International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 1, Issue 3, July 2014, PP 33-38
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
www.arcjournals.org

Detection of Malicious URLs by Correlating the Chains of Redirection in an Online Social Network (Twitter)

[1]MD.Sabeeha, SK.Karimullah, [2]P.Babu

[1]PG Schalor,CSE, Quba college of engineering and technology
[2,] Associate professor, QCET, NELLORE

ABSTRACT: *Twitter is prone to malicious tweets containing URLs for spam, phishing, and malware distribution. Conventional Twitter spam detection schemes utilize account features such as the ratio of tweets containing URLs and the account creation date, or relation features in the Twitter graph. These detection schemes are ineffective against feature fabrications or consume much time and resources. Conventional suspicious URL detection schemes utilize several features including lexical features of URLs, URL redirection, HTML content, and dynamic behavior. However, evading techniques such as time-based evasion and crawler evasion exist. In this paper, we propose WARNINGBIRD, a suspicious URL detection system for Twitter. Our system investigates correlations of URL redirect chains extracted from several tweets. Because attackers have limited resources and usually reuse them, their URL redirect chains frequently share the same URLs. We develop methods to discover correlated URL redirect chains using the frequently shared URLs and to determine their suspiciousness. We collect numerous tweets from the Twitter public timeline and build a statistical classifier using them. Evaluation results show that our classifier accurately and efficiently detects suspicious URLs. We also present WARNINGBIRD as a near real-time system for classifying suspicious URLs in the Twitter stream.*

Keywords: *Suspicious URL, Twitter, URL redirection, conditional redirection, classification*

1. INTRODUCTION

URL shortening has evolved into one of the main practices for the easy dissemination and sharing of URLs. URL shortening services provide their users with a smaller equivalent of any provided long URL, and redirect subsequent visitors to the intended source. Although the first notable URL shortening service, namely tinyURL [3], dates back to 2002, today, users can choose from a a wide selection of such services. 1 The recent popularity of shortening services is a result of their extensive usage in Online Social Networks (OSNs). Services, like Twitter, impose an upper limit on the length of posted messages, and thus URL shortening is typical for the propagation of content. While short URL accesses represent a small fraction of the *"web hits"* a site receives, they are rapidly increasing by as much as 10% per month according to Alexa [1]. Despite this rapid growth, there is, to the best of our knowledge, no other large-scale study in the literature that sheds light onto the characteristics and usage patterns of short URLs. We feel that understanding their usage has become important for several reasons, including: i) Short URLs are widely used in specialized communities and services such as Twitter, as well as in several Online Social Networks and Instant Messaging (IM) systems. A study of URL shortening services will provide insight into the interests of such communities as well as a better understanding of their characteristics compared to the broader web browsing community. ii) Some URL shortening services, such as bit.ly have grown so much in popularity, that they now account for as much as one percent of the total web population per day [1]. If this trend continues, URL shortening services will become part of the web's critical infrastructure, posing challenging questions regarding its performance, scalability, and reliability. We believe that answering these questions and defining the proper architectures for URL shortening services without understanding their access patterns is not feasible.

To understand the nature and impact of URL shortening services, we perform the first large-scale crawl of URL shortening services and analyze the use of short URLs across different applications. Our study is based on traces of short URLs as seen from two different perspectives: i) collected through a large-scale crawl of URL shortening services, and ii) collected by crawling Twitter

messages. The first trace provides insights for a general characterization on the usage of short URLs. The second trace moves our focus onto how certain communities use shortening services.

Phishing costs Internet users billions of dollars a year. Using various data sets collected in real-time, this paper analyzes various aspects of phisher modi operandi. We examine the anatomy of phishing URLs and domains, registration of phishing domains and time to activation, and the machines used to host the phishing sites. Our _ndings can be used as heuristics in _ltering phishing-related emails and in identifying suspicious domain registrations.

2. EXISTING SYSTEM

In the existing system attackers use shortened malicious URLs that redirect Twitter users to external attack servers. To cope with malicious tweets, several Twitter spam detection schemes have been proposed. These schemes can be classified into account feature-based, relation feature-based, and message feature based schemes. Account feature-based schemes use the distinguishing features of spam accounts such as the ratio of tweets containing URLs, the account creation date, and the number of followers and friends. However, malicious users can easily fabricate these account features. The relation feature-based schemes rely on more robust features that malicious users cannot easily fabricate such as the distance and connectivity apparent in the Twitter graph. Extracting these relation features from a Twitter graph, however, requires a significant amount of time and resources as a Twitter graph is tremendous in size. The message feature-based scheme focused on the lexical features of messages. However, spammers can easily change the shape of their messages. A number of suspicious URL detection schemes have also been introduced.

2.1. Disadvantages of Existing System

- Malicious servers can bypass an investigation by selectively providing benign pages to crawlers.
- For instance, because static crawlers usually cannot handle JavaScript or Flash, malicious servers can use them to deliver malicious content only to normal browsers.
- A recent technical report from Google has also discussed techniques for evading current Web malware detection systems.
- Malicious servers can also employ temporal behaviors— providing different content at different times-to evade an investigation

3. PROPOSED SYSTEM

In this paper, we propose WARNINGBIRD, a suspicious URL detection system for Twitter. Instead of investigating the landing pages of individual URLs in each tweet, which may not be successfully fetched, we considered correlations of URL redirect chains extracted from a number of tweets. Because attacker's resources are generally limited and need to be reused, their URL redirect chains usually share the same URLs. We therefore created a method to detect correlated URL redirect chains using such frequently shared URLs. By analyzing the correlated URL redirect chains and their tweet context information, we discover several features that can be used to classify suspicious URLs. We collected a large number of tweets from the Twitter public timeline and trained a statistical classifier using the discovered features.

3.1. Advantages Of Proposed System

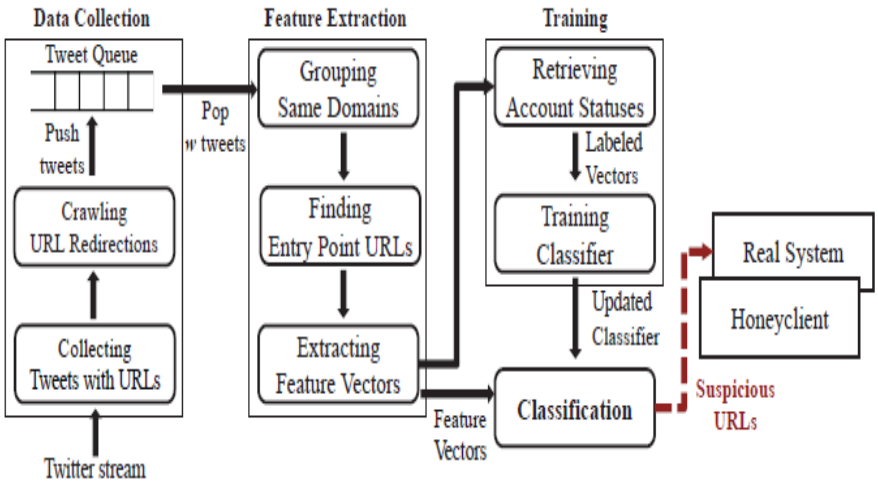
The trained classifier is shown to be accurate and has low false positives and negatives. The contributions of this paper are as follows:

a. We present a new suspicious URL detection system for Twitter that is based on the correlations of URL redirect chains, which are difficult to fabricate. The system can find correlated URL redirect chains using the frequently shared URLs and determine their suspiciousness in almost real time.

b. We introduce new features of suspicious URLs: some of which are newly discovered and while others are variations of previously discovered features.

c. We present the results of investigations conducted on suspicious URLs that have been widely distributed through Twitter over several months.

4. SYSTEM ARCHITECTURE



System overview

5. SYSTEM CONFIGURATION

5.1. Hardware Configuration

Processor - Pentium –IV

RAM - 512MB

Hard disk - 80GB

5.2. Software Configuration

Operating System : Windows XP

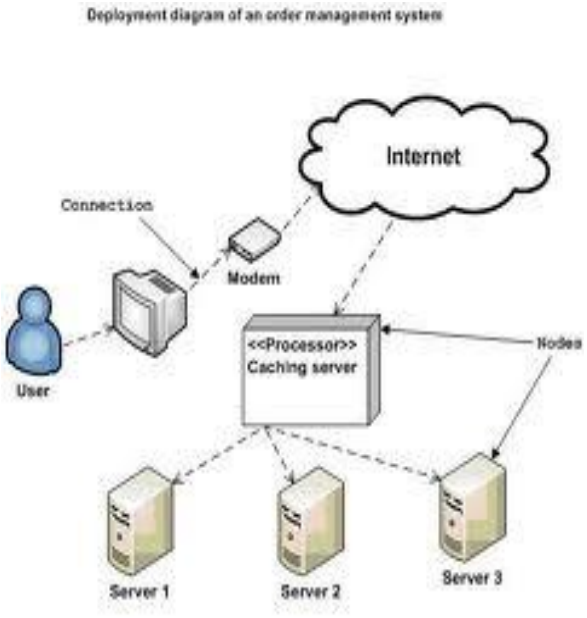
Programming Language : JAVA

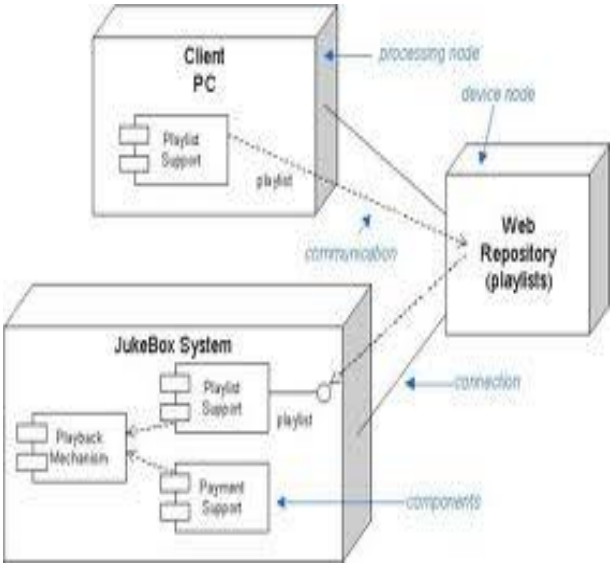
Frontend : JSP, Servlets

Backend : oracle10g

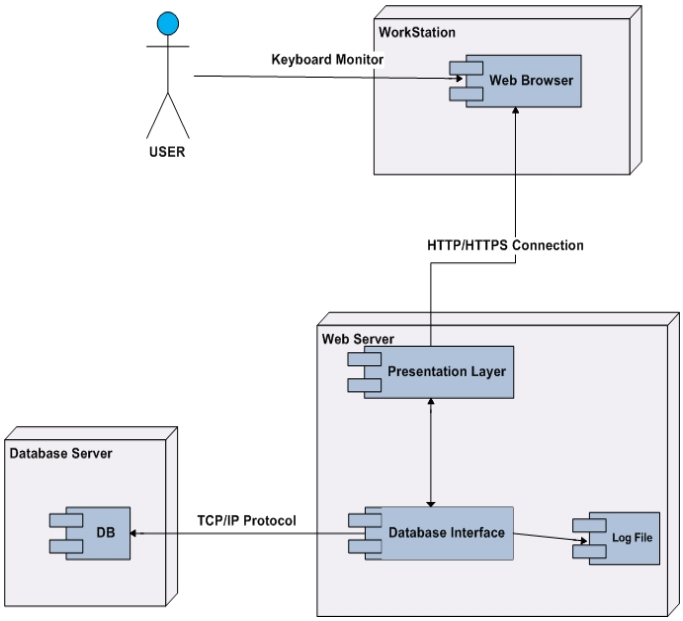
IDE : my eclipse 8.6

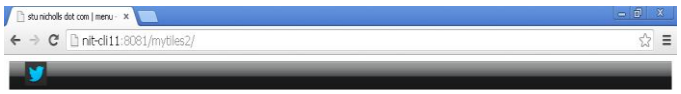
6. COMPONENT DIAGRAM

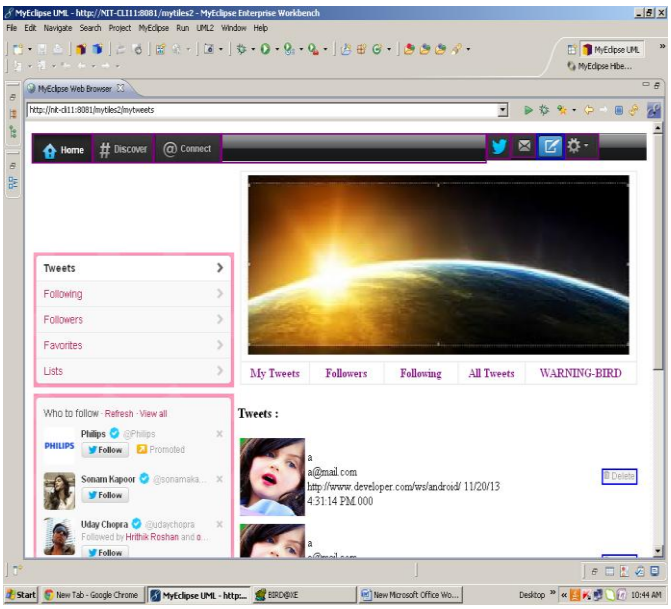




6.1. Deployment Diagram







7. CONCLUSION

Conventional suspicious URL detection systems are ineffective in their protection against conditional redirection servers that distinguish investigators from normal browsers and redirect them to benign pages to cloak malicious landing pages. In this paper, we proposed a new suspicious URL detection system for Twitter, called WARNINGBIRD.

Unlike the conventional systems, WARNINGBIRD is robust when protecting against conditional redirection, because it does not rely on the features of malicious landing pages that may not be reachable. Instead, it focuses on the correlations of multiple redirect chains that share the same redirection servers. We introduced new features on the basis of these correlations, implemented a near real-time classification system using these features, and evaluated the system's accuracy and performance. The evaluation results show that our system is highly accurate and can be deployed as a near real-time system to classify large samples of tweets from the Twitter public timeline. In the future, we will extend our system to address dynamic and multiple redirections. We will also implement a distributed version of WARNINGBIRD to process all tweets from the Twitter public timeline.

REFERENCES

[1] S. Lee and J. Kim, "WarningBird: Detecting Suspicious URLs in Twitter Stream," Proc. 19th Network and Distributed System Security Symp. (NDSS), 2012.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What Is Twitter, a Social Network or a News Media?" Proc. 19th Int'l World Wide Web Conf. (WWW), 2010.

[3] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E.P. Markatos, and T. Karagiannis, "we.b: The Web of Short URLs," Proc. 20th Int'l World Wide Web Conf. (WWW), 2011.

[4] D.K. McGrath and M. Gupta, "Behind Phishing: An Examination of Phisher Modi Operandi," Proc. First USENIX Workshop Large-Scale Exploits and Emergent Threats (LEET), 2008.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who Is Tweeting on Twitter: Human, Bot, or Cyborg?" Proc. 26th Ann.Computer Security Applications Conf. (ACSAC), 2010.

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

[7] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@spam: The Underground on 140 Characters or Less," Proc. 17th ACM Conf. Computer and Comm. Security (CCS), 2010.

[8] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru, "Phi.sh/$oCiaL: the Phishing Landscape through Short URLs,"Proc. Eighth Ann. Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS), 2011.

[9] F. Klien and M. Strohmaier, "Short Links under Attack: Geographical Analysis of Spam in a URL Shortener Network," Proc. 23rd ACM Conf. Hypertext and Social Media (HT), 2012.

[10] K. Lee, J. Caverlee, and S. Webb, "Uncovering Social Spammers: Social Honeypots þ Machine Learning," Proc. 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, 2010.
[11] A. Wang, "Don't Follow Me: Spam Detecting in Twitter," Proc. Int'l Conf. Security and Cryptography (SECRYPT), 2010.
[12] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers on Twitter," Proc. Seventh Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS), 2010.
[13] J. Song, S. Lee, and J. Kim, "Spam Filtering in Twitter Using Sender-Receiver Relationship," Proc. 14th Int'l Symp. Recent Advances in Intrusion Detection (RAID), 2011.
[14] C. Yang, R. Harkreader, and G. Gu, "Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers," Proc. 14th Int'l Symp. Recent Advances in Intrusion Detection (RAID), 2011.
[15] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. Choudhary, "Towards Online Spam Filtering in Social Networks," Proc. 19th Network and Distributed System Security Symp. (NDSS), 2012
[16] J. Ma, L.K. Saul, S. Savage, and G.M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs," Proc. 15th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD), 2009.
[17] J. Ma, L.K. Saul, S. Savage, and G.M. Voelker, "Identifying Suspicious URLs: An Application of Large-Scale Online Learning," Proc. 26th Int'l Conf. Machine Learning (ICML), 2009.
[18] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler: A Fast Filter for the Large-Scale Detection of Malicious Web Pages," Proc. 20th Int'l World Wide Web Conf. (WWW), 2011.
[19] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and Evaluation of a Real-Time URL Spam Filtering Service," Proc. IEEE Symp. Security and Privacy (S&P), 2011.
[20] C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale Automatic