Computer Communications 54 (2014) 48–57

Contents lists available at ScienceDirect

Computer Communications

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

Early filtering of ephemeral malicious accounts on Twitter ☆

Sangho Lee *, Jong Kim

Department of Computer Science and Engineering, Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea

ARTICLE INFO

Article history:
Received 18 January 2014
Received in revised form 7 August 2014
Accepted 9 August 2014
Available online 19 August 2014

Keywords:
Online social network
Twitter
Spam
Malicious account

ABSTRACT

Cybercriminals exploit a large number of ephemeral malicious accounts for conducting large-scale simple attacks such as spam distribution on online social networks. However, conventional detection schemes relying on account or message information take a considerable time to collect such information before running detection algorithms so criminals utilize their accounts until suspension and exploit others again. In this paper, we propose a new detection scheme to filter potentially malicious account groups around their creation time. Our scheme utilizes the differences between algorithmically generated account names and human-made account names to identify malicious accounts generated using the similar algorithms. For accounts created within a short period of time, we apply a clustering algorithm to group accounts sharing similar name-based features and a classification algorithm to classify malicious account clusters. As a case study, we analyze 4.7 million accounts collected from Twitter. Even though our scheme only relies on account names and their creation time, it achieves reasonable accuracy. Therefore, we can use it as a fast filter against malicious account groups to selectively conduct an in-depth analysis.

1. Introduction

Online social networks (OSNs), which allow people to establish and manage social relationships on Internet, suffer from malicious accounts. Many famous OSNs, such as Facebook and Twitter, allow any people to create accounts of their services to attract more users. They have a simple procedure for account creation, such as requiring an email address, because people dislike sophisticated verification processes. However, the simplified procedure allows cybercriminals to create malicious accounts for various attacks including spam, phishing, and malware distribution.

Among many OSNs, we focus on Twitter because it is one of the largest OSNs and its openness attracts many cybercriminals. Twitter has more than 140 million active users and 340 million messages created each day [1]. Due to this popularity, cybercriminals aim at exploiting Twitter to distribute malicious messages to its users. Some cybercriminals also develop tools for automated account creation and messages distribution; Twitter recently bring a lawsuit against them [1]. Detection of malicious Twitter accounts is therefore a crucial problem that demands countermeasures.

Researchers propose various *reactive* schemes to detect malicious social accounts and messages. Some schemes rely on the features of malicious accounts, such as a large number of messages containing URLs and a small number of friends [2–8]. However, these schemes require a time to collect account features *before* detection because they can observe the features *only after* the malicious accounts have initiated several malicious activities. In contrast, other schemes that inspect the malice of texts [9] or URLs [10,11] contained in a message do not require a time to collect information. However, these schemes also need to wait *until* malicious accounts post at least one distinguishable malicious tweet. Moreover, they demand more resources than the account-based detection schemes because they rely on complicated methods and the number of messages is fairly larger than the number of accounts.

Cybercriminals take two opposite approaches to bypass the existing detection schemes: (i) preparing advanced accounts to *evade* detection methods and (ii) creating a tremendous number of ephemeral accounts to *ignore* detection methods. First, criminals create and maintain advanced malicious accounts with human assistance. Recent studies [12,13] verify that criminals exploit crowdsourcing systems for acquiring human-assisted malicious accounts. The lifetime of the advanced malicious accounts is long because detection systems may not distinguish them from normal accounts, and, thus, criminals utilize them for long-term sophisticated attacks. However, because the costs of the advanced malicious accounts are high, criminals may not utilize them for large-scale simple attacks due to the trade-offs between costs and benefits. The recent countermeasure against the advanced malicious accounts also relies on humans for accurate detection [14].

☆ This research was supported by the MSIP, Korea, under the ITRC support program supervised by the NIPA (NIPA-2013-H0301-13-3002).

* Corresponding author. Tel.: +82 54 279 2915; fax: +82 54 279 1805.
E-mail addresses: sangho2@postech.ac.kr (S. Lee), jkim@postech.ac.kr (J. Kim).

http://dx.doi.org/10.1016/j.comcom.2014.08.006
0140-3664/

Second, cybercriminals create a huge number of *ephemeral malicious accounts* to ignore detection. The account-based and human-assisted detection schemes require a time to collect enough evidence for detection, so that *criminals can exploit their accounts until account suspension and other accounts again*. Message-based detection schemes can be a solution against this attack; however, they demand many resources and are weak against message obfuscation and conditional URL redirection [15]. To conduct this attack, criminals implement simple account creation programs and distribute them to botnets for avoiding IP address-based filtering of account creation. These ephemeral malicious accounts are suitable for large-scale simple attacks such as spam distribution. Thomas et al. [16] verify that most suspended Twitter accounts are created for the sole purpose of spam distribution. Many of them are the ephemeral malicious accounts we focus on.

We must consider a countermeasure against ephemeral malicious accounts: *finding potentially malicious accounts around their creation time*. An addressable characteristic of ephemeral malicious accounts, which we can observe around their creation time, is that they have algorithmically generated account names differing from human-made names. Twitter has a tremendous number of users and they already occupy a huge number of account names, so algorithmically creating account names that look like human-made names and not yet registered with Twitter is not an easy task. A similar problem exists on domain flux botnets which rely on algorithmically generated domain names, so that a number of studies inspect name characteristics to detect malicious domain names [17–19].

In this paper, we propose a new scheme to filter potentially malicious account groups around their creation time, which consists of (i) clustering newly created accounts that share similar name-based features and (ii) classifying the clustered accounts. To cluster accounts, we compute the log-likelihood that the account names are generated given a Markov chain and the lengths of the names. We create the Markov chain using verified account names and use an agglomerative hierarchical clustering algorithm. To classify account clusters, we train a Support Vector Machine (SVM) classifier using the name-based features of suspicious account groups. Evaluation results show that our scheme can cluster distinguished account names and classify benign and suspicious account clusters with reasonable accuracy. Therefore, our scheme is *a fast filter* to selectively conduct in-depth testing and monitoring of potentially malicious accounts.

The main contributions of this study are as follows:

- To the best of our knowledge, this is the first attempt to find potentially malicious account groups around the time of their creation. Our scheme can infer suspiciousness of malicious account groups before they initiate malicious activities.
- We propose a clustering method for proactively and accurately grouping accounts that seem to belong to the same campaigns.
- We derive a number of features to distinguish malicious account groups from benign account groups and develop an accurate classifier for account groups using these features.

The rest of this paper is as follows. In 2 we explain the manner in which our dataset is collected and analyze its characteristics. In 3 we briefly explain our filtering scheme. In 4 we introduce our clustering method used for grouping account names that share similar features. In 5 we explain the manner in which clustered account names can be classified. In 6 we discuss a possible evasive method against our scheme and the accuracy problem. In 7 we introduce related work. Lastly, we conclude this paper in 8.

2. Data collection and analysis

2.1. Data collection

We explain our dataset sampled from Twitter public timeline. Our dataset consists of 4,687,345 Twitter accounts created between April 2011 and October 2011 (214 days). We attempt to classify whether the collected accounts are benign or malicious according to whether they are active or suspended in March 2012 as a previous study [16] does. The number of active accounts is 3,618,096 (77.19%) and that of suspended accounts is 1,069,249 (22.81%). Twitter announced that 572,000 new accounts created on March 12, 2011 [20], so that, the number of accounts we focus on is approximately 3.83% of all the new accounts created during the 214 days.

We also collect the names of the verified Twitter accounts to consider them as ground truth of human-made account names. We obtain 18,289 verified Twitter accounts from Twitter's official accounts `@verified` on February 25, 2012.

2.2. Data analysis

We analyze our dataset to identify the characteristics of malicious accounts and their groups to utilize them for implementing a detection system.

Active and suspended accounts. We first analyze how many active and suspended accounts are created each hour using our dataset (Fig. 1). On average (median), 912 accounts are created each hour of which 704 are active (77.19%) and 208 are suspended (22.81%).

We then investigate *abnormal periods of time* in which the number of suspended accounts exceeds that of active accounts, among periods of 10, 30, and 60 min. Table 1 shows that the fractions of abnormal periods are fairly small (around 1%) in comparison with that of all suspended accounts (22.81%). Even we use a short period of 1 min, the fraction becomes only 4.14%. Therefore, without specialized methods, the low-rate creation of malicious accounts will go undetected.

Interval between account creation and first tweet. We measure the time interval between the creation of Twitter accounts suspended and their first tweets (also known as a dormancy period [16]) to estimate the *time advantage* of early filtering over the reactive detection schemes. We first analyze our dataset and find 50,572 first tweets from suspended accounts. We then measure the interval between the time of account creation and the time of first tweet, and obtain a complementary CDF (Fig. 2). On average (median), the suspended accounts post their first tweets 550 min

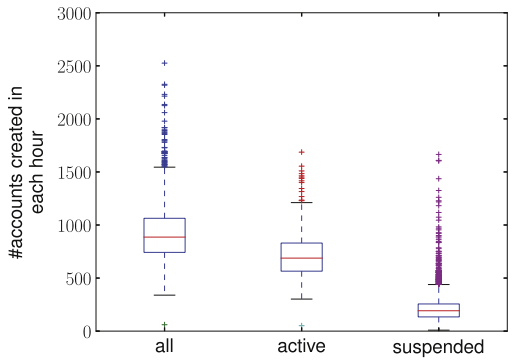

Fig. 1. Statistics of the number of active and suspended accounts created in each hour between April 2011 and October 2011 (boxplot).