A Review on Enhanced Machine Learning Approach for Detection of Malicious Urls and Spam in Social Network

Ms. Roshani K. Chaudhari, Prof. D. M. Dakhane

Abstract— Nowadays online social network such as Twitter, Face Book plays a vital role in daily life. Twitter can suffer from malicious tweets. The tweets contain suspicious URLs for phishing, spam and malware distribution. Conventional Twitter spam detection techniques have used features of account such as the ratio of tweets comprising URLs and relation features or the account creation date in the Twitter graph. However, the account features can easily fabricated by malicious users. Furthermore extraction of relation features from the Twitter graph is time as well as resource consuming. Conventional techniques for the detection of suspicious URL have categorised URLs using different features comprising lexical features of URLs, HTML content, dynamic behaviour and URL redirection. The aim of this paper related to detect the suspicious URLs by using the system, known as WARNINGBIRD. In this paper, we have concentrated on support vector machines (SVMs), the supervised learning models with associated learning algorithms which are used to analyze data used for classification and regression analysis. As WARNINGBIRD does not depend on the features of malicious landing pages that may not be reachable, it is robust while protecting against condition redirection. Specifically the WARNINGBIRD focuses on the correlations of multiple redirect chains which share the same redirection servers.

Index Terms— Twitter, Threats on Twitter, Support vector machines (SVM), Types of Spammer.

I. INTRODUCTION

Online social networks have become very popular beyond the last few years. People share the information, news by using the social network. Generally, users get messages published by the users they are connected to in the form of tweets, wall post or status updates.

As popularity of Twitter is growing, malicious users often try ro detect a way to attack it. The most common web attack such as scam, spam, phishing and malware distribution attacks have appeared on Twitter. As tweets are short in length attackers use shortened malicious URLs that redirect Twitter users to external attack servers [1], [2], [3], [4].

A. Twitter as an OSN

Twitter is a social network service launched in March 21, 2006 [5]. Approximately 500 million active users [5] till date who share information. The logo of Twitter is a chirping bird and hence the name Twitter. Users can access it to exchange the frequent information called 'tweets' which are messages of up to 140 characters long. Hence anyone can send or read. By default the tweets are public and visible to all those who are following the tweeter. These tweets are shared by users which may contain news, opinions, photos, videos, links, and messages. The standard terminology used in Twitter and relevant to our work is as follows:

1. *Tweets* [6]: A message on Twitter comprising maximum length of 140 characters.
2. *Followers & Followings* [6]: Followers are the users who are following a particular user and followings are the users whom user follows.
3. *Retweet* [6]: A tweet that has been reshared with all followers of a user.
4. *Hash tag* [6]: The # symbol is used to tag topics or keywords in a tweet to make it easily distinguishable for search purposes.
5. *Mention* [6]: Tweets can include replies and mentions of other users by preceding their usernames with @ sign.
6. *Lists* [6]: Twitter provides a mechanism to list users you follow into groups.
7. *Direct Message* [6]: Also called a DM represents Twitter's direct messaging system for private communication amongst users.

B. Threats on Twitter

1. *Spammed Tweets* [7]: Twitter allows its users to post tweets of maximum 140 characters but having no regard of the character limit. Cybercriminals have found a way to actually use this limitation to their advantage by creating short but compelling tweets with links for promotions for free vouchers or job advertisement posts or other promotions.
2. *Malware downloads* [7]: Cyber criminals have used Twitter to spread posts with links to malware download pages. Backdoor[7] and FAKEAV applications are the examples of Twitter worm that sent direct messages, and even malware that affected both operating system such as Windows and Mac. The KOOBFACE [7] is a most tarnished social media malware, which targeted both Twitter and Facebook.
3. *Twitter bots* [7]: Cybercriminals burn to use Twitter to control and manage botnets. These botnets control the users' accounts and pose a threat to their security and privacy.

ISSN: 2278 – 1323

II. LITERATURE REVIEW

H. Kwak, C.Lee, H.Park, and S.Moon proposed a work in 2010 [8] which mainly focuses on Twitter, a social networking service ,more than 41 million users of twitter as of July 2009 and is growing fast. Twitter users tweet about any topic within the 140- character limit. Twitter offers an Application Programming Interface (API) that is easy to crawl and collect data. Twitter tracks words, phrases and hash (#) tags that are most often mentioned and post them under the title of "trending topics" regularly. A hash tag is a convention among Twitter users to create and follow a thread of discussion by prefixing a word with a '#' character. In order to identify influential on Twitter. Juan Chen and chuan xiongguo described online detection of prevention of phishing attacks and phishing attacks [9].

McCord et.al. [10] used user based features like number of friends, number of followers and content based features like number of URLs, replies/mentions, retweets, hash tags of collected database. Classifiers namely Random Forest, Support Vector Machine (SVM), Naive Bayesian and K-Nearest Neighbor have been used to identify spam profiles in Twitter. Method has been validated on 1000 users with 95.7% precision and 95.7% accuracy using the Random Forest classifier and this classifier gives the best results followed by the SMO, Naive Bayesian and K-NN classifiers. Limitation of this approach is that for considered dataset reputation feature has been showing wrong results i.e. it is not able to differentiate spammers and non-spammers, unbalanced dataset has been used so Random Forest is giving best results as this classifier is generally used in case of unbalanced dataset, and finally the approach has been validated on less dataset.

Lin et. al. [11] detected long-surviving spam accounts in Twitter on the basis of two different features that are URL rate and interaction rate. Most of the papers have used lot many features for detection of spam accounts like numbers of followers, numbers of following, followers/following ratio, tweet content, no of hash tags, URL links etc. But as per this paper all these features are not so effective in detecting spammers so only simple yet effective features like URL rate and interaction rate have been used for detection purpose. URL rate is the number of tweets with URL / total number of tweets and interaction rate is the number of tweets interacting / total number of tweets. 26,758 International Journal of Computer Applications (0975 – 8887) Volume 85 – No 10, January 2014 31 accounts have been crawled using Twitter API and 816 long surviving accounts have been analyzed J48 classifier with 86% precision. Limitation of the approach is that only two features have been used for spam profile detection and if spammers keep low URL rate and low interaction rate then this technique will not work as intended.

G. Stringhini, C. Kruegel, and G.Vigna in 2010 [12] used account features such as Friend-Follower ratio, URL ratio and message similarity to differentiate spam tweets. This paper ascertains to which extent spam has entered social network and how spammers who target social networking sites operate. For collecting the spamming activity data, a large and diverse set of "honey-profiles" are created on three large social networking sites and then analyzed the collected data and identified aberrant behavior of users who contacted honey-profiles. Features are developed based on the analysis of this behavior which is used for detection.

A. Wang in 2010 [13] modeled Twitter as directed graph where vertices represent user accounts and the edge direction determines the type of relationship between users, friend or follower. In this paper, detection mechanism is depend on graph based features such as in-degree and out-degree of nodes and content based features such as presence of HTTP links and Relevant topics in tweets. This work applies machine learning methods to instinctively differentiate spam accounts from normal ones. A Web crawler is developed relying on the API methods provided by Twitter for extracting public available data on Twitter website. Lastly, a system is established to evaluate the detection method. J. Song, S. Lee, and J. Kim in 2011[14] viewed Twitter as an undirected graph and made use of Menger"s theorem to evaluate the values of message features such as distance as well as connectivity between nodes in order to perform detection.

III. ANALYSIS OF PROBLEM

In the existing work following problems were discovered:-

1. Malicious servers can bypass an investigation by discriminatively providing benign pages to crawlers. For instance, because static crawlers usually are unable to handle JavaScript or Flash, malicious servers can use them to deliver malicious content only to normal browsers.
2. A recent technical report from Google has also discussed techniques for evading current Web malware detection systems.
3. Malicious servers can also employ temporal behaviors— providing different content at different times-to evade an investigation.

To facilitate the problems which are discovered in existing system we proposed a new system. The contributions of proposed system are as follow:

1. A new suspicious URL detection system for Twitter that is based on the correlations of URL redirect chains, which are difficult to invent. The system can find correlated URL redirect chains using the often shared URLs and determine their suspiciousness in almost real time.
2. New features of suspicious URLs are introduced: some of which are newly discovered and while others are variations of previously discovered features.
3. The results of investigations conducted on suspicious URLs that have been widely distributed through Twitter over several months have are presented.

ISSN: 2278 – 1323

IV. System Details

WARNINGBIRD is composed of four major components:

- Data collection,
- Feature extraction,
- Training and
- Classification

1. *Data collection***:** The data collection component has two subcomponents: the gathering of tweets with URLs and crawling for URL redirections. For collecting tweets with URLs and their context information from the Twitter public timeline, this component uses Twitter Streaming APIs [8].

2. *Feature extraction***:** There are three subcomponents of feature extraction: grouping identical domains, finding entry point URLs, and extraction of feature vectors. This component monitors the tweet queue to check whether a sufficient number of tweets have been collected.

3. *Training*: The training component has two subcomponents: retrieval of account statuses and the training classifier. As we use an offline supervised learning algorithm, the feature vectors for training are relatively existing values than feature vectors for classification. For labeling the training vectors, we use the Twitter account status; URLs from suspended accounts are considered malicious and URLs from active accounts are considered benign. We periodically update our classifier by using labeled training vectors.

4. *Classification***:** The classification component executes our classifier using input feature vectors to distinguish suspicious URLs. When the classifier returns a number of malicious feature vectors, this component flags the corresponding URLs and their tweet information as suspicious. These URLs, detected as suspicious, will be delivered to security experts or more sophisticated dynamic analysis environments for in-depth investigation.

V. Support vector MACHINES (SVM)

The support vector machine (SVM) is a training algorithm for learning classification and regression rules from data, for example to learn polynomial, radial basis function (RBF) and multi-layer perception (MLP) classifiers the SVM can be used. SVMs were first recommended by Vapnik in the 1960s for the classification and have recently become an area of strained research owing to developments in the techniques and theory coupled with extensions to regression and density estimation. SVMs emerged from statistical learning theory; the aim being to evaluate only the problem of interest without solving a more troublesome problem as an intermediate step. SVMs are based on the structural risk minimization principle closely related to regularization theory. This fundamental essence incorporates capacity control to prevent over fitting and thus is an incomplete solution to the bias-variance trade-off dilemma. In the implementation of SVM two key elements are the methods of mathematical programming and kernel functions. The parameters are found by solving quadratic programming questions with linear equality and inequality constraints; rather than by solving a non-convex, unconstrained optimization problem. The pliability of kernel functions allows the SVM to search a wide variety of hypothesis spaces.

VI. Types of Spammers

Spammers are the malicious users who contaminate the information presented by legitimate users and in turn pose a risk to the privacy and security of social networks. Spammers belong to one of the following categories [15]:

1. *Phishers*: Phishers are the users who behave like a normal user to acquire personal data of other genuine users.
2. *Fake Users*: These are the users who impersonate the profiles of genuine users to send spam content to the friends of that user or other users in the network.
3. *Promoters*: These are the ones who send malicious links of advertisements or other promotional links to others so as to obtain their personal information.

Motives of Spammers:
a) Disseminate pornography
b) Spread viruses
c) Phishing attacks
d) Compromise system reputation.

VII. Conclusion

After reviewing the work of different researches it has been observed that the conventional suspicious URL detection systems are ineffectual in their protection in contrast conditional redirection servers that discriminate investigators from normal browsers and redirect them to mild pages to cloak malicious landing pages. The WARNINGBIRD is a system for Twitter to detect the suspicious URLs. As WARNINGBIRD does not depend on the features of malicious landing pages that may not be reachable, it is robust while protecting against condition redirection. Specifically it focuses on the correlations of multiple redirect chains which share the same redirection servers.

Acknowledgment

I would like to thank my guide Prof. D. M. Dakhane for fulfilling my research work on Twitter. Moreover I thank for the facilities provided by Sipna College of Engineering and Technology, Amravti for providing me necessary article for

ISSN: 2278 – 1323

completing my study on this topic.

REFERENCES

[1]. Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. *Who is tweeting on Twitter: Human, bot, or cyborg? In Annual Computer Security Applications Conf. (ACSAC), 2010.*

[2] C. Grier, K. Thomas, V. Paxson, and M. Zhang. *@spam: The underground on 140 characters or less. In ACM Conf. Computer and Communications Security (CCS), 2010.*

[3] D. K. McGrath and M. Gupta. *Behind phishing: An examination of phisher modi operandi. In USENIX Workshop Large-Scale Exploits and Emergent Threats (LEET), 2008.*

[4] G. Stringhini, C. Kruegel, and G. Vigna. *Detecting spammers on social networks. In Annual Computer Security Applications Conf. (ACSAC), 2010.*

[5*]*. http://en.wikipedia.org/wiki/Twitter-Information of Twitter.

[6] Anshu Malhotra, Luam Totti, Wagner Meira Jr., Ponnurangam Kumaraguru, Virgılio Almeida, *Studying User Footprints in Different Online Social Networks ,International Conference on Advances in Social Networks Analysis and Mining, 2012, IEEE/ACM.*

[7].http://aboutthreats.trendmicro.com/us/webattack- Information regarding Twitter threats.

[8] S. Lee and J. Kim," Warning Bird: *A near Real-Time Detection System for Suspicious URLs in Twitter Stream "IEEE transactions on dependable and secure computing, vol. 10, no. 3, may/june 2013.*

[9] *Ollman, G.(2004) The phishing Guie-Understanding and Preventing , White paper , Next Generation Security software Ltd.*

[10] M. McCord, M. Chuah, *Spam Detection on Twitter Using Traditional Classifiers, ATC'11,* Banff, Canada, Sept 2-4, 2011, IEEE.

[11].Po-Ching Lin, Po-Min Huang, *A Study of Effective Features for Detecting Long-surviving Twitter Spam Accounts, Advanced Communication Technology (ICACT), 15th International Conference on 27-30 Jan. 2013, IEEE.*

[12]G. Stringhini, C. Kruegel, and G. Vigna, *"Detecting Spammers on Social Networks,"Proc. 26th Ann. Computer Security Applications Conf. (ACSAC), 2010.)*

[13]A. Wang, "*Don"t Follow Me: Spam Detecting in Twitter," Proc. Int'l Conf. Securityand Cryptography (SECRYPT), 2010.*

[14] J. Song, S. Lee, and J. Kim, "*Spam Filtering in Twitter Using Sender-ReceiverRelationship," Proc.14th Int'l Symp. Recent Advances in Intrusion Detection.(RAID), 2011.*

[15] Kyumin Lee, James Caverlee, Steve Webb, *Uncovering Social Spammers: Social Honeypots + Machine Learning, Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, Pages 435–442, ACM, New York (2010).*

Roshani K. Chaudhari: perceiving the masters in Computer Engineering from Sipna College of Engineering and Technology, Amravati. Completed the graduation in Computer Science and Engineering. (India)

Prof. D. M. Dakhane, Associate Professor, department of computer Science and Engineering at Sipna College of Engineering and Technology, Amravati. (India)

ISSN: 2278 – 1323