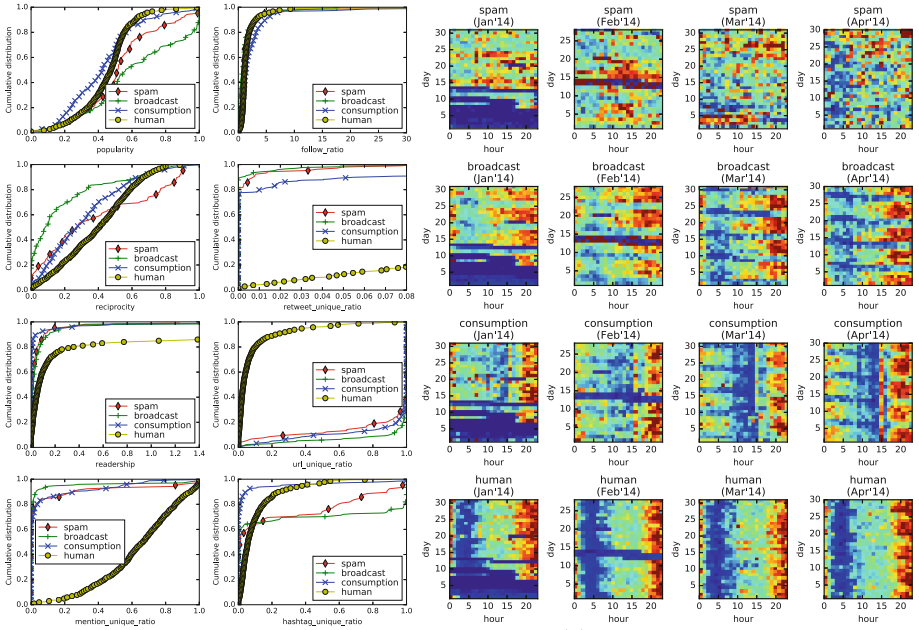
Table 1. Distribution of our Twitter dataset

Labeled data				Unlabeled data
Consumption bot	Broadcast bot	Spam bot	Human account	
313	171	105	1,024	158,111

Total no. of labeled data = 1,613; Total no. of data = 159,724

To identify bots, we first checked active accounts who tweeted at least 15 times within the month of April 2014. We then manually labeled these accounts and found 589 bots. As many more human users are expected in the Twitter population, we randomly sampled the remaining accounts, manually checked them, and identified 1,024 human accounts. In total, we have 1,613 labeled accounts,

[3] https://dev.twitter.com/overview/.



(a) Cumulative distribution functions (b) Temporal dynamics

Fig. 3. Statistics of humans and bots in our labeled Twitter data

as summarized in Table 1. The labeling was done by four volunteers, who were carefully instructed on the definitions in Sect. 3. The volunteers agree on more than 90 % of the labels, and any labeling differences in the remaining accounts are resolved by consensus. Also, if an account exhibits both human and bot characteristics, we determine the label based on the majority posting patterns.

Exploratory analysis. We conducted a preliminary study on our 1,613 labeled data to get a glimpse of the activity patterns of bots as well as human accounts. Figure 3(a) shows the cumulative distribution functions (CDF) of several key attributes. An early increase in CDF value means a more skewed distribution. We focus on key attributes that reflect a user's social and posting patterns: $popularity = \frac{|F|}{|E|+|F|}$, $follow\_ratio = \frac{|E|}{|F|}$, $reciprocity = \frac{|E \cap F|}{|E \cup F|}$, $retweet\_unique\_ratio = \frac{|R|}{|T|}$, $url\_unique\_ratio = \frac{|U|}{|T|}$, $mention\_unique\_ratio = \frac{|M|}{|T|}$, $hashtag\_unique\_ratio = \frac{|H|}{|T|}$, where E, F, R, T, U, M, H are the set of followees, followers, retweets, tweets, URLs, user mentions, and hashtags for a given account, respectively. We also define $readership = \frac{retweeted}{|T|}$, where $retweeted$ is the number of times a user's tweets get retweeted (by others). Figure 3(b) shows heatmaps of tweet counts $|T|$ for different days and hours over 4 months.

How do humans compare with bots and how do bots differ from one another? The *popularity*, *follow_ratio*, and *reciprocity* results in Fig. 3(a) suggest that

bots (except for consumption bots) generally have more followers than followees, but are less reciprocal (i.e., follow each other) than humans. Based on the *retweet_unique_ratio* and *readership* results, humans are more likely to reshare contents from others and have their contents reshared than bots, respectively. Similarly, the *mention_unique_ratio* result suggests that humans are more likely to mention (i.e., talk to) others than bots. Meanwhile, the *url_unique_ratio* and *hashtag_unique_ratio* results show the bots tend to include more diverse web links and topics than humans, respectively. Finally, comparisons among the three bot types show that broadcast bots are the most popular and post the most diverse URLs and hashtags, but they are the least reciprocal and rarely mention others. A plausible reason is that broadcast bots are typically used by organizations solely for information dissemination, and not for interaction with others.

How do activities of humans and bots change over time? Figure 3(b) shows that seasonality exists in the tweet activities of human and bot accounts[4]. That is, humans seldom tweet in early morning (from 2am to 7am) and post moderately from 7am to 8pm. Afterwards, their tweet traffic increases significantly between 8pm and midnight, suggesting that Singapore users are more active after dinner time and before they sleep. Meanwhile, consumption bots tweet more actively than humans from 3am to 7am (i.e., sleep hours), but are less active from 9am to 3pm (i.e., busy working/school hours). Also, consumption bots are less active in the weekends than in the weekdays. While broadcast bots have generally similar patterns to consumption bots, the former is less active during sleep hours (3am–7am) whereas the latter during busy hours (9am–3pm). We can attribute this to the intuition that broadcast bots aim to reach a wider audience during their non-sleep hours. Lastly, unlike broadcast and consumption bots, spam bots are active all days/hours, and they exhibit very random timings. In summary, different bots serve different purposes and their temporal signatures reflect these.

5 Profiling Framework

We develop a systematic profiling framework to facilitate comprehensive analyses of bots. Below we describe each component of the framework in turn.

Database. Our framework takes as input three types of database: *profile*, *tweet*, and *follow* databases. The profile database contains user information such as the Twitter user id, screenname, location, and profile description. The tweet database contains all the tweets posted by different users, which may include various entities such as hashtags, URLs, user mentions, videos/images, retweet information, and tweet sources/devices. We collectively refer to these as *tweet entities*. Finally, the follow database contains the snapshots of users' relationship network over time, which include both followers and followees of the users at different time periods. We collectively call these *follow entities*.

[4] The exceptionally low tweet frequencies in the first week of January and 12-14 February are due to major downtime of our servers.

Feature extraction. This component serves to construct a *feature vector* that represents a Twitter account. It takes three types of feature: *numeric*, *categorical*, and *series*. We describe the extraction steps for each type below:

- For **numeric features**, we perform *standarization* by scaling each feature to a unit range $[0, 1]$. This would allow us to mitigate feature scaling issues, particularly for classification methods that rely on some distance metric. Examples of numeric features are count and ratio attributes (see Table 2).
- For **categorical features**, we first select the top K categories based on their frequencies in each data point, and then filter out the remaining categories. Next, we perform *one-hot encoding* by transforming the top K categories into a binary vector with K elements. For example, a categorical attribute with four possible values: "A", "B", "C", and "D" is encoded as $[1, 0, 0, 0]$, $[0, 1, 0, 0]$, $[0, 0, 1, 0]$, and $[0, 0, 0, 1]$, respectively.
- For **series features**, we first count the frequency of every (discrete) number in the series. For instance, given a series $[a, a, b, a, c, b, c, a, b]$, we can compute the histogram bins: $(a, 4), (b, 3), (c, 2)$. To ensure a moderate feature size, we keep only top 100 bins with the highest count frequencies. Subsequently, we normalize the frequencies such that they sum to 1, thus forming a probability distribution. For the previous histogram bins $(a, 4)$, $(b, 3)$, $(c, 2)$, the normalization will result in $(a, \frac{4}{9})$, $(b, \frac{3}{9})$, $(c, \frac{2}{9})$.

Classifier bank. Finaly, to learn the association between the extracted features and different bot types (or human), our framework includes a classifier bank that comprises a rich collection of classification algorithms. In our study, we employ four prominent classifiers: *naïve Bayes* (NB) [6], *random forest* (RF) [3], and two instances of generalized linear model, i.e., *support vector machine* (SVM) and *logistic regression* (LR) [7]. These algorithms represent the state-of-the-art methods previously used for (malicious) bot classification. For instance, RF was utilized in [4,5,8,13], while SVM and NB were used in [5,21].

6 Feature Engineering

We have crafted a rich set of features based on the feature extraction component in our bot profiling framework. Our feature set consists of three groups: *tweet*, *follow* and *profile* features. For tweet features, we also distinguish between *static* (i.e., time-independent) and *dynamic* (i.e., time-dependent) tweet features. Table 2 provides a listing of all the features used in our empirical study.

Static tweet features. We generate static tweet features based on the combination of entities and statistical metrics, as shown in Table 2. For instance, to generate the hashtag features of a user, we treat each hashtag as a "bag" and count how many times the word occurs in all of x's tweets. This yields a bag-of-hashtag vector, from which we can compute first-order statistics (i.e., *count*, *unique_count*, *mean*, *median*, *min*, and *max*) as well as second-order metrics (i.e., standard deviation (*std*) and Shannon entropy [16] (*entropy*)). We note

Table 2. List of features used in our bot classification task

Group	Entity	Features
Static tweet features	Tweet_word	Count (N), unique_count (N), unique_ratio (N), basic_stats (N)
	Retweet	Retweeted (N), readership (N), count (N), unique_count (N), ratio (N), unique_ratio (N), basic_stats (N)
	Hashtag	Count (N), unique_count (N), ratio (N), unique_ratio (N), basic_stats (N)
	Mention	Count (N), unique_count (N), ratio (N), unique_ratio (N), basic_stats (N)
	Url	Count (N), unique_count (N), ratio (N), unique_ratio (N), basic_stats (N)
	Media	Count (N), unique_count (N), ratio (N), unique_ratio (N), basic_stats (N)
	Source	Sources (S)
Dynamic tweet features	Tweet	Hours (S), days (S), weekdays (S), timeofdays (S), extended_stats (N)
	Retweet	Hours (S), days (S), weekdays (S), timeofdays (S), extended_stats (N)
	Hashtag	Hours (S), days (S), weekdays (S), timeofdays (S), extended_stats (N)
	Mention	Hours (S), days (S), weekdays (S), timeofdays (S), extended_stats (N)
	Url	Hours (S), days (S), weekdays (S), timeofdays (S), extended_stats (N)
	Media	Hours (S), days (S), weekdays (S), timeofdays (S), extended_stats (N)
Follow features	Followees_count	basic_stats (N)
	Followers_count	Basic_stats (N)
	Mutual_count	Basic_stats (N)
	Reciprocity	Basic_stats (N)
	In_reciprocity	Basic_stats (N)
	Out_reciprocity	Basic_stats (N)
	Popularity	Basic_stats (N)
	Follow_ratio	Basic_stats (N)
Profile features	Profile	Is_geo_enabled (C), lang (C), time_zone (C), account_age (N), favourites_count (N), listed_count (N), statuses_count (N), utc_offset (N)

Basic_stats: set of statistical metrics {mean, median, min, max, std, entropy}
Extended_stats: Cartesian product of {timegap, hour, day, weekday, timeofday} and basic_stats
N: numeric feature, C: categorical feature, S: series feature

that the second-order metrics serve to quantify the *diversity* of the entities. We also compute the $ratio = \frac{count}{|T|}$ and $unique\_ratio = \frac{unique\_count}{|T|}$, where $|T|$ is the total number of tweets posted by a user. For the retweet entity, we additionally consider *retweeted* and *readership* features, as described in Sect. 4. Finally, we consider a series feature to represent the source entity, whereby each source maps to a histogram bin containing the normalized frequency of the source.

Dynamic tweet features. For these features (cf. Table 2), we introduce additional time dimensions that capture the dynamics of tweet activities, namely: $hours \in \{0, \ldots, 23\}$, $days \in \{1, \ldots, 31\}$, $weekdays \in \{Monday, \ldots, Sunday\}$, $timeofdays \in \{morning$ (4am–12pm), $afternoon$ (12pm–5pm), $evening$ (5pm–8pm), $night$ (8pm–4am)$\}$, and *timegaps*. The timegap dimension refers to the gap (in milliseconds) between two *consecutive* entity timestamps, e.g., for N tweets

posted by a user x, we can compute a timegap vector with length $(N-1)$. For each time dimension, we can then generate the series features based on the histogram binning described in Sect. 5, as well as compute the statistical metrics such as *mean*, *median*, *min*, *max*, *std* and *entropy*.

Follow features. These features are derived by computing metrics that summarize snapshots of the follow network at different time points (cf. Table 2). Let E and F be the set of followees and followers of a given user. In turn, we compute the $followees\_count = |E|$, $followers\_count = |F|$, $mutual\_count = |E \cap F|$. as well as ratio metrics such as $reciprocity = \frac{|E \cap F|}{|E \cup F|}$, $in\_reciprocity = \frac{|E \cap F|}{|F|}$, $out\_reciprocity = \frac{|E \cap F|}{|E|}$, $popularity = \frac{|F|}{|E|+|F|}$, and $follow\_ratio = \frac{|E|}{|F|}$. We calculate these metrics for every snapshot of the follow network at a given time point, and then compute the statistics *mean*, *median*, *min*, *max*, *std* and *entropy* to summarize the metrics over all time points.

Profile features. Finally, we also consider several basic user profile features, as per Table 2. Here, *account_age* refers to the lapse between the time a user first joined Twitter and the current reference time. Further details on the definitions of the other profile features can be found in https://dev.twitter.com/.

7 Results and Findings

This section elaborates our empirical study on bots. We first describe our experiment setup, and then address several research questions in Sects. 7.1–7.3.

Evaluation metrics. To evaluate our classifiers, we utilize three metrics popularly used in information retrieval [14]: *Precision*, *Recall* and $F1$. We report, for each class $c \in \{broadcast, consumption, spam, human\}$, the $Precision(c) = \frac{TP(c)}{TP(c)+FP(c)}$, $Recall = \frac{TP(c)}{TP(c)+FN(c)}$, and $F1(\mathrm{c}) = \frac{2Precision(c)Recall(c)}{Precision(c)+Recall(c)}$, where $TP(c)$, $FP(c)$ and $FN(c)$ are the true positives, false positives, and false negatives respectively. Based on these, we also report the macro-averaged $Precision = \frac{1}{4}\sum_{c=1}^{4} Precision(c)$, $Recall = \frac{1}{4}\sum_{c=1}^{4} Recall(c)$, and $F1 = \frac{1}{4}\sum_{c=1}^{4} F1(\mathrm{c})$.

Experiment protocols. In this work, we consider two sets of experiment:

- **Experiment** E_1: This set of experiment involves evaluation on our **1,613** *labeled data* (see Table 1). For this evaluation, we use a *stratified* 10-fold cross-validation (CV), whereby we split the labeled data into 10 mutually exclusive groups, each retaining the class proportion as per the original data. This stratification serves to ensure that each fold is a good representative of the whole, i.e., it retains the (unbalanced) class distribution as in the original data. For each CV iteration f, we then use group f (10 %) for testing and the remaining groups $f' \neq f$ (90 %) for training. We report the results averaged over 10 iterations, which include $Precision(c)$, $Recall(c)$ and $F1(c)$ for each class c, as well as the macro-averaged *Precision*, *Recall* and $F1$.

– **Experiment** E_2: This set of experiment serves to evaluate predictions on the remaining **158,111** *unlabeled data* (see again Table 1). Based on this, we can infer the behavioral traits of bots in a larger Twitter population. For this experiment, we are unable to compute *Recall*, as we would have to manually verify one by one a large number of unlabeled data. Instead, we evaluate based on *Precision* at top K for each class ($K \ll 158,111$).

Model parameters. We configured our classifier bank as follows: For the NB classifier, we use the smoothing parameter $\alpha = 1$. For RF, we use $N = 100$ decision trees. Finally, for SVM and LR, we set the cost parameter $C = 1$ and `class_weight` = "balanced"; the latter is for automatically handling the imbalanced class distribution. We performed grid search to determine all these parameters, which give the optimal performances for each classifier. In particular, we varied the NB parameter from the range $\alpha \in \{0.1, 1, 10\}$. For RF, we tried $N \in \{10, 20, \ldots, 100\}$, and for SVM and LR, we tried $C \in \{0.01, 0.1, 1, 10, 100\}$.

Significance test. Finally, we use *Wilcoxon signed-rank test* [22] to test for the statistical significance of our results. When comparing between two performance vectors, we look at the p-value at a significance level of 0.01. If the p-value is less than 0.01, we say that the performance difference is indeed significant.

7.1 How Well Can the Classifiers Predict for Bots?

To answer this research question, we first conduct a sensitivity study by varying the time duration for which features (cf. Table 2) are generated. For this study, we use the CV procedure on our labeled data (i.e., Experiment E_1), whereby the classifiers were trained using all features listed in Table 2. Figure 4 shows the macro-averaged *Precision*, *Recall*, and $F1$ over 10 CV folds, with the duration varied from 1 week, 2 weeks and 1 month to 2 months and 4 months (up to 30 April 2014). Based on the $F1$ results, we can conclude that 2 weeks is the best duration and that LR outperforms the other classifiers. In this case, RF gives higher *Precision* than LR, but its *Recall* is much lower, and so is its $F1$. It is also shown that a tradeoff exists in choosing the duration; an overly short duration degrades the performance, which can be attributed to data scarcity. The same goes for an overly long duration, due to inclusion of outdated data.

Table 3 shows further breakdown of the CV results for the best time duration (i.e., 2 weeks). Overall, LR and SVM give the best results, and outperform the more complex RF and simpler NB methods (except for *Precision* of the "spam" class). For spam bots, RF yields higher *Precision*, but much lower *Recall* and $F1$ than LR and SVM. While SVM and LR perform very similarly, we decided to use LR as our main classifier for two reasons: (i) LR outputs more meaningful probabilitic scores than the unbounded decision scores in SVM; and (ii) LR is more robust than SVM against variation in time duration, as we saw in Fig. 4.

Based on the individual $Precision(c)$, $Recall(c)$ and $F1(c)$ of each class c, we can conclude that, among the bots, consumption bots are the easiest to detect, followed by broadcast and spam bots. This is expected, owing to the imbalanced

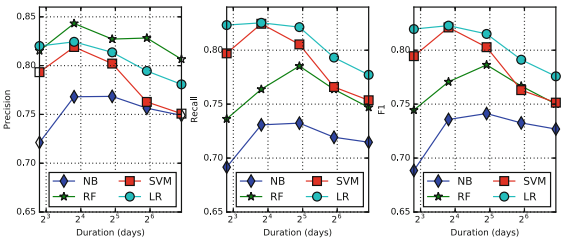


Fig. 4. Classification results for varying durations

Table 3. Breakdown of 10-fold cross-validation results using 2-week training data

Metric	Method	Class label				Macro average
		Broadcast	Consumption	Spam	Human	
Precision	NB	0.6519 (−)	0.7206 (−)	0.7069 (+)	0.9929	0.7681 (−)
	RF	0.5880 (−)	**0.9462**	**0.8636** (+)	0.9750 (−)	**0.8432** (+)
	SVM	**0.6952**	0.9278	0.6574 (−)	**0.9961**	0.8191
	LR	0.6798	0.9366	0.6869	0.9942	0.8244
Recall	NB	0.6901 (−)	**0.8818** (+)	0.3905 (−)	0.9609(−)	0.7308 (−)
	RF	**0.8596** (+)	0.8435	0.3619 (−)	0.9902	0.7638 (−)
	SVM	0.7602(−)	0.8626	**0.6762** (+)	**0.9990**	0.8245
	LR	0.8070	0.8498	0.6476	0.9971	**0.8254**
F1-score	NB	0.6705 (−)	0.7931 (−)	0.5031 (−)	0.9767 (−)	0.7358 (−)
	RF	0.6983 (−)	0.8919	0.5101 (−)	0.9826 (−)	0.7707 (−)
	SVM	0.7263	**0.8940**	**0.6667**	**0.9976**	0.8211
	LR	**0.7380**	0.8911	**0.6667**	0.9956	**0.8228**

NB: näive Bayes, SVM: support vector machine, LR: logistic regression, RF: random forest
(−): significantly worse than LR at 0.01, (+): significantly better than LR at 0.01

class distribution as per Table 1. We can also compare the results of our classifiers with that of a random guess[5]. Based on the statistics in Table 1, the expected $F1$ scores of a random guess for broadcast bot, consumption bot, spam bot, and human classes are 10.6 %, 19.40 %, 6.51 % and 63.49 %, respectively. Our four classifiers thus outperform the random guess baseline by a large margin.

For spam bots, several studies [4,8,13] have reported high classification accuracies, while our results are modest by comparison, largely due to the lack of spam bot accounts in our data. However, it must be noted that these works focused largely on distinguishing between (malicious) bots vs. other accounts, whereas our study deals with a much more challenging and fine-grained categorization of broadcast, consumption and spam bots. Also, the lack of spam bots in our data can be attributed to several factors, such as our relatively strict definition of spam bot (whereby the majority of its postings need to have malicious

[5] Random guess w.r.t. a class c refers to a classifier that assigns a proportion p_c% of the instances to class c, and $(1-p_c)$% to classes other than c. In this case, $Precision(c) = Recall(c) = F1(c) = p_c$, where $p_c = \frac{P(c)}{P(c)+N(c)} = \frac{TP(c)+FN(c)}{TP(c)+FN(c)+TN(c)+FP(c)}$.

or irrelevant contents), or our data collection process that begins with popular seed users and their connections (thus possibly missing unpopular spam bots). Nevertheless, our main focus is to analyze benign bots, which has been largely ignored in the past studies. Further studies on less prominent spam bots that post malicious contents at a sparse rate is beyond the scope of our current study.

7.2 Which Features Are the Most Indicative of Each Bot Type?

In light of this research question, we trained our best classifier (i.e., LR) using all 1,613 labeled data, and look at the weight coefficients $w_{i,c}$ of each class in the trained LR. Here we use the raw weights $w_{i,c}$ instead of the absolute values $|w_{i,c}|$ or squared values $w_{i,c}^2$, as the raw weights allow us to distingush between features that correlate positively with a class label (which are our main interest) and those that correlate negatively. Figure 5 shows the top 15 positively-correlated features for each class. In general, we find that the top features are dominated by the *source* (i.e., where the tweets come from) and *entropy-based dynamic tweet* features. Below we elaborate our feature analysis for each class further.

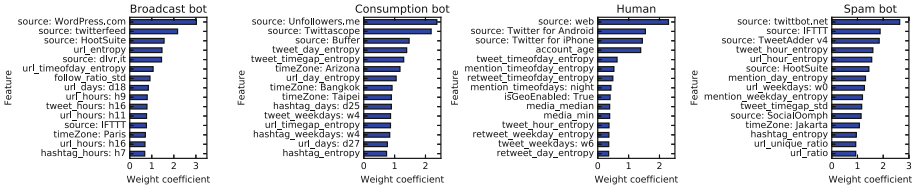


Fig. 5. Top discriminative features for each label in bot classification task

Broadcast bots. Among the top features for broadcast bots, certain sources that are popularly used for blogging (such as WordPress and Twitterfeed) or brand management (such as HootSuite) are found to be highly indicative. It is also shown that the entropy-based features for the url entity correlate strongly with broadcast bots. Recall from Sect. 6 that entropy is a second-order metric that quantifies how diverse a distribution is. Accordingly, as broadcast bots generally aim to disseminate information about certain sites/brands, we can expect that they would have more concentrated url distribution (i.e., low entropy). We will further verify this in Sect. 7.3. Figure 5 also suggests that certain critical timings of the url postings are highly indicative of broadcast bots.

Consumption bots. From Fig. 5, we firstly find that the top three sources for consumption bots (i.e., Unfollowers, Twittascope, and Buffer) are service apps that allow users to track their followers/followees status, horoscope readings, and scheduled postings, respectively. Secondly, we discover that the diversity (entropy) of tweet postings is a strong indicator for consumption bots. Lastly, Fig. 5 shows that certain timezones and timings (weekday and day) of the hashtag

and url activities constitute yet another important set of indicators. All these led us to conclude that consumption bots post tweets in a way that follows certain timings/schedules. We will further analyze this in Sect. 7.3.

Spam bots. The result in Fig. 5 suggests that there are certain sources that can be exploited by spammers to post irrelevant or unsolicited tweets. For example, TwittBot is an application that allows multiple users (and thus spammers) to post to a single Twitter account. In addition, the timing diversities of the url, mention, tweet and hashtag activities are found to be the key signatures of spam bots. As also shown in Fig. 3(b) (of Sect. 4), the temporal patterns of spam bots are highly irregular. Altogether, these suggest that spam bots have highly diverse timings (i.e., high entropy), which we will again verify in Sect. 7.3.

Humans. The top three features in Fig. 5 suggest that human accounts typically use credible sources such as "web" (i.e., Twitter website) and the official Twitter mobile apps. Next, the *account_age* and *isGeoEnabled* features suggest that human accounts have lived relatively long in Twitter and usually have his/her tweets' location enabled, respectively. Also, high timing diversity (entropy) of the tweet, retweet and mention activities are indicative of human accounts, although it is not as high as that of spam bots. Again, Sect. 7.3 analyzes this further. Lastly, the *media_median* and *media_mean* features suggest that human accounts like to attach media files (e.g., photos) in their tweets.

7.3 What Can We Tell About Bots in a Larger Twitter Population?

To address this question, we performed Experiment E_2 by deploying our trained LR classifier to predict for the unlabeled 158,111 accounts. We then picked the top K accounts with the highest probability scores for each class, and manually assessed the class assignments of these accounts. The assessment results can be found in Appendix A (Table 4). We found that the prediction results generally match well with our manual judgments. Based on this, we can make inference on the behavior of bots in a larger Twitter population, i.e., the entire population of Singapore Twitter users. We focus our analyses on the entropy-based dynamic tweet features, which dominate the top features as shown in Fig. 5. That is, we analyze the entropy distributions of the tweet, retweet, mention, hashtag and url activities. The complete distributions can be found in Appendix A (Fig. 6), which reveals several interesting insights as elaborated below.

Tweet patterns. We first compared the distributions of the tweet timings, and discovered that consumption and spam bots exhibit higher diversity (entropy) than that of humans. In contrast, broadcast bots were found to have more concentrated timings. These suggest that broadcast bots post tweets at more specific timings than humans and other types of bots. We also found that consumption and spam bots are very similar in terms of daily timings (i.e., weekday and day entropies), but the former is less diverse than the latter in terms of hourly timings. We can thus conclude that consumption and spam bots tweet equally regularly on a daily basis, but the latter tend to post at random hours.

Retweet and mention patterns. Retweet and mention activities can be used to gauge how much a bot (or human) cares about other accounts. Comparing the distributions of the retweet and mention timings in Fig. 6, we can see again that spam bots have the most random patterns compared to humans and other bot types. But unlike the results for tweet timings, consumption bots have the lowest diversity in terms of daily and hourly timings for the retweet and mention activities. This suggests that consumption bots reshare contents and mention other users at more specific timings, respectively. Such regularity makes sense, especially for consumption bots that provide update services to their users, e.g., Unfollowers and Twittascope (cf. Sect. 7.2).

Hashtag patterns. In Twitter, a hashtag can be viewed as representing a topic of interest. As shown in Fig. 6, humans and consumptions bots have very similar diversities of hashtag timings. It is also shown that spam bots have the most diverse hashtag timings (as expected), whereas broadcast bots exhibit very focused hashtag timings. The latter suggests that broadcast bots tend to talk about different topics at more regular time intervals. This is intuitive, especially if we consider the nature of the account owners of broadcast bots (e.g., news/blogger sites), which aim to disseminate various information on a regular basis.

URL patterns. For the URL timings, we find that in general humans and broadcast bots use URLs at more specific timings than consumption and spam bots. Interestingly, however, we observe that consumption bots exhibit higher diversity in daily timings than spam bots, but the reverse is true for hourly timings. This suggests that consumption bots use URLs on a more regular daily basis than spam bots, but the latter post URLs at more random hours.

Comparisons. It is also interesting to see how our results in Figs. 5 and 6 put little emphasis on the importance of the follow network features in the classification task. This is different from previous studies on (malicious) bots [4,5,13,17,20], whereby the follow features play a key role. We can attribute this to the evolution of bot activities as well as stricter regulations imposed by Twitter (especially for spam bots). Also, to our best knowledge, no attempt has been made in the previous works to infer on a larger population. Thus, our work offers more comprehensive insights on the behavioral traits of bots.

8 Conclusion

In this paper, we present a new categorization of bots, and develop a systematic bot profiling framework with a rich set of features and classification methods. We have carried out extensive empirical studies to analyze on broadcast, consumption and spam bots, as well as how they compare with regular human accounts. We discovered that the diversities of timing patterns for posting activities (i.e., tweet, retweet, mention, hashtag and url) constitute the key features to effectively identify the behavioral traits of different bot types.

This study hopefully will benefit social science studies and help create better user services. In the future, we plan to examine the prevalence of our findings across multiple countries, beyond our current Singapore data. We also wish to study information diffusion and user interaction in Twitter with the aid of bots.

Acknowledgments. This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative.

A Predictions on Unlabeled Twitter Accounts

To facilitate our study on a larger Twitter population, we first examined how well our best classfier (i.e., LR) can predict for unlabeled data that it never sees in the (labeled) CV data. Table 4 summarizes the top K prediction results, whereby we varied K from 10 to 50 to verify the robustness of the predictions. For each class, we computed the number of correctly predicted instances (TP) as well as precision at top K, i.e., $Precision = \frac{TP}{K}$.

Table 4. Top K predictions on unlabeled 158,111 Twitter accounts

Label	$K = 10$		$K = 20$		$K = 30$		$K = 40$		$K = 50$	
	TP	Precision	TP	Precision	TP	Precision	TP	Precision	TP	Precision
Broadcast bot	9	0.80	18	0.90	27	0.90	34	0.85	38	0.76
Consumption bot	10	1.00	20	1.00	30	1.00	38	0.95	48	0.96
Spam bot	4	0.40	9	0.45	12	0.43	19	0.475	23	0.48
Human	10	1.00	20	1.00	30	1.00	40	1.00	40	1.00

TP: number of true positives

As shown in Table 4, our LR classifier produces fairly accurate and consistent predictions across different K values. With respect to human accounts, our LR classifier achieved perfect *Precision* for all K values. Unsurprisingly, we can expect that human accounts constitute the largest proportion of the Twitter population, and thus they should be the easiest to classify. We also obtained good results for the broadcast and consumption bots, with precision scores greater than 75 % and 95 % respectively. On the other hand, we observe rather modest *Precision* scores for spam bots (i.e., 40–47.5 %). We can attribute this to the insufficient number of instances for spam bots, which form only $\frac{105}{1{,}613} = 6.51\,\%$ of our labeled data (cf. Table 1). This may (again) be due to our data collection procedure that involved popular users as seeds and/or due to our relatively strict criteria for the characterization of spam bot accounts (cf. Sect. 7.1). Nevertheless, the *Precision* scores of 40–47.5 % remain relatively good, if we compare with that of a random guess for our labeled data (i.e., 6.51 %).

All in all, we find our top K predictions on unlabeled data to be satisfactory. Based on this, we can use our predictions to infer the behavioral profiles of bots in a larger Twitter population, which in this case spans the overall Singapore users.

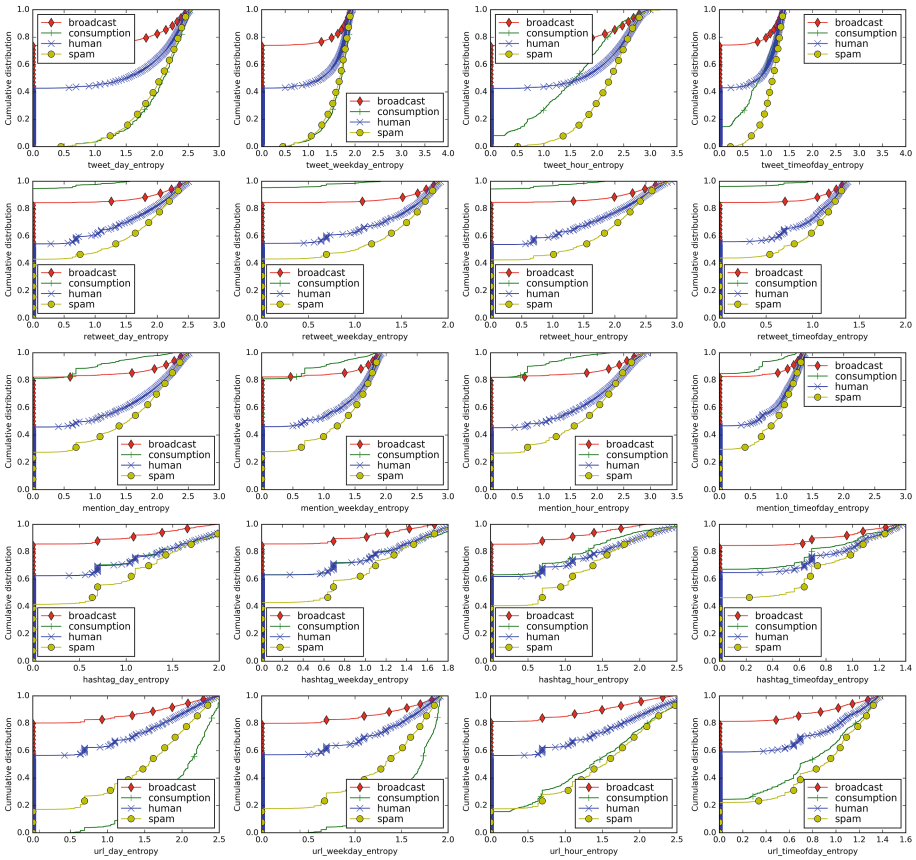


Fig. 6. Distribution of entropy-based features for 158,111 Twitter accounts

In particular, we analyze the entropy-based dynamic tweet features, namely the entropy distributions of the tweet, retweet, mention, hashtag and url activities, which constitute the majority group of the top discriminative features in Fig. 5. Figure 6 presents the cumulative distribution functions of these features. The detailed analysis of the distributions can be found in Sect. 7.3.

References

1. Abokhodair, N., Yoo, D., McDonald, D.W.: Dissecting a social botnet: growth, content and influence in Twitter. In: CSCW (2015)
2. Boshmaf, Y., Muslukhov, I., Beznosov, K., Ripeanu, M.: Design and analysis of a social botnet. Comput. Netw. **57**(2), 556–578 (2013)
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Detecting automation of Twitter accounts: are you a human, bot, or cyborg? IEEE Trans. Dependable Secure Comput. **9**(6), 811–824 (2012)

5. Dickerson, J.P., Kagan, V., Subrahmanian, V.: Using sentiment to detect bots on Twitter: are humans more opinionated than bots? In: ASONAM (2014)
6. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. **29**(2–3), 103–130 (1997)
7. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: a library for large linear classification. JMLR **9**, 1871–1874 (2008)
8. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Commun. ACM **59**(7), 96–104 (2016)
9. Freitas, C., Benevenuto, F., Ghosh, S., Veloso, A.: Reverse engineering socialbot infiltration strategies in Twitter. In: ASONAM, pp. 25–32 (2015)
10. 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: WWW, pp. 61–70 (2012)
11. Hu, X., Tang, J., Zhang, Y., Liu, H.: Social spammer detection in microblogging. In: IJCAI, pp. 2633–2639 (2013)
12. Hwang, T., Pearce, I., Nanis, M.: Socialbots: voices from the fronts. Interactions **19**(2), 38–45 (2012)
13. Lee, K., Eoff, B.D., Caverlee, J.: Seven months with the devils: a long-term study of content polluters on Twitter. In: ICWSM, pp. 185–192 (2011)
14. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
15. Mitter, S., Wagner, C., Strohmaier, M.: A categorization scheme for socialbot attacks in online social networks. In: ACM Web Science (2013)
16. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
17. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: ACSAC (2010)
18. Subrahmanian, V., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F., Waltzman, R., Stevens, A., Dekhtyar, A., Gao, S., Hogg, T., Kooti, F., Liu, Y., Varol, O., Shiralkar, P., Vydiswaran, V., Mei, Q., Huang, T.: The DARPA Twitter bot challenge. IEEE Comput. **49**(16), 38–46 (2016)
19. Tavares, G., Faisal, A.A.: Scaling-laws of human broadcast communication enable distinction between human, corporate and robot Twitter users. PloS One **8**(7), e65774 (2013)
20. Wagner, C., Mitter, S., Körner, C., Strohmaier, M.: When social bots attack: modeling susceptibility of users in online social networks. In: MSM (2012)
21. Wang, A.H.: Detecting spam bots in online social networking sites: a machine learning approach. In: DBSec, pp. 335–342 (2010)
22. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bull. **1**(6), 80–83 (1945)