Alghamdi, Bandar Abdulrahman A, Watson, Jason, & Xu, Yue
(2016)
Toward detecting malicious links in online social networks through user behavior.
In Ferilli, S, Reformat, M, & Belkasim, S (Eds.) *Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW).*
Institute of Electrical and Electronics Engineers Inc., United States of America, pp. 5-8.

https://doi.org/10.1109/WIW.2016.014

Toward Detecting Malicious Links in Online Social Networks through User Behavior

Bandar Alghamdi, Jason Watson, Yue Xu
Faculty of Science and Engineering
Queensland University of Technology
Brisbane, Australia
bandar.alghamdi@hdr.qut.edu.au, {ja.watson, yue.xu}@qut.edu.au

***Abstract*—It is becoming increasingly difficult to ignore the importance of using online social networks (OSNs) for various purposes such as marketing, education, entertainment, and business. However, OSNs also open the door for harmful activities and behaviors. Committing financial fraud and propagating malware and spam advertisements are very common criminal actions that people engage in by accessing uniform resource locators (URLs). It has been reported that advanced attackers tend to exploit human flaws rather than system flaws; thus, users are targeted in social media threats by hour.**

This research aims to understand the state of literature on detecting malicious URLs in OSNs, with a focus on two major aspects: URL and OSN objects. Although the literature presents these two aspects in a different context, this paper will mainly focus on their features in relation to malicious URL detection using classification methods. We firstly review three features of URLs: lexical features, hosts, and domains, then we discuss their limitations. We then introduce social spam analytics and detection models using combined features from both URLs and OSNs, particularly the use of user profiles and posts together with URL features, to enhance the detection of malicious behavior. This combination can help in understanding the interests of the user either explicitly, by stating choices in the profile, or implicitly, by analyzing the post behavior, as the spammers do not maintain a regular interest and tend to exploit events or top trend topics.

Keywords—URL; social media; malicious URL detection.

I. Introduction

It is becoming increasingly difficult to ignore the importance of using online social networks (OSNs) for various purposes such as marketing, education, entertainment, and business. OSNs facilitate the way that information is communicated and shared between people, and they have been a tremendously successful route for doing so. For example, Facebook has more than 900 million active users on average, with a 17% increase every year[1]. Twitter has 320 million monthly active users,[2] and 500 million tweets are posted every day.[3]

However, OSNs also open the door for harmful activities and behaviors. Committing financial fraud and propagating malware and spam advertisements are very common criminal actions that people engage in by accessing uniform resource locators (URLs) [1]. A report released in 2016 by Proofpoint—a leading information security company—states that advanced attackers tend to exploit human flaws rather than system flaws; thus, people are targeted in social media threats by hour [2]. With this huge growth in the amount of OSN data, it is challenging to distinguish malicious links from non-malicious links that use dynamic features that do not evolve over time.

This paper will focus on two major aspects: URLs and OSNs. Although the literature presents these two aspects in a different context, this paper will mainly focus on their features in relation to malicious URL detection.

II. Classification Using URL Features

This section presents the existing extracted URL features that are related to the host, domain, and lexical characteristics. These features sometime overlap, especially the host and domain features, and also can be used together. This section first describes each feature separately, and then analyses and discusses the drawbacks of these features.

A. Lexical Features

Lexical features reflect some characteristic of a URL as a string; for example, the length of the URL, the length of the host name, and the number of dots present in the URL. Researchers primarily use lexical features to identify websites, blogs, and URLs. One advantage of using lexical features is that the content from the entire web page is not needed in order to analyze it [3]. Therefore, it can be efficiently used in real-time detection. [1] used lexical features for the aforementioned reason to classify phishing URLs. They argued that phishing links tend to have a certain pattern of URL length that differs from legitimate URLs. The lexical features were used initially by [3] and [4]; however, [1] added more features extracted from the host name and the URL path. These features are strings, delimited by '/', '?', '.', '=', '-', and '_', to classify the URL. The Markov model used in this study to model these textual properties then different classifying algorithms were used resulting in accuracy of 95%.

Very similar work was done by [5] using the same features; however, they included a bigram language model to characterize the host name portion of each URL. As Feroz and Mengel (2014) noted, the key point of using the bigram is that the model has the ability to capture the randomness of the string in a particular URL. This classifier has an accuracy of 97%.

[1].http://investor.fb.com/releasedetail.cfm?ReleaseI D=908022
[2] https://about.twitter.com/company
[3] http://www.internetlivestats.com/twitter statistics/

The approach used by [1] was adapted by [6] with the aim of increasing the accuracy of the classification of phishing URLs using additional obfuscation-resistant lexical features. They extracted obfuscation from the URL by using hyphens in the domain information instead of dots. They also included the directory, sub-directory, page name, length of the file name and the number of dots presented in URL. Finally, the Whois query that provides information about the website's owner were used with a combination of lexical characteristics mentioned in four classification algorithms; as a result, the accuracy of the experiments ranged from 96% to 98%. Their experiments showed that using obfuscation resistant lexical features successfully improved the overall accuracy, and enabled them to filter more phishing URLs using only lexical features.

B. *Host-based Features*

Typically, host-based features are used with lexical features to enhance the detection algorithm and improve the classification accuracy. The classifiers used to distinguish malicious URLs from legitimate URLs are more accurate when the most relevant features are extracted. The host-based features of any URL has rich information about the website that hosts the URL [7], and can be extracted by a simple query known as Whois. This query can provide information about the registrar, and who the registrant is, as well as data about the registration, updates, expiration, and other information.

Fette, Sadeh, and Tomasic published a paper in which they described how to detect phishing URLs in an email [8]. They used the IP (Internet Protocol) for a URL, as they assumed that the phisher might store the website on a normal personal computer (PC) that did not have domain name system (DNS) entries. They also included the domain age and compared the registration data with the email that was sent. If the elapsed time was less than 60 days, they labelled the email as a phishing email. Additional features were used with the host-based features mentioned previously, and this study achieved an accuracy of 99%.

It is worth mentioning that the study on lexical features conducted by [1] used a combination of host and lexical features to enhance the detection rate. They used host features such as IP, connection speed, and registrar information together with the lexical features discussed in Section IIA and achieved accuracy with 95%.

C. *Domain-based Features*

Domain features and host features can partly overlap since they provide valuable information about the underlying infrastructure of a particular website. Based on the domain information such as IP, domain age and some DNS queries, a wide range of blacklist lookup services can be used to detect malicious URLs. These include Google Safe Browsing, Virus Total, Spamhaus and Web of Trust.

Several studies have utilized domain information to detect malicious URLs. [9] used the page rank, domain name, and lexical features as the main features to classify phishing URLs. Page rank is the numeric value, ranging from 0 to 10, which determines the importance of a given web page in relation to other web pages. Based on this ranking they argued that phishing pages have short lifespans and thus have a lower page rank. They used a logistic regression classifier and achieved an accuracy of 97.3%. That study's findings were supported by [10] who, after monitoring a group of suspicious domains for five months, found that more than 80% of them no longer existed. This indicated that the domains were created for malicious purposes. Thus, [10] included domain age as one of the features in the different classification algorithms, and they achieved 86.41% accuracy with the Random Forest classifier.

[11] also focused on domain features to determine the efficiency of the blacklist based on the domain name. They focused on `.com,` the top-level domain (TLD), so they could access the historic server information using a DNS zone file to predict malicious use of domain registration features. They started using a small group of known bad domains, and they were able to infer quite a large number of other bad domains. Using name server (NS) features and registration information from the domains, they achieved an accuracy of 93%.

D. *Issues Related to URL Features*

All the URL features introduced in this section can be used in a classification algorithm either separately or in combination with one another. Although using URL features has been shown to result in a high percentage of overall accuracy, attackers use different evasion techniques, making it useless to detect URLs based on existing features.

- Obfuscation of the lexical features: when attackers use URL shortener services in OSNs they are be able to reduce the length of a URL causing fewer lexical features to be captured [1]. This drawback was also noted by [12] who found that attackers use the URL shortener services extensively to obfuscate phishing URLs.
- Domain and host hacking: for the host- and domain-based features [13] mentioned that 80% of phishing link attacks use hacked domains that actually refer to legitimate users. Consequently, using the host and domain registration information could fail if the classifier depends heavily on the registration information because it appears as a legitimate domain. Thus, an increase in the percentage of attacked domains will also result in the same percentage of classification failures.
- Availability of the services: The Whois query, which all the previously mentioned studies used to collect registration information, might not be always available.

III. URL Classification using OSN Features

This section looks at the user profile and post features. The reason for investigating user profiles and posts can be attributed to the idea that a person who subscribed to a particular social networking site will have a profile based on the standards of the site. According to a study by [14] it is possible to detect spammers or attackers on social sites simply by using their profile as a guide. Regarding the posts which are generated by users to share information on social sites, simply it is the way of propagation of malicious attacks and can help inferring some information about the behavior of the attacker. We will discuss the current techniques used to identify the attacker's behavior based on the profile and post characteristics.

Finally, we will discuss how the combination of URL features and OSN features can help improve malicious classification.

A. User Profile-based Features

The profile is a page that contains the personal information of a particular user's account such as name, gender, photo, etc. A profile can be used to differentiate individual users in OSNs and help with understanding social behavior in relation to posts or messages sent by a particular user. A compromised account on a social networking site is an interesting possible source of malicious URLs. A compromised account is a legitimate account that has been taken by an attacker to leverage the trust relationship that the account has. In the research conducted by [15] compromised accounts were shown to be very useful to spammers and malicious entities to hack into the network. In this manner, the individuals connected to the compromised accounts can maintain their trust without realizing the requesting account is maliciously compromised. This allows the attackers to efficiently distribute spam ads, phishing links, and malware, all of which can include malicious URLs. This confirms the assumptions made by Stringhini [14] whereby a profile in a social network can become a platform for malicious content distribution.

In a study made by Hua and Zhang [16] it was noted that, particularly for Twitter, spam and non-spam profiles overlap in numbers of aspects, making it challenging to identify malicious contents across the network. However, certain characteristics are prominent among spam profiles, including young accounts, tweeting at a higher rate, having more statuses, containing spam words, and rarely being Twitter verified.

[17] placed emphasis on the creation of spammer accounts in Twitter and Myspace. In Twitter, spam accounts have links embedded into their profiles that contain pornographic URLs, advertisements for marketing, duplicate usernames, and phishing URLs. In contrast, Myspace spam classification has some details on the content profile of the accounts, and many of these spam profiles are of a sexual and advertising nature, which can become a very good platform for the inclusion of malicious URLs.

B. Post-based Features

URL sharing is a significant part of social media systems in which users exchange URLs. In one study [18] it was found that up to 25% of social medial posts contain URLs. Here, Cao found that forwarding messages is one reason for the propagation of malicious URLs on OSNs, and that users are more likely to forward malicious URLs when they come from their friends [18]. Since OSN forwarded messages are based on trust, it may be hard to prevent URL based malicious attacks. Therefore, Cao proposed training a detection model for malicious URLs within the bounds of OSNs using post account, intensity, and standard deviation. In the same study, it was inferred that behavioral signals are more difficult to manipulate than message content, thus are more robust in identifying the malicious URL that is being shared.

Another study by [19] points out that the behavior of users on a social network, particularly on Twitter, depends on many different factors such as interests, followers, time, and comments from others . However, there are different types of accounts that can be observed in the Twitter networking space. This study identified at least four account categories, namely individual accounts, news platform accounts, advertising accounts and robot accounts. Individual accounts are maintained by private individuals who use Twitter as a mode of communication. These people post updates on their accounts and even use URLs for their published texts. However, most of users are considered random, including their communications with friends, reading news, or posting messages to friends. In this case, the inclusion of a URL could also signify the possible propagation of malicious contents.

In contrast, news based organizations are publishing news feeds that link back to their websites. Meanwhile, a marketing user uses the Twitter platform in order to promote their products and services, and robot users publish content automatically from third party organizations [19]. By using multi scale entropy, it was found that Twitter users have complex posting behavior. This is very similar to the findings of Cao, in the aforementioned study, where the behavior of users in OSNs is more difficult to analyze, especially if the concern is detecting malicious URLs. Individuals posting on Twitter are harder to measure than other types of users because the random characteristics do not differentiate posts that contain malicious URLs from those that do not.

IV. A Suggested Model for the Detection of Malicious URLs

As can be seen, OSNs have many advantages in identifying malicious users in a real dynamic environment. To improve classifications that can scale efficiently and handle URLs in OSNs there is a need to combine the relevant features of URLs and OSNs. Many researchers, such as [20, 21] and [22] have proposed this combination; however, little attention has been paid to the interests of users of OSNs, which can be seen by their profile settings, their posts or messages, and the URLs that are being posted at any given time. Since attackers intend to propagate malicious URLs, their interests change frequently to exploit any event that has a high trend or active users. Therefore, there is a need to have a deeper understanding of users' interests. There is also a need to apply related social theories to obtain a comprehensive and conceptual understanding of users' interests.

In this paper we bridge this gap and provide a more detailed understanding of the user's behavior, and use this to detect malicious URLs. The suggested model is driven by a user's interest. This approach is inspired by online personalization, where recommender systems focus mainly on the user's interest in order to recommend relevant topics or products to them. In OSNs, involving in any activities is completely driven by user interest and social habit. This assumption is in different studies to detect compromised profiles in OSNs, such as [15]; however, this research will use it to detect the malicious URLs rather than compromised accounts based on a user's changes of interest. The model has the following components:

- **Data Collection** Including 1) Profile: where the explicit information about a user and the user's interests can be gathered. This helps to extract the features that can infer anything about the user. 2) Messages: where the implicit information about the user can be derived. The framework will look at the messages posted by the user that contain URLs as our main target is to detect malicious URLs.

- **Feature Extraction** Including 1) Profile: the user's features need to be carefully extracted from the profile. These features can found from the profile setting or from historical information. 2) Messages: the user's features can be extracted from messages using text mining techniques.

- **URL Classification** Classifiers will be constructed based on features extracted from user profiles and messages together with features of URLs and OSNs with the hope of identifying malicious URLs with a lower rate of false positives.

Unlike the approach adapted by [23], where two different classifiers are used to deal with the profile and the message separately before combining them, we propose the consideration of both profile and message features to find the link between what users do and what they define in their profiles. We are interested in understanding the user's features through the user's profile and posts and by answering the following questions: 1) How to identify the user's features based on the user's demographical profile; 2) How to identify the user's features based on user generated messages (posts); 3) How to identify URLs' features in online social media. Answering these questions will let us explore further opportunities for accurately detecting malicious URLs in social networks.

V. Conclusion and Future Work

The purpose of this paper was to look at the application of classification methods to the field of malicious URL detection in online social networks. It has shown that there are a number of features and characteristics applied, using classification methods, to detect malicious URLs. Although many have been successful, studies show that malicious URLs still appear in OSNs and more work needs to be done to address this. The nature of OSNs imposes difficulties, which need an efficient approach for malicious URL detection. The current studies on malicious URL detection show that there are still problems with aspects such as features extraction, efficiency, and scalability. This paper also confirms that much more work is needed to combine the common URL features with OSN features to handle large and dynamic content. This research suggests the integration of both post and profile features in addition to the URL features. This integration aims to understand users' behavior and interest through implicit and explicit information in order to identify malicious URLs. Our future work is to build the suggested model by applying text-mining techniques to identify users' behavior and applying classification techniques to identify malicious URLs in OSNs. The result will be published in near future.

References

[1] Ma, J.S., Lawrence K Savage, Stefan Voelker, Geoffrey M. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. in the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009. New York, USA: ACM.

[2] Proofpoint. The Human Factor 2016. [Report] 2016; Available from: https://www.proofpoint.com/us/human-factor-2016

[3] Kolari, P.F., Tim Joshi, Anupam. SVMs for the Blogosphere: Blog Identification and Splog Detection. in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.

[4] McGrath, D.K. and M. Gupta, Behind Phishing: An Examination of Phisher Modi Operandi. LEET, . 8: p. 4.

[5] Feroz, M.N.M., Susan. Examination of data, rule generation and detection of phishing URLs using online logistic regression. in Big Data (Big Data), 2014 IEEE International Conference on. 2014. Washington,DC: IEEE.

[6] Le, A.M., Athina Faloutsos, Michalis. Phishdef: Url names say it all. in INFOCOM, 2011 Proceedings IEEE. 2011. Shanghai: IEEE.

[7] Thomas, K.G., Chris Ma, Justin Paxson, Vern Song, Dawn. Design and evaluation of a real-time url spam filtering service. in Security and Privacy (SP), 2011 IEEE Symposium on. 2011. Berkeley, CA: IEEE.

[8] Fette, I.S., Norman Tomasic, Anthony. Learning to detect phishing emails. in the 16th international conference on World Wide Web. 2007. Berkeley, CA: ACM.

[9] Garera, S.P., Niels Chew, Monica Rubin, Aviel D. A framework for detection and measurement of phishing attacks. in the 2007 ACM workshop on Recurring malcode. 2007. New York, USA: ACM.

[10] Gupta, N.A., Anupama Kumaraguru, Ponnurangam. bit. ly/malicious: Deep Dive into Short URL based e-Crime Detection. in Electronic Crime Research (eCrime), 2014 APWG Symposium on. 2014. Birmingham, AL: IEEE.

[11] Felegyhazi, M., C. Kreibich, and V. Paxson, On the Potential of Proactive Domain Blacklisting. LEET, 10: p. 6-6.

[12] Chhabra, S.A., Anupama Benevenuto, Fabricio Kumaraguru, Ponnurangam. Phi. sh/$ oCiaL: the phishing landscape through short URLs. in the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference. 2011. New York, USA: ACM.

[13] Blum, A.W., Brad Solorio, Thamar Warner, Gary. Lexical feature based phishing URL detection using online learning. in the 3rd ACM Workshop on Artificial Intelligence and Security. 2010. New York, USA: ACM.

[14] Stringhini, G.K., Christopher Vigna, Giovanni. Detecting spammers on social networks. in the 26th Annual Computer Security Applications Conference. 2010. New York, USA: ACM.

[15] Ruan, X., et al., Profiling Online Social Behaviors for Compromised Account Detection. Information Forensics and Security, IEEE Transactions on, 2016. 11(1): p. 176-187.

[16] Hua, W.Z., Yanqing. Threshold and Associative Based Classification for Social Spam Profile Detection on Twitter. in Semantics, Knowledge and Grids (SKG), 2013 Ninth International Conference on. 2013. Beijing: IEEE.

[17] Lee, K.C., James Webb, Steve. Uncovering social spammers: social honeypots+ machine learning. in the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010. New York, USA: ACM.

[18] Cao, C.C., James. Behavioral detection of spam URL sharing: Posting patterns versus click patterns. in Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. 2014. Beijing: IEEE.

[19] He, S.W., Hui Jiang, Zhi Hong. Identifying user behavior on Twitter based on multi-scale entropy. in Security, Pattern Analysis, and Cybernetics (SPAC), 2014 International Conference on. 2014. Wuhan: IEEE.

[20] Robertson, M.P., Yin Yuan, Bo. A social approach to security: Using social networks to help detect malicious web content. in Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on. 2010. Hangzhou: IEEE.

[21] Thomas, K.G., Chris Song, Dawn Paxson, Vern. Suspended accounts in retrospect: an analysis of twitter spam. in the 2011 ACM SIGCOMM conference on Internet measurement conference. 2011. New York, USA: ACM.

[22] Chen, C.-M., D. Guan, and Q.-K. Su, Feature set identification for detecting suspicious URLs using Bayesian classification in social networks. Information Sciences, 289: p. 133-147.

[23] D. Wang, Analysis and detection of low quality information in social networks, in Proc. Ph.D. Symp. 30th IEEE Int. Conf. Data Engineering (ICDE 2014), Chicago, IL, United States (2014).