Journal of Information Security and Applications 35 (2017) 75–84

Contents lists available at ScienceDirect

Journal of Information Security and Applications

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

Identification of phishing webpages and its target domains by analyzing the feign relationship

Gowtham Ramesh*, Jithendranath Gupta, P.G. Gamya

Dept. of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

ARTICLE INFO

Article history:

Keywords:
Phishing
Anti-phishing framework
E-commerce security
Target Domain Identification

ABSTRACT

Phishing is the act of stealing personal information from the online users by impersonating as a statutory source in the cyberspace. Phishers often bait online users to visit their forged webpages to acquire users sensitive information. Most of the anti-phishing techniques today, endeavor to identify the legitimacy of the webpages the user visits and warn them with a phishing label when the webpage is a phish. But, these warnings generated by the anti-phishing tools are generic and does not provide any assistance for the users to safely navigate to the legitimate webpages. Any anti-phishing technique will be incomplete and incompetent without having a victimized domain identification in place. The method proposed in this paper addresses this lacuna by automatically identifying the victimized domain (target domain) of every successfully distinguished phishing webpage. This method initially identifies the possible target domains of the webpage by analyzing the feign relationships which exist between the webpage and its associated domains through the in-degree link associations. Further, a novel Target Validation (TVD) algorithm is used to ensure the correctness of the identified target domain which in turn helps in reducing the false target predictions of the system. The legitimacy of the webpage is further confirmed using the identified target domain. The experiment results show that this method is efficient in protecting users from the online identity attacks and also in identifying victimized domain with over 99% accuracy.

1. Introduction

Electronic spamming involves sending unsolicited bulk messages to users using electronic messaging systems. This minor annoyance has indiscriminately evolved into a new and dangerous form known as phishing where phishers try to lure internet users into revealing their personal and financially sensitive information such as credit card numbers, usernames, and passwords. Criminals spam thousands of online users with social engineered emails along with links to phished webpages. The phishing webpages imitate the look and feel of the popular legitimate websites to deceive the users into revealing their sensitive information [1].

The phishing attack has grown tremendously and has evolved in its strategies and targets over the period of time. Phishing attack in 2016 primarily targeted five industries which include financial industries, cloud storage services, webmail services, payment services, and e-commerce companies. The attack volume on these targets is increased by an average of 33% in 2016 compared to 2015. Similarly, the phishing attack has increased over 300% against government tax departments since 2014. Most of the phishing websites launched today are created using phishing kits. This makes it easier for the attackers to create fake webpages with little technical knowledge. According to PhishLab's research report, more than 29,000 phish kits are sold in the black market. Most of these phish kits create phishing webpages with techniques to evade the simple phishing detection methods. These kits are the primary reason for the bountiful increase in phishing webpages. More than 170,000 unique phishing domains are identified to have hosted phishing webpages in 2016 which are 23% higher than 2015 [2]. These attacks not only lead to financial losses of the organizations and individuals but also indirectly undermines the reputation of the organizations and trust on the Internet. This statistic shows the need for robust approaches to restrain and prevent such increasing phishing attacks [3].

Various countermeasures have been developed to mitigate the inimical effects of the phishing attack which are commonly categorized into technical and non-technical solutions. The non-technical solutions mostly emphasize on educating the novice online users through awareness programs, training, and workshops to correctly identify the phishing emails and websites [4]. The policies adopted by the Law Enforcement Agencies (LEA) against phishing attacks includes restrictions on the domain registrations and issue of punishments such as imprisonment and fines [5,6]. The technical

* Corresponding author.
E-mail address: r_gowtham@cb.amrita.edu (G. Ramesh).

http://dx.doi.org/10.1016/j.jisa.2017.06.001
2214-2126/

solutions are developed with the intention of better classification and also, to overcome the human flaws or ignorance in the view of detecting the phishing webpages. This is an important alternate to the non-technical solutions as it is not as expensive as compared to the human training and also due to the feasibility of implementation at all the times. The technical solutions are typically grouped into black-list and white-list based approaches, heuristics based approaches, visual similarity approaches and multifaceted methods. The black-list and white-list based methods maintain the list of URLs either locally or globally. These methods are inefficient in protecting users from zero-day phishing attacks mainly because 47% to 83% of phishing URL were blacklisted after 12 h [7], on the other hand, all the webpages which are not in the whitelist are irrationally labeled as suspicious. The heuristics based approaches detect the phishing webpages based on the set of characteristics present in it. This method is efficient in protecting users from a zero-day phishing attack, but are merely subjected to the presence of characteristics considered in the webpage. Visual similarity approaches identify phishing webpages by analyzing a set of visual features extracted from suspicious webpages [8]. These approaches require a robust method to retrieve website's visual content, any distortion in retrieving the content of the webpage leads to misclassification. The multifaceted approaches detect phishing webpages using any of the techniques or combinations of techniques in computational informatics. These methods commonly apply techniques like link relationship, ranking relationship, and text similarity relationship on the suspicious webpages to confirm its legitimacy. The anti-phishing techniques developed using multifaceted approaches guarantees relatively reliable results compared to any other anti-phishing techniques [9,10].

Most of the anti-phishing methods today primarily attempt to identify the legitimacy of the suspicious webpages, but lack techniques to identify the victimized legitimate webpage that the phishing webpages mimic, where, the legitimate webpage is referred to as the phishing target [11]. However, any anti-phishing technique would be incomplete without identification of the phishing target, as it plays a crucial role in assisting the users to safely navigate to the legitimate webpages. At times, when the phishers attack less popular or newly created webpages it becomes tough to find the target webpage. Also when phishers use masquerading techniques, detection of the target webpage becomes a challenge. Masquerading techniques include creating a webpage using only embedded objects like images and scripts without using any content that could provide us a clue.

The method proposed in this paper detects phishing webpages and its target domain efficiently by working on all the anticipated lacunas. Moreover, this method identifies legitimacy of the suspicious webpages without depending completely on the external information repositories such as search engines, and other third party data sources. Here, we take the suspicious webpage under scrutiny; visit its links up to level two to check for the possible number of domains that can be reached. This domain count value determines the method to be followed in generating the Target Domain Set. We then formulate a cost matrix based on the relationships that exist between the domains in the Target Domain Set. This cost matrix in turn exposes the strength of feigning relationships which exist between the domains in the Target Domain Set and webpage the user visits. The domain with higher in degree feign relationship will be considered as a target domain. The target domain is further validated using Target Validation (TVD) algorithm to ensure its correctness. Finally, the legitimacy of the webpage will be determined by comparing it with the confirmed target domain. Thus, as the content of the suspicious webpage is the only subject on which our proposed methodology is built on, neither prior knowledge about the site is required nor does it require the training data.

An overview of literature review and related work presented in Section 2. Section 3 covers the architecture of the overall system. In Section 4, we have explained the Target Domain Identification module of our system. Section 5 explains the Target Domain validation and Phishing detection methods. The implementation details, evaluation methodology, experimental results and the limitations of the proposed work are discussed in Section 6. Finally, conclusions are presented in Section 7.

2. Related work

In the recent years, many countermeasures have been developed to overcome the phishing attacks. With the development of phishing techniques over the years, meticulous efforts have been taken in the quest to find an efficient method to determine the legitimacy of the suspicious webpages and forewarn the users about the phishing attack. The current anti-phishing approaches depict many drawbacks that need to be addressed are highlighted in this section. This motivated us to propose an anti-phishing method which attempts to overcome some of the limitations of the existing schemes.

2.1. Whitelist based anti-phishing approaches

The whitelist based anti-phishing approaches maintain a list of safe websites along with necessary information. Any website which is not a part of the whitelist is considered as suspicious and such a website is further scrutinized to detect its legitimacy.

Han et al. [12] developed a whitelist based approach which records the well-known legitimate websites of the user rather than maintaining a universal legitimate sites list. In this approach, every URL which the user visits is recorded along with its LUI (Login User Interface) information and the IP addresses. The users are warned when the account information submitted to the website does not match with the corresponding details that are present in the whitelist. But, this method identifies every legitimate webpage as suspicious when the user visits the page for the first time.

2.2. Blacklist based approaches

In contrast to whitelist based approaches, blacklist approaches maintain a list of known phishing sites along with their corresponding necessary details. The blacklist entries are typically compiled from multiple data sources which include spam traps, user posts, or verified phishes compiled by third parties such as takedown vendors.

Prakash et al. [13] developed a system PhishNet which uses the approximate matching algorithm to check if the URL of the suspicious webpage is in the blacklist maintained. Along with this, five heuristics were also proposed to identify new phishing URLs from entries in the blacklist.

Zhang et al. [14] proposed a system which yields customized blacklists to individuals by using relevance ranking scheme and severity score generated for the user. The ranking scheme uses attackers history and users recent log data to measure how closely they are related, and this value is fused with the severity metric to construct the individualized blacklist. But these blacklist based approaches needs frequent updates from their sources and the rapid growth of the list demands need of massive system resources.

2.3. Heuristic-based approaches

The heuristic-based approaches decide the legitimacy of a suspicious webpage by analyzing the webpage content and using the information from the external and internal repositories.

He et al. [15] proposed a method which identifies the phishing webpages when it detects the claimed identities of the suspicious webpage to be fake. In this method, twelve heuristics are used to correlate the behavior of the suspicious webpage to its claimed identities. This method confirms the phishing attack based on the abnormalities observed in the webpage identity.

Prevost et al. [16] developed an anti-phishing method which uses twenty heuristics to detect the phishing webpages by inspecting the presence of abnormal characteristics. In this research, the authors have also analyzed the effectiveness of the heuristics considered in detecting phishing and legitimate webpages.

Marchal et al. [17] developed a heuristic based anti-phishing and target identification system "Know Your Phish". It extracts 212 features from the webpage most of them from its URL. These feature results are fed as input to the supervised classifier model to determine the webpage legitimacy. The target of the phishing webpage is detected based on the brand or service related keywords extracted from the suspicious webpage. These keywords are used in forming the search query and fed as input to the search engine. The webpages returned by the search engine are grouped based on the domain name and the most frequently repeated domain will be identified as a primary target domain. This method also returns secondary target domains in view of avoiding the possibility of missing the actual target domain.

Mahmood et al. [18] proposed rule-based method to determine the legitimacy of the webpages by assessing the relationship existing between the content and its URL. This method extracts set of features using an approximate string matching algorithm to evaluate the identity of the resources and protocols used in the webpage. These feature results are further fed as input to the classifier to determine the webpage legitimacy. The authors have developed set of rules based on the inferred knowledge obtained through the classification system and have implemented it in the web browser extension.

Xiang et al. [19] proposed an approach CANTINA+ which detects phishing webpages using fifteen heuristics and a machine learning algorithm. Along with these heuristics, they have also deployed two pre-filters namely, hash-based near-duplicate page remover and Login form detector. The first filter is a precompiled list of known phishing URLs and the second filter checks the presence of login forms in the webpage and determines the necessity of further processing. These filters are deployed mainly to reduce false positives and improve the speedup of the system.

Gowtham et al. [20] developed a method which detects phishing webpages by analyzing its fifteen features using machine learning algorithm. Along with these heuristics, they have deployed preapproved site identifier module and login form finder module to reduce the false positives of the system without compromising on the false negatives.

2.4. Multifaceted approaches

Wenyin et al. [21] proposed a method to identify the phishing webpages. This method not only determines the authenticity of a webpage but also determines the target page of the phishing webpage by analyzing its implicit association relationship with other webpages in the web. It uses the concept of the Semantic Link Network (SLN) constructed from the suspicious webpage links to resolve its legitimacy.

Wenyin et al. [10] proposed another anti-phishing approach which constructs a webgraph for the suspicious webpage with its associated webpages to identify the parasitic community. The parasitic community is identified by applying min-cut and max-flow algorithms on the constructed webgraph which in turn helps to narrow down the target webpage. The suspicious webpage is then compared against the target webpage to decide its legitimacy.

Swapan et al. [22] developed a web browser plug-in Virtual Browser Extension (VBEx). It prevents users from accessing phishing websites by establishing a trusted channel between the web browser and a legitimate webpage. This is achieved by verifying the authoritative name server of the webpage the user visits.

Tan et al. [23] developed an anti-phishing system PhishWHO which identifies the legitimacy of the queried suspicious webpage and its associated target domain in three phases. The first phase focuses on identifying the potential uni-gram and multi-gram keywords from the content of the suspicious webpage. These keywords are utilized as a search term and fed as an input to the search engine to identify the possible target domains. A set of features is extracted from each of the webpage resulted by the search engine to pinpoint the target domain. Finally, the system decides the legitimacy of the queried webpage from the identified target domain and the domain name of the queried webpage using a three tire identity matching system.

Varshney et al. [24] proposed a search engine based client side anti-phishing method Lightweight Phish Detector (LPD). This method solely depends on the search engine results to decide upon the webpage legitimacy. A search query is constructed by concatenating the domain name extracted from the URL along with the page title for every webpage the user visits. This search query is given as an input to the search engine and the search results are compared against the domain name of the URL the user visits to decide webpage legitimacy. The user is informed with a green webpage background when the webpage is legitimate otherwise warned with a red background.

Volkamer et al. [25] developed a method TORPEDO, to assist users in detecting malicious links embedded in the phishing emails. This method provides just-in-time and just-in-place tooltips of the URLs when the mouse hovers the hypertexts. These tooltips display the actual URL embedded in the hypertexts and highlights the domain name as well. This helps the users to decide whether to visit the link or to skip the links embedded in the email.

The work proposed in this paper is also motivated by our earlier multifaceted approach of Gowtham et al. [26] which detects phishing webpages and its target webpage. This method considers all the directly and indirectly associated domains of the suspicious webpage to narrow down the possible target domain. The domain name of the suspicious webpage is compared to that of the target webpage domain to detect the phishing attack. But, this method completely relies on the search engine for its operations (retrieving the links which are related to the suspicious webpage). So, the computation time and accuracy of this method depends on the search engines response time and its search results. At times, this method may result in false positive outputs just because the targeted webpage is not listed in the top results of the search engines. This is mainly because the search query may be formed from improper keywords extracted from the suspicious webpage. By considering all aforesaid factors, we have proposed a method described in this paper that does not completely rely on the search engines in determining the legitimacy of the suspicious webpage. Instead, it tries to identify the possible target domains by examining the traces presents in the suspicious webpage itself. Similarly, the correctness of the identified target page is verified before confirming the phishing attack.

The proposed method is also compared with other related methods with respect to six different features as shown in Table 1. This includes ability to detect zero-day phishing webpages, language independence detection method, the level of dependency on the third party resources, protection against pharming attack detection, the capability of detecting phishing webpages hosted in the compromised domains, and automatic validation of the target domain.

Table 1
Comparison of related works with proposed work.

Work	Zero-day phishing detection	Language independence	Third party dependence	Pharming attack detection	Compromised domains	Target Detection and Validation
Han et al.[12]	No	Yes	No	Yes	No	No
Prakash et al. [13]	No	Yes	High	No	No	No
Zhang et al. [14]	No	Yes	High	No	No	No
He et al. [15]	Yes	No (English)	High	No	No	No
Marchal et al. [17]	Yes	Yes	High	No	No	Yes (No validation, Multiple target domains)
Mahmood et al. [18]	Yes	Yes	No	No	No	No
Xiang et al. [19]	Yes	Yes	High	No	No	No
Gowtham et al. [20]	Yes	No	Medium	No	No	No
Wenyin et al. [21]	Yes	Yes	No	No	No	Yes (No validation)
Wenyin et al. [10]	Yes	Yes	High	No	No	Yes (No validation)
Gowtham et al. [26]	Yes	Yes	High	Yes	No	Yes (No validation)
Swapan et al. [22]	No	Yes	No	Yes	No	No
Tan et al. [23]	Yes	Yes	High	No	No	Yes
Varshney et al. [24]	Yes	No	High	No	No	No
Volkamer et al. [25]	Yes	Yes	No	No	No	No
Our work	Yes	Yes	Low	Yes	Yes	Yes

3. System overview

Fig. 1 shows the overall system design. The proposed system identifies phishing websites based on the following certainty that for a phishing website the target will be a legitimate site, whereas, for a genuine website, the system will point to the genuine site itself as its target. On this stand, we identify the phishing webpage by comparing the suspicious webpage with its target.

For a given suspicious webpage, our method initially inspects the existence of at least one login form [20] using the pre-compiled list of login keywords (e.g., passcode, customer number, email, etc.). If login form is present in the suspicious webpage, we proceed to the next step of phishing detection; otherwise, the phishing detection is not required, as the users do not have a threat of revealing their private details.

The phishing detection technique initially checks for the number of unique domains that can be reached from the suspicious page by crawling it up to the depth of two levels. The number of distinct level two domains determines the possibility of reaching the target domain from the suspicious webpage; if the number of domains is less than the threshold value it indicates that the links in the suspicious page may be populated by the attackers and there is a less possibility to reach the target domain merely from the content of the suspicious webpage. Otherwise, the count indicates that the target page can be identified from the content of the suspicious page itself. Based on the count value we determine the method to construct the Target Domain Set. We then analyze the feign relationship that existing between domains in the Target Domain Set and suspicious webpage which is represented in the numeric value. Based on the feign relationship values the target domain will be selected and further verified using TVD algorithm. The legitimacy of the suspicious webpage will be determined based on the identified target webpage.

4. Target Domain Identification

The method detailed in this section at first determines the set of possible target domains from the suspicious webpage. Further, these possible target domains are evaluated and one which has strong feigning relationship with the suspicious webpage is identified as the target domain.

4.1. Threshold limit validation

In this module, we check the number of domains that can be reached from a suspicious webpage by visiting its links up to link depth two. This crawling starts by considering only the own domain links of the suspicious webpage especially links from the login forms. In the subsequent levels, all the links in the crawled webpages are also considered. This domain count is compared against the threshold value to determine the method of constructing the Target Domain Set. When the domain count is higher than the threshold value it indicates that the target domain could be identified from the links of the suspicious webpage itself. Here, the Target Domain Set will be constructed only from the suspicious webpage. Similarly, when the domain count is less than the threshold value it indicates that the links in the suspicious webpage may be populated by the attackers to bypass the anti-phishing tools. In this case, we use the search engine results to construct Target Domain Set instead of depending only on the suspicious webpage.

4.2. Target Domain Set Construction

In this Section, we have discussed our approach to determine the Target Domain Set (T), consisting of domains which are either directly or indirectly associated with the suspicious webpage. In

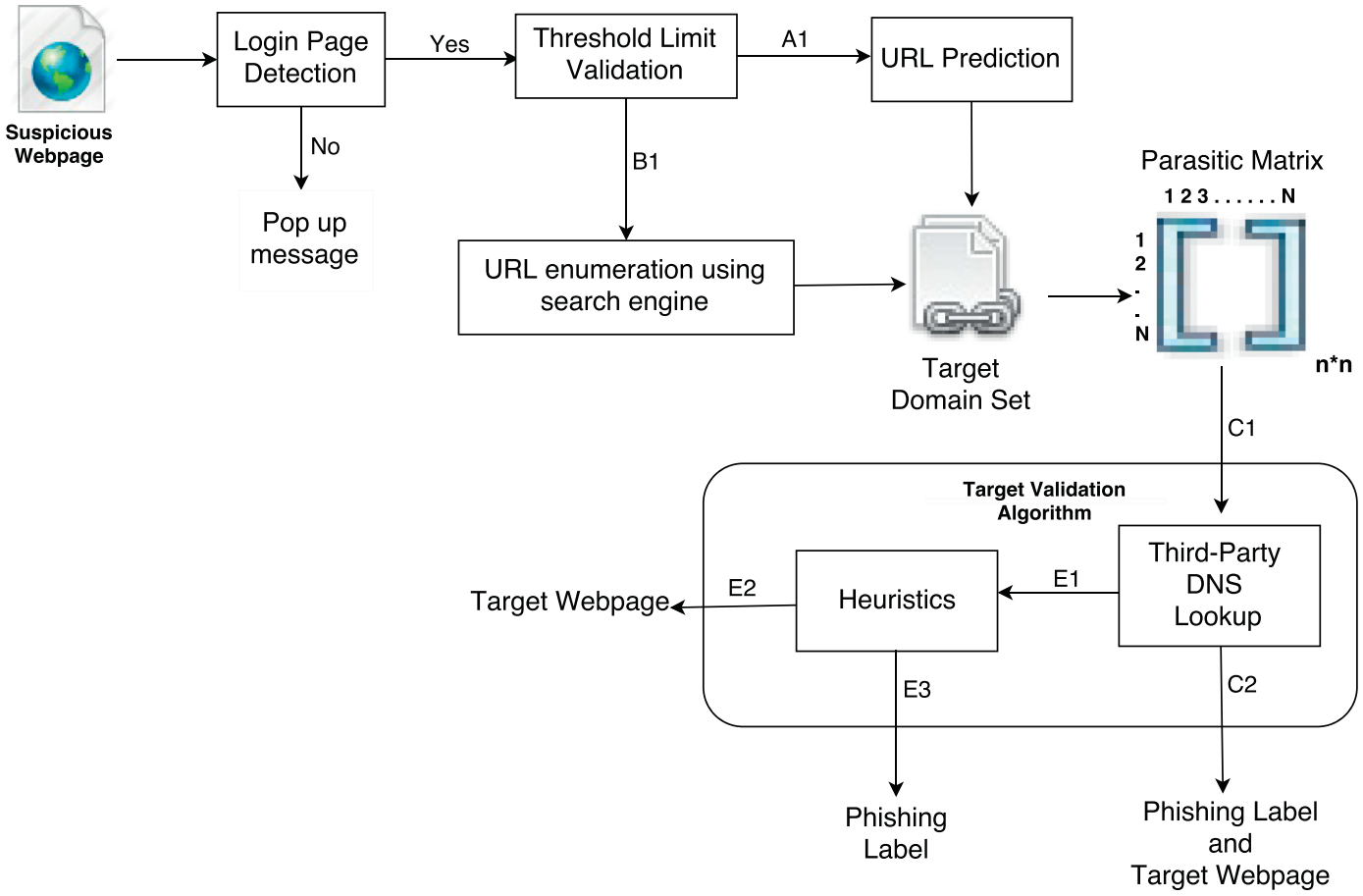

Fig. 1. System overview (A1: Number of level two domains greater than the threshold value; B1: Number of level two domains less than the threshold value; C1: Domain(s) those have highest indegree count; C2: Domain names of the suspicious webpage and the target domain are equal. E1: Domain names of the suspicious webpage and the target domain are different; E2-E3: Heuristics confirms the legitimacy of the suspicious webpage and validates the target domain).

the Target Domain Set Construction, we have used two different methods namely URL enumeration and URL enumeration using the search engine; the exact method would be selected based on the number of domains that can be reached from the suspicious webpage.

4.2.1. URL enumeration

This method is preferred only when the number of domains reached from the suspicious webpage is higher than the threshold value. It gives the maximum possibility for the phishing target identification by analyzing the links present in the suspicious webpage rather than depending on the external repositories for adjudging the legitimacy of the webpage and detecting its target.

The Target Domain Set is constructed by considering URL of the suspicious webpage as well as URLs extracted from the DOM objects of the suspicious webpage. These URLs are specifically extracted from href, src, and alt attribute of the anchor, image, area, script and link objects. Along with these URLs, the URLs predicted from the suspicious webpage is also considered. All these URL's are stored in set *L* and further these links are grouped by domains which result in a target domain set (T).

URL Prediction. Most of the time, phisher generate deceived URLs to steal users private details by applying simple changes to the legitimate URL. These deceived URLs appear like a legitimate URL though they are not. In this module, we apply three simple heuristics on the URL of the suspicious webpage to determine the possible deceived URLs which are active. These heuristics are applied in the chronological order as mentioned below in this Section. On the success of any one of the heuristics; all other heuristics following it are skipped. The active URLs identified are included in the URL set and finally grouped by the domain names and are used to generate Target Domain Set (T). These URLs can be either a legitimate URL deceived by the attacker or another phishing URL. Both these types of URLs helps this system in predicting the legitimacy of the suspicious webpage. The keywords are extracted and terms set is generated from the suspicious URL prior to applying the heuristics.

1. *Brand name check*: This heuristics checks the words in the terms set to find the existence of any brand names maintained in our precompiled list. If it exists, then the corresponding URL maintained in our list will be included in the URL set. This precompiled list is manually populated with 250 pairs of the commonly targeted brand name along with their corresponding URLs.
2. *Domain name in the path of the URL*: Some of the phishing URLs add domain name of the legitimate website within the path segment of the URL to cheat the user. This heuristic extracts the domain name from the path segment if it exists. The domain name is identified by checking the presence of Top-level domains (e.g., net, in), Second-level domains (e.g., co.in, ac.in) or any of its combinations (e.g., com.br, ac.be). If present, it is concatenated with every entry in the term set and added to the URL set. The precompiled list used in this heuristics is gathered from Public Suffix List [27].
3. *Discovering URLs*: This heuristic uses brute force method of combining every word in the terms set with every entry in our

Public Suffix precompiled list (Top-level domains, Second-level domains, and combinations) to generate new URLs. These URLs are further checked for its existence based on its HTTP response code and the active URLs are included in the Target Domain Set.

4.2.2. URL enumeration using search engine

This method is chosen by our system when the number of domains that can be reached from the suspicious page is less than the threshold value. This case occurs only when the links in the webpage are constructed with malicious intentions. Therefore the detection of phishing target from the webpage content alone is not guaranteed. To overcome this problem, webpages that are indirectly associated with the suspicious webpage are retrieved using the search engine. Here, we have used the term frequency-inverse document frequency (tf-idf) scheme to extract keywords from the suspicious webpage. These keywords are extracted from various portions of the webpage which includes title, meta (meta description, meta keywords), and body tags. For every word extracted from the webpage the tf-idf scheme assigns a score, based on its significance to the document. A word that appears often in the document but rarely in the corpus will score high, while the other combinations will not. Finally, top seven keywords are retrieved from tf-idf result based on its score [28]. These keywords are supplied as input to the search engine to identify webpages those are indirectly associated. The top ten links of the search engine result along with the links extracted from suspicious webpage are stored in set L and further grouped according to their domain name which results in the Target Domain Set T.

4.3. Feign relationship Matrix Construction

Phishing webpages are often associated with their target to imitate its behavior and mislead the user to believe that they are communicating with the actual site. This association is possible only through hyperlinks of the suspicious webpage which links to the resources of the legitimate page such as images, CSS and other objects. The feign nature of the suspicious webpage is exploited using the adjacency matrix constructed in this Section to identify the phishing target. This adjacency matrix is constructed from the webpages in the Target Domain Set, where, every member of a Target Domain Set is considered as vertices and the hyperlinks between these webpages are considered as edges. This adjacency matrix maintains a count value for every intersection of a row and column domains which indicates the number of links pointing from a domain in the row to the column domain. The row sum and column sum of this matrix represent out-degree and in-degree links of the domains respectively. The column sum of a domain represents the level of association that exists between the suspicious webpage and the respective domain. The domain with the highest column sum value is considered to be the target host of the suspicious webpage.

The set L contains links which are either gathered from the suspicious webpage or extracted from the search engine results. The links in L, point from the suspicious webpage P to domains in the Target Domain Set T. A new link set L_{di} is a subset of L which is constructed by selecting all the links with the domain name d_i as shown in Eq. (2). Similarly, the L_{di} will be constructed for every domain in T.

$$T = \{d_1, d_2, \ldots, d_n\} \tag{1}$$

$$L_{di} \subset L \quad \text{where } 1 \leq i \leq n \tag{2}$$

The hyperlinks are extracted from the webpages by visiting each of the entry in L_{di} and grouping it based on the domain

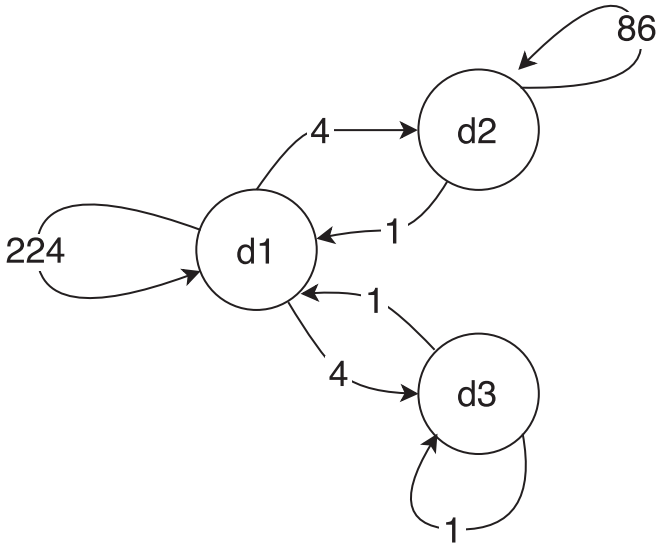

Fig. 2. Links between domains in T.

Table 2
Cost matrix constructed for T.

Domains in T	**d1**	**d2**	**d3**
d1	185.92	0.332	0.332
d2	0.083	7.138	0
d3	0.083	0	0.083
$\sum_{i=1}^{n} C_{i,j}$	186.086	7.47	0.415

names present in the Target Domain Set *T*. We calculate the number of links in every group $N_{i,j}$ where, *i* is currently visiting domain and *j* is a domain in *T*. These count values ($C_{i,j}$) are normalized and then correspondingly reflected in the adjacency matrix. The normalization is done by multiplying the corresponding weight of domain *j* with the count value $N_{i,j}$ as shown in Eq. (3). This minimizes distortion in the count value when links from domain *i* point to any of the non-target domain in *T*. Here, the weight is calculated for every domain by considering its membership in the links set L as shown in Eq. (4).

$$C_{i,j} = N_{i,j} \times W_j \tag{3}$$

$$\text{where} \quad W_j = \frac{|L_{dj}|}{|L|} \tag{4}$$

For example, let us consider the graph constructed from the http://www.ebay.com/ which is shown in Fig. 2, and the corresponding adjacency cost matrix constructed for the webpage is shown in Table 2. The column sum represents the in-degree of each domain in *T*.

5. Target Domain Validation and Phishing Detection

In this section, we check the legitimacy of the suspicious webpage and validate the target domain identified by applying heuristics on the possible target domain(s). The in-degree domain set is constructed by grouping domains which have highest in-degree cost in the adjacency matrix. Any of the domain in the in-degree domain set will be identified as a target domain of the suspicious webpage.

The legitimacy of the suspicious webpage is identified by comparing its domain name with the target domain detected. In this comparison, we have used IP-based comparison instead of string matching to avoid discrepancies in the domain name. Third party

Input: Suspicious Webpage, In-Degree Domain Set

Output: Phishing Label, Validated Target Domain

1. When the domain name of the suspicious webpage matches with the target domain name the webpage will be declared as legitimate
2. If the domain name of the suspicious page does not match with the domain name of the target domain (TD)
   (a) The Keywords extracted from the suspicious page are compared with keywords extracted from the TD if they match, the suspicious page will be concluded as a Phish and target as TD
   (b) If in case the maximum number of links in the suspicious page point to the target domain TD then the suspicious webpage will be considered as a Phish and target as TD
   (c) If the ratio of inactive links is less than 3% then the suspicious webpage is considered as legitimate
   (d) Otherwise, the webpage will be concluded as Phishing. If so, then
      i. Find domain d from the suspicious webpage which has maximum numbers of active links
      ii. If d and TD are same then the target of the suspicious webpage is TD. Otherwise, the possible targets of the suspicious webpage are d and TD
3. Repeat the above steps 1 and 2 for every TD

Fig. 3. Target Validation algorithm.

DNS lookup is used to map the domain names to its IP addresses. The suspicious webpage will be identified as legitimate when its IP address matches with any one of the IP addresses of the target domain. Otherwise, the legitimacy and target domain of the suspicious webpage will be determined using the TVD algorithm as shown in Fig. 3.

- The keywords set extracted from the suspicious webpage domain (d) is compared with keywords set extracted from the target domain (TD). If there a is maximum match then the suspicious webpage is identified as a phishing page and TD will be declared as a target page, whereas if there is minimum or no match between the keywords sets the algorithm proceeds to the next step. This case occurs only when the keywords of the suspicious webpage are explicitly masked by the attacker using scripts and images or the target identified by this method may not be a valid target.
- Most of the phishing websites are created by stealing the source code from the legitimate webpages and customizing it according to their need. This method of creating phishing websites is simple and also gives the look and feel of the legitimate site. Most of the links in these phishing pages point to the legitimate webpage that is being mimicked. In this step, we check links in the suspicious webpage and when most of it points to the target domain TD the legitimacy of the page will be decided as a phishing and target domain as TD. But, this test can be circumvented when attacker manipulates the links in the phishing page. In such cases, the algorithm proceeds to the next step of the detection process.
- When the suspicious webpage contains less than 3% of inactive links [29] and maximum number of URLs pointing to its own domain (d) then the webpage will be recognized as legitimate. On the other hand, if the suspicious webpage contains many numbers of inactive links then it is considered to be a phish. Here, the domain d and TD are recognized as possible target domains of the suspicious webpage. The hyperlinks in the webpage will be classified as inactive links when it points to the web resources that is either permanently unavailable, undesignated (null links, blank links, and void links), or redirects to the same page.

6. Implementation and evaluation

The proposed system is implemented in Java platform standard edition 8. This method takes the URL of the suspicious webpage as input and outputs the legitimacy of the webpage. Here, patterns are used to extract keywords (Popular brand names), words with TLDs and IP addresses from the path segment and subdomains

section of the URL which in turn is used to form the new URLs. The Jsoup HTML parser is used to extract directly associated links present in the webpage. In addition to Jsoup, pattern matching techniques are applied to extract links from the webpages those are not well formed. Guava library is used to extract the domain names from the links retrieved from the webpages.

The threshold value used in Section 4.1 was computed by crawling through the links of 1000 known phishing and legitimate webpages respectively up to level two. On reviewing our URL testing dataset, we observed that in the legitimate pages domain count increased exponentially at every level. As far as the phishing webpages are concerned the links of it either pointed to itself or to the legitimate webpage. The observation on the phishing webpage made us select the threshold value as 21. But, this count may vary with the advancement of the phishing techniques. Thereby as an extension, we are currently working on the method to determine this threshold value automatically.

The keywords extracted from the suspicious webpage are fed as input to a search engine to identify its indirectly associated webpages and for this, we have used Googles Custom Search JSON/Atom API. In this system, OpenDNS is used to map the domain names of the suspicious webpage and the target domain to its corresponding IP address. These IP addresses are further compared to check its equality to determine the legitimacy of the suspicious webpage. The execution time of our system is measured using hrtlib.jar timing library which is capable of observing even a sub-millisecond timing spent.

6.1. Metrics used in evaluation

In our experiments, we have used three metrics to evaluate the performance of the system namely false positive rate (FPR), true positive rate (TPR) and accuracy (ACC).

The false Positive rate (FPR) measures the rate of legitimate webpages that are incorrectly classified as phishing webpages which is computed as shown in Eq. (5).

$$FPR = \frac{N_{L\to P}}{N_L} = \frac{N_{L\to P}}{N_{L\to L} + N_{L\to P}} \tag{5}$$

The true positive rate (TPR) is computed as shown in Eq. (6). This measures the rate of correctly classified phishing webpages in relation with all the existing phishing webpages.

$$TPR = \frac{N_{P\to P}}{N_P} = \frac{N_{P\to P}}{N_{P\to P} + N_{P\to L}} \tag{6}$$

The accuracy (ACC) measures the overall rate of correctly detected phishing and legitimate webpages as shown in Eq. (7).

$$ACC = \frac{N_{L\to L} + N_{P\to P}}{N_{L\to L} + N_{L\to P} + N_{P\to P} + N_{P\to L}} \tag{7}$$

where, the notations $N_{L\to L}$, $N_{P\to P}$ represents correctly classified webpages and $N_{L\to P}$, $N_{P\to L}$ notations represents incorrectly classified webpages by our system.

6.2. Description of test data

Our dataset consists of 3675 active phishing and legitimate websites collected from various sources on the Internet over the period of 10 months from February 2016 to December 2016. Specifically, the corpus consists of 1546 legitimate pages and 2129 unique phishing pages.

The legitimate pages considered for the evaluation are obtained from the three sources as shown in Table 3. Most of these webpages are popular and commonly targeted by the attackers.

The phishing webpages considered in the evaluation are downloaded from two sources listed in Table 4.

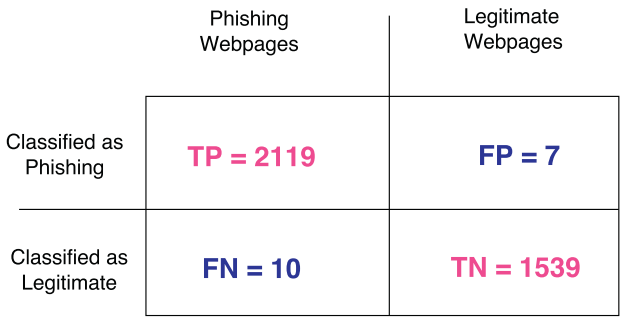

Fig. 4. Confusion matrix of the experimental results.

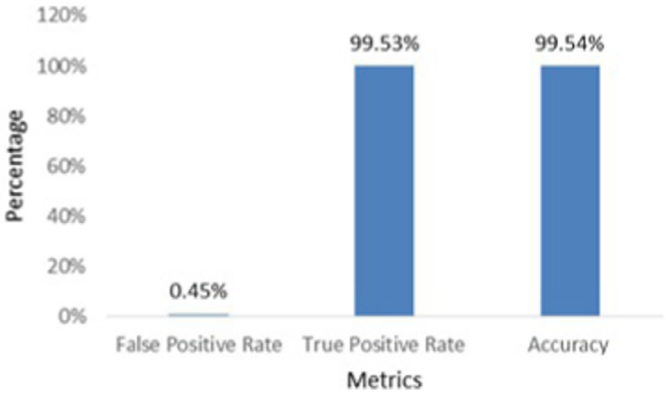

Fig. 5. Assessment of experiments.

6.3. Detection accuracy

The experiment results are as shown in Fig. 4. The true positive rate of the proposed method is 99.53%, the false positive rate is 0.45% and accuracy is 99.54% as shown in Fig. 5. These results clearly show that our system could identify the phishing webpages more precisely with less false positives. In addition to this, our method also identifies the valid target page for every successfully classified webpage.

Though the system seems to be correct in its arena, the key reason for the false positive is that for some of the phishing webpages the number of level two domains reached is higher than the threshold value (discussed in Section 4.1). In this condition, our system predicts the target domain only from the links in the suspicious webpage. But, some of the phishing webpages contains all its links pointing to arbitrary legitimate webpages which are irrelevant to the context of the phishing page. These links neither point to the target page nor to the pages of its own domain. These irrelevant links of the phishing page lead to a false positive prediction of our system. Similarly the false negative occur when legitimate webpages use shortened URLs links to refer the associated domains.

The test result of our method is compared to those of Wenyin et al. [10,21] and Gowtham et al. [26] methods. All these methods considered for the comparison are capable of detecting the phishing webpages along with its target webpage. As shown in Table 5, our method is more advantageous over other methods considered, as our method has a low false positive rate, higher accuracy, and high target detection. The results of other methods (except Gowtham et al. [26] method) are collected from the respective papers and the testing dataset for each method is different.

Table 3
Legitimate webpage data sources.

Source	Sites	Link
Googles top 1000 most-visited sites	704	http://adwords.google.com/da/DisplayPlanner/Home
Alexas Top Sites	393	http://www.alexa.com/topsites
Netcrafts Most Visited Sites	115	http://toolbar.netcraft.com/stats/topsites/
Millersmiles of top targeted sites	334	http://www.millersmiles.co.uk

Table 4
Phishing data sources.

Source	Sites	Link
Phishtank's open database	1565	http://www.phishtank.com/
Reasonable-Phishing webpages list	564	http://antiphishing.reasonables.com/BlackList.aspx

Table 5
Comparison of our experimental results with other anti-phishing methods.

Anti-phishing methods	No. of legitimate pages (L)	No. of phishing pages(P)	Total no. of webpages	TPR	FPR	Accuracy	Target detection
Wenyin et al. [21]	1000	1000	2000	83.4%	13.8%	84.8%	83.4%
Wenyin et al. [10]	–	–	1000 × 10 (random)	–	0.9%	99.2%	92.1%
Gowtham et al. [26]	1546	2129	3675	99.48%	0.51%	99.42%	99.34%
Our Method	1546	2129	3675	99.53%	0.45%	99.54%	99.54%

Table 6
Average runtime and standard deviation of three module.

Modules	Average runtime (In milliseconds)
Threshold limit validation	15,745 ± 16,544
Target Domain Set Construction	176,332 ± 28,487
Target identification module	5766 ± 8653

6.4. Runtime analysis

The experiments were carried out on a computer with a 2.4 GHz processor, 4GB RAM with the internet connection of 1GB bandwidth. In the mentioned environment, our system decides the legitimacy of the suspicious webpages with the average runtime of 29,233ms ± 31,541ms (SD, n=3675). Along with the average runtime, we have also measured the time spent by our system on each of three major modules which include, Threshold limit validation, Target Domain Set Construction, Page validation as shown in Table 6. In threshold limit validation module, we measure the time spent by our system in crawling the suspicious webpage links up to link depth two and domain name count computation. The Target Domain Set Construction module includes time spent in extracting the associated domains of the suspicious webpage and constructing the feign matrix. The target identification module includes turnaround time of DNS query and time taken to check the validity of the target domain detected. Among these three modules the Target Domain Set Construction module has a huge standard deviation. This is mainly because of the method used in enumerating entries of the Target Domain Set.

6.5. Limitations of our approach

Although our approach has been efficient in finding the target domain if links or keywords cannot be extracted from the suspicious webpage, the resulting Target Domain Set will have no or limited domains that are from the target webpage thus decreasing the scope of finding the exact target domain. This is because our approach requires either links or keywords of a suspicious webpage to proceed in the detection of the target domain.

Furthermore, the accuracy of the prediction also relies upon the trustworthiness of the keywords extracted from the suspicious webpage. Thereby in some exceptional cases when the algorithm fails to extract correct keyword, it may lead to erroneous classification.

The prediction time of our system would be proportional to the delay caused by any of the external sources which would, in turn, increase the target prediction time proportionally. But with, today's high-speed internet and availability of alternate sources this bottleneck problem is eliminated. Also, we have used Google Custom Search, which is though efficient, has query limit as 100 making its usage extremely restricted.

7. Conclusion

The system proposed in this paper has an efficiency to identify the phishing websites along with its victimized domain that most of the anti-phishing methods lack. Also, our approach detects newest phishing websites hosted in any language. We have convincing results which show that our system has 99.54% of accuracy in classifying webpages and target domains of the rightly classified webpages are also identified with 100% accuracy. This high detection accuracy is possible only because of the method we have implemented to narrow down the possible target domains from the suspicious webpage, identification of the target domain by exploiting the feigning relationship, and target validation method deployed. Our system also has a future outlook to set the threshold value, which is dynamically constructed for each webpage under our consideration depending on the number of the first level and second level domains.

References

[1] Williams R. Cybercrime costs global economy $445bn annually. Proof point, http://www.proofpointcom/about-us/security-compliance-and-cloud-news/articles/study-assesses-the-full-costs-of-cybersecurity-failures-569572, 20142014.
[2] 2017 phishing trends and intelligence report, 2017. PhishLabs, https://pagesphishlabscom/ Visited: March 2017.
[3] Ferreira A, Lenzini G. An analysis of social engineering principles in effective phishing. In: Socio-technical aspects in security and trust (STAST), 2015 Workshop on. IEEE; 2015. p. 9–16.
[4] Sheng S, Holbrook M, Kumaraguru P, Cranor LF, Downs J. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM; 2010. p. 373–82.
[5] Anti-phishing act of 2005. Cybercrime law, http://www.govtrackus/congress/bills/109/hr1099/text, Visited: Feb 2017.

[6] State of spam and phishing report. Symantec Global Intelligence Network 2010, http://evalsymanteccom/mktginfo/enterprise/other_resources/b-state_of_spam_and_phishing_report_02-2010en-uspdf, Visited: Feb 2017.
[7] Sheng S, Wardman B, Warner G, Cranor LF, Hong J, Zhang C. An empirical analysis of phishing blacklists. In: Proceedings of sixth conference on email and anti-spam (CEAS); 2009.
[8] Hara M, Yamada A, Miyake Y. Visual similarity-based phishing detection without victim site information. In: Computational intelligence in cyber security, 2009. CICS'09. IEEE Symposium on. IEEE; 2009. p. 30–6.
[9] Shahriar H, Zulkernine M. Trustworthiness testing of phishing websites: a behavior model-based approach. Future Gener Comput Syst 2012;28(8):1258–71.
[10] Wenyin L, Liu G, Qiu B, Quan X. Antiphishing through phishing target discovery. IEEE Internet Comput 2012;16(2):52–61.
[11] Xiang G, Hong JI. A hybrid phish detection approach by identity discovery and keywords retrieval. In: Proceedings of the 18th international conference on World wide web. ACM; 2009. p. 571–80.
[12] Cao Y, Han W, Le Y. Anti-phishing based on automated individual white-list. In: Proceedings of the 4th ACM workshop on Digital identity management. ACM; 2008. p. 51–60.
[13] Prakash P, Kumar M, Kompella RR, Gupta M. Phishnet: predictive blacklisting to detect phishing attacks. In: INFOCOM, 2010 proceedings IEEE. IEEE; 2010. p. 1–5.
[14] Zhang J, Porras PA, Ullrich J. Highly predictive blacklisting. In: USENIX security symposium; 2008. p. 107–22.
[15] He M, Horng S-J, Fan P, Khan MK, Run R-S, Lai J-L, et al. An efficient phishing webpage detector. Expert Syst. Appl. 2011;38(10):12018–27.
[16] Gastellier-Prevost S, Granadillo GG, Laurent M. Decisive heuristics to differentiate legitimate from phishing sites. In: Network and information systems security (SAR-SSI), 2011 Conference on. IEEE; 2011. p. 1–9.
[17] Marchal S, Saari K, Singh N, Asokan N. Know your phish: Novel techniques for detecting phishing sites and their targets. In: Distributed cmputing systems (ICDCS), 2016 IEEE 36th International conference on. IEEE; 2016. p. 323–33.
[18] Moghimi M, Varjani AY. New rule-based phishing detection method. Expert Syst Appl 2016;53:231–42.
[19] Xiang G, Hong J, Rose CP, Cranor L. Cantina+: a feature-rich machine learning framework for detecting phishing web sites. ACM Trans Inf Syst Secur(TISSEC) 2011;14(2):21.
[20] Gowtham R, Krishnamurthi I. A comprehensive and efficacious architecture for detecting phishing webpages. Comput Secur 2014;40:23–37.
[21] Wenyin L, Fang N, Quan X, Qiu B, Liu G. Discovering phishing target based on semantic link network. Future Gener Comput Syst 2010;26(3):381–8.
[22] Purkait S. Preventing phishing attacks with virtual browser extension. IUP J Inf Technol 2013;9(3):7.
[23] Tan CL, Chiew KL, Wong K, et al. Phishwho: phishing webpage detection via identity keywords extraction and target domain name finder. Decis Support Syst 2016;88:18–27.
[24] Varshney G, Misra M, Atrey PK. A phish detector using lightweight search features. Comput Secur 2016;62:213–28.
[25] Volkamer M, Renaud K, Reinheimer B, Kunz A. User experiences of torpedo: tooltip-powered phishing email detection. Comput Secur 2017.
[26] Ramesh G, Krishnamurthi I, Kumar KSS. An efficacious method for detecting phishing webpages through target domain identification. Decis Support Syst 2014;61:12–22.
[27] Public suffix list. Mozilla Foundation, http://publicsuffixorg (2007), Visited: March 2017.
[28] Zhang Y, Hong JI, Cranor LF. Cantina: a content-based approach to detecting phishing web sites. In: Proceedings of the 16th international conference on world wide web. ACM; 2007. p. 639–48.
[29] Nelson ML, Allen BD. Object persistence and availability in digital libraries. D-Lib Mag 2002;8(1).