Soc. Netw. Anal. Min. (2016)6:23
DOI 10.1007/s13278-016-0331-3

ORIGINAL ARTICLE

An empirical study of socialbot infiltration strategies in the Twitter social network

Carlos Freitas[1] · Fabrício Benevenuto[1] · Adriano Veloso[1] · Saptarshi Ghosh[2]

Received: 17 December 2015 / Revised: 5 March 2016 / Accepted: 8 April 2016

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 human 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, 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 Twitter 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 others' tweets or by generating their own synthetic tweets. Then, we employ a 2^k factorial design experiment to quantify the infiltration performance of different socialbot strategies, and examine the effectiveness of individual profile and activity-related attributes of the socialbots. Our analysis is the first of a kind, and reveals what strategies make socialbots successful in the Twitter sphere.

✉ Saptarshi Ghosh
sghosh@cs.iiests.ac.in

[1] Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[2] Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology Shibpur, Howrah 711103, India

Keywords Twitter social network · Socialbots · Infiltration strategies · Factorial design experiment

1 Introduction

Online social networks (OSNs) have become popular communication mediums where people post about a wide variety of topics, ranging from day-to-day conversations to their opinions about noteworthy events. The large amounts of social interactions and user-generated content on these sites make them a lucrative framework for researchers of various disciplines, including sociology, network science, different sub-disciplines of computer science, such as data mining, natural language processing, artificial intelligence and machine learning, and so on.

Specifically for artificial intelligence (AI) designers, one of the key ambitions 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 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 (wikipedia-bot 2015). There are companies that develop chatbots for those interested in advertising using interactive and friendly AI entities or in providing virtual assistance for specific

Published online: 02 May 2016

services (pandora-bots 2015). Particularly, the Twitter OSN is becoming a suitable place for the proliferation of socialbots (Chu et al. 2012; 20M-fake-users-twitter 2013) with objectives that are as diverse as attempts to influence political campaigns (reuters-botsban 2014), spamming (Benevenuto et al. 2010; Lee et al. 2011), launching Sybil attacks (Viswanath et al. 2010a), 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.

However, socialbots are often used in ways that are harmful to the other users or the OSN itself, such as degrading the services and creating a skewed perception of who (or what content) is influential. For instance, consider that users of a Twitter-based service might be interested in knowing what others think about a certain political candidate, to formulate their own opinion. In this scenario, socialbots could be used to post tweets to dishonestly improve or damage the public perception about this candidate, as an attempt to manipulate public opinion.

Because of the potential risks associated with socialbots, Twitter's Trust and Safety team regularly seeks to eliminate automated accounts. Some means of identifying bots in Twitter have been proposed (Lee et al. 2011; Ferrara et al. 2014), such as incomplete profiles, skewed follower/following ratio, frequent posting of quotes and URLs, and so on. However, distinguishing socialbots from legitimate Twitter users is proving to be a challenging 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 (Messias et al. 2013; Aiello et al. 2012). 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. There are many intriguing questions related to socialbots infiltration in Twitter. For instance, *Can socialbots really infiltrate Twitter easily?*, *What are the characteristics of socialbots that would enable them to evade current Twitter defenses? What strategies could be more effective to gain followers and influence? What automatic posting patterns could be deployed by socialbots without being detected?* and so on.

In this paper, we take an early step towards answering these questions. 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 (Jain 1991) to qufrom the sample set of documentsantify the extent to which each bot strategy performs according to different social acceptance metrics.

Note that this work is an extension of our prior work (Freitas et al. 2015). Compared to (Freitas et al. 2015), this work contains a much more detailed analysis of the impact of each individual attribute—gender, tweeting strategy, activity level, and type of target users—on the infiltration performance of the socialbots (see Sect. 6). Additionally, while Freitas et al. (2015) studied the infiltration performance only at the end of the one-month duration, the present work analyzes the performance of different attributes on each day throughout the experiment.

Our findings raise an alert about the vulnerability of many existing Twitter-based services. 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 (Klout 2015). Our quantitative analysis shows that higher activity (such as following users and tweeting) and the type of users targeted are the two most important factors in determining how successful a socialbot is in infiltrating the network. Specifically, the activity level is the most important attribute 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 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 on online social media platforms. 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.

The rest of the paper is organized as follows. The next section briefly surveys related work. In Sect. 3, we present the methodology used to create the socialbots, and the various strategies/attributes that we analyze. Section 4 checks to what extent socialbots can gain popularity and social engagement in the Twitter social network. Sect. 5 and 6 analyze the impact of the various strategies/attributes in the socialbots' infiltration performance. Specifically, Sect. 5 describes a 2^k factorial design experiment to quantitatively assess the relative importance of various attributes in socialbot infiltration strategies, while Sect. 6 analyzes the performance of each individual attribute throughout the experiment. Finally, Sect. 7 discusses the implications of our findings to future defense mechanisms and directions of future work.

2 Related work

Most of the prior research related to socialbots in OSNs take one of two directions: (i) demonstrating vulnerability of various social systems to bot infiltration, and (ii) creating counter mechanisms to detect bots. This section summarizes some recent studies in these directions. We also discuss how the present work differs from most of the prior work.

2.1 Vulnerability of social systems to bot infiltration

Viability of creating socialbots in OSNs: We begin by describing some recent attempts to create socialbots in OSNs. Boshmaf et al. (2011) 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. Aiello et al. (2012) created a bot that becomes highly connected in a social network for book lovers. Similarly, Messias et al. (2013) 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 such as the Realboy project (Coburn and Marra 2008) or the Web Ecology project (web-ecology 2015). 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.

Predicting users' susceptibility to bot attacks: Some studies (Wagner et al. 2012; Wald et al. 2013) have attempted to predict users' susceptibility to bot attacks, depending on various network and linguistic characteristics of the users. Wagner et al. (2012) 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 (Wald et al. 2013) 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 the above efforts attempted to investigate and compare different socialbot strategies, which is the goal of the present work.

2.2 Detecting bots and malicious accounts in OSNs

Detecting bots in social networks: There have been several attempts for *detecting bots* in OSNs. A recent effort (Ferrara et al. 2014) 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 (Chu et al. 2012) 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.

There have also been a few studies on detecting *influence bots* that attempt to influence discussion on a particular topic. For instance, there have been recent reports that terrorist groups are using bot accounts in online social media to spread radicalism (Shane and Hubbard 2014). To counter such mechanisms, DARPA recently organized a challenge to develop methodologies to identify influence bots on Twitter (Subrahmanian et al. 2016). Specifically, the teams participating in the challenge were required to detect influence bots that were supporting a pro-vaccination discussion.

Detecting trustworthy/untrustworthy nodes in a social network: The success of socialbots in infiltrating a social network depends on the trust that is implicit in a social system. There have been many studies on identifying which nodes in a social network are trustworthy. For instance, (Chandra et al. 2012) proposed a random walk-based methodology to identify a subset of trustworthy nodes with respect to a particular user. There have also been similar attempts on massive multiplayer online games (MMOs), e.g., to identify trustworthy users (Ahmad et al. 2011) and players with illicit behavior (Roy et al. 2012). Another line of work has been on detecting Sybil accounts in social networks; see the study by Viswanath et al. (2010b) for a review on Sybil detection methods.

2.3 Two perspectives of studying bots/malicious accounts in social systems

Overall, a large majority of the studies on socialbots are from the *perspective of those who build defense mechanisms* such as developing methodologies to detect bots. However, to effectively counter malicious activity in online systems, it is also necessary to conduct studies from the *perspective of the malicious users*, such as spammers. Such studies (from the perspective of malicious users) essentially attempt to reverse engineer the strategies of spammers to gain insights which can help to develop better defenses.

Most of the studies conducted from the perspective of spammers have been on e-mail spam, and spam in the Web. For instance, Pitsillidis et al. (2010) attempted to filter spam e-mails by exploiting the perspective of the spammers—they instantiated botnet hosts in a controlled environment, and monitored spam e-mails as they were created, and thus inferred the underlying template used to generate such e-mails. Stone-Gross et al. (2011) studied a large-scale botnet from the perspective of the botmaster, and analyzed the methodologies used in orchestrating spam e-mail campaigns. Gyöngyi and Garcia-Molina (2005) studied link farms on the Web, which are groups of interconnected web pages which attempt to boost the rankings of particular web pages. Specifically, they investigated how multiple web pages can be interconnected to optimize rankings.

With respect to socialbots, most of the prior studies have been from the perspectives of the social media sites and have focused on designing defense mechanisms (as described earlier in this section). To the best of our knowledge, there is no previous study attempting to analyze the strategies of socialbots *from the perspective of the bot-creators themselves*. This is the motivation of the present study—to reverse engineer socialbot strategies in the Twitter OSN. We believe that this is complementary to all the aforementioned studies on socialbots in social media, and can offer a novel perspective to building more effective defense mechanisms against bot accounts in the future.

3 Methodology

This study aims to reverse engineer socialbot strategies in Twitter, and analyze how various strategies of the socialbots impact their infiltration performance. For this, it is necessary to create a set of socialbots in Twitter, which would attempt to infiltrate the network, and then observe their behavior and infiltration performance. This section discusses the methodology used to create the socialbot accounts in Twitter, and the characteristics/strategies of the various socialbots.

3.1 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 (Coburn and Marra 2008). 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, we monitored their interactions with other users over a period of 30 days.

3.1.1 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. In addition, 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.

3.1.2 Activity settings of socialbots

Our socialbots can perform a set of basic actions to interact with other users: (i) follow users, (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 (Benevenuto et al. 2010; Lee et al. 2011), our socialbots follow back non-targeted users only if those users have their number of followers greater than half the number of their followings.

3.2 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, which intuitively seem important in determining how successful a socialbot is in infiltrating the network: (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.

3.2.1 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).

3.2.2 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, based on the interval between two consecutive instants when a bot becomes 'active' and engages in following users and posting tweets:

(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. In addition, all socialbots 'sleep' between 22:00 and 09:00 Pacific time zone, simulating the expected downtime of human users.

3.2.3 Tweet generating strategy

One of the key challenges to make a socialbot to look like a real user is to employ *automated* methodologies of generating tweets with relevant, interesting content. Our bots can employ two different approaches:

(i) *Reposting*: This approach consists of reposting tweets that were originally posted by another user, as if they were one's own. A socialbot employing this strategy simply reposting tweets drawn from the 1 is provided publicly by Twitter. However, since a very large fraction of posts in Twitter are merely conversational (Wagner et al. 2012; Ghosh et al. 2013), blindly reposting *any* random tweet would not seem interesting to the target users (whom the socialbot intends to interact with). Thus, we adopted the following approach to increase the odds that the tweets reposted by our bots have content relevant to the target users. 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 stopwords). The bot considers a tweet for reposting only if it contains at least one of these top 20 terms.

(ii) *Generating synthetic tweets*: This approach synthetically generates tweets using a *Markov generator* (Barbieri et al. 2012; Jurafsky and Martin 2000)—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. 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". A possible text generated by the Markov generator in

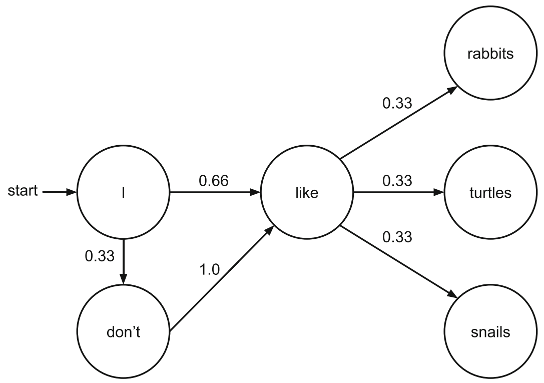


Fig. 1 Example of a bigram Markov chain—to demonstrate the approach used to synthetically generate tweets posted by the socialbots.

Fig. 1 is "I don't like rabbits". The reader is referred to (Barbieri et al. 2012; Jurafsky and Martin 2000) 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 the Markov generator. We use a trigram Markov generator, since trigrams showed the best results when compared to *n*-grams of any other order. We initially extract the empirical probability of occurrence of each trigram in the sample set, then generate a Markov generator from the obtained set of trigrams, and finally randomly generate tweets using this 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 1 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.

Table 1 Examples of tweets synthetically generated by the Markov generator.

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
Night y'all
take me to fernandos and you'll see

3.2.4 *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. We decided to focus on a group of software developers; hence, we selected users from the Twitter random sample, who have posted at least one 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 subgraph 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.

To bring out the differences among the three groups of target users (selected as described above), we conducted a brief characterization of each group. Figure 2 shows distributions of the users in the three target groups according to: (i) the age of their accounts, (ii) the total number of tweets posted during their lifetime, and (iii) their number of followers. We found that users in group 1 have relatively newer accounts than the other groups (Fig. 2a); however, they are more active in posting tweets (Fig. 2b). Further, users in group 3 are slightly more influential in the social network than the other groups, i.e., have a greater number of followers (Fig. 2c).

3.3 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 reposted 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. In addition, 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.

4 Can socialbots engage socially in Twitter?

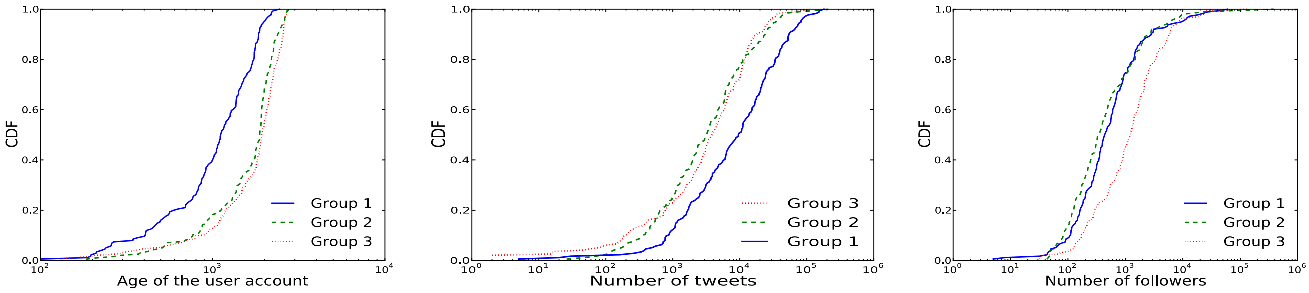
We now check to what extent the socialbot accounts could socially engage other users in Twitter. A successful socialbot needs to: (i) evade detection by Twitter's defenses which regularly detect and suspend automated accounts (twitter-shut-spammers 2012), and (ii) acquire popularity/influence in the social network by interacting with other users. In this section, we investigate how successful the socialbots were with respect to the above objectives.

4.1 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 30 days during which the experiment was carried out, 38 out of the 120 socialbots were suspended. 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 spam defense mechanisms.

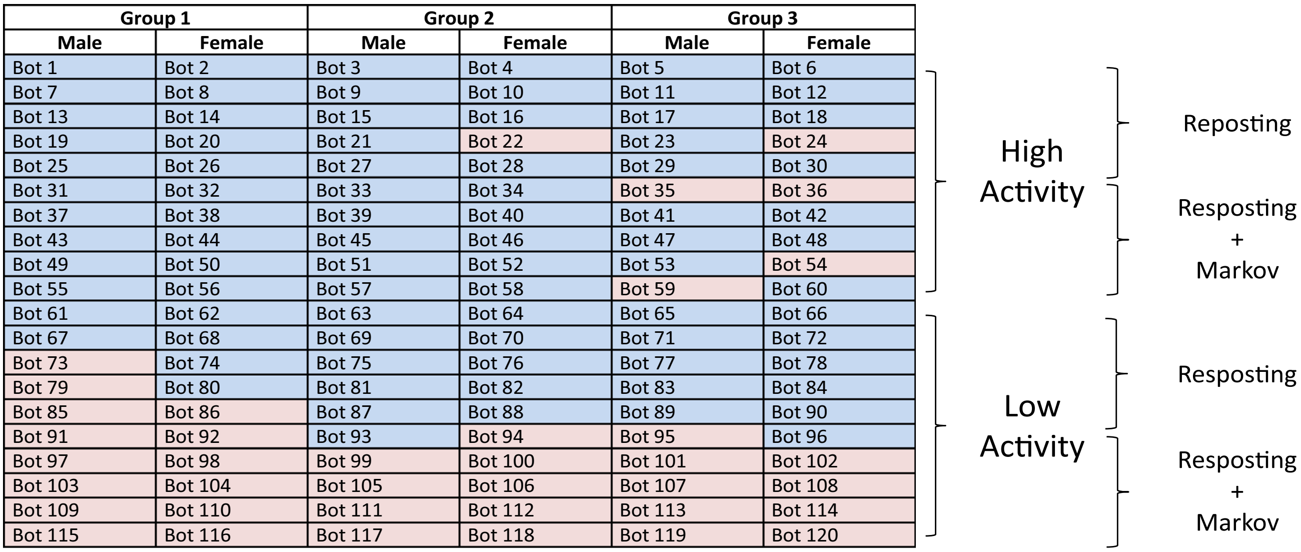
We now analyze which of the 120 socialbots could be detected by Twitter. Figure 3 shows the distribution of the four attributes—gender, activity, tweeting, and target group—among the 120 socialbots. The socialbots are indicated by numeric identifiers in the chronological order in which they were created, 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 socialbots which could not be detected by Twitter are shown 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 bias towards suspension of accounts created later is probably explained as follows. Recalling developed in (from Sect. 3, we used 12 distinct IP addresses to create the 120 socialbots, i.e., 10 accounts were operated from each IP address. Hence, by the time the last few accounts



(a) Age of user account **(b)** Number of tweets posted **(c)** Number of followers

Fig. 2 Comparing the three groups of target users: CDFs for (i) age of the user accounts, (ii) number of tweets posted by the users, and (iii) number of followers of the users in the three target groups (Color figure online)



Group 1		Group 2		Group 3	
Male	Female	Male	Female	Male	Female
🟦 Bot 1	🟦 Bot 2	🟦 Bot 3	🟦 Bot 4	🟦 Bot 5	🟦 Bot 6
🟦 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
🟦 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
🟦 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
🟥 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. 3 Distribution of attributes of the 120 socialbots, numbered in the chronological 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 repost others' tweets (Color figure online)

were created, Twitter's defenses had probably become suspicious of several accounts being created from the same block of IP addresses. In addition, 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 of the socialbots which were created early, and which simply reposted others' tweets.

Note that since we ensured that our socialbots do *not* engage in any spam activity (as stated in Sect. 3), Twitter is justified in not suspending the accounts since their rules (twitter-rules 2015) 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. The danger is that, though such socialbots are not violating the Twitter rules, they might be used for malicious objectives like influencing political campaigns (Orcutt 2012).

4.2 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 (Cha et al. 2010). As stated in Sect. 3, 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.

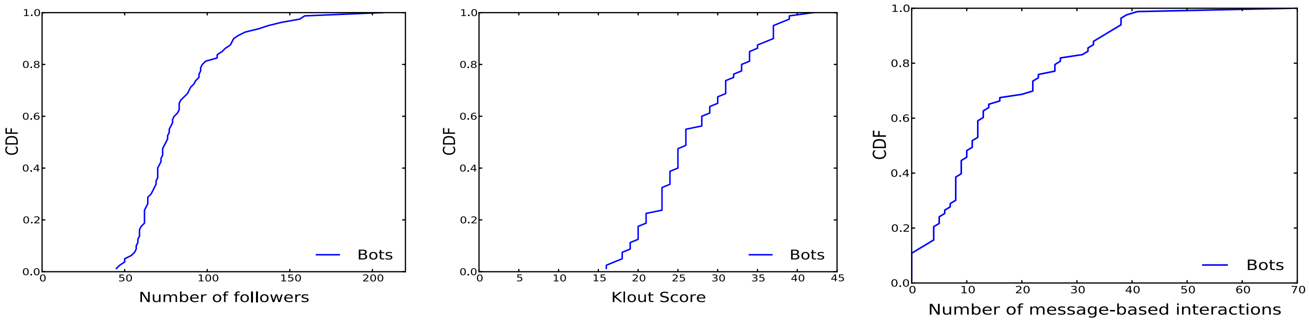
(2) Klout score: Klout score (klout 2015) is a popular measure for online influence. Though the exact algorithm for the metric is not known publicly, the Klout score for a given user is known to consider various data points from Twitter (and other OSNs, if available), such as the number of followers and followings of the user, retweets, membership of the user in Lists, how many spam/dead accounts are following the user, how influential are the people who retweet/mention the user, and so on (klout-wiki 2015). 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. This metric estimates the *social engagement* of the bot, which is defined as the extent to which a user participates in a broad range of social roles and relationships (William and Avison 2007).

Over the duration of the experiment, our 120 socialbots received in total 4601 follows from 1952 distinct users, and 1991 message-based interactions from 1187 distinct users.

Table 2 Distribution of various types of message-based interactions received by the socialbots.

Sl. No.	Type of message-based interaction	Percentage (%)
1	A tweet posted by a socialbot got retweeted	29.15
2	A tweet posted by a socialbot got favorited	33.19
3	A tweet posted by a socialbot got replied to	7.26
4	A socialbot got @mentioned in a tweet	12.94
5	A socialbot received a directed message	6.95
6	A tweet @mentioning a socialbot got retweeted	5.13
7	A tweet @mentioning a socialbot got favorited	5.37



(a) Number of followers

(b) Klout Score

(c) Number of interactions

Fig. 4 Performance of our socialbots: CDFs for (i) number of followers acquired, (ii) Klout Score, and (iii) number of message-based interactions with other users

More specifically, the socialbots received seven distinct types of message-based interactions, as shown in Table 2. The table also reports the distribution of various types of message-based interactions received by the socialbots. Most of the interactions were due to tweets posted by the socialbots being retweeted (29.15 %) and favorited (33.2 %). Additionally, the socialbots were mentioned in a tweet (posted by some other user) in about 13 % of the cases. Out of the various types of message-based interactions, some can be considered to be more active than the others. For instance, actually posting a tweet mentioning a certain user can be considered a more active form of interaction than just favoriting a tweet posted by that user. The numbers in Table 2 show that socialbots are able to obtain not only passive interactions, but also various types of active message-based interactions supported in Twitter.

Note that, in this article, we only considered the first five types of interactions listed in Table 2. Though we did not consider the other two types of interactions in our analysis, all the interactions are included in the dataset that we make publicly available.

Figure 4 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 (Fig. 4a); it can be noted that 46 % of all users in Twitter have less than 100 followers (46 % of Twitter users 2014).

Figure 4b shows that 20 % of the socialbots acquired Klout scores higher than 35 within only one month. We focus on the three socialbots that acquired the highest Klout scores. We note that these three socialbots have common characteristics—gender specified as 'female', highly active, used only reposting as the mechanism for tweeting, and followed Group 2 of target users. Table 3 compares the Klout scores acquired by these three socialbots with the Klout scores of some well-known researchers in Computer Science, who are also active Twitter users. We find that within just one month, our socialbots could achieve 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 (Messias et al. 2013).

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

Table 3 Comparison of Klout scores of some of our socialbots with well-known researchers and bots developed in the study by Messias et al. (2013).

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* (Messias et al. 2013)	*37.5*
fepessoinha	*Bot developed in* (Messias et al. 2013)	*12.3*

The socialbots developed as part of the present study are indicated in bold, while the socialbots developed in Messias et al. (2013) are indicated in italics

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 (Ghosh et al. 2012).

5 Assessing effectiveness of socialbot configurations

The previous section showed that a large fraction of the socialbots could successfully infiltrate the Twitter social network. This section analyzes the effectiveness/impact of the different strategies or attributes (gender, tweet posting method, activity level, target group) on the infiltration performance of the socialbots. Note that the results stated in this section (and the next) consider only those socialbots which were not suspended by Twitter during the experiment (as described in Sect. 4).

We present a *factorial design experiment* to assess the relative impact of the different infiltration strategies. We begin by briefly describing how we designed our experiments, and then discuss the obtained results.

5.1 2^k Factorial experiment

We here include a brief description of the theory of 2^k factorial experiments; we refer the reader to (Jain 1991) for a comprehensive description.

An experimental design strategy is usually necessary in scenarios with a large number of factors, as an attempt to reduce the number of factors that will be part of the experiment. Particularly, 2^k designs refer to experimental designs with k factors where each factor has the minimal number of levels, just two. As an illustrative example, suppose an experimental performance scenario in which three factors—memory, disk, and CPU of a machine—can potentially affect the performance of an algorithm. Suppose now that each experiment takes about one day to run and there are 10 possible types of memory, 10 types of disks, and 10 types of CPUs to be tested. Running an experiment with all possibilities would take $10 \times 10 \times 10 = 1000$ days. Instead of running all possibilities, a 2^k design would consider two (usually extreme) types of memory, two types of disk, and two types of CPUs to compare, which would result in only $2^3 = 8$ days of experiments. The theory of factorial experiments (Jain 1991) would then allow one to estimate how much each factor impacts on the final result, a key information to help decide on which factors an experiment should focus.

Note that, differently of the above example, our goal here is *not* primarily to reduce the number of experiment scenarios. Instead we use a 2^k design to infer how much a factor—which, in our case, correspond to attributes like gender, activity level, and posting method—impacts the different infiltration metrics.

5.2 Factorial experiment on socialbot configuration

For certain applications, the objective of socialbots might be to infiltrate a particular target group of users. Hence, we here individually consider the success of our socialbots in infiltrating each of the three target groups (which were described in Sect. 3). For each target group, we consider the three infiltration metrics stated earlier—the number of followers acquired, the number of message-based interactions and the Klout score. Then, for each metric and each target group, we executed a 2^3 design considering the attributes and their values as described in Table 4, resulting in $3 \times 3 \times 2^3 = 216$ experiments. We performed experiments that associates $+1$ or -1 for the strategies employed for each attribute. All experimental configurations for all datasets were averaged over 5 results, which is the number of socialbots in each configuration.

The basic idea of the factorial design model consists of formulating y, the infiltration impact, as a function of a number of factors and their possible combinations, as defined by Eq. 1. Here, GP, AP, AG, and GAP account for

Table 4 Factors used in the factorial design experiment for analyzing the infiltration performance of the socialbots.

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

all possible combinations among the factors. 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’).

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

where $F = \{\mathrm{G, A, P, GA, GP, AP, GAP}\}$ and x_i is defined as follows.

$$x_G = \begin{cases} -1 & \text{if female} \\ +1 & \text{if male} \end{cases}$$

$$x_A = \begin{cases} -1 & \text{if low activity} \\ +1 & \text{if high activity} \end{cases}$$

$$x_P = \begin{cases} -1 & \text{if repost} \\ +1 & \text{if repost + Markov} \end{cases}$$

and the x_is for the feature combinations (e.g., AG, GP) are defined from the values of x_G, x_A, and x_P following the standard way described in the study by Jain (1991) (details omitted for brevity).

In the above equation, Q_i is the infiltration performance (according to a certain metric like number of followers, or Klout score) when strategy $i \in F$ is applied, and Q_0 stands for the average infiltration performance, averaged over all possible features and their combinations. By empirically measuring y according to different feature combinations (which, in our case, refer to the various socialbot strategies), 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 infiltration performance.

Instead of presenting results for all possible values of Q_i, we focus on the *variations of Q_i due to changes in the features (or their combinations)*, which helps to estimate the importance of a particular factor to the final result. As an example, if we find that a factor accounts for only 1 % of total variation on the results, we can infer that this attribute is unimportant for infiltrating Twitter with a socialbot.

As proposed in the study by Jain (1991), the importance of the various factors can be quantitatively estimated by assessing the *proportion of the total variation in the final result that is explained by each factor*. To compute this variation, we first consider the variation of y (as defined by Eq. 1) 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 as the *variation only due to the changes on factor i*, which can be computed similar to SS_T, but considering only those runs in which the values of the factor i were changed. Finally, we calculate the fraction of variation due to factor i as $\frac{\mathrm{SS}_i}{\mathrm{SS}_T}$. We now use this metric to compute the impact of each attribute for different infiltration metrics and groups of target users.

5.3 Analyzing bot configurations

Table 5 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 it 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.

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.

Table 5 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)

	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

The statistics which have been specifically discussed in the text (Sect. 5.3) are shown in bold

6 Evaluating individual infiltration strategies

In this section, we perform a more fine-grained analysis of the impact of each individual attribute on the infiltration performance of the socialbots. Different from the previous section, here we compare the performance of each distinct value of an attribute (e.g., ‘male’ and ‘female’ for the attribute gender, or ‘high activity’ and ‘low activity’ for the activity attribute) Moreover, while Sect. 5 studied only the infiltration performance at the end of the one-month duration of the experiment, here we study how the performance of different attributes evolves over time.

6.1 Gender

We start by analyzing the impact of the gender of the socialbots in our experiments. Figure 5a, b, c, respectively, shows the mean number of followers, the Klout score, and the number of message-based interactions acquired by the male and female socialbots over each day during our experiment. In these figures, the curves represent the mean values considering all the socialbots of a particular gender (on a given day during the experiment), and the error bars indicate the 95 % confidence intervals of the mean values.

We find that there is no significant difference in the popularity acquired by socialbots of different genders. Note, however, that here we are considering all the target users together. In the previous section, where we separately analyzed the performance of socialbots in infiltrating each group of target users, we saw that the gender is significant for target group 3, but not for the other groups. Thus, we conclude that the gender specified in the account profile can affect the infiltration performance for certain groups of target users, but not for others.

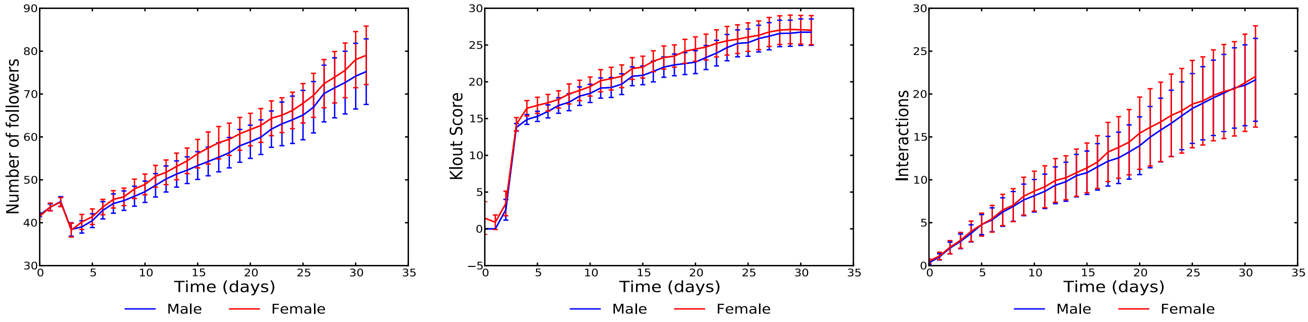
Also note the slight fall in the mean number of followers of the socialbots during the first 2–3 days of the experiment (Figure 5a). This initial fall in the number of followers will be observed in the later analysis as well, and is explained as follows. During the first few days, our socialbots were followed by some unknown user accounts which presumably were other bots/automated accounts. However, as described in Sect. 3, our socialbots did *not* follow back or interact with these accounts. Hence, these accounts unfollowed our socialbots after few days, which resulted in a temporary fall in the number of followers of our socialbots. A similar phenomenon of bot accounts following and unfollowing Twitter users has been observed in prior works (Lee et al. 2010, 2011).

6.2 Activity level

We next study the impact of the socialbots’ activity levels, which we define as *low* or *high* based on how frequently a socialbot posts tweets and follows users.

Figure 6a, b, respectively, shows the mean number of followers and mean Klout scores of the socialbots having two different levels of activity, on each day during the experiment. We can see that socialbots with higher activity levels achieve significantly more popularity and Klout score than the less active socialbots. Figure 6c shows the mean number of message-based interactions of socialbots with other users in Twitter. Again, the more active socialbots achieved much more interactions.

Thus, we find that the more active are the bots, the more likely they are to be successful in infiltration tasks, as well as in gaining popularity in the social network. This is expected, since the more active a bot is, the higher is the likelihood of its being visible to other users. However, it must also be noted that the more active a bot is, the more likely it is to be detected by Twitter’s defense mechanisms.



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

Fig. 5 Infiltration performance of socialbots of different genders through the duration of the experiment: (i) mean number of followers acquired, (ii) mean Klout score acquired, and (iii) mean number of message-based interactions with other users. The *curves* represent the mean values, which the *error bars* indicated the 95 % confidence intervals (Color figure online)

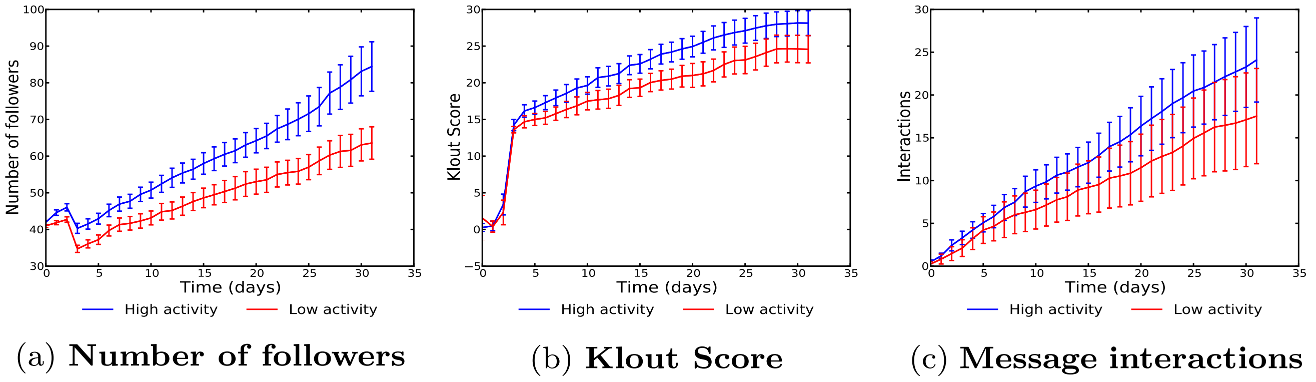


Fig. 6 Infiltration performance of socialbots having different activity levels: (i) mean number of followers, (ii) mean Klout score, and (iii) mean number of message-based interactions with other users (Color figure online)

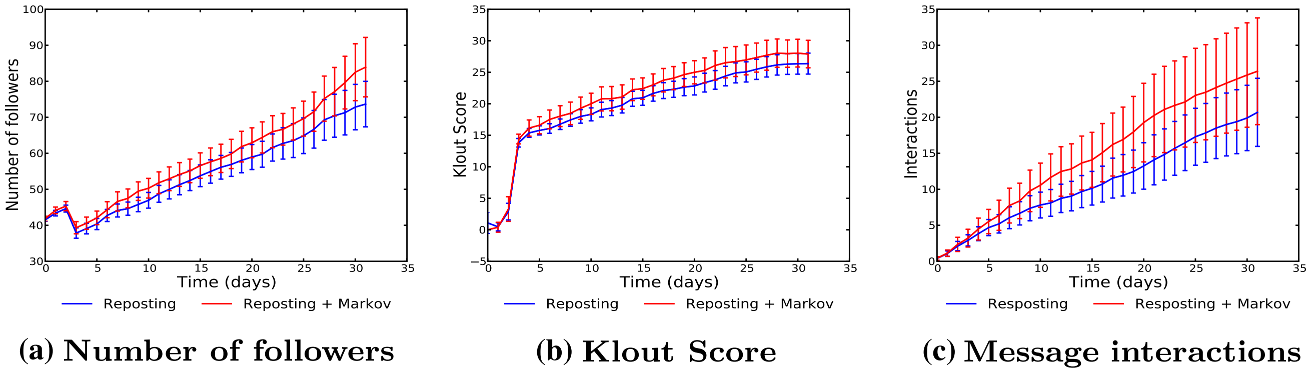


Fig. 7 Infiltration performance of socialbots employing different methodologies to generate tweets: (i) mean number of followers acquired, (ii) mean Klout score, and (iii) mean number of message-based interactions with other users (Color figure online)

6.3 Tweet generating method

We next analyze the impact of the tweet generating method used by the socialbots. Recall from Sect. 3 that half of our socialbots only reposted tweets written by other users (strategy denoted as ‘reposting’), while the other half reposted tweets as well as synthetically generated tweets using a Markov generator, with equal probability (strategy denoted as ‘reposting + Markov’).

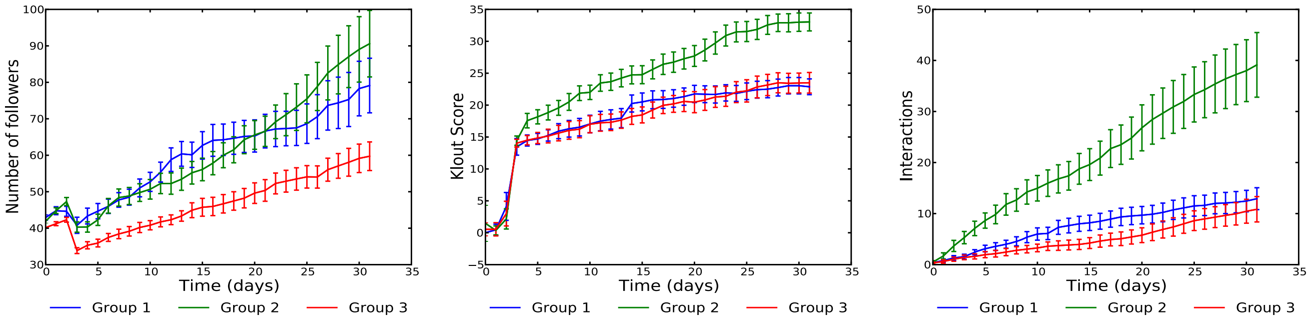
Figure 7a, b, c, respectively, shows the mean number of followers, mean Klout scores, and the mean number of message-based interactions acquired by the socialbots employing the two posting strategies (on each day during the experiment). It is seen that the socialbots employing the ‘reposting + Markov’ strategy acquired marginally higher levels of popularity (number of followers and Klout scores), and much higher amount of interactions (social engagement) with other users.

The fact that socialbots which automatically generated about half of their tweets achieved higher social engagement is surprising, since it indicates that users in Twitter are not able to distinguish between (accounts which post) human-generated tweets and automatically generated tweets using simple statistical models. This is possibly because a large fraction of tweets in Twitter are written in an informal, grammatically incoherent style (Kouloumpis et al. 2011), so that even simple statistical models can produce tweets with quality similar to those posted by humans in Twitter.

6.4 Target users

Finally, we analyze the infiltration performance of socialbots who were assigned different sets of target users to follow. Recall from Sect. 3 that the socialbots were divided into three groups based on the target set—Group 1 followed users selected at random, Group 2 followed target users who post tweets on a specific topic (related to software development), and Group 3 of socialbots followed target users who not only post tweets on the specific topic but are also socially well connected among themselves.

Figure 8a shows the average number of followers acquired by each group of socialbots throughout the experiment. It is seen that the socialbots in Group 3 had the



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

Fig. 8 Infiltration performance of socialbots which followed different sets of target users: (i) mean umber of followers acquired, (ii) mean Klout score, and (iii) mean number of message-based interactions with other users (Color figure online)

lowest number of followers, while those in Group 2 had a significantly higher number of followers. Figure 8b shows the average values of Klout score achieved by our socialbots over time. Again, the socialbots in Group 2 have the highest Klout scores, while the other groups have a similar performance. Figure 8c shows the average number of message-based interactions of each group of socialbots (with other Twitter users) over time. Again, we find that socialbots in Group 2 got significantly more interactions with other users, and those in Group 3 got the least interactions.

These observations lead to some interesting insights. From the perspective of socialbots, following a set of users who post tweets on a specific common topic (e.g., software development) is a more promising approach than following random users (as done by Group 1). However, although the target users for both Group 2 and Group 3 post tweets on a common topic, the socialbots in Group 2 achieved significantly higher popularity and engagement—this implies that infiltrating into *interconnected* groups of users (Group 3) is far more difficult than engaging with users without any relation among themselves (Group 2).

Summary: The analysis in this section gives some interesting insights on the impact of various attributes on the infiltration performance of socialbots. While certain attributes—such as the gender mentioned in the profile, and the tweet posting strategy—do *not* significantly affect infiltration performance, other attributes such as the activity level and the choice of the target users have large impact upon infiltration performance.

7 Concluding discussion

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 (Chu et al. 2012; 20M-fake-users-twitter 2013), and various forms of spam attacks—such as link farming (Ghosh et al. 2012), search spam (Benevenuto et al. 2010) and phishing (Chhabra et al. 2011)—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. Specifically, we created 120 socialbots in the Twitter social network, and quantified the extent to which different socialbot strategies impact their social acceptance in Twitter.

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 (twitter-shut-spammers 2012)] 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 reposting 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 to prevent bots from becoming influential.

Note that in this work, we studied four features which intuitively affect how successful a socialbot is in infiltrating the social network. However, other features might also determine the success of socialbots, and we leave it as a potential future work to extend the study to other features.

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 (Ratkiewicz et al. 2011). Particularly, there are numerous concerns that socialbots may influence political campaigns, such as trying to change the “trending topics” during elections (Orcutt 2012). In fact, Reuters even launched an internet campaign for political candidates to not use socialbots (reuters-botsban 2014). 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.

Acknowledgments This work was partially supported by the project FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web, process number APQ-01400-14, and individual grants from CNPq, CAPES, and Fapemig.

References

46 % of Twitter users have less than 100 followers-Simplify360 (2014) http://simplify360.com/blog/46-of-twitter-users-have-less-than-100-followers/. Accessed 1 May 2014

Ahmad MA, Ahmed I, Srivastava J, Poole MS (2011) Trust me, i’m an expert: trust, homophily and expertise in mmos. In: International conference on privacy, security, risk and trust (passat) and international conference on social computing (socialcom), pp 882–887

Aiello LM, Deplano M, Schifanella R, Ruffo G (2012) People are strange when you’re a stranger: impact and influence of bots on social networks. In: Proceedings of AAAI international conference on web and social media (ICWSM)

Barbieri G, Pachet F, Roy P, Esposti MD (2012) Markov constraints for generating lyrics with style. In: Proceedings of European conference on artificial intelligence

Benevenuto F, Magno G, Rodrigues T, Almeida V (2010) Detecting spammers on twitter. In: Proceedings of annual collaboration, electronic messaging, anti-abuse and spam conference (CEAS)

Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M (2011) The socialbot network: when bots socialize for fame and money. In: Proceedings of annual computer security applications conference (ACSAC)

Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in twitter: the million follower fallacy. In: Proceedings of AAAI international conference on web and social media (ICWSM)

Chandra J, Scholtes I, Ganguly N, Schweitzer F (2012) A tunable mechanism for identifying trusted nodes in large scale distributed networks. In: Proceedings of IEEE international conference on trust, security and privacy in computing and communications (TRUSTCOM), pp 722–729

Chhabra S, Aggarwal A, Benevenuto F, Kumaraguru P (2011) Phi.sh/SoCiaL: The phishing landscape through short URLs. In: Proceedings of collaboration, electronic messaging, anti-abuse and spam conference (CEAS)

Chu Z, Gianvecchio S, Wang H, Jajodia S (2012) Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Trans Dependable Secure Comput 9(6):811–824

Coburn Z, Marra G (2008) Realboy: believable twitter Bots. http://ca.olin.edu/2008/realboy/. Accessed 1 Dec 2015

Creating a bot on wikipedia (2015) http://en.wikipedia.org/wiki/Wikipedia:Creating_a_bot. Accessed 1 Dec 2015

Edwards J (2013) 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. Accessed 1 Dec 2013

Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2014) The rise of social bots. arXiv:1407.5225

Freitas C, Benevenuto F, Ghosh S, Veloso A (2015) Reverse engineering socialbot infiltration strategies in twitter. In: Proceedings of ACM/IEEE international conference on advances in social networks analysis and mining (ASONAM)

Ghosh S, Viswanath B, Kooti F, Sharma NK, Korlam G, Benevenuto F, Gummadi, K. P. (2012) Understanding and combating link farming in the twitter social network. In: Proceedings of World Wide Web Conference (WWW)

Ghosh S, Zafar MB, Bhattacharya P, Sharma N, Ganguly N, Gummadi K (2013) On sampling the wisdom of crowds: random vs. expert sampling of the twitter stream. In: Proceedings of ACM conference on information knowledge management (CIKM)

Gyöngyi Z, Garcia-Molina H (2005) Link spam alliances. In: Proceedings of international conference on very large data bases (VLDB)

Jain R (1991) The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling. Wiley, London

Jurafsky D, Martin JH (2000) Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 1st edn. Prentice Hall PTR, Englewood Cliffs

Klout—The standard for influence (2015) http://klout.com/. Accessed 1 Dec 2015

Klout—wikipedia (2015) http://en.wikipedia.org/wiki/Klout. Accessed 1 Dec 2015

Kouloumpis E, Wilson T, Moore J (2011) Twitter sentiment analysis: the good, the bad and the OMG! In: Proceedings of AAAI international conference on web and social media (ICWSM)

Lee K, Caverlee J, Webb S (2010) Uncovering social spammers: social honeypots + machine learning. In: Proceedings ACM SIGIR conference on research and development in information retrieval (SIGIR)

Lee K, Eoff BD, Caverlee J (2011) Seven months with the devils: a long-term study of content polluters on twitter. In: Proceedings of AAAI international conference on web and social media (ICWSM)

Let’s make candidates pledge not to use bots (2014) http://blogs.reuters.com/great-debate/2014/01/02/lets-make-candidates-pledge-not-to-use-bots/. Accessed 1 Dec 2015

Messias J, Schmidt L, Rabelo R, Benevenuto F (2013) You followed my bot! Transforming robots into influential users in Twitter. First Monday 18(7) http://firstmonday.org/ojs/index.php/fm/article/view/4217

Orcutt M (2012) Twitter mischief plagues Mexico’s election. http://www.technologyreview.com/news/428286/twitter-mischief-plagues-mexicos-election/. Accessed 1 Dec 2015

Pandora Bots (2015) http://www.pandorabots.com/. Accessed 1 Dec 2015

Pitsillidis A, Levchenko K, Kreibich C, Kanich C, Voelker GM, Paxson, V, Weaver N, Savage S (2010) Botnet judo: fighting spam with itself. In: Proceedings of symposium on network and distributed system security (NDSS), San Diego, CA

Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Patil S, Flammini A, Menczer F (2011) Truthy: mapping the spread of astroturf in microblog streams . In: Proceedings of World Wide Web Conference (WWW)

Roy A, Ahmad MA, Sarkar C, Keegan B, Srivastava J (2012) The ones that got away: false negative estimation based approaches for gold farmer detection. In: International conference on privacy, security, empirical study of socialbots in twitter 27 risk and trust (passat) and international conference on social computing (socialcom), pp 328–337)

Shane S, Hubbard B (2014) ISIS displaying a deft command of varied media. http://www.nytimes.com/2014/08/31/world/middleeast/isis-displaying-a-deft-command-of-varied-media.html. Accessed 1 Dec 2015

Shutting down spammers (2012) https://blog.twitter.com/2012/shutting-down-spammers. Accessed 1 Dec 2015

Stone-Gross B, Holz T, Stringhini G, Vigna G (2011) The underground economy of spam: a Botmaster's perspective of coordinating large-scale spam campaigns. In: Proceedings of USENIX conference on large- scale exploits and emergent threats (LEET)

Subrahmanian VS, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, Waltzman R et al. (2016) The DARPA twitter bot challenge. arXiv:1601.05140

The twitter rules—twitter help center (2015) https://support.twitter.com/articles/18311#. Accessed 1 Dec 2015

Viswanath B, Post A, Gummadi KP, Mislove A (2010a) An analysis of social network-based Sybil defenses. ACM SIGCOMM Comput Commun Rev 40(4):363–374

Viswanath B, Post A, Gummadi KP, Mislove A (2010b) An analysis of social network-based sybil defenses. SIGCOMM Comput Commun Rev 40(4):363–374

Wagner C, Liao V, Pirolli P, Nelson L, Strohmaier M (2012) It's not in their tweets: modeling topical expertise of twitter users. In: Proceedings of AASE/IEEE international conference on social computing (SocialCom)

Wagner C, Mitter S, Körner C, Strohmaier M (2012) When social bots attack: modeling susceptibility of users in online social networks. In: Proceedings of workshop on making sense of microposts (with WWW)

Wald R, Khoshgoftaar TM, Napolitano A, Sumner C (2013) Which users reply to and interact with twitter social bots? In: Proceedings of IEEE conference on tools with artificial intelligence (ICTAI)

Web Ecology Project (2015) http://www.webecologyproject.org/

William R, Avison JDM (eds) BAP (2007) Mental health, social mirror. Springer, Berlin