International Journal For Technological Research In Engineering
Volume 5, Issue 7, March-2018 *ISSN (Online): 2347 - 4718*

DETECTION AND ANALYSIS OF MALICIOUS URLs

Amit Kumar[1], Prof. Daya Shankar Pandey[2], DR. Varsha Nam Deo[3]
Research scholar[1], Research Guide[2], Head of Department[3]

Abstract: Social Network Sites (SNS) is the soul of the Internet. It has become a global phenomenon with enormous social as well as economic importance within a few years of their launch. Because of larger user space SNS has become popular day by day. Information exploitation popularity in SNS has attracted not only novice users but also spammers. In SNS spammers are using evolving technology and they safely trading their illegal activities by phishing through e-mails, Social Reverse Engineering (SRE), by posting some incite messages. The novice users often become victim to this malicious activity which impacts them both socially and economically. The study shows that because of this illegal activity the SNS organizers and users are losing $2 million for three months. In this thesis we exploited the security gap that many popular SNS services like Twitter, Facebook do not provide to its users. We have collected a large scale of long URLs and short URLs from multiple sources of SNS which are checked against malicious and non-malicious detectors and we analyses their features to classify the URLs. Our result shows that Native Bayes classifier performs better than other classifier algorithm with accuracy 95.4%

Keyword: Social Network Sites (SNS), Short URLs, Malicious, Phishing,

I. INTRODUCTION

From ancient time man is called as a social animal. From his beginning man has maintained a social relation with nature, animals and with a fellow human being. It was this social relationship that helps him to have a close relationship in the universe with one another. In modern times with increase in population the SNSs has become an easy and a much efficient platform in maintaining social relationships. Online Social Network sites like Facebook, Twitter, Linkedln, MySpace or Google+ has become popular sites in Internet platform. They have attracted of all ages from technicians to novice users. In the wide area sphere like research, industries, business, working Office, news media, organization, entrepreneurship SNS have become a daily practice in use. Mostly SNS have mainly used for information sharing and to express on common interest views example political view.SNS are basically a web-based application which usually allows the individual to construct the semi-public data with in a closed system, articulate a list of users to whom an individual can connect, share information, express their common view in a common platform. SNS allow the individual to meet the strangers of common interest, and view and traverse the list of one's individual connection. Though SNS vary from one to the other in terms of their nomenclature in connection and service provision to its users, the basic principality is to share information.

II. SECURITY ISSUE IN SNSs

For the past few decades the popularity in SNS such as Facebook, Twitter, and Google+ has increased rapidly. Though it has attracted all age groups, but the youngster has out-set among the other groups. SNSs have become an important communication platform in social life, with increasing security concern over a period of time. Some of the security breaches may be like viral marketing, network, structural attacks malware attacks. Some of them are explained briefly as follow.

- ***Privacy Breach Attack***: SNS allow the individual to construct their SNS network with semi-public data like date-of-birth, current address, photo, videos. Such ready available personal information can mark for privacy Breaches.
- ***Breaches by Service Provider***: This readily available data may be used by the one's service provider for advertisement purpose to benefit them in multi-ways. As such the data may fall into the hands of untrustworthy person.
- ***Breach form Third Party***: To have more functionality the user in the SNSs may use the trusted third party application. To use such application the user must have to accept or compromise some privacy issues by accepting theirs term and conditions.

III. SURVEYS OF DIFFERENT URLs

Information Security is of growing interest of policy makers as society become more dependent on secure communication. Anderson and Moore in their research work have briefly explained about the security concern in economic perspective how this malicious content have impact the economic issue. Despite this large malicious activity, information about the malicious content and the losses done by such crimes has largely remained hidden from public crime. The reasons may be as follow. First fear of negative impact on public which arises if incident are openly discard and dis-cussed Second some argues that disclose of information about the incident actually aid attackers more than it helps defenders. Ransbitham has observed that vulnerability in open source software are more frequently exploited by attackers in open software than in closed software. The spammer uses number of techniques to find any vulnerability in the website. The way they can employ is by scanner. Scanner is the technique where the spammers scan the other website and such for some loop holes and inject some rootkits. A rootkit is a stealth type of software, typically malicious design, to hide the existence of certain process from normal method of detection and help to find some loopholes from which they access the privilege of the server. By applying rootkits they compromise the other end

 Copyright 2018.All rights reserved.

software. Then they exploit the machine for their own purpose, and sell to the third party.

3.1 Analysis of Malicious Long URLs

Benevento et al. collected a large scale of URLs of nearly 2 billion and identify the features of the URLs which detect spam on Twitter. And after the manual labeling of features they classified and achieved 70% accuracy. They have founded that a fraction of tweets particularly of some hot topics contains more number of URLs then other. This clearly highlights to what extend the attackers have been using. Though the research has come up the most efficient popular blacklisting in detecting spam but it has been observed that their evaluation technique has not suitable for the detecting spam in the Twitter when the user employs the short URLs technique for the particular long URL as they are in obfuscate in nature. By using these services the attacker has taken more advantage in trading their illicit con-tent. In such case in both Twitter and Facebook by using such services, the spammer has complicated the process of detecting and applying the multiple chain re-directions. Wanget al. in research they used the Click rate measure as a feature and concluded that the rate of spam in Twitter is (0.13%) is higher amount of spam that spread through spam e-mails.

3.2 Analysis of Malicious Short URLs

Kandylas et al. research confirmed that the attackers have used short URLs to trade their illicit work and they found that the malicious short URL by clicked based method. And they concluded that the usage of the short URLs in trading malicious URLs is more then by long URL. They also found that duration of the short URLs is less than the long URLs by which they evade the security check. The dataset they have collected by crawling the webpage, the dataset mainly consist of two domains short URLs and reveal that 50% of the short URLs exceeds 100 days. They have analyzed that the usage of the short URL services is because of the space reduction. Concentrated mainly on the malicious short URLs in the emails and highlighted the privacy and security concern by the short URLs service over the SNSs. They found a lot of private user information traces associated with short URLs and observed a low spam detection rate for 16 shortening services they analyzed. For a particular short URL domain they found that 57% of them are bit.ly.

IV. TYPES OF URLS EXPLOITATION

4.1 Long URLs Data Collection

Our long URL data collection contains the URLs, where some of the URLs land the user the to the intended landing page and few others redirects from one page to the other page. Such a way the attackers employ multiple redirection technique to evade the security scanners. These multiple redirection had made the security observers very difficult to detect the attackers. In some instances the attackers has used the some popular long URLs and modified them to trade their malicious URLs. In some cases they have used the unpopular domain and get registered their URLs whose domain information do not have with the security observer. The other technique the attackers employ is that they registered themselves with the domain for a short period of life span which makes them to evade without any security checking. We have also concentrated on lexical features to detect the malicious of Long URLs, we have used these feature form the standard red tag words, these words are the already available tag words which are used to detect the suspicious features of the URLs. In our research, we have used nearly 13 features which are as below.

4.2 Short URLs Data Collection

Because of the evolution of the emerging technology, the attackers have used the most efficient methods to trade their illicit content on the SNSs, one of the such technology is the usage of the short URLs services. Due to the characters limit in social networks (for example 140 characters in Twitter) the users used the social network services as a space, reducing technology in SNSs. Due to the increasing popularity in usage of the short URLs services which comes by obfuscate behavior in SNS it has not only attracted the genuine users but also the attackers to use them as a safe trading methodology for their malicious data trading

V. DETECTION METHODOLOGY

We have used wide range of online detection methodology Figure5.6 to detect whether the collected URLs are malicious or non-malicious. The following fig explain the method we have employ to detect the URLs are malicious or benign . Virus Total stores all the analyses it performs, this allows users to search for reports given an MD5, SHA1, SHA256 or URL. Search responses return the latest scan performed on the re-source of interest. Virus Total also allows you to search through the comments that users post on files and URLs, inspect our passive DNS data and retrieve threat intelligence details regarding domains and IP addresses. Learn more about searching with Virus Total. Phish Tank Phishing is a fraudulent attempt usually employ by the attackers through email, to steal one’s personal information for their benefit. The best to protect from such type of email is to know the behavior of the emails. Phishing emails generally appear that they are delivered by the well-known organizations and they personally entice one individual to theft their personal information such as credit card number, social security number, account number, or password. Most often such mails are received from the sender where the receiver does not have any account with them. Often the email of the one individual data is sold to the attackers by the third party. One should keep in mind that the legitimate organizer does not ask for the personal credential through insecure email. Google Safe Browsing is an online detector service available that provides the registered users API which enable the user to scan against the URLs. In return it checks the whether the query URLs are malicious and non-malicious against their frequently updated list of URLs. It checks against the suspected phishing, mal-ware and unwanted software.

VI. CONCLUSIONS

Malicious URL detection plays a critical role for many cyber

 Copyright 2018.All rights reserved.

security applications, and clearly machine learning approaches are a promising direction. In this article, we conducted a comprehensive and systematic survey on Malicious URL Detection using machine learning techniques. In particular, we offered a systematic formulation of Malicious URL detection from a machine learning perspective, and then detailed the discussions of existing studies for malicious URL detection, particularly in the forms of developing new feature representations, and designing new learning algorithms for resolving the malicious URL detection tasks. In this survey, we categorized most, if not all, the existing contributions for malicious URL detection in literature, and also identified the requirements and challenges for developing Malicious URL Detection as a service for real-world cyber security applications. Finally, we highlighted some practical issues for the application domain and indicated some important open problems for further research investigation. In particular, despite the extensive studies and the tremendous progress achieved in the past few years, automated detection of malicious URLs using machine learning remains a very challenging open problem. Future directions include more effective feature extraction and representation learning (e.g., via deep learning approaches), more effective machine learning algorithms for training the predictive models particularly for dealing with concept drifts (e.g., more effective online learning) and other emerging challenges (e.g., domain adaption when applying a model to a new domain), and finally a smart design of closed-loop system of acquiring labeled data and user feedback (e.g., integrating an online active learning approach in a real system).

REFERENCE

[1] FacebookDeveloper.https://developers.facebook.com/docs/ apps/.
[2] TwitterDeveloper.http://https://dev.twitter.com/.
[3] BitlyDeveloper.http://dev.bitly.com/index.html.
[4] VirusTotalDeveloper.https://www.virustotal.com/en / community/.
[5] GoogleSafeBrowsing.https://developers.google.com / safe-browsing/.
[6] DaronAcemoglu, Munther A Dahleh, IlanLobel, and AsumanOzdaglar.Bayesian learning in social networks. The Review of Economic Studies, 78(4):1201–1236, 2011.
[7] DaronAcemoglu, AsumanOzdaglar, and Ali ParandehGheibi. Spread of (mis) information in social networks. Games and Economic Behavior, 70(2):194–227, 2010.
[8] Ross Anderson and Tyler Moore.The economics of information security.Sci-ence, 314(5799):610–613, 2006.
[9] Fabr´ıcioBenevenuto, Tiago Rodrigues, Meeyoung Cha, and Virg´ılio Almeida. Characterizing user behavior in online social networks. In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, pages 49–62. ACM, 2009.
[10] Andre Bergholz, Jeong Ho Chang, Gerhard Paaß, Frank Reichartz, and Siehyun Strobel .Improved phishing detection using model-based features. In CEAS, 2008.
[11] Leyla Bilge, EnginKirda, Christopher Kruegel, and Marco Balduzzi. Exposure: Finding malicious domains using passive dns analysis. In NDSS, 2011.
[12] Aaron Blum, Brad Wardman, Thamar Solorio, and Gary Warner. Lexical feature based phishing url detection using online learning. In Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security, pages 54–60. ACM, 2010.
[13] Chia-Mei Chen, DJ Guan, and Qun-Kai Su. Feature set identification for detect-ing suspicious urls using bayesian classification in social networks. Information Sciences, 289:133–147, 2014.
[14] Sidharth Chhabra, Anupama Aggarwal, Fabricio Benevenuto, and Ponnurangam Kumaraguru. Phi.sh$ocial: the phishing landscape through short urls. In Pro-ceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, pages 92–101.ACM, 2011.
[15] Hyunsang Choi, Bin B Zhu, and Heejo Lee. Detecting malicious web links and identifying their attack types. In Proceedings of the 2nd USENIX conference on Web application development, pages 11–11. USENIX Association, 2011.
[16] Nicole B Ellison et al. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1):210–230, 2007.
[17] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In Proceedings of the 2005 ACM workshop on Privacy in the electronic society, pages 71–80. ACM, 2005.
[18] DJ Guan, Chia-Mei Chen, and Jia-Bin Lin. Anomaly based malicious urldetec-tion in instant messaging. In Proceedings of the joint workshop on information security (JWIS), 2009.
[19] Huaping Hu and Jianli Wei. Instant messaging worms propagation simulation and countermeasures. Wuhan University Journal of Natural Sciences, 12(1):95–100, 2007.
[20] Emily M Jin, Michelle Girvan, and Mark EJ Newman. Structure of growing social networks. Physical review E, 64(4):046132, 2001.
[21] Vasileios Kandylas and Ali Dasdan.The utility of tweeted urls for web search. In proceedings of the 19th international conference on World wide web, pages 1127–1128. ACM, 2010.
[22] Florian Klien and Markus Strohmaier. Short links under attack: geographical analysis of spam in a urlshortener network. In proceedings of the 23rd ACM conference on Hypertext and social media, pages 83–88. ACM, 2012.
[23] Rui Li, Kin Hou Lei, Ravi Khadiwala, and KC-C Chang. Tedas: A twitter-based event detection and analysis system. In Data engineering (icde), 2012 ieee 28th international conference on, pages 1273–1276. IEEE, 2012.
[24] Federico Maggi, Alessandro Frossi, Stefano Zanero,

 Copyright 2018.All rights reserved.

GianlucaStringhini, Brett Stone-Gross, Christopher Kruegel, and Giovanni Vigna. Two years of short urls internet measurement: Security threats and countermeasures. In proceedings of the 22nd international conference on World Wide Web, pages 861–872. Interna-tional World Wide Web Conferences Steering Committee, 2013.

[25] Niels Provos, Dean McNamee, Panayiotis Mavrommatis, Ke Wang, Nagendra Modadugu, et al. The ghost in the browser analysis of web-based malware. In Proceedings of the first conference on First Workshop on Hot Topics in Under-standing Botnets, pages 4–4, 2007.

[26] Sam Ransbotham and Gerald C Kane. Membership turnover and collabora-tion success in online communities: Explaining rises and falls from grace in wikipedia. MIS Quarterly-Management Information Systems, 35(3):613, 2011.

[27] Lisa Singh and Justin Zhan.Measuring topological anonymity in social networks.In Granular Computing, 2007.GRC 2007. IEEE International Conference on, pages 770–770. IEEE, 2007.

[28] PravinSoni, ShamalFirake, and BB Meshram. A phishing analysis of web based systems. In Proceedings of the 2011 International Conference on Communication, Computing & Security, pages 527–530. ACM, 2011.

[29] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting spam-mers on social networks. In Proceedings of the 26th Annual Computer Security Applications Conference, pages 1–9. ACM, 2010.

[30] Brian K Tanner, Gary Warner, Henry Stern, and Scott Olechowski. Koobface: The evolution of the social botnet. In eCrime Researchers Summit (eCrime), 2010, pages 1–10. IEEE, 2010.

[31] Kurt Thomas. The koobface botnet and the rise of social malware. 2010.

[32] Kurt Thomas, Chris Grier, and David M Nicol. unfriendly: Multi-party privacy risks in social networks. In Privacy Enhancing Technologies, pages 236–252. Springer, 2010.

[33] Bimal Viswanath, Ansley Post, Krishna P Gummadi, and Alan Mislove. An analysis of social network-based sybil defenses. ACM SIGCOMM Computer Communication Review, 41(4):363–374, 2011.

[34] De Wang, Shamkant B Navathe, Ling Liu, DaneshIrani, AcarTamersoy, and CaltonPu. Click traffic analysis of short url spam on twitter. In Collaborative Computing: Networking, Applications and Work sharing (Collaboratecom), 2013 9th International Conference Conference on, pages 250–259. IEEE, 2013.

[35] David Weiss. The security implications of url shortening services, 2009.

[36] PengYali and Yu Min. Research of intrusion detection technology and its for-mal modeling. International Journal of Information Technology and Computer Science (IJITCS), 1(1):33, 2009.

 Copyright 2018.All rights reserved.