The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race

Stefano Cresci [†‡]
s.cresci@iit.cnr.it

Roberto Di Pietro [§¶†]
roberto.di_pietro@nokia-bell-labs.com

Marinella Petrocchi [†]
m.petrocchi@iit.cnr.it

Angelo Spognardi [‖]
angsp@dtu.dk

Maurizio Tesconi [†]
m.tesconi@iit.cnr.it

[†] Institute for Informatics and Telematics, National Research Council, Italy
[‡] Department of Information Engineering, University of Pisa, Italy
[§] Nokia Bell Labs, France
[¶] Department of Mathematics, University of Padua, Italy
[‖] DTU Compute, Denmark

ABSTRACT

Recent studies in social media spam and automation provide anecdotal argumentation of the rise of a new generation of spambots, so-called *social spambots*. Here, for the first time, we extensively study this novel phenomenon on Twitter and we provide quantitative evidence that a paradigm-shift exists in spambot design. First, we measure current Twitter's capabilities of detecting the new social spambots. Later, we assess the human performance in discriminating between genuine accounts, social spambots, and traditional spambots. Then, we benchmark several state-of-the-art techniques proposed by the academic literature. Results show that neither Twitter, nor humans, nor cutting-edge applications are currently capable of accurately detecting the new social spambots. Our results call for new approaches capable of turning the tide in the fight against this raising phenomenon. We conclude by reviewing the latest literature on spambots detection and we highlight an emerging common research trend based on the analysis of collective behaviors. Insights derived from both our extensive experimental campaign and survey shed light on the most promising directions of research and lay the foundations for the arms race against the novel social spambots. Finally, to foster research on this novel phenomenon, we make publicly available to the scientific community all the datasets used in this study.

Keywords

Social spambots; Social networks security; Twitter.

1. INTRODUCTION

The widespread availability and ease of use of Online Social Networks (OSN) have made them the ideal setting for the proliferation of fictitious and malicious accounts [29]. Indeed, recent work uncovered the existence of large numbers of OSN accounts that are purposely created to distribute unsolicited spam, advertise events and products of doubtful legality, sponsor public characters and, ultimately, lead to a bias within the public opinion [16, 24]. Moreover, the plague of such spammers and bots leads to an ingenious and lucrative "underground economy", where account vendors, their customers, and oblivious victims play a piece staging since the very introduction of social networks [34, 36, 38].

One of the most fascinating peculiarities of spambots is that they "evolve" over time, adopting sophisticated techniques to evade early-established detection approaches, such as those based on textual content of shared messages [25], posting patterns [35] and social relationships [18]. As evolving spammers became clever in escaping detection, for instance by changing discussion topics and posting activities, researchers kept the pace and proposed complex models, such as those based on the interaction graphs of the accounts under investigation [44, 21].

Noticeably, spambots evolution still goes on. Recent investigations anecdotally highlight how new waves of *social spambots* are rising [16, 48]. In this paper, we target these new waves, finding evidence of the difficulties for OSN users to distinguish between genuine and malicious accounts. We also highlight the difficulties for OSN administrators to take appropriate countermeasures against the takeover of evolving spambots. Remarkably, a large number of tools and techniques have been proposed by Academia to detect OSN spambots [16, 24]. Until recently, such tools have proved to be valid allies for spambots timely detection. Unfortunately, the characteristics of the new wave of social spambots are such that standard classification approaches, where a single account is evaluated according to a set of established features tested over known datasets, are no longer successful. In this work, we demonstrate this claim by investigating the performances of several state-of-the-art tools techniques when struggling against the latest wave of social spambots. The

			statistics			
dataset	description		accounts	tweets	year	used in section
`genuine accounts`	verified accounts that are human-operated		3,474	8,377,522	2011	3.1, 3.2
`social spambots #1`	retweeters of an Italian political candidate		991	1,610,176	2012	3.1, 3.2
`social spambots #2`	spammers of paid apps for mobile devices		3,457	428,542	2014	3.1, 3.2
`social spambots #3`	spammers of products on sale at *Amazon.com*		464	1,418,626	2011	3.1, 3.2
`traditional spambots #1`	training set of spammers used by Yang *et al.* in [44]		1,000	145,094	2009	3.1
`traditional spambots #2`	spammers of scam URLs		100	74,957	2014	3.1
`traditional spambots #3`	automated accounts spamming job offers		433	5,794,931	2013	3.2
`traditional spambots #4`	another group of automated accounts spamming job offers		1,128	133,311	2009	3.2
`fake followers`	simple accounts that inflate the number of followers of another account		3,351	196,027	2012	3.1
`test set #1`	mixed set of 50% `genuine accounts` + 50% `social spambots #1`		1,982	4,061,598	–	4, 5
`test set #2`	mixed set of 50% `genuine accounts` + 50% `social spambots #3`		928	2,628,181	–	4, 5

Table 1: Statistics about the datasets used for this study.

unsatisfactory results of the surveyed techniques call for new approaches capable of turning the tide of this long-lasting fight.

Interestingly, we assist to a paradigm-shift in modeling and analyzing online accounts. Independently from each other, new research efforts were born, which leverage characteristics of groups of accounts – rather than those of a single account – as a red flag for anomalous behaviors. We provide a review of these prominent research directions, highlighting the new dimensions to sound out for successfully fighting against this novel generation of spambots.

Contributions. Our main contributions are:

- We provide empirical evidence of the existence of a novel wave of Twitter spambots, which, up to now, has been just theorized [16].

- We evaluate if, and to which extent, state-of-the-art detection techniques succeed in spotting such new spambots.

- We critically revise an emerging stream of research, which adopt features tied to groups of accounts rather than individual accounts features.

- We leverage results of a crowdsourcing spambot detection campaign for drawing new guidelines for the annotation of datasets comprising social spambots.

- Finally, we publicly release to the scientific community an annotated dataset[1], consisting of genuine accounts, traditional spambots, and – for the first time – the novel social spambots.

2. DATASETS

We describe the different Twitter datasets that constitute the real-world data used in our experiments. Table 1 reports the name of the datasets, their brief description, and the number of accounts and tweets they feature. The year represents the average of the creation years of the accounts that belong to the dataset.

The `genuine accounts` dataset is a random sample of genuine (human-operated) accounts. Following a hybrid crowd-sensing approach [3], we randomly contacted Twitter users by asking them a simple question in natural language. All

the replies to our questions were manually verified and all the 3,474 accounts that answered were certified as humans. The accounts that did not answer to our question were discarded and are not used in this study.

The `social spambots #1` dataset was created after observing the activities of a novel group of social bots that we discovered on Twitter during the last Mayoral election in Rome, in 2014. One of the runners-up employed a social media marketing firm for his electoral campaign, which made use of almost 1,000 automated accounts on Twitter to publicize his policies. Surprisingly, we found such automated accounts to be similar to genuine ones in every way. Every profile was accurately filled in with detailed – yet fake – personal information such as a (stolen) photo, (fake) short-bio, (fake) location, etc. Those accounts also represented credible sources of information since they all had thousands of followers and friends, the majority of which were genuine users[2]. Furthermore, the accounts showed a tweeting behavior that was apparently similar to those of genuine accounts, with a few tweets posted every day, mainly quotes from popular people, songs, and *YouTube* videos. However, every time the political candidate posted a new tweet from his official account, all the automated accounts retweeted it in a time span of just a few minutes. Thus, the political candidate was able to reach many more accounts in addition to his direct followers and managed to alter Twitter engagement metrics during the electoral campaign. Amazingly, we also found tens of human accounts who tried to engage in conversation with some of the spambots. The most common form of such human-to-spambot interaction was represented by a human reply to one of the spambot tweets quotes.

We also discovered a second group of social bots, which we labeled `social spambots #2`, who spent several months promoting the `#TALNTS` hashtag. Specifically, *Talnts* is a mobile phone application for getting in touch with and hiring artists working in the fields of writing, digital photography, music, and more. The vast majority of tweets were harmless messages, occasionally interspersed by tweets mentioning a specific genuine (human) account and suggesting him to buy the VIP version of the app from a Web store.

Further, we uncovered a third group of social bots, `social spambots #3`, which advertise products on sale on *Amazon.com*. The deceitful activity was carried out by spam-

[2]This was made possible also by the adoption of social engineering techniques, such as the photo of a young attractive woman as the profile picture and the occasional posting of provocative tweets.

dataset	accounts			
	total	alive	deleted	suspended
genuine accounts	3,474	3,353 (96.5%)	115 (3.3%)	6 (0.1%)
social spambots #1	994	946 (95.2%)	2 (0.2%)	46 (4.6%)
social spambots #2	3,457	3,322 (96.1%)	1 (0.1%)	134 (3.8%)
social spambots #3	467	465 (99.6%)	2 (0.4%)	0 (0.0%)
traditional spambots #1	1,000	889 (88.9%)	25 (2.5%)	86 (8.6%)
traditional spambots #2	100	1 (1.0%)	0 (0.0%)	99 (99.0%)
fake followers	3,351	851 (25.4%)	38 (1.1%)	2,462 (73.5%)

Table 2: Statistics about alive, deleted, and suspended accounts, for different groups of genuine and malicious accounts.

ming URLs pointing to the advertised products. Similarly to the retweeters of the Italian political candidate, also this family of spambots interleaved spam tweets with harmless and genuine ones.

We exploited a Twitter crawler to collect data about all the accounts we suspected to belong to the three groups of social spambots. All the accounts collected in this process have then undergone an internal manual verification phase to certify their automated nature. Among all the distinct retweeters of the Italian political candidate, 50.05% (991 accounts) were certified as spambots. Similarly, 94.50% (3,457 accounts) of the accounts who tweeted the `#TALNTS` hashtag resulted as spambots. Finally, 89.29% (464 accounts) of the accounts that tweeted suspicious *Amazon.com* URLs were also certified as spambots. The three sets of accounts represent our ground truth of novel social spambots.

Our internal manual annotation has been carried out by comparing every account to all the others, in order to highlight possible similarities and common behaviors. This is in contrast with the typical annotation process where accounts are labeled one-by-one and by solely exploiting the characteristics of the account under investigation.

In addition to genuine users and social spambots, we also collected several datasets of traditional spambots. Such datasets are used throughout the paper as a strong baseline. The `traditional spambots #1` dataset is the training set used in [44], kindly provided to us by the authors of that work. In [44], the dataset has been used to train a machine learning classifier for the detection of evolving Twitter spambots. Accounts belonging to the `traditional spambots #2` dataset are rather simplistic bots that repeatedly mention other users in tweets containing scam URLs. To lure users into clicking the malicious links, the content of their tweets invite the mentioned users to claim a monetary prize. The `traditional spambots #3` and `traditional spambots #4` datasets are related to 2 different groups of bots that repeatedly tweet about open job positions and job offers.

Fake followers are another kind of malicious accounts that recently gained interest both from platform administrators and from the scientific world [13]. Given that fake followers are rather simplistic in their design and functioning, they can serve as a weak baseline against which to compare social spambots. In April, 2013, we bought 3,351 fake accounts from three different Twitter online markets, namely *fastfollowerz.com*, *intertwitter.com*, and *twittertechnology.com*. All the accounts acquired in this way have been merged in order to obtain the `fake followers` dataset used in this study.

By considering a diverse set of spammer accounts we have captured many of the different dimensions currently exploited by spambots and tamperers to perpetrate their illicit activ-

ities. In detail, we have considered (i) fake follower frauds, (ii) retweet frauds, (iii) hashtag promotion, (iv) URL spamming, (v) scamming, and (vi) spam of generic messages.

3. REAL-WORLD EXPERIMENTATION

3.1 Twitter monitoring

A first assessment of the extent and the severity of Twitter social spambots problem can be obtained by measuring Twitter's capacity of detecting and removing them from the platform. This section thus answers the research question:

RQ1 − *To what extent is Twitter currently capable of detecting and removing social spambots?*

Interesting insights can be gained by comparing the rate at which Twitter accounts are removed, for different types of malicious accounts. The intuition is that accounts that are easily identified as malicious can be rapidly removed by platform administrators. Thus, in this experiment, we let different types of accounts behave for a rather long amount of time (i.e., years). Then, we check whether Twitter managed to identify such accounts as malicious and to remove them from the platform. We perform this experiment on our set of genuine accounts, on our 3 groups of social spambots, on 2 groups of traditional spambots, and on the group of fake followers.

In order to perform this experiment, we exploited Twitter's responses to API calls and, particularly, the Twitter error codes. Given a query to a specific account, Twitter's API replies with information regarding the status of the queried account. Specifically, accounts that are suspected to perform malicious activities get suspended by Twitter. API queries to a suspended account result in Twitter responding with the error code 63. API queries to accounts that have been deleted by their original owner result in Twitter responding with the error code 50. Instead, for accounts that are neither suspended nor deleted, Twitter replies with the full metadata information of the account, without issuing error codes. By exploiting this response mechanism, we were able to measure the *survivability* of the different groups of accounts. Results of this experiment are reported in Table 2 and are pictorially depicted in Figure 1.

As shown in Table 2, `genuine accounts` feature a very high survival rate (96.5%). In addition, among the no longer available accounts, the vast majority have been deleted by the original owner, rather than suspended by Twitter. These results are quite intuitive, by considering that legitimate accounts rarely perform any kind of malicious activity. Conversely, the simplest kind of malicious accounts, `fake followers`, have mostly been detected and suspended by Twitter. The same also applies to one of the two groups of
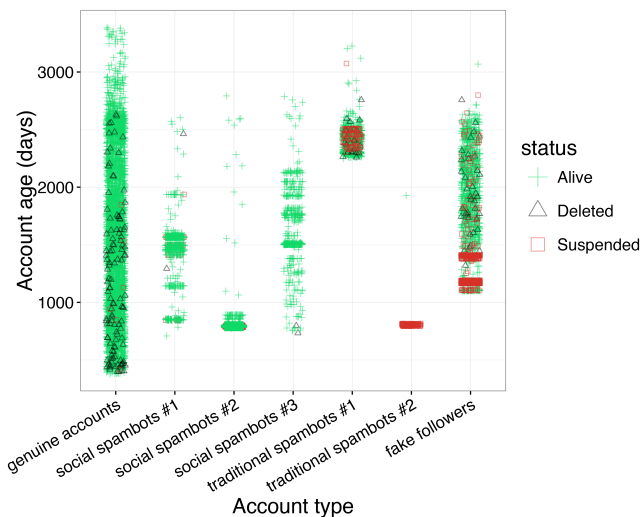
Figure 1: Survival rates for different types of accounts.

traditional spambots, identified as `traditional spambots #2` in Table 2, which features a suspension rate as high as 99%. The most interesting results are however related to those kinds of malicious accounts that better mimic human behaviors. So far, `traditional spambots #1` have largely managed to evade suspension, despite dating back to 2009. Indeed, only 8.6% of the bots have been suspended, while 88.9% of them are still alive. This seems to suggest that Twitter's spambot detection mechanisms are still unable to accurately identify such accounts, while recent solutions proposed by Academia have succeeded in this task [44]. Twitter's performance in suspending malicious accounts is even worse if we consider social spambots. All the 3 groups of social spambots feature very high survival rates, respectively 95.2%, 96.1%, and 99.6%. Even if the difference between the survival rate of social spambots and that of `traditional spambots #1` is marginal, these results nonetheless suggest an increased difficulty for the detection of social spambots. Table 3 also reports the results of a comparison between the ratios of alive, deleted, and suspended accounts between spambots and `genuine accounts`. As shown, social spambots feature very small differences with respect to genuine accounts ($\sim \pm 3\%$). Some of these differences are not even statistically significant, according to a chi-square test. `Traditional spambots #1` have differences $\sim \pm 8\%$ that are highly significant ($p < 0.01$) for alive and suspended accounts. Instead, `traditional spambots #2` and `fake followers` show massive differences: $\sim \pm 96\%$ and $\sim \pm 72\%$, respectively.

Figure 1 shows results of the survivability experiment, with respect to the account age[3]. This can allow to understand if temporal patterns exist in the way malicious accounts are created, and if Twitter's mechanisms for suspending malicious accounts are related to an account's age. For instance, Twitter might be better in detecting and suspending older accounts than newer ones. However, this hypothesis can be ruled out by considering that 99% of `traditional spambots #2` accounts have been suspended despite

---

[3] Account age is computed as the number of days between the account's creation date and the day we performed the experiment.

dataset	accounts		
	alive	deleted	suspended
`social spambots #1`	$-1.3\%^{*}$	$-3.1\%^{***}$	$+4.5\%^{***}$
`social spambots #2`	-0.4%	$-3.2\%^{***}$	$+3.7\%^{***}$
`social spambots #3`	$+3.1\%^{***}$	$-2.9\%^{***}$	-0.1%
`traditional spambots #1`	$-7.6\%^{***}$	-0.8%	$+8.7\%^{***}$
`traditional spambots #2`	$-95.5\%^{***}$	-3.3%	$+98.9\%^{***}$
`fake followers`	$-71.1\%^{***}$	$-2.2\%^{***}$	$+73.4\%^{***}$

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 3: Effect size and statistical significance of the difference between the survivability results of malicious accounts with respect to those of `genuine accounts`.

being younger than most of the social spambots. Overall, an analysis of Figure 1 shows that account suspensions seem to depend on the type of the account, its design and behavior, rather than on its age.

Results reported in this first experiment already reveal interesting differences between social spambots, traditional spambots, and fake followers. Notably, social spambots appear to be more similar to genuine accounts than to traditional spambots, with regards to Twitter suspensions.

3.2 Crowdsourcing: tasks and results

This section addresses the following research questions:

RQ2 − *Do humans succeed in detecting social spambots in the wild?*

RQ3 − *Do they succeed in discriminating between traditional spambots, social spambots, and genuine accounts?*

Even if Twitter users were generally capable of distinguishing between traditional spambots and genuine accounts, they might still find it difficult to spot social spambots in the wild. If confirmed, this would provide additional evidence of the evolutionary step characterizing the new social spambots with respect to traditional ones.
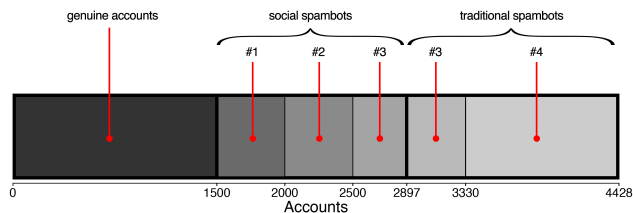

Figure 2: Dataset composition for the crowdsourcing experiment.

To answer these research questions, we asked a large set of real-world users to classify the accounts in our datasets. To obtain a large and diverse set of users, we recruited contributors from the CrowdFlower[4] crowdsourcing platform. Figure 2 shows the distribution of the 4,428 accounts that we have employed for this crowdsourcing experiment, picked up from the datasets in Section 2. Contributors were asked to assign to each account one of the following classes: (i) spambot, (ii) genuine, and (iii) unable to classify. The latter class (iii) has been inserted to deal with Twitter accounts

---

[4] https://www.crowdflower.com/

Num. accounts to classify	4,428
Min. contributors per account	3
Max. answers per contributor	100
Num. test questions	25
Min. accuracy threshold	70%
Reward	0.1 US$ per 5 accounts classified

Table 4: Crowdsourcing campaign settings.

|	
Instructions clear	4.0 / 5
Test questions fair	3.5 / 5
Ease of job	3.5 / 5
Pay	3.8 / 5
Overall	3.7 / 5

Table 5: Contributors' evaluation of our campaign.

possibly getting deleted, suspended, or protected[5] while our crowdsourcing task was ongoing.

Notably, our experiment marks a difference with those typically carried out with crowdsourcing. In fact, crowdsourcing tasks are typically aimed at creating a ground truth (i.e., labeled) dataset for later use. For instance, crowdsourcing is often used to create large training-sets for machine learning algorithms. Here, instead, the datasets are labeled in advance. Thus, by asking contributors to (re-)classify our datasets, we are actually evaluating their ability to spot the different types of accounts.

Enforcing results reliability. We only recruited contributors who were tech-savvy and Twitter users themselves, in order to be reasonably sure about their knowledge of Twitter and its dynamics. Furthermore, we required each account to be classified by at least 3 different contributors, with the final class decided by majority voting. We also fixed to 100 the upper threshold of the number of accounts that a single contributor could classify. In this way, we have obtained redundant results from a broad set of contributors. Then, in order to further guarantee the reliability of our crowdsourcing results, we designed a set of "test" (or "gold") questions aimed at evaluating the quality of contributors' answers. A test question is one for which the correct answer is already known by the system. Within the crowdsourcing platform, such questions are indistinguishable from standard ones and are randomly mixed among all the questions, so that contributors cannot know whether they are answering to a test or to a standard question. Contributors' answers to test questions were checked against the known correct answers. Only the trusted contributors who answered correctly to more than the 70% of the test questions have been considered in our study. Our test questions consist of accounts whose nature is "easily" recognizable, and specifically: (i) a set of traditional spambots sampled from the dataset of Yang *et al.* [44], (ii) a subset of genuine accounts, and (iii) a set of suspended, deleted, and protected accounts. Notably, by designing test questions with traditional spambots and genuine accounts, and by enforcing the policy of at least 70% correct answers, we can guarantee that all our trusted contributors are typically able to detect traditional spambots and to distinguish them from genuine accounts. This further strengthens the results of their classification of the novel social spambots.

---

[5] *Protected* accounts are those accounts whose tweets and timeline are not publicly visible.
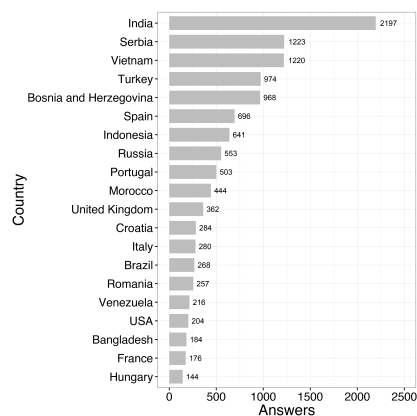
(a) Answers per country (top 20).
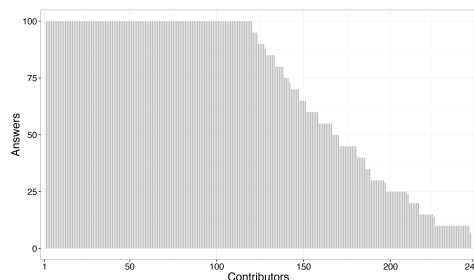
(b) Answers per contributor.

Figure 3: Distribution of crowdsourcing results.

Table 4 shows a recap of the settings used in our crowdsourcing campaign. The thorough description of our campaign, with the complete set of instructions, a list of example accounts, and the task preview, is available online[6]. The campaign completed when each of the 4,428 accounts was classified by 3 different trusted contributors.

Results of the crowdsourcing campaign. Overall, we collected 13,284 answers given by 247 trusted contributors from 42 different countries. Figure 3(a) shows the distribution of answers per country, while Figure 3(b) depicts the distribution of answers per contributor. CrowdFlower also gives contributors the possibility to evaluate crowdsourcing campaigns for: (i) clarity of instructions, (ii) fairness of the test questions, (iii) ease of the task, and (iv) appropriateness of the payment. Out of the 247 participating contributors, 60 of them ($\sim 24\%$) evaluated our campaign, leading to a convincing aggregated score of 3.7/5, as shown in detail in Table 5. Our campaign costed us 410 US$ in total.

The most interesting results of our crowdsourcing campaign are undoubtedly related to the detection performance of our human contributors. As reported in Table 6, overall, the human annotators obtained an accuracy of less than 0.24 on the social spambots, with more than 1,000 False Negatives (FN), meaning that contributors classified more than 1,000 accounts as genuine, when they actually belonged to the dataset of the last generation of spambots. Human detection performances for the two other groups of accounts, namely traditional spambots and genuine accounts, are in-

---

[6] http://wafi.iit.cnr.it/fake/fake/crowdflower/instructions/

type	accounts $^\sharp$	detection results					Fleiss' kappa (κ)
		TP	TN	FP	FN	Accuracy	
traditional spambots	1,516	1,385	0	0	131	0.9136	0.007
social spambots	1,393	328	0	0	1,065	0.2355	0.186
genuine accounts	1,377	0	1,267	110	0	0.9201	0.410

$^\sharp$: The total number of accounts considered is 4,286 instead of 4,428 because 142 accounts (3.2%) got deleted, suspended, or protected during our campaign.

Table 6: Results of the crowdsourcing campaign on spambots detection.

stead quite satisfactory, with an accuracy of 0.91 and 0.92 respectively. These important results further highlight the existence of a striking difference between traditional and social spambots. More worryingly, they also suggest that humans might not be able to detect social spambots in the wild, and to distinguish them from genuine accounts.

Given that each account under investigation has been classified by at least 3 different contributors, we have also computed the Fleiss' kappa (κ) inter-rater agreement metric [20]. All inter-rater agreement metrics measure the level of agreement of different annotators on a task. The level of agreement can be also interpreted as a proxy for the difficulty of a task. In our experiment, human contributors showed a decent agreement for the classification of genuine accounts, with $\kappa = 0.410$. Instead, they showed very little agreement while classifying traditional spambots, as represented by $\kappa = 0.007$. This interesting result shows that, overall, the human contributors were able to correctly detect traditional spambots, as shown by the 0.91 accuracy, but also that contributors rarely agreed on the class. Surprisingly, we measured a slightly higher agreement for the classification of social spambots than for traditional ones, with $\kappa = 0.186$. These results imply that humans generally failed in classifying social spambots (accuracy = 0.2355) and, furthermore, that they also were more in agreement on this mistake than they were when (correctly) classifying traditional spambots.

Annotation guidelines for spambots detection. Despite the recent advances in machine learning-based detection systems, manual verification of accounts to assess their degree of automation is still carried out by platform administrators [16]. In [42], it is reported that human experts "consistently produced near-optimal results" on a dataset of traditional spambots. However, the results of our crowdsourcing experiment confirmed that the traditional "account-by-account" annotation process used by human workers to evaluate social media accounts is no longer viable when applied to the detection of the novel wave of social spambots. Given the importance of manual annotation for the creation of ground-truth datasets and for double-checking suspicious accounts on social networking platforms, we call for the adoption of new annotation methodologies that take into account the similarities and synchronized behaviors of the accounts. We have adopted a practical implementation of this methodology to annotate our datasets of social spambots. In particular, we have compared the timelines of large groups of accounts, in order to highlight tweeting similarities among them. By comparing the behaviors of different accounts, rather than by analyzing them one by one, we were able to spot the social spambots among all the collected accounts, as thoroughly described in Section 2. Therefore, we envisage the possibility to adopt this methodology, as well as similar

ones, in order to safeguard the manual annotation process from elusive social spambots.

4. ESTABLISHED TECHNIQUES

So far, we have demonstrated that neither Twitter nor human operators are currently capable of identifying novel social spambots. Here, we investigate whether established tools and techniques are able to succeed in this task. Thus, our research question is:

RQ4 − *Are state-of-the-art scientific applications and techniques able to detect social spambots?*

The *BotOrNot?* service. BotOrNot? is a publicly-available service[7] to evaluate the similarity of a Twitter account with the known characteristics of social spambots [15]. It has been developed by the Indiana University at Bloomington and it was released in May 2014. Claimed capable of detecting social spambots [16], at the time of writing it was the only publicly-available social spambot detection system. BotOrNot? leverages a supervised machine-learning classifier that exploits more than 1,000 features of the Twitter account under investigation. Specifically, it employs off-the-shelf supervised learning algorithms trained with examples of both humans and bots behaviors, based on the Texas A&M dataset [26] with 15,000 examples of each class and millions of tweets. Similarly to most already established techniques, BotOrNot? performs its analyses on an account-by-account basis. Despite being specifically designed for the detection of social spambots, authors state that the detection performances of BotOrNot? against evolved spambots might be worse than those reported in [15]. Here, we aim at evaluating this point by querying the BotOrNot? service with our sets of genuine and social spambot accounts. As shown in Table 7, BotOrNot? achieves rather unsatisfactory results for the accounts of both `test set #1` and `test set #2` (such datasets are described in Table 1). Its detection performances are particularly bad for the accounts of `test set #1` – where the spambots are from the `social spambots #1` group. The low values of F-Measure and Mathews Correlation Coefficient (MCC), respectively 0.288 and 0.174, are mainly due to the low Recall. In turn, this represents a tendency of labeling `social spambots #1` as genuine accounts.

Supervised spambot classification. Among the many supervised classification approaches to spambot detection proposed in recent years by Academia, we decided to experiment with the one presented by C. Yang *et al.* in [44], since it focuses on the detection of *evolving* Twitter spambots. Thus, it is interesting to evaluate if the system recently presented in [44] is actually able to detect the sophisticated social spambots. This supervised system provides a machine

---

[7] http://truthy.indiana.edu/botornot/

968

technique	type	detection results					
		Precision	Recall	Specificity	Accuracy	F-Measure	MCC
test set #1							
Twitter countermeasures	mixed	**1.000**	0.094	**1.000**	0.691	0.171	0.252
Human annotators	manual	0.267	0.080	0.921	0.698	0.123	0.001
BotOrNot? [15]	supervised	0.471	0.208	0.918	0.734	0.288	0.174
C. Yang *et al.* [44]	supervised	0.563	0.170	0.860	0.506	0.261	0.043
Miller *et al.* [31]	unsupervised	0.555	0.358	0.698	0.526	0.435	0.059
Ahmed *et al.* [2] ♯	unsupervised	0.945	0.944	0.945	0.943	0.944	0.886
Cresci *et al.* [14]	unsupervised	0.982	**0.972**	0.981	**0.976**	**0.977**	**0.952**
test set #2							
Twitter countermeasures	mixed	**1.000**	0.004	**1.000**	0.502	0.008	0.046
Human annotators	manual	0.647	0.509	0.921	0.829	0.570	0.470
BotOrNot? [15]	supervised	0.635	**0.950**	0.981	0.922	0.761	0.738
C. Yang *et al.* [44]	supervised	0.727	0.409	0.848	0.629	0.524	0.287
Miller *et al.* [31]	unsupervised	0.467	0.306	0.654	0.481	0.370	-0.043
Ahmed *et al.* [2] ♯	unsupervised	0.913	0.935	0.912	0.923	**0.923**	0.847
Cresci *et al.* [14]	unsupervised	**1.000**	0.858	**1.000**	**0.929**	0.923	0.867

♯: Modified by employing *fastgreedy* instead of *MCL* for the graph clustering step.

Table 7: Comparison among the spambot detection techniques, tools, and algorithms surveyed in this study. For each test set, the highest values in each evaluation metric are shown in bold.

learning classifier that infers whether a Twitter account is genuine or spambot by relying on account's relationships, tweeting timing and level of automation. We have reproduced such a classifier by implementing and computing all the features proposed in [44], and by training the classifier with its original dataset. Results in Table 7 show that the system fails to correctly classify the novel social spambots. Similarly to the results of the BotOrNot? service, the worst results of this system in both `test set #1` and `test set #2` are related to the Recall metric. This means that also this classifier labeled social spambots as genuine accounts.

Unsupervised spambot detection via Twitter stream clustering. Our initial claim, supported by preliminary work [16, 48], is that social spambots might be so sophisticatedly designed to make it very difficult to distinguish them from genuine accounts, if observed one by one. If demonstrated, this claim would imply that supervised classification approaches are intrinsically worse than unsupervised ones for the detection of social spambots. For this reason, we have also experimented with unsupervised approaches for spambot detection. The approach in [31] considers vectors made of 126 features extracted from both accounts and tweets as input of modified versions of the DenStream [8] and StreamKM++ [1] clustering algorithms, to cluster feature vectors of a set of unlabeled accounts. We have implemented the system proposed in [31] to cluster the accounts of our 2 test sets. As shown in Table 7, this achieved the worst performances among all those that we have benchmarked in this study. Low values of both Precision and Recall mean incomplete and unreliable spambot detection. Among the 126 features, 95 are based on the textual content of tweets. However, novel social spambots tweet contents similar to that of genuine accounts (e.g., retweets of genuine tweets and famous quotes). For this reason, an approach almost solely based on tweet content will not be able to achieve satisfactory results.

Unsupervised spambot detection via graph clustering. The approach in [2] exploits statistical features related to URLs, hashtags, mentions and retweets. Feature vectors generated in this way are then compared with one another

via an Euclidean distance measure. Distances between accounts are organized in an adjacency matrix, which is later used to construct an undirected weighted graph of the accounts. Then, graph clustering and community detection algorithms are applied in order to identify groups of similar accounts. Graph clustering is done by employing the *Markov cluster algorithm* (*MCL*) [39]. We fully implemented this solution and we experimented with our datasets. However, the approach failed to identify 2 distinct clusters, since accounts of both our test-sets were assigned to a single cluster. We also performed a grid search simulation in order to test the best parameter configuration for MCL^8, but to no avail. To achieve effective detection results, instead of the *MCL*, we adopted the *fastgreedy* community detection algorithm [11]. As reported in Table 7, our modified implementation proved effective in detecting social spambots, with an MCC = 0.886 for `test set #1` and MCC = 0.847 for `test set #2`.

5. EMERGING TRENDS

As shown in Table 7, the established works benchmarked in Section 4 largely failed to detect the new wave of social spambots, one notable exception being the unsupervised approach proposed in [2]. These results call for novel analytic tools able to keep pace with the latest evolutionary step of spambots. Thus, in this section we revise the most recent literature on spambots detection aiming to answer the research question:

RQ5 − *Is it possible to find new dimensions over which to fight and overcome the novel social spambots?*

An answer to this question might be uncovered by reviewing the evolution of research efforts towards the detection of malicious accounts in OSNs. Such analysis highlights that traditional spambot detection systems typically relied on the application of off-the-shelf machine learning algorithms on the accounts under investigation. Indeed since 2009, most of the works in this field were focused on designing machine-learning features capable of maximizing detection performances of well-known algorithms, such as SVM, Decision Trees, Random Forests, and more. However, since 2013, a number of research teams independently started to formalize

---

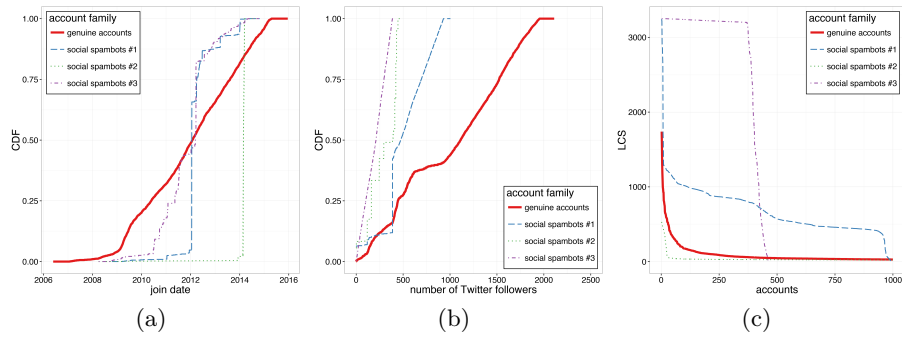[8] *MCL* admits 2 fundamental parameters: *inflation* and *expansion*.

Figure 4: **Distribution of join date, number of followers and LCS for genuine and social spambots accounts.**

established work		emerging trends	
Yardi, Romero *et al.* [46]	2009	Beutel, Faloutsos *et al.* [6, 22, 19, 23]	2013-16
Benevenuto *et al.* [5, 4, 18]	2009-12	Cao *et al.* [10]	2014
K. Lee, Caverlee *et al.* [25, 26]	2010-11	Yu *et al.* [47]	2015
Stringhini *et al.* [35, 34, 36]	2010-13	Viswanath, Mislove, Gummadi *et al.* [40]	2015
Viswanath, Mislove, Gummadi *et al.* [41]	2011	Cresci *et al.* [14]	2016
Stein *et al.* [33]	2011		
Thomas *et al.* [37]	2011		
Gao *et al.* [17]	2012		
Cao *et al.* [9]	2012		
Xie *et al.* [43]	2012		
C. Yang *et al.* [44]	2013		
Wang *et al.* [42]	2013		
S. Lee *et al.* [27, 28]	2013-14		
Z. Yang *et al.* [45]	2014		
Liu *et al.* [30]	2014		
Paradise *et al.* [32]	2014		
Cresci *et al.* [12, 13]	2014-15		
Ferrara *et al.* [7, 15]	2014-16		

This table does not aim to be complete, but rather to testify the emergence of a new trend of research.

Table 8: Recent work in spambot detection.

work	key concept
Beutel, Faloutsos *et al.* [6, 22]	detection of lockstep behaviors
Beutel, Faloutsos *et al.* [19, 23]	anomalies in synchronicity and normality
Cao *et al.* [10]	detection of loosely synchronized actions
Yu *et al.* [47]	detection of latent group anomalies in graphs
Viswanath, Mislove, Gummadi *et al.* [40]	distance between distributions of reputation scores
Cresci *et al.* [14]	similarity between digital DNA sequences

Table 9: Key concepts of emerging trends.

new approaches, also from the algorithmic point of view, for detecting the coordinated and synchronized behavior that characterizes groups of automated malicious accounts [6]. In Table 8 we grouped such techniques and we labeled them as *emerging trends*. Despite being based on different key concepts, all these systems propose novel algorithmic solutions and investigate groups of accounts as a whole, marking a significant difference with the previous literature. Table 9 reports on the new concepts introduced by this emerging body of work.

In order to derive even more insights into the characteristics that might make these emerging solutions successful against current, and possibly future, spambots, in the following we discuss and experiment with the systems proposed in [40, 14].

Tamper Detection in Crowd Computations. The contribution by Viswanath *et al.* in [40] checks whether a given group of accounts (e.g., retweeters of another account, reviewers of a venue on *Yelp*) contains a subset of malicious accounts. The intuition behind the methodology is that the statistical distribution of reputation scores (e.g., number of friends and followers) of the accounts participating in a tam-

pered computation significantly diverge from that of untampered ones. The detection of a tampered computation is performed by computing the Kullback-Leibler distance between the statistical distribution of a given reputation score for the computation under investigation with that of a reference – untampered – computation. If such a distance exceeds a given threshold, the computation under investigation is labeled as tampered. To test this technique against social spambots, we have computed the statistical distribution of the two reputation scores used in [40] (join date and number of followers) for the genuine and the social spambots accounts of our datasets. The results are in figures 4(a) and 4(b). Whereas the genuine accounts feature distributions that almost uniformly span across the possible range of values, social spambots have anomalous distributions. Thus, the technique proposed in [40] is capable of spotting the differences between groups of genuine accounts and the new wave of social spambots. However, the technique cannot *directly* spot the tampering accounts, and, thus, detection and removal of the single accounts must be performed using a separate methodology.

Digital DNA for social spambots detection. Similarly to [40], the technique in [14] analyses a group of accounts, for detecting possible spambots among them. Authors introduced a bio-inspired technique to model online users behaviors by so-called "digital DNA" sequences. Extracting digital DNA for an account means associating that account to a string that encodes its behavioral information. Digital DNA sequences are then compared between one another to find anomalous similarities among sequences of a subgroup of accounts. The similarity among digital DNA sequences is

computed in [14] by measuring the Longest Common Substring (LCS) – that is, the longest DNA substring shared by all the accounts of the group. Accounts that share a suspiciously long DNA substring are then labeled as spambots. Notably, although working at group level, [14] is capable of spotting single spambot accounts. For this reason, we have been able to compare this technique with the ones previously benchmarked in Table 7. Applying the technique to our datasets, the similarity curve of genuine accounts is significantly different from that of social spambots, as shown in Figure 4(c). More specifically, as measured by the LCS metric, `social spambots #1` and `#3` feature a level of similarity much higher than that of genuine accounts. Results reported in Table 7 demonstrate that the digital DNA-based technique [14] achieves excellent detection performances.

Spambot detection: the way ahead. The other systems listed in Table 9, similarly to [40, 14], also focus on those group characteristics that are particularly suitable for discriminating between malicious and genuine accounts, like, e.g., measuring the *synchronicity* and the *normality* of such groups [19, 23]. Focusing on groups has the advantage that, no matter how sophisticated a single spambot can be, a large enough group of spambots will still leave traces of automation, since they do have a common goal (e.g., increasing someone's reputation score). By performing analyses at group level, this emerging trend might be able to significantly raise the bar for social spambots to evade detection. In addition, a further property – and advantage – of this recent research wave is that it proposes ad-hoc detection algorithms, rather than adopting generic and off-the-shelf machine learning algorithms.

The compelling features of the emerging techniques listed in this section represents a fertile ground for fighting the novel social spambots. We can observe a paradigm-shift for research and development of spambot detection systems, which may exploit the new concepts to achieve better resilience and robustness and to withstand the next evolution of social media spambots.

6. CONCLUDING REMARKS

Our long-lasting experiment on malicious accounts survival rate in Twitter demonstrated that spambot detection is still an open issue. Moreover, the already difficult problem to detect spambots in social media is bound to worsen, since the emergence of a new wave of so-called social spambots. By accurately mimicking the characteristics of genuine users, these spambots are intrinsically harder to detect than those studied by Academia in the past years. In our experiments, neither humans nor state-of-the-art spambot detection applications managed to accurately detect the accounts belonging to this new wave of spambots. Indeed, our experiments highlighted that the majority of existing automated systems, as well as crowdsourcing, erroneously label social spambots as genuine (human-operated) accounts. We demonstrated the need for novel analytic tools capable of turning the tide in the arms race against such sophisticated spambots. One promising research direction stems from the analysis of collective behaviors. We highlighted a few emerging approaches that analyze groups as a whole, rather than individuals. The promising outcome of these novel approaches clearly indicates that this is a favorable research avenue.

8. REFERENCES

[1] M. R. Ackermann, M. Märtens, C. Raupach, K. Swierkot, C. Lammersen, and C. Sohler. StreamKM++: A clustering algorithm for data streams. *Experimental Algorithmics*, 17:2–4, 2012.

[2] F. Ahmed and M. Abulaish. A generic statistical approach for spam detection in online social networks. *Computer Communications*, 36(10):1120–1129, 2013.

[3] M. Avvenuti, S. Bellomo, S. Cresci, M. N. La Polla, and M. Tesconi. Hybrid crowdsensing: A novel paradigm to combine the strengths of opportunistic and participatory crowdsensing. In *WWW companion*. ACM, 2017.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on Twitter. In *CEAS*, 2010.

[5] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *SIGIR*. ACM, 2009.

[6] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos. CopyCatch: stopping group attacks by spotting lockstep behavior in social networks. In *WWW*. ACM, 2013.

[7] Bot or not? `http://truthy.indiana.edu/botornot/`, 2014. Accessed: 2016-10-24.

[8] F. Cao, M. Ester, W. Qian, and A. Zhou. Density-Based Clustering over an Evolving Data Stream with Noise. In *SDM*. SIAM, 2006.

[9] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *NSDI*. USENIX, 2012.

[10] Q. Cao, X. Yang, J. Yu, and C. Palow. Uncovering large groups of active malicious accounts in online social networks. In *CCS*. ACM, 2014.

[11] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[12] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. A Criticism to Society (as seen by Twitter analytics). In *DASec*. IEEE, 2014.

[13] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, 80:56–71, 2015.

[14] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. DNA-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, 5(31):58–64, 2016.

[15] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. BotOrNot: A system to evaluate social bots. In *WWW companion*. ACM, 2016.

[16] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Commun. ACM*, 59(7):96–104, 2016.

[17] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. N. Choudhary. Towards online spam filtering in social networks. In *NDSS*. Internet Society, 2012.

[18] 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 *WWW*. ACM, 2012.

[19] M. Giatsoglou, D. Chatzakou, N. Shah, A. Beutel, C. Faloutsos, and A. Vakali. ND-Sync: Detecting synchronized fraud activities. In *PAKDD*, pages 201–214. Springer, 2015.

[20] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[21] X. Hu, J. Tang, and H. Liu. Online social spammer detection. In *Artificial Intelligence*. AAAI, 2014.

[22] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Inferring lockstep behavior from connectivity pattern in large graphs. *Knowledge and Information Systems*, pages 1–30, 2015.

[23] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Catching synchronized behaviors in large networks: A graph mining approach. *ACM Trans. Knowl. Discov. Data*, 10(4):35:1–35:27, 2016.

[24] M. Jiang, P. Cui, and C. Faloutsos. Suspicious behavior detection: Current trends and future directions. *IEEE Intelligent Systems*, 31(1):31–39, 2016.

[25] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots + machine learning. In *SIGIR*. ACM, 2010.

[26] K. Lee, B. D. Eoff, and J. Caverlee. Seven months with the devils: A long-term study of content polluters on Twitter. In *ICWSM*. AAAI, 2011.

[27] S. Lee and J. Kim. WarningBird: A near real-time detection system for suspicious URLs in Twitter stream. *IEEE Trans. Depend. Sec. Comput.*, 10(3):183–195, 2013.

[28] S. Lee and J. Kim. Early filtering of ephemeral malicious accounts on Twitter. *Computer Communications*, 54:48–57, 2014.

[29] H. Liu, J. Han, and H. Motoda. Uncovering deception in social media. *Social Network Analysis and Mining*, 4(1):1–2, 2014.

[30] Y. Liu, B. Wu, B. Wang, and G. Li. SDHM: A hybrid model for spammer detection in Weibo. In *ASONAM*. IEEE/ACM, 2014.

[31] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang. Twitter spammer detection using data stream clustering. *Information Sciences*, 260:64–73, 2014.

[32] A. Paradise, R. Puzis, and A. Shabtai. Anti-reconnaissance tools: Detecting targeted socialbots. *IEEE Internet Computing*, 18(5):11–19, 2014.

[33] T. Stein, E. Chen, and K. Mangla. Facebook immune system. In *SNS*. ACM, 2011.

[34] G. Stringhini, M. Egele, C. Kruegel, and G. Vigna. Poultry markets: on the underground economy of Twitter followers. In *WOSN*. ACM, 2012.

[35] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *ACSAC*. ACM, 2010.

[36] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao. Follow the green: growth and dynamics in Twitter follower markets. In *IMC*. ACM, 2013.

[37] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and evaluation of a real-time URL spam filtering service. In *S&P*. IEEE, 2011.

[38] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson. Trafficking fraudulent accounts: The role of the underground market in Twitter spam and abuse. In *Security Symposium*. USENIX, 2013.

[39] S. M. Van Dongen. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141, 2008.

[40] B. Viswanath, M. A. Bashir, M. B. Zafar, S. Bouget, S. Guha, K. P. Gummadi, A. Kate, and A. Mislove. Strength in numbers: Robust tamper detection in crowd computations. In *COSN*. ACM, 2015.

[41] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove. An analysis of social network-based sybil defenses. *Computer Communication Review*, 41(4):363–374, 2011.

[42] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B. Y. Zhao. Social Turing tests: Crowdsourcing sybil detection. In *NDSS*. Internet Society, 2013.

[43] Y. Xie, F. Yu, Q. Ke, M. Abadi, E. Gillum, K. Vitaldevaria, J. Walter, J. Huang, and Z. M. Mao. Innocent by association: early recognition of legitimate users. In *CCS*. ACM, 2012.

[44] C. Yang, R. Harkreader, and G. Gu. Empirical evaluation and new design for fighting evolving Twitter spammers. *IEEE Trans. Inform. Forens. Sec.*, 8(8):1280–1293, 2013.

[45] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai. Uncovering social network sybils in the wild. *ACM Trans. Knowl. Discov. Data*, 8(1):2, 2014.

[46] S. Yardi, D. Romero, G. Schoenebeck, et al. Detecting spam in a twitter network. *First Monday*, 15(1), 2009.

[47] R. Yu, X. He, and Y. Liu. GLAD: Group anomaly detection in social media analysis. *ACM Trans. Know. Discov. Data*, 10(2):1–22, 2015.

[48] J. Zhang, R. Zhang, Y. Zhang, and G. Yan. The rise of social botnets: Attacks and countermeasures. *arXiv preprint arXiv:1603.02714*, 2016.