Decision Support Systems 80 (2015) 56–71

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Fame for sale: Efficient detection of fake Twitter followers

Stefano Cresci [a,b], Roberto Di Pietro [b,c], Marinella Petrocchi [a], Angelo Spognardi [a,d,*], Maurizio Tesconi [a]

[a] IIT-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy
[b] Bell Labs, Alcatel-Lucent, Paris, France
[c] University of Padua, Maths Dept., Padua, Italy
[d] DTU Compute, Technical University of Denmark, Richard Petersens Plads, 2800 Kgs Lyngby, Denmark

ARTICLE INFO

Article history:
Received 20 June 2014
Received in revised form 29 July 2015
Accepted 13 September 2015
Available online 24 October 2015

Keywords:
Twitter
Fake followers
Anomalous account detection
Baseline dataset
Machine learning

ABSTRACT

Fake followers are those Twitter accounts specifically created to inflate the number of followers of a target account. Fake followers are dangerous for the social platform and beyond, since they may alter concepts like popularity and influence in the Twittersphere—hence impacting on economy, politics, and society. In this paper, we contribute along different dimensions. First, we review some of the most relevant existing features and rules (proposed by Academia and Media) for anomalous Twitter accounts detection. Second, we create a baseline dataset of verified human and fake follower accounts. Such baseline dataset is publicly available to the scientific community. Then, we exploit the baseline dataset to train a set of machine-learning classifiers built over the reviewed rules and features. Our results show that most of the rules proposed by Media provide unsatisfactory performance in revealing fake followers, while features proposed in the past by Academia for spam detection provide good results. Building on the most promising features, we revise the classifiers both in terms of reduction of overfitting and cost for gathering the data needed to compute the features. The final result is a novel *Class A* classifier, general enough to thwart overfitting, lightweight thanks to the usage of the less costly features, and still able to correctly classify more than 95% of the accounts of the original training set. We ultimately perform an information fusion-based sensitivity analysis, to assess the global sensitivity of each of the features employed by the classifier.
The findings reported in this paper, other than being supported by a thorough experimental methodology and interesting on their own, also pave the way for further investigation on the novel issue of fake Twitter followers.

1. Introduction

Originally started as a personal microblogging site, Twitter has been transformed by common use to an information publishing venue. Statistics reported about a billion of Twitter subscribers, with 302 million monthly active users.[1] Twitter annual advertising revenue in 2014 has been estimated to around $480 million.[2] Popular public characters, such as actors and singers, as well as traditional mass media (radio, TV, and newspapers) use Twitter as a new media channel.

Such a versatility and spread of use have made Twitter the ideal arena for proliferation of anomalous accounts, that behave in unconventional ways. Academia has mostly focused its attention on *spammers*, those accounts actively putting their efforts in spreading malware, sending spam, and advertising activities of doubtful legality [1–4]. To enhance their effectiveness, these malicious accounts are often armed with automated twitting programs, as stealthy as to mimic real users, known as *bots*. In the recent past, media have started reporting that the accounts of politicians, celebrities, and popular brands featured a suspicious inflation of followers.[3] So-called *fake followers* correspond to Twitter accounts specifically exploited to increase the number of followers of a target account. As an example, during the 2012 US election campaign, the Twitter account of challenger Romney experienced a sudden jump in the number of followers. The great majority of them has been later claimed to be fake.[4] Similarly, before the last general Italian elections (February 2013), online blogs and newspapers had reported statistical data over a supposed percentage of fake followers of major candidates.[5] At a first glance, acquiring fake followers could seem a practice limited to foster one's vanity—a maybe questionable, but

* Corresponding author. Tel.: +39 050 315 3376.
E-mail addresses: stefano.cresci@iit.cnr.it (S. Cresci), roberto.di_pietro@alcatel-lucent.com (R. Di Pietro), marinella.petrocchi@iit.cnr.it (M. Petrocchi), angsp@dtu.dk (A. Spognardi), maurizio.tesconi@iit.cnr.it (M. Tesconi).

[1] C. Smith, By the Numbers: 150+ Amazing Twitter statistics, http://goo.gl/o1lNi8 — June 2015. Last checked: July 23, 2015.

[2] Statistic Brain, Twitter statistics, http://goo.gl/XEXB1 — March 2015. Last checked: July 23, 2015.

[3] Corriere Della Sera (online Ed.), Academic Claims 54% of Grillo's Twitter Followers are Bogus, http://goo.gl/qi7Hq — July 2012. Last checked: July 23, 2015.

[4] New York Times (online Ed.), Buying Their Way to Twitter Fame, http://goo.gl/VLrVK, — August 2012. Last checked: July 23, 2015.

[5] The Telegraph (online Ed.), Human or bot? Doubts over Italian comic Beppe Grillo's Twitter followers, http://goo.gl/2yEgT — July 2012. Last checked: July 23, 2015.

http://dx.doi.org/10.1016/j.dss.2015.09.003
0167-9236/

harmless practice. However, artificially inflating the number of followers can also be finalized to make an account more trustworthy and influential, in order to stand from the crowd and to attract other genuine followers [5]. Recently, banks and financial institutions in the U.S. have started to analyze Twitter and Facebook accounts of loan applicants, before actually granting the loan.[6] Thus, to have a "popular" profile can definitely help to augment the creditworthiness of the applicant. Similarly, if the practice of buying fake followers is adopted by malicious accounts, as spammers, it can act as a way to post more authoritative messages and launch more effective advertising campaigns [6]. Fake followers detection seems to be an easy task for many bloggers, that suggest their "golden rules" and provide a series of criteria to be used as red flags to classify a Twitter account behavior. However, such rules are usually paired neither with analytic algorithms to aggregate them, nor with validation mechanisms. As for Academia, researchers have focused mainly on spam and bot detection, with brilliant results characterizing Twitter accounts based on their (non-)human features, mainly by means of machine-learning classifiers trained over manually annotated sets of accounts.

To the best of our knowledge, however, despite fake followers constitute a widespread phenomenon with both economical and social impacts, in the literature the topic has not been deeply investigated yet.

1.1. Contributions

The goal of this work is to shed light on the phenomenon of fake Twitter followers, aiming at overcoming current limitations in their characterization and detection. In particular, we provide the following contributions. First, we build a baseline dataset of Twitter accounts where humans and fake followers are known a priori. Second, we test known methodologies for bot and spam detection on our baseline dataset. In particular, we test the Twitter accounts in our reference set against algorithms based on: (i) single classification rules proposed by bloggers, and (ii) feature sets proposed in the literature for detecting spammers. The results of the analysis suggest that fake followers detection deserves specialized mechanisms: specifically, algorithms based on classification rules do not succeed in detecting the fake followers in our baseline dataset. Instead, classifiers based on features sets for spambot detection work quite well also for fake followers detection. Third, we classify all the investigated rules and features based on the cost required for gathering the data needed to compute them. Building on theoretical calculations and empirical evaluations, we show how the best performing features are also the most costly ones. The novel results of our analysis show that data acquisition cost often poses a serious limitation to the practical applicability of such features. Finally, building on the crawling cost analysis, we design and implement lightweight classifiers that make use of the less costly features, while still being able to correctly classify more than 95% of the accounts of our training dataset. In addition, we also validated the detection performances of our classifiers over two other sets of human and fake follower accounts, disjoint from the original training dataset.

1.2. Road map

The remainder of this paper is structured as follows. Section 2 considers and compares related work in the area of Twitter spam and bot detection. Section 3 describes our baseline dataset. In Section 4, we evaluated a set of criteria for fake Twitter followers detection promoted by Social Media analysts using our baseline dataset. In Section 5, we examine features used in previous works for spam detection of Twitter accounts. In Section 6 we compute the cost for extracting the features our classifiers are based on. A lightweight and efficient classifier is also provided, attaining a good balance between fake followers detection capability and crawling cost. Finally, Section 7 concludes the paper.

2. Related work

Quoting from [7], "A fake Twitter account is considered as one form of deception (i.e., deception in both the content and the personal information of the profiles as well as deception in having the profile follow others not because of personal interest but because they get paid to do so)." The second characterization for deception is exactly the one we deal with in our paper. We specifically consider *fake followers* as those Twitter accounts appropriately created and sold to customers, which aim at magnifying their influence and engagement to the eyes of the world, with the illusion of a big number of followers.

So defined fake followers are only an example of anomalous accounts which are spreading over Twitter. Anomalies have been indeed identified in the literature as either spammers (i.e. accounts that advertise unsolicited and often harmful content, containing links to malicious pages [8]), or bots (i.e., computer programs that control social accounts, as stealthy as to mimic real users [9]), or cyborgs (i.e., accounts that interweave characteristics of both manual and automated behavior [10]). Finally, there are fake followers, accounts massively created to follow a target account and that can be bought from online accounts markets.

2.1. Grey literature and online blogs

Before covering the academic literature, we briefly report on online documentation that presents a series of intuitive fake follower detection criteria, though not proved to be effective in a scientific way. The reason why we cite this work is twofold: on the one hand, online articles and posts testify the quest for a correct discrimination between genuine and fake Twitter followers; on the other hand, we aim at assessing in a scientific manner whether such criteria could actually be employed for fake followers detection.

As an example, a well-known blogger in [11] indicates as possible bots-like distinctive signals the fact that bots accounts: 1) have usually a huge amount of following and a small amount of followers; 2) tweet the same thing to everybody; and, 3) play the follow/unfollow game, i.e. they follow and then unfollow an account usually within 24 h. Criteria advertised by online blogs are mainly based on common sense and the authors usually do not even suggest how to validate them.

A series of reports published by the firm *Digital evaluations* [12] have attracted the attention of Italian and European newspapers and magazines, raising doubts on the Twitter popularity of politicians and leading international companies. A number of criteria, inspired by common sense and denoting *human* behavior, are listed in the reports and used to evaluate a sample of the followers of selected accounts. For each criterion satisfied by a follower, a *human* score is assigned. For each not fulfilled criterion, either a *bot* or *neutral* score is assigned. According to the total score, Twitter followers are classified either as humans, as bots or as neutral (in the latter case, there is not enough information to assess their nature), providing a quality score of the effective influence of the followed account. The results in [12], however, lack a validation phase.

Finally, some companies specialized in social media analysis offer online services to estimate how much a Twitter account is *genuine* in terms of its followers [13–15]. However, the criteria used for the analysis are not publicly disclosed and just partially deductible from information available on their web sites. Moreover, as demonstrated in our previous work [16], these analyses are affected by several biases like small and statistically unsound sampling strategies.

2.2. Academic literature

In recent years, spam detection on Twitter has been the matter of many investigations, approaching the issue from several points of view. As an example, a branch of research focused on the textual

[6] Le Monde (online Ed.), Dis-moi combien d'amis tu as sur Facebook, je te dirai si ta banque va t'accorder un prêt, http://goo.gl/zN3PJX — Sept. 2013. Last checked: July 23, 2015.