Information Sciences 372 (2016) 332–346

Contents lists available at ScienceDirect

Information Sciences

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

# Twitter turing test: Identifying social machines☆

Abdulrahman Alarifi[a,*], Mansour Alsaleh[a], AbdulMalik Al-Salman[b]

[a] Computer Research Institute, King Abdulaziz City for Science and Technology, P.O. Box 6086, Riyadh 11442, KSA
[b] Computer Science Department, King Saud University, Riyadh, KSA

ARTICLE INFO

*Article history:*
Received 2 April 2015
Revised 27 July 2016
Accepted 11 August 2016
Available online 20 August 2016

*Keywords:*
Content spam
Fake user account
Social network
Sybil account
Twitter
Web spam

ABSTRACT

Many machine-controlled Twitter accounts (also called "Sybils") are created each day to provide services, flood out messages for astroturf political campaigns, write fake product reviews, or produce an underground marketplace for purchasing Twitter followers, retweets, or URL advertisements. In addition, fake identities and user accounts in online communities are resources used by adversaries to spread malware, spam, and harmful links over social networks. In social networks, Sybil detectors rely on the assumption that Sybils will find it harder to befriend real users; thus, Sybils that are connected to each other form strongly connected subgraphs, which can be detected using the graph theory. However, a majority of Sybils have actually successfully integrated themselves into real social media user communities (such as Twitter and Facebook). In this study, we compared the current methods used for detecting Sybil accounts. We also explored the detection features of various types of Twitter Sybil accounts in order to build an effective and practical classifier. To evaluate our classifier, we collected and manually labeled a dataset of Twitter accounts, including human users, bots, and hybrids (i.e., tweets posted by both human and bots). We consider that this Twitter Sybils corpus will help researchers to conduct high-quality measurement studies. We also developed a browser plug-in, which utilizes our classifier and warns the user about possible Sybil accounts before accessing or following them after clicking on a Twitter account.



## 1. Introduction

The Internet is now a fundamental source of information and a vital tool for individuals and businesses, and thus the desire to automatically generate and spread influential content has increased dramatically. This capability can be applied to various objectives, including: (1) providing advertisements; (2) promoting politically oriented views and opinions; (3) promoting financial trends; (4) generating product reviews; (5) spreading malware, spam, and harmful links; (6) influencing search engine results such that particular links are shown first; (7) affecting voting results; (8) generating news feeds; and (9) creating an underground marketplace for purchasing social media followers. Some of these objectives are benign (e.g., news and information accounts), but the majority of these uses can be categorized as fraud. A common and effective method for exploiting this capability is utilizing social media tools via machine-controlled accounts with the hope that these accounts will be perceived as humans.

☆ This manuscript is an extension of an earlier conference version [4].

* Corresponding author. Fax: +966114814553.
*E-mail addresses:* aarifi@kacst.edu.sa (A. Alarifi), maalsaleh@kacst.edu.sa (M. Alsaleh), salman@ksu.edu.sa (A. Al-Salman).

http://dx.doi.org/10.1016/j.ins.2016.08.036
0020-0255/

At present, billions of people use social networks, such as Facebook, Twitter, LinkedIn, and Google Plus, in their daily lives for different purposes [10,24,34]. Twitter is a social network where users can publish a post as a "micro-blog" or "tweet", but it is limited to 140 characters per tweet. According to the United States Securities and Exchange Commission, the number of monthly Twitter active users exceeds 230 million and over 100 million daily active users generate about 500 million tweets daily [40]. Given these huge volumes, Twitter is a target for spammers and hackers who prey on the less technologically capable. In fact, numerous machine-controlled Twitter accounts (called "Sybils") are created each day to provide services, send spam and harmful links, or to fuel the black market for buying Twitter followers or retweets. We cannot ignore the impact that these accounts have on the online community because they give a false sense of credibility to the poster as well as increasing security and privacy concerns among users of the Web (e.g., because most links posted by a Sybil account tend to be malicious).

In this study, we aimed to mitigate these user concerns by implementing a system to determine whether the user of a Twitter account is a human or a bot. Thus, before following or interacting with a Twitter user, people can know whether the user is a human or a bot, thereby helping to avoid opening tweets or links posted by an account that could be harmful, spam, or inappropriate. However, machine-controlled accounts are constantly evolving to prevent detection such as by randomizing their tweeting times, posting human-like interactions, or avoiding communication with other bots. Therefore, we aimed to find new detection features for detecting bots regardless of the semantic content of the tweets.

The main contributions of this study are as follows.

1. Establishing Twitter Sybils corpus. We collected and manually labeled a dataset of Twitter accounts (including human, bot, and hybrid accounts) to evaluate our proposed classifier. We have also made this dataset available to the public to help other researchers in this area[1].
2. Analyzing detection features. We analyzed various types of Sybil account detection features to obtain an appropriate set of detection features, which facilitate reasonably accurate detection. To the best of our knowledge, we determined several novel features for Twitter Sybil account detection.
3. Twitter Sybils classifier. We built a classifier for detecting Twitter Sybil accounts by using supervised machine learning techniques.
4. Browser detection plug-in. We developed a browser plug-in referred to as the Twitter Sybils Detector (TSD), which utilizes our classifier and warns users about Sybil accounts before accessing them after clicking on a Twitter account[2].

The remainder of this paper is organized as follows. Section 2 describes how we built our Twitter Sybils corpus. Section 3 explains the feature extraction and selection process. In Section 4, we first describe our classifier and we then present an evaluation of its performance based on our dataset. Section 5 considers the system architecture and components. Section 6 presents an overview of related research. In Section 7, we provide our conclusions.

## 2. Building our corpus

**Lack of absolute ground truth (AGT) datasets.** The reliable evaluation of anomaly detection systems is challenging owing to the difficulty of validating the detection results and the lack of AGT datasets [5]. An AGT dataset is considered to be an ideal reference, where all of the true positives and true negatives are identified correctly. Some studies have investigated the detection of Sybil accounts on Twitter, but no known dataset can be treated as an AGT. Different approaches can be employed for establishing an AGT, but most of these approaches either misrepresent the sample space or they require a rigorous and costly process. For example, a simulation approach for generating a labeled dataset might not provide a realistic representation of the actual data collected from the source. Another approach is to slowly label a real dataset by using various approaches and rater groups, and to make them publicly available for a period of time for testing and validation by the community.

**Deriving a ground truth reference (GTR) dataset.** Given the challenges of establishing an AGT dataset and the lack of labeled datasets with an AGT, we derived a GTR dataset (denoted as $GTR_1$) that uses a reliable labeling mechanism, but it is less rigorous than those usually employed for deriving an AGT. In addition to the development of new dataset features for capturing the up-to-date tactics employed to lure users to Sybil accounts, we consider that our dataset can be used by other researchers as a GTR to evaluate and benchmark potential detection systems. A two-step process was followed to collect the data for $GTR_1$. First, we randomly collected a large set of tweets by using the Twitter Streaming API to provide random samples of recently published tweets. The collected tweets were then used by the Twitter REST API to gather account information for the users who posted the tweets, which comprised nearly 1.8 million accounts. The account information comprised the user's name, account creation date, account description, profile picture, tweet text, tweet submission time and date, tweet source, number of favorites and retweets for each tweet, number of tweets, number of followees and followers, names and profile pictures of followees and followers, number of times the user had favorited, number of times the user had been favorited, number of lists the user had created, and the number of lists of which the user was a member, which were used later to allow us to differentiate between human, Sybil, or hybrid accounts. Our account was not on Twitter's

[1] Our dataset of Twitter accounts can be downloaded from the following link: https://github.com/TwitterSybilDetector/TwitterSybilDetector

[2] The Chrome browser plug-in can be download from: https://github.com/TwitterSybilDetector/TwitterSybilDetector