2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Reverse Engineering Socialbot Infiltration Strategies in Twitter

Carlos Freitas*, Fabricio Benevenuto*, Saptarshi Ghosh†‡ and Adriano Veloso*
*Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
†Max Planck Institute for Software Systems, Kaiserslautern–Saarbruecken, Germany
‡Indian Institute of Engineering Science and Technology Shibpur, India

***Abstract*—Online Social Networks (OSNs) such as Twitter and Facebook have become a significant testing ground for Artificial Intelligence developers who build programs, known as socialbots, that imitate actual users by automating their social-network activities such as forming social links and posting content. Particularly, Twitter users have shown difficulties in distinguishing these socialbots from the human users in their social graphs. Frequently, legitimate users engage in conversations with socialbots. More impressively, socialbots are effective in acquiring human users as followers and exercising influence within them. While the success of socialbots is certainly a remarkable achievement for AI practitioners, their proliferation in the Twitter-sphere opens many possibilities for cybercrime. The proliferation of socialbots in the Twitter-sphere motivates us to assess the characteristics or strategies that make socialbots most likely to succeed. In this direction, we created 120 socialbot accounts in Twitter, which have a profile, follow other users, and generate tweets either by reposting messages that others have posted or by creating their own synthetic tweets. Then, we employ a 2^k factorial design experiment in order to quantify the infiltration effectiveness of different socialbot strategies. Our analysis is the first of a kind, and reveals what strategies make socialbots successful in the Twitter-sphere.**

I. Introduction

One of the key ambitions for Artificial Intelligence (AI) designers is to build computer systems that are capable of interacting with humans in a way that they are indistinguishable from real humans. This is a classical AI task which is gaining considerable popularity in online social media, mainly because the emergence of *socialbots*. These are computer programs designed to use social networks by simulating how humans communicate and interact with each other, and are becoming pervasive in Online Social Networks (OSNs), being highly effective in convincing users that they are actually humans.

Socialbots can have many applications, with good or malicious objectives. Like any software, they can automate tasks and perform them much faster than humans, like automatically posting news or change a template on Wikipedia of all pages in a category [1]. There are companies that develop chatbots for those interested in advertising using interactive and friendly AI entities or in providing virtual assistance for specific services [2]. Particularly, the Twitter OSN is becoming a suitable place for the proliferation of socialbots [3], [4] with objectives that are as diverse as attempts to influence political campaigns [5], spamming [6], [7], launching Sybil attacks [8], or simply to push out useful information like weather updates, and sports scores.

Independent of their goals, the proliferation of socialbots in the Twitter-sphere is certainly a remarkable achievement for AI practitioners. Intelligent machines that can pass for humans have long been dreamed of. However, since socialbots are often used in ways that are harmful to the other users or the OSN itself (e.g., degrading the services and creating a skewed perception of who (or what content) is influential), Twitter's Trust and Safety team regularly seeks to eliminate automated accounts. Although there are some accepted means of identifying bots in Twitter [7], [9] – such as, incomplete profile, skewed follower / following ratio, frequent posting of quotes and URLs, etc – distinguishing socialbots from legitimate Twitter users is proving to be a hard task as socialbot strategies are becoming smarter. Some recent efforts have demonstrated that socialbots can acquire social links and even become influential like celebrities in Twitter [10], [11]. Although these efforts suggest that it is possible to make socialbots pass for humans, it is still unclear which automated strategies are most likely to make socialbots succeed.

In this paper, we take the first step in this direction. Our methodology consists of creating 120 socialbot accounts with different characteristics and behaviors (e.g., gender specified in the profile, how active they are in interacting with users, the method used to generate their tweets, the type of users they attempt to interact with), and investigating the extent to which these bots are socially accepted in the Twitter social network over the duration of a month. More specifically, we quantitatively analyze which socialbot strategies are more successful in acquiring followers and provoking interactions (such as retweets and mentions) from other Twitter users. For this, we perform a 2^k factorial design experiment [12] to quantify the extent to which each bot strategy performs according to different social acceptance metrics.

We find that out of the 120 socialbot accounts, only 31% could be detected by Twitter after a period of one month of executing only automated behavior. This indicates that creating socialbots in the scale of hundreds is feasible with the current Twitter defense mechanisms for detecting automated accounts. We also show that socialbots employing simple automated mechanisms can acquire large number of followers and trigger hundreds of interactions from other users, making several bots to become relatively highly influential according to metrics like Klout score [13]. Our quantitative analysis shows that higher activity (such as following users and tweeting) is the most important factor towards successful infiltration when bots target a random group of users. Other factors, such as the gender and the profile picture, may gain importance when

ASONAM '15, August 25-28, 2015, Paris, France

ACM ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2809292

socialbots are concentrated on interacting with a particular group of users.

We hope our effort can open a new avenue for the AI community interested in developing AI entities in social environments and we also hope our observations may impact the design of future defense mechanisms. As a final contribution, we make our dataset available to the research community at `http://homepages.dcc.ufmg.br/~fabricio/asonam2015/`. The dataset (anonymized) consists of the timeline of activities and performance of infiltration of each of the 120 socialbots during the 30 days of experimentation. To the best of our knowledge, this dataset is the first of its kind, and will potentially allow researchers to explore new aspects of socialbots in Twitter.

II. Related work

Broadly speaking, most prior research related to socialbots in OSNs take one of two directions: (i) demonstrating vulnerability of social systems to bot infiltration, and (ii) creating counter mechanisms to detect bots. This section summarizes some recent studies in these directions.

We begin by describing some recent attempts that describe the creation of socialbots in OSNs. [14] designed a social network of bot accounts to infiltrate the Facebook OSN, and showed that, depending on users' privacy settings, a successful infiltration can result in privacy breaches of users' data, where more users' data are exposed compared to a purely public access. [11] created a bot that become highly connected in a social network for book lovers. Similarly, [10] created a bot that interacted with users on Twitter. Their bot, which described itself as a Brazilian journalist, achieved significant influence in the network according to influence metrics such as Klout and Twitalyzer (http://twitalyzer.com). There are also open source initiatives for the development of socialbots in Twitter [15], [16]. Overall, these efforts demonstrate that it is relatively easy to launch a socialbot, especially in Twitter, and it is possible to have it highly connected or even make it to be considered influential. Complementarily, our effort consists of measuring which strategies should be deployed to make socialbots more social accepted.

There have also been attempts for *detecting bots* in OSNs. A recent effort [9] characterized several aspects that can differentiate between content posted by certain types of social bots and humans, and created a tool that incorporated their findings into a machine learning model. A similar effort [3] used machine learning techniques to classify between three types of accounts in Twitter − users, bots and cyborgs (users assisted by bots). They showed that the regularity of posting, the fraction of tweets with URLs and the posting medium used (e.g., external apps), provide evidence for the type of the account. Complementary to the detection of bots, [17] created a machine learning model to predict user's susceptibility to bot attacks, using network, behavior and linguistic characteristics of the users. Their results indicate that users who are more "open" to social interactions are more susceptible to attacks. A similar study [18] found that the Klout score, number of followers and friends, are good predictors of whether a user will interact with bots. To the best of our knowledge, none of these efforts attempted to investigate and compare different socialbot strategies. Thus, our effort is also complementary to them.

III. Methodology

There are a vast number of characteristics and behaviors of socialbots that can impact their social acceptance. Particularly, they are devised to engage socially with legitimate Twitter users; in other words, the objectives of socialbots are that the users follow the bot, and socially engage with the bot by mentioning the bot or retweeting / favoriting the tweets posted by the bot. In order to analyze how various strategies of the socialbots impact their performance in terms of social engagement, it is necessary to create a set of socialbots, and then observe how successful their strategies are. This section discusses the methodology used to create the socialbot accounts in Twitter, and the characteristics / attributes of the various socialbots.

A. Creation of socialbots

We created a set of 120 socialbot accounts on Twitter. The socialbots were implemented based on the open-source Realboy project which is an experimental effort to create 'believable' Twitter bots [15]. The 120 bots were created over a period of 20 days, using 12 distinct IP addresses (10 bots were operated from each IP address). Subsequently, starting from 10 days after the creation of the last bot, we monitored their interactions with other users over a period of 30 days.

1) Profile settings of socialbots: To make socialbots look similar to legitimate users, we took the following steps while creating their accounts. Each socialbot was given a customized profile, which includes a name, a biography, a profile picture, and a background. The gender of the bot was set to 'male' or 'female' using a name from public lists of common female and male names and a suitable public profile picture obtained from the Web. Human volunteers carefully chose pictures that look like 'typical student profile pictures' (and not celebrity photos).

Further, to ensure that when other users see our bot accounts, they do not see a totally 'empty' profile, the socialbots were initially set to have a few followers and followings. As detailed later in this section, the 120 socialbots are divided into groups based on the set of target users they are assigned to follow. Each bot initially followed a small number (randomly selected between one and seven) of the most popular users among the target users assigned to it. Also, all socialbots assigned to the same target-set followed each other, so that every bot account had some followers to start with. Finally, every socialbot posted 10 tweets before attempting to interact with other Twitter users.

2) Activity settings of socialbots: Our socialbots can perform a set of basic actions to interact with other users: (i) follow them, (ii) post tweets, and (iii) retweet posts of users they follow. A socialbot becomes 'active' at pre-defined instants of time; the gap between two such instants of activity is chosen randomly (as detailed later in this section). Once a socialbot becomes active, it performs the following two actions: (i) with equal probability, the socialbot either posts a new tweet, or retweets a post that it has received from its followings, and (ii) the socialbot follows a random number

(between one and five) of the target users assigned to it, and follows some of the users who have followed it (if any) since the last instant of activity.

Note that we attempt to ensure that our bots do not link to spammers or other fake accounts, which could make Twitter's spam defense suspicious and lead to suspension of our bot accounts. For this, our bots only follow users from their respective target-set, and some selected users from among those who have followed them. Since spammers in Twitter usually have far less number of followers than the number of followings [6], [7], our socialbots follow back non-targeted users only if those users have their number of followers greater than half the number of their followings.

B. Attributes of the socialbots

There are a number of attributes of a Twitter user-account which could potentially influence how it is viewed by other users. Since analyzing the impact of all possible attributes would involve a high cost, we decided to focus on the following four specific attributes of the socialbot accounts: (i) the gender mentioned in the bot's profile, (ii) the activity level, i.e., how active the bot is in following users and posting tweets, (iii) the strategy used by the socialbot to generate tweets, and (iv) the target set of users whom the socialbot links with.

We set the bot accounts such that they have diverse characteristics with respect to these four attributes, and then attempt to measure whether any of these attributes can make a bot more successful in interacting with other users. The rest of this section describes these attributes, and how they are assigned to the 120 socialbots created.

1) Gender: Of the 120 socialbots, half are specified as male, and the other half as female. Setting the gender of a socialbot involves using an appropriate name and profile picture (as discussed above).

2) Activity level: Here we aim to investigate whether more active bots are more likely to be successful in acquiring interactions. Note that while more active bots are more likely to be visible to other users, they are also more likely to be detected as a bot; hence there is a trade-off in deciding the activity level of socialbots. For simplicity, we create socialbots with only two levels of activity:
(i) High activity: For these socialbots, the intervals between two consecutive actions are chosen randomly between 1 and 60 minutes.
(ii) Low activity: For these, intervals between two consecutive actions are chosen randomly between 1 and 120 minutes.

Half of our 120 socialbots exhibit high activity, while the other half exhibit low activity. Also, all socialbots 'sleep' between 22:00 and 09:00 Pacific time zone, simulating the expected downtime of human users.

3) Tweet generating strategy: Making a socialbot look like a legitimate user requires automated approaches for generating well-written tweets with relevant content. Our bots can employ two different approaches:

(i) Re-posting: This approach consists of re-posting tweets that were originally posted by another user, as if they were one's own. A socialbot employing this strategy simply re-posts tweets drawn from the 1% random sample of the Twitter stream that is provided publicly by Twitter. However, since a very large fraction of posts in Twitter are merely conversational [19], [20], blindly re-posting *any* random tweet would not seem interesting to the target users (whom the socialbot intends to interact with). To guard against this, for a particular bot, we extracted the top 20 terms that are most frequently posted by the target users of that bot (after ignoring a common set of English stop-words). The bot considers a tweet for re-posting only if it contains at least one of these top 20 terms.

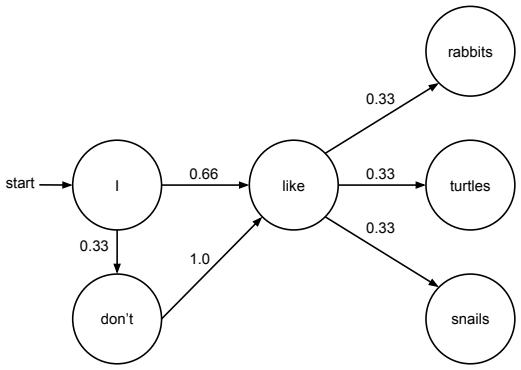


Fig. 1: Example of a bigram Markov chain – to demonstrate the approach used to synthetically generate tweets posted by the socialbots.

(ii) Generating synthetic tweets: This approach synthetically generates tweets using a *Markov generator* [21], [22] – a mathematical model used to generate text that looks similar to the text contained in a sample set of documents. Figure 1 shows an example of a bigram Markov generator, extracted from the sample set of documents {"I like turtles", "I like rabbits" and "I don't like snails"}. The weight of an edge $w_i \rightarrow w_j$ denotes the probability that the word w_j immediately follows word w_i, as measured from the sample documents.[1] A possible text generated by the Markov generator in Figure 1 is "I don't like rabbits" (see [21], [22] for details of the method).

To increase the likelihood that the tweets generated by a socialbot are considered relevant by its target users, we use a set of tweets recently posted by the target users of that socialbot, as the sample set to create a trigram Markov generator. The advantage of this approach is that, since it generates text containing the representative terms of the sample documents, the tweets generated by the socialbots are likely to be on the topics of interest of the target group. However, the textual quality of the tweets may be low (e.g., some tweets may be unfinished sentences). Moreover, because of the way that the method has been implemented, it is unable to generate tweets containing user-mentions or URLs. Table I shows some example tweets generated by the Markov generator used in our experiment.

Half of our socialbots use only the reposting approach, while the other half uses both the above approaches, where each approach has an equal probability to generate the next tweet.

[1] For instance, there is an edge of weight $\frac{2}{3}$ between the nodes "I" and "like" since, out of the three occurrences of the word "I" in the sample documents, two occurrences are immediately followed by "like".

I don't have an error in it :)
The amount of content being published this week :: the number of people who've finished this website but it makes it easier to argue that
Why isn't go in the morning! night y'all
take me to fernandos and you'll see

TABLE I: Examples of tweets synthetically generated by the Markov generator.

4) Target users: Another factor which potentially affects how socialbots are able to engage socially is the set of target users with whom the socialbot attempts to interact. For instance, we wanted to check whether it is easier for socialbots to interact with randomly selected users, or users who are similar to each other in some way (e.g., users who are interested in a common topic, or users who are socially connected among themselves).

As stated earlier, we wished to ensure that our socialbots do not link to other fake accounts. Hence, we consider a user-account as a potential target user, only if (i) it is controlled by a human (as manually judged from the account's profile and the nature of the tweets posted), (ii) it posts tweets in English (so that they understand the tweets of our bots), and (iii) it is active (i.e., has posted at least one tweet since December 2013). We considered the following three groups of target users:

Group 1: Consists of 200 users randomly selected from the Twitter random sample, and verified that they meet the above mentioned criteria.

Group 2: Consists of 200 users who post tweets on a specific topic. Deciding to focus on software developers, we selected users from the Twitter random sample, who have posted a tweet containing any of the terms "jQuery", "javascript" or "nodejs". Subsequently, we randomly selected 200 accounts from among these users, after verifying that they meet the criteria stated above. Note that though we focus on software developers, the study could be conducted on groups of users interested in any arbitrary topic.

Group 3: Consists of 200 users who post tweets on a specific topic (same as above), and are also socially connected among themselves. As the topic, we again focus on software developers. Here, we started with the 'seed user' @*jeresig* (an influential software developer on Twitter, and creator of 'jQuery') and collected the 1-hop neighborhood of the seed user. From among these users, we extracted 200 users whose profiles show that they are software developers, who satisfy the criteria stated above, and whose social links form a dense sub-graph in the Twitter social network.

The justification behind our choices of target users is as follows. First, we intend to check whether it is easier for socialbots to engage socially with heterogeneous groups of users (Group 1), or a set of users having common interests (e.g., software developers, as in Group 2 and Group 3). Second, we wish to compare the relative difficulty in interacting with a group of users who are socially well-connected among themselves (Group 3), versus users who are not socially connected (Group 1 and Group 2). Out of the 120 socialbots, 40 were assigned to each group of target users.

C. Ethical considerations of the study

In the course of this study, a set of 120 socialbot accounts were created, which created a few thousand social links in the Twitter social network, and posted tweets as described earlier. We believe that the few thousand links created by the socialbots have negligible effect on a large social network like Twitter. Further, the socialbots only re-posted tweets which are already public or automatically generated tweets from models that combine words from public tweets. Because of the way that we generated tweets, we ensure that none of our socialbots posted spam or malicious content as bots are unable to generate tweets containing user-mentions or URLs. Also, the users who follow the bots could decide whether or not to follow the socialbots, and they could unfollow if they disliked the content they receive in their timelines. All socialbot accounts were deleted after one month of experimentation and we will ensure that the usernames of the socialbot accounts or the users who interacted with them are not publicly revealed in the future.

IV. Can socialbots engage socially in Twitter?

We now check to what extent the socialbot accounts could socially engage other users in Twitter. A successful socialbot needs to (i) evade detection by Twitter's defenses which regularly detect and suspend automated accounts [23], and (ii) acquire popularity / influence in the social network by interacting with other users. In this section, we investigate how successful the socialbots were with respect to the above objectives.

A. Socialbots can evade Twitter defenses

We start by checking how many of the 120 socialbots created by us could be detected by Twitter. Over the duration of the experiment (30 days), 38 out of our 120 socialbots were suspended by Twitter. Thus, though all our socialbots actively posted tweets and followed other users during this period, as many as 69% of the socialbots could not be detected by Twitter defenses.

We now analyze those socialbots that were detected by Twitter. Figure 2 shows the distribution of the four attributes – gender, activity, tweeting, and target group – among the 120 socialbots which are indicated by numeric identifiers in the order of their creation (i.e., Bot1 was created first and Bot120 was created last). The socialbots which were detected by Twitter are indicated in red color, while the others are indicated in blue color.

We find that the large majority of the suspended socialbots were the ones which were created at the end of the account creation process (with IDs between 90 and 120). This is probably because by the time these accounts were created, Twitter's defenses had become suspicious of several accounts being created from the same block of IP addresses.[2] Also, socialbots which used the Markov-based posting method were more likely to be suspended. This is expected, since their synthetically generated tweets are likely to be of low textual quality. However, Twitter could detect only a small fraction

[2]As stated in Section III, we used 12 distinct IP addresses to create the 120 socialbots (i.e., 10 accounts were operated from each IP address).

Group 1		Group 2		Group 3			
Male	Female	Male	Female	Male	Female		
Bot 1	Bot 2	Bot 3	Bot 4	Bot 5	Bot 6	High Activity	Reposting
Bot 7	Bot 8	Bot 9	Bot 10	Bot 11	Bot 12		
Bot 13	Bot 14	Bot 15	Bot 16	Bot 17	Bot 18		
Bot 19	Bot 20	Bot 21	Bot 22	Bot 23	Bot 24		
Bot 25	Bot 26	Bot 27	Bot 28	Bot 29	Bot 30		
Bot 31	Bot 32	Bot 33	Bot 34	Bot 35	Bot 36		Resposting + Markov
Bot 37	Bot 38	Bot 39	Bot 40	Bot 41	Bot 42		
Bot 43	Bot 44	Bot 45	Bot 46	Bot 47	Bot 48		
Bot 49	Bot 50	Bot 51	Bot 52	Bot 53	Bot 54		
Bot 55	Bot 56	Bot 57	Bot 58	Bot 59	Bot 60		
Bot 61	Bot 62	Bot 63	Bot 64	Bot 65	Bot 66	Low Activity	Resposting
Bot 67	Bot 68	Bot 69	Bot 70	Bot 71	Bot 72		
Bot 73	Bot 74	Bot 75	Bot 76	Bot 77	Bot 78		
Bot 79	Bot 80	Bot 81	Bot 82	Bot 83	Bot 84		
Bot 85	Bot 86	Bot 87	Bot 88	Bot 89	Bot 90		
Bot 91	Bot 92	Bot 93	Bot 94	Bot 95	Bot 96		Resposting + Markov
Bot 97	Bot 98	Bot 99	Bot 100	Bot 101	Bot 102		
Bot 103	Bot 104	Bot 105	Bot 106	Bot 107	Bot 108		
Bot 109	Bot 110	Bot 111	Bot 112	Bot 113	Bot 114		
Bot 115	Bot 116	Bot 117	Bot 118	Bot 119	Bot 120		

Fig. 2: Distribution of attributes of the 120 socialbots, numbered in the order in which they were created. Socialbots detected by Twitter are shown in red, while those shown in blue could not be detected by Twitter. Twitter could not detect most of the socialbots which were created early, and those which simply re-post others' tweets.

of the socialbots which were created early, and which simply re-posted others' tweets.

Note that since we ensured that our socialbots do *not* engage in any spam activity (see Section III), Twitter is justified in not suspending the accounts since their Terms of Service are not violated. However, these observations indicate that creating socialbots in the scale of hundreds is feasible with current Twitter defense mechanisms which are of limited efficacy in detecting socialbots employing simple but intelligent strategies for posting tweets and linking to other users.

B. Socialbots can become influential in Twitter

We next check to what extent socialbots can gain popularity and influence in the Twitter social network. We use the following metrics (measured at the end of the duration of the experiment) to quantify how successful a socialbot is.

(1) Number of followers acquired: This is a standard metric for estimating the popularity of users in Twitter [24]. As stated in Section III, each of our socialbots is followed by some of our other socialbots (those which are assigned the same set of target-users). However, while counting the number of followers of a socialbot, we do *not* consider follows from other socialbots.

(2) Klout score: Klout score [13] is a popular measure for online influence. Although the exact algorithm is not known publicly, The Klout score for a user considers various aspects, including the number of followers and followings of the user, retweets, how many spam and dead accounts are following the user, how influential are the people who retweet and mention the user, and so on [25]. Klout scores range from 1 to 100, with higher scores implying a higher online social influence of a user.

(3) Number of message-based interactions with other users: We measure the number of times other users interact with a socialbot through messages (tweets), such as when some user @mentions the bot, or replies to the bot, or retweets or favorites a tweet posted by the bot.[3] This metric is a direct measure of social engagement, i.e., the extent to which the bot participates in a broad range of social roles and relationships [26].

User	Description	Klout
ladamic	Data scientist at Facebook	48
vagabondjack	Data Scientist at LinkedIn	46
emrek	Senior researcher at Microsoft Research	44
Bot 28	**Socialbot in this study**	**42**
wernergeyer	Data Scientist at IBM Research	40
Bot 4	**Socialbot in this study**	**39**
Bot 16	**Socialbot in this study**	**39**
scarina	*Bot developed in [10]*	*37.5*
fepessoinha	*Bot developed in [10]*	*12.3*

TABLE II: Comparison of Klout scores of some of our socialbots with well-known researchers and bots developed in [10].

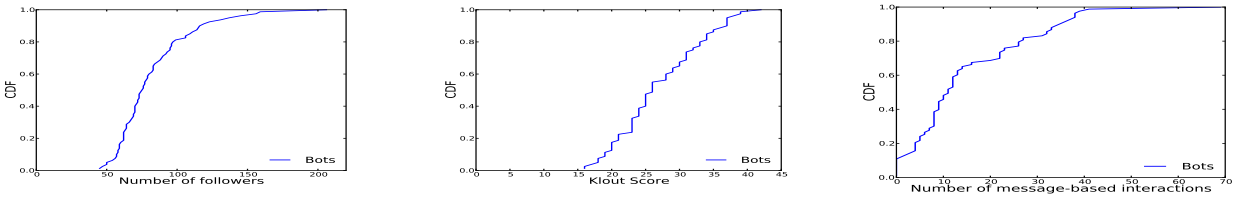
Over the duration of the experiment, our 120 socialbots received in total 4,601 follows from 1,952 distinct users, and 1,991 message-based interactions from 1,187 distinct users. Figure 3 shows the distribution of the number of followers, the Klout score and the number of message-based interactions acquired by the socialbots at the end of the experiment. It is evident that a significant fraction of the socialbots acquire relatively high popularity and influence scores. Within just one month (the duration of the experiment), more than 20% of the socialbots acquired more than 100 followers (Figure 3(a)); it can be noted that 46% of all users in Twitter have less than 100 followers [27].

Figure 3(b) shows that 20% of the socialbots acquired Klout scores higher than 35 within only one month. Table II compares the Klout scores acquired by the three socialbots that acquired the highest Klout scores[4] with some well-known researchers in Computer Science, who are also active Twitter users. We find that within just one month, our socialbots achieved Klout scores of the same order of these well-known academicians (who have accumulated influence over several years). Additionally, these socialbots also acquired higher Klout scores than the two bots developed in the prior study [10].

Thus, we find that socialbot accounts can not only evade the existing Twitter defense mechanisms, but also successfully engage with users in the social network and acquire high scores according to standard influence / popularity metrics. These observations also imply that influence metrics such as Klout score and number of followers are susceptible to manipulation by socialbots, and advocates use of influence metrics that are more resilient to activities such as link farming [28].

[3]Some of the bots encountered some other types of interactions, such as a tweet in which the bot was mentioned getting retweeted or favorited. Though we did not consider these interactions in our analysis, all the interactions are included in the dataset that we make publicly available.

[4]The three socialbots which acquired the highest Klout scores have common characteristics – gender specified as 'female', highly active, used only reposting as the mechanism for tweeting, and followed Group 2 of target users.



(a) Number of followers **(b) Klout Score** **(c) Number of interactions**

Fig. 3: Performance of our socialbots: CDFs for (i) number of followers acquired, (ii) Klout Score, and (iii) number of message-based interactions with other users.

Factor	-1	$+1$
Gender (**G**)	Female	Male
Activity Level (**A**)	Low activity	High activity
Posting Method (**P**)	Repost	Repost+Markov

TABLE III: Factors used in the factorial experiment for the socialbot infiltration study.

V. Assessing Bot Configuration Effectiveness

The previous section showed that a significant fraction of the socialbots are able to infiltrate and gain popularity in the Twitter social network. In this section, we quantify which attribute configuration (gender, activity, tweeting, and target users) has the greatest impact in the performance of the socialbots.[5] We conduct a 2^k factorial design experiment [12] to assess the relative impact of the different configurations, as described next.

A. Factorial experiment on socialbot configuration

We individually consider the performance of our socialbots in terms of social engagement, considering the three target groups (which were described in Section III), where the performance is measured by the (i) number of followers, (ii) Klout score, and (iii) number of message-based interactions. For each of the three engagement measures and for each of the three target groups, we executed a 2^3 design considering the three attributes Gender (G), Activity level (A) and Posting approach (P) whose values are described in Table III. This results in $3 \times 3 \times 2^3$ experiments. All experimental configurations for all datasets were averaged over the performance of all 5 socialbots in each attribute configuration.

The basic idea of our factorial design model consists of formulating y, the social engagement performance, as a function of a number of factors and their possible combinations (GP, AP, AG, and AGP)[6], as defined by Eqn. 1:

$$y = Q_0 + \sum_{i \in F} Q_i \cdot x_i \quad (1)$$

where $F = \{G, A, P, GA, GP, AP, GAP\}$ and x_i is defined as follows: x_G is -1 if the gender is specified as Female, and $+1$ if Male. Similarly, x_A is -1 if the socialbot has low activity, and $+1$ if high activity, and x_P is -1 if the posting method is Repost and $+1$ if Repost + Markov. The x_i's for combinations (e.g., AG, GP) are defined from the values of x_G, x_A, and x_P following the standard way described in [12]. In Eqn. 1, Q_i is the infiltration performance (number of followers, Klout score, or number of message-based interactions) when attribute $i \in F$ is applied, and Q_0 stands for the average performance, over all possible attribute configurations. By empirically measuring y according to different combinations (which, in our case, refer to the various socialbot attributes), we can estimate the values of the different Q_i and Q_0. This allows us to understand by how much each factor impacts the final socialbot performance.

As proposed in [12], the importance of a particular factor (i.e., socialbot attribute) can be quantitatively estimated by assessing the proportion of the total variation in y that is explained by that factor. To compute this, we consider the value of y across all runs, and then compute SS_T as the sum of the squared difference between each measured value of y and the mean value of y. Then, we compute SS_i, the variation only due to factor i, which is computed similarly to SS_T, but considering only those runs in which the value of the factor i were changed. Finally, we calculate the fraction of variation due to factor i as $\frac{SS_i}{SS_T}$. We now use this metric to compute the impact of each attribute for different performance measures and groups of target users.

B. Analyzing Bot Configurations

Table IV shows the percentage variation in (i) the number of followers, (ii) number of interactions, and (iii) Klout score acquired by the socialbots who followed each of the three target groups, as explained by each possibility in F. We note that the activity level (A) of a socialbot is the most important factor impacting its popularity. For instance, for Group 1 of target users (random users), the activity level is 61.9% responsible for deciding the number of followers acquired by a socialbot. This is expected, since the more active a socialbot is, (i.e., the more frequently it posts tweets or creates social links), the higher is the likelihood of its being visible to other users. However, note that the more active a bot is, the more likely it is to be detected by Twitter's defense mechanisms.

The second most important attribute is the posting method (P), which accounts for 16.9% of the variation on the number of followers for Group 1. The combination of these two factors (AP) also leads to a high variation in the number of followers (14.3%) and number of interactions (37.6%) for Group 1.

[5]Note that the analysis in this section consider only those socialbots which were not suspended by Twitter during the experiment.

[6]For instance, the experiments for 'GP' attempts to measure the impact of a certain combination of the attributes Gender (G) and Posting method (P) (e.g., 'Female and Repost', or 'Male and Repost+Markov').

	G	A	P	GA	GP	AP	GAP
Percentage variation in the number of followers							
Group 1	4.2	**61.9**	**16.9**	2.6	0.1	**14.3**	0.0
Group 2	4.0	**72.6**	2.8	4.4	3.5	2.8	9.9
Group 3	**20.5**	**49.3**	2.0	2.4	5.4	12.7	7.7
Percentage variation in the number of message-based interactions							
Group 1	0.4	**41.6**	17.3	1.1	1.4	**37.6**	0.6
Group 2	0.0	**40.6**	7.3	20.7	19.4	6.3	5.8
Group 3	**12.7**	**43.2**	4.5	19.6	8.2	1.2	10.6
Percentage variation in the Klout score							
Group 1	0.5	**40.2**	23.9	0.0	0.5	**34.9**	0.0
Group 2	7.6	**32.2**	12.6	17.0	15.6	8.8	6.2
Group 3	12.1	**29.3**	17.3	13.3	14.1	2.6	11.4

TABLE IV: Percentage variation in (i) number of followers, (ii) number of message-based interactions, and (iii) Klout score, explained by each attribute or combination of attributes (G: gender, A: activity level, P: posting method).

Also note that impact of some of the attributes varies significantly according to the group of users targeted by the socialbots. For instance, the gender attribute has a great impact in the experiments with target users from Group 3, being responsible for 20.5% of the variation in the number of followers and 12.7% of variation in interactions when the target users are from this group. We found that the users in Group 3 were more likely to follow and interact with socialbots having female profiles. However, the gender does not have much influence on the other target groups.

C. Evaluating the impact of individual attributes

Finally, we individually analyze the impact of the four attributes on the popularity and influence acquired by our socialbots. For brevity, we only report statistics for the number of followers (and considering all three target-groups together); analyses on the Klout score and message-based interactions yielded very similar results. Figure 4 shows the mean number of followers acquired by the socialbots over each day during our experiment. In these figures, the curves represent the mean values considering all the socialbots employing a particular strategy configuration on a given day, and the error bars indicate the 95% confidence intervals of the mean values.

It is evident that in general, the gender (Fig. 4(a)) and the posting method (Fig. 4(b)) have very little impact on the popularity of the socialbots. The low impact of the tweet posting method is especially surprising, since it indicates that Twitter users are not able to distinguish between re-posted human-generated tweets and automatically generated tweets using statistical models. This is possibly because a large fraction of posts in Twitter are written in an informal, grammatically incoherent style [29], so that even simple statistical models can produce tweets with quality similar to those posted by humans.

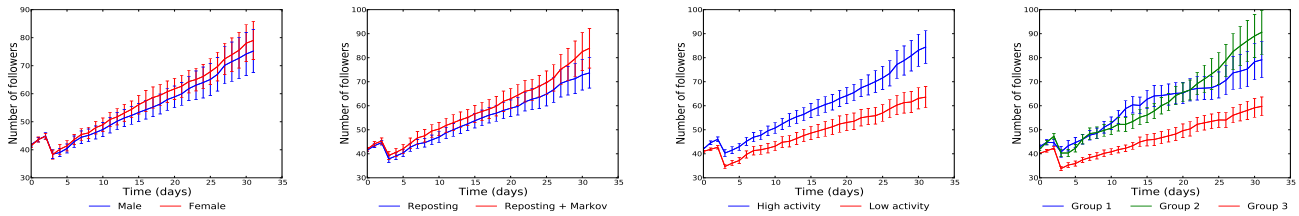
On the other hand, the activity level (Fig. 4(c)) and the target group of users (Fig. 4(d)) have large effect on the popularity acquired by the socialbots. Figure 4(c) shows that socialbots with higher activity levels achieve significantly more popularity than less active socialbots (as also seen in the 2^k factorial experiment). Figure 4(d) shows the number of followers gained by socialbots which followed different target-groups. Socialbots in Group 2 acquired a significantly higher number of followers, while those in Group 3 acquired less followers. This implies that following users who post tweets on a specific common topic (as done by Group 2) is a more promising approach for socialbots, than following random users (as done by Group 1). However, interacting with interconnected groups of users (as attempted by Group 3) is far more difficult than engaging with users without any relation among themselves (Group 2).

VI. Concluding Discussion

This work presented a reverse engineering study of socialbot strategies in the Twitter social network. Socialbots can potentially be used in OSNs with good as well as malicious intentions. For instance, several conferences today employ automated bot accounts to enhance the publicity of the conference. On the other hand, malicious socialbots also abound in Twitter [3], [4], and various forms of spam attacks – such as link-farming [28], search spam [6] and phishing [30] – can use socialbots to first infiltrate and acquire influence, making the attacks much harder to detect. The issue of socialbots in OSNs is a clear adversarial fight, or as is usually called, a cat and mouse fight. In this study, we put ourselves in the mouse's shoes (i.e., assumed the perspective of socialbot-developers) as an attempt to bring to the research community a novel perspective to the problem.

We exposed Twitter's vulnerability against large-scale socialbot attacks that can affect both Twitter itself and services built on crowd-sourced data gathered from Twitter. For instance, we show that Twitter users are *not* good at distinguishing tweets posted by humans and tweets generated automatically by statistical models; hence, relying on user-generated reports for identifying bots (as done by Twitter today [23]) may not be effective. Again, standard influence metrics such as Klout score and number of followers are susceptible to socialbot attacks. We also showed that re-posting others' tweets is a simple and effective strategy for socialbots. On the other hand, it is comforting that to achieve high social acceptance in a short time, socialbots need to be highly active, e.g., they need to post tweets and follow users almost every hour. Thus, it might be sufficient to monitor active accounts in order to prevent bots from becoming influential.

We show that it is possible to create a large number of bots in Twitter today and we quantitatively show what can make them influential or not. As socialbots can be created in large numbers, they can potentially be used to bias public opinion. There are already evidences of the use of socialbots to create an impression that emerging political movements are popular and spontaneous [31]. Particularly, there are numerous concerns that that socialbots may influence political campaigns, such as trying to change the "trending topics" during elections [32]. In fact, Reuters even launched an internet campaign for political candidates to not use socialbots [5]. This scenario only gets worse when we consider the existence of socialbot sale services (such as http://www.jetbots.com/). Thus, ultimately, our effort calls for an attention to the validity of any service that utilizes Twitter data without attempting to differentiate socialbots from real users, and calls for more secure mechanisms for creating online identities.



(a) Gender **(b) Tweeting method** **(c) Activity level** **(d) Target user-group**

Fig. 4: Performance of socialbots, for different attributes: (a) gender, (b) tweet generating method, (c) activity level, (d) target user-group. The curves represent the mean values, which the error bars indicated the 95% confidence intervals.

As a final contribution, we make our (anonymized) dataset – containing the timeline of activities and infiltration performance of each of the 120 socialbots during the 30 days of experimentation – available to the research community at `http://homepages.dcc.ufmg.br/~fabricio/asonam2015/`.

Acknowledgements: The work is partially supported by grants from CNPq, CAPES, and Fapemig. Additionally, S. Ghosh is supported by a postdoctoral fellowship from the Alexander von Humboldt Foundation.

REFERENCES

[1] "Creating a bot on Wikipedia," http://en.wikipedia.org/wiki/Wikipedia:Creating_a_bot.

[2] "Pandora Bots," http://www.pandorabots.com/.

[3] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans. Dependable Secur. Comput.*, vol. 9, no. 6, pp. 811–824, Nov. 2012.

[4] J. Edwards, "There Are 20 Million Fake Users On Twitter, And Twitter Can't Do Much About Them – Business Insider," http://tinyurl.com/twitter-20M-fake-users, Apr 2013.

[5] "Let's make candidates pledge not to use bots," http://blogs.reuters.com/great-debate/2014/01/02/lets-make-candidates-pledge-not-to-use-bots/.

[6] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers on Twitter," in *Proc. Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[7] K. Lee, B. D. Eoff, and J. Caverlee, "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter," in *Proc. AAAI International Conference on Web and Social Media (ICWSM)*, 2011.

[8] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, "An analysis of social network-based sybil defenses," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 4, pp. 363–374, Aug. 2010.

[9] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The Rise of Social Bots," *ArXiv e-prints*, Jul. 2014, http://arxiv.org/abs/1407.5225.

[10] J. Messias, L. Schmidt, R. Rabelo, and F. Benevenuto, "You followed my bot! Transforming robots into influential users in Twitter," *First Monday*, vol. 18, no. 7, July 2013.

[11] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo, "People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks," in *Proc. AAAI International Conference on Web and Social Media (ICWSM)*, 2012.

[12] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley and Sons, INC, 1991.

[13] "Klout — The Standard for Influence," http://klout.com/.

[14] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money," in *Proc. Annual Computer Security Applications Conference (ACSAC)*, 2011.

[15] Z. Coburn and G. Marra, "Realboy: Believable Ttwitter Bots," 2008, http://ca.olin.edu/2008/realboy/.

[16] "Web Ecology Project," http://www.webecologyproject.org/.

[17] C. Wagner, S. Mitter, C. Körner, and M. Strohmaier, "When social bots attack: Modeling susceptibility of users in online social networks," in *Proc. Workshop on Making Sense of Microposts (with WWW)*, 2012.

[18] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, "Which Users Reply to and Interact with Twitter Social Bots?" in *Proc. IEEE Conference on Tools with Artificial Intelligence (ICTAI)*, Nov 2013.

[19] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier, "It's not in their tweets: modeling topical expertise of Twitter users," in *Proc. AASE/IEEE International Conference on Social Computing (SocialCom)*, 2012.

[20] S. Ghosh, M. B. Zafar, P. Bhattacharya, N. Sharma, N. Ganguly, and K. Gummadi, "On Sampling the Wisdom of Crowds: Random vs. Expert Sampling of the Twitter Stream," in *Proc. ACM Conference on Information & Knowledge Management (CIKM)*, 2013.

[21] G. Barbieri, F. Pachet, P. Roy, and M. D. Esposti, "Markov Constraints for Generating Lyrics with Style," in *Proc. European Conference on Artificial Intelligence*, 2012.

[22] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed. Prentice Hall PTR, 2000.

[23] "Shutting down spammers," Apr 2012, https://blog.twitter.com/2012/shutting-down-spammers.

[24] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *Proc. AAAI International Conference on Web and Social Media (ICWSM)*, 2010.

[25] "Klout – Wikipedia," http://en.wikipedia.org/wiki/Klout.

[26] J. D. M. William R. Avison and B. A. P. (Eds.), *Mental Health, Social Mirror*. Springer, 2007.

[27] "46% of Twitter users have less than 100 followers - Simplify360," http://simplify360.com/blog/46-of-twitter-users-have-less-than-100-followers/.

[28] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and Combating Link Farming in the Twitter Social Network," in *Proc. World Wide Web Conference (WWW)*, 2012.

[29] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis: The Good, the Bad and the OMG!" in *Proc. AAAI International Conference on Web and Social Media (ICWSM)*, 2011.

[30] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru, "Phi.sh/$oCiaL: The Phishing Landscape through Short URLs," in *Proc. Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2011.

[31] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Truthy: Mapping the Spread of Astroturf in Microblog Streams," in *Proc. World Wide Web Conference (WWW)*, 2011.

[32] M. Orcutt, "Twitter mischief plagues mexico's election," http://www.technologyreview.com/news/428286/twitter-mischief-plagues-mexicos-election/, Jun. 2012.