International Journal of Computer Trends and Technology (IJCTT) – volume 12 number 5 – Jun 2014

On URL Classification

Anjali B. Sayamber, Arati M. Dixit

Padambhooshan Vasantdada Patil Institute of Technology.
Bavdhan, Pune 21

Abstract:

Web-based malware attacks are growing threat to today's Internet security. These types of Attacks are common and lead to serious security cost. Malicious links are used as source to the distribution channels to propagate malware all over the Web. Due to which victim systems get easily infected and systems fall in the control of attackers, who can utilize them for various cyber crimes such as stealing credentials, spamming, and phishing, denial-of-service attacks. Security technologies such as browsers, blacklists and popup blockers, firewalls and intrusion detection systems have only limited ability to diminish this new problem. That requires fast and precise systems with the ability to detect new malicious content. This paper introduces various aspects associated with the URL (Uniform Resource Locator) classification to identify whether the target website is a malicious or benign. It introduces classification models learning methods and their approaches. And datasets are used for training purpose.

I.INTRODUCTION

The Internet is a source to unlimited knowledge and information which can be easily used by any person throughout the World Wide Web at any time regardless of the time zone and place issues. Internet usage is an essential part of the modern life people, which takes advantage of what was going to be only used by the scientists and military from the Internet. It is still growing very fast, yet it has taken a control on people's mind from children to elders, by fascinating and making them more dependent on it. The interesting fact is that people have no idea what they would do without it.

The World Wide Web (WWW) is a collection of all existing technologies which are constructed upon the Internet. It simplifies the delivery of a wide range of services to any user from simple services such as reading the news to complicated and classified services like online military services. In order to provide these services to the customers, different technologies are applied, enabling web browsers to become one of the most important communication techniques in this regard. Web browsers play an important role in allowing users to easily interact with the World Wide Web by traversing, retrieving and finally presenting the related topics to them. While, one may say the Internet is a powerful resource to gain knowledge, yet the Internet has another side as well. Many people benefit from using the Internet since they can simply access huge amount of information in little time. This is one of the strongest advantages of internet. However, internet is a reflection of containing both good and bad impacts. The most important issues in using the Internet are related to the user's security. Although for user security various concepts are defined and they might have different levels of obtaining it, yet one common aspect between all is *how to provide it*, especially while they are using online services. Without security, the user might or even might not encounter a threat which can somehow result in gaining access to the user's belongings without his notice and thus, allows the attack to support his system or to simply carry out another different kind of attack that result in losing everything which possesses great value like bank accounts. One type of attack among various existing attacks is malware which is installed and spread easily.

Before using a particular URL if one could inform users that it was dangerous to visit, much of this problem could be solved. To solve these problems the security community has developed blacklisting services, appliances and search engines that provide accurate feedback. The blacklists are particularly human feedbacks that are highly accurate yet time-consuming. Blacklisting [3] is effective only for known malicious URLs. Predictably, many malicious sites are not blacklisted either because they are too new, were never evaluated, or were evaluated incorrectly. To find out solution to this problem, some client-side systems analyze the content or behavior of a Web link when it is visited.

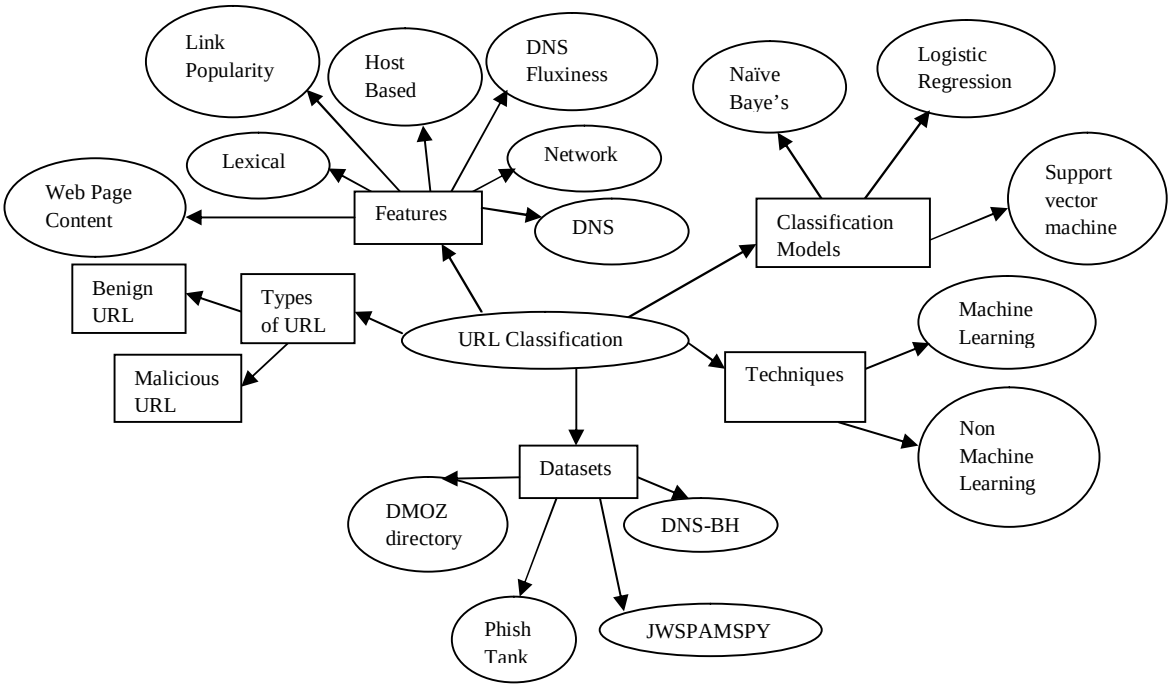


Fig. 1 Classification of URLs

In this paper the *classification of websites* is described with respect to heterogeneous aspects associated with it like type, features, learning models, datasets, models, etc. The rest of the paper is organized as follows: Section II proposes a classification of URL framework which includes datasets. Section III which includes Attack Types as a URL classifier datasets, Section IV describes about the Techniques as a URL classifier. Section V discusses the feature based classification of websites, this includes the list of features, collecting the training and testing datasets. The variety of classification models like support vector machine(SVM), Naïve Bayes, and Logistic regression used for the URL classification are the part of discussion in Section VI. The section VII concludes the paper.

II. CLASSIFICATION OF URLS

A URL classification framework as shown in fig.1 is proposed in this section so as to provide a mechanism for successfully and clearly classify the URLs. The various parameters considered for the proposed classification includes- type of URL, features, datasets, learning approaches, models and attack types. The classification of URLs on basis of '*type*' parameter involves two types - benign and malicious URLs. The malicious URLs are further categorized on basis of attack types of malicious URLs. The variety of attack types are: Spamming, phishing, malware, attack page, Gumblar, sql injection, Fastflux and denial of service etc. Beginning with an overview of the classification problem, for which trained datasets are used as a collection of URLs, followed by a discussion of the learning approaches used for classification on basis of features, and finally support vector machine, Naïve Bayes, and Logistic regression are discussed in detail which are used for the URL classification. URLs status is treated as a binary classification problem where positive examples are malicious URLs and negative examples are benign URLs. For the principle study of classification of URLs trained data is created which is collected from many sources as discussed following.

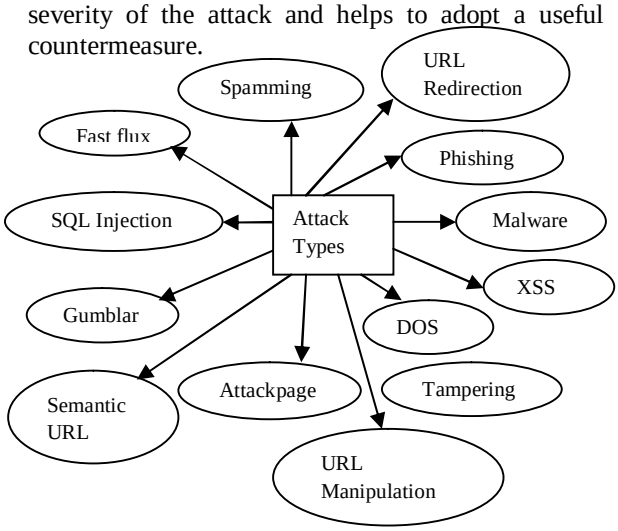
DATA SETS

Benign URLs were collected from two sources 1) DMOZ Open Directory, 2) Yahoo!'s directory [24]. Malicious URLs were collected from these sources: The spam URLs were acquired from jwSpamSpy [10] which is known as an e-mail spam. The phishing URLs were acquired from Phish Tank [16], it is a free community site where anyone can submit, verify,

track and share phishing data. The malware URLs were obtained from DNS-BH [7].

III.TYPES AS A URL CLASSIFIER

The 'type' focuses on the nature of URL, which when accessed is either harmful or not. The URLs, which when accessed do not pose any sort of security threats can be defined as benign URLs. The URLs, which when accessed pose significant amount of threat can be defined as malignant or malicious ones. There are number of attacks on URLs some of them are mentioned here. If attack identifier is familiar with the type of a threat [23] it enables evaluation of severity of the attack and helps to adopt a useful countermeasure.



Some of the attack types are:

- **Spamming-** *Spamming is term applied for sending bulk of unwanted emails like the advertising emails or email bombardments to hold up a product.*
- **Phishing-** *phishing typically involves sending an email seemingly from a trustworthy source to trick people to click a URL contained in the email that links to a fake webpage.* Phishing is used to acquire confidential information such as usernames, passwords, and credit card details by hiding its own identity behaving as a trustworthy entity in electronic communication.
- **Malware-**It is short for malicious software*; it can be in the form of code, scripts, active content, and other software.* Malware is a universal term used to refer variety of forms like unfriendly or interfering software. Malware includes computer viruses like worms, Trojan horses, spyware, adware, and other malicious programs. Malware threats are usually worms or trojans rather than viruses.
- **Denial of Service (DOS)-***In computing, to make a machine or network resource unavailable to its proposed users is an effort of DoS or distributed denial-of-service (DDoS) attack* . Targets of a DoS attack may vary, but it generally consists of efforts for an indefinite period interrupts or suspends services of a host connected to the Internet.
- **Attackpage-** *It is a page, in any namespace, that exists primarily to disparage or threaten its subject; or biographical material which is entirely negative in tone and outsourced.*
- **Gumblar-** *It is a malicious Javascript trojan horse file that redirects a user's Google searches, and then installs rogue security software.*
- **SQL injection-** *It is a code injection technique, used to attack data-driven applications, in which malicious SQL statements are inserted into an entry field for execution.*
- **Fast flux** - *It is a DNS technique used by botnets to hide phishing and malware delivery sites behind an ever-changing network of compromised hosts acting as proxies.*
- **Semantic URL** - *a client manually adjusts the parameters of its request by maintaining the URL's syntax but altering its semantic meaning.*
- **URL Manipulation** - *by manipulating certain parts of a URL, a hacker can get a web server to deliver web pages he is not supposed to have access to.*
- **Tampering attack** -*The parameter modification of form fields can be considered a typical example of Web Parameter Tampering attack.*
- **URL Redirection Attack** - *is a kind of vulnerability that redirects you to another page freely out of the original website when accessed, usually integrated with a phishing attack.*
- **Cross-site scripting (XSS)** - *is a type of computer security vulnerability typically found in Web applications. XSS enables attackers to inject client-side script into Web pages viewed by other users.*

IV. TECHNIQUES AS A URL CLASSIFIER

URLs can be classified into two categories that are malicious links and benign links depending on the techniques used to build the classifier. They are popularly fall in one of the following categories:

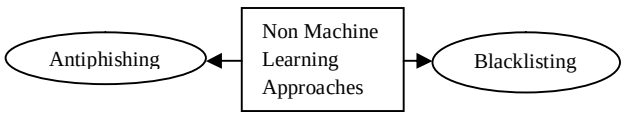
1. Machine learning methods which use machine learning approaches to build classifiers, and
2. Other Non-machine learning methods which build classifiers with a priori knowledge.

1. Machine Learning Approaches

The data observations, measurements, etc. are labeled with pre-defined classes. It is like that a "teacher" who gives the classes. It tests the model using unseen test data to assess the model accuracy. Machine learning approaches are reliable with high accuracy. Ntoulas et al. [13] proposed to detect spam Web pages through content analysis. They used site dependent heuristics, such as words used in a page or title and fraction of visible content. Fette et al. [8] Used statistical methods to classify phishing emails. They used a large publicly available corpus of legitimate and phishing emails. Provos et al. [17] analyzed the maliciousness of a large collection of web pages using a machine learning algorithm as a pre-filter for VM-based analysis. Whittaker et al. [21] proposed a phishing website classifier to update Google's phishing blacklist automatically. They used several features obtained from domain information and page contents. The classification model of Ma et al. can detect spam and phishing URLs. They described a method of URL classification using statistical methods on lexical and host based properties of malicious URLs.

2. Non-Machine Learning Approaches

In Non Machine Learning Approaches Class labels of the data are unknown. They are based on simple Blacklist to block malicious URLs was one of the most popular approaches. A number of websites provide blacklists such as jwSpamSpy [10], Phish-Tank [16], and DNS-BH [7]. McAfee's Site Advisor [12], WOT Web of Trust [19], Trend Micro Web Reputation Query Online System [20], and Cisco Iron Port Web Reputation [4]. A blacklist is a list or register of entities or people who, for some reasons are being denied a particular service. Blacklisting is still one of most popular technique. Whittaker et al. [3] he works offline and analyzed millions of pages daily from the Google's phishing blacklist. Their main contribution was achieving maximum classification accuracy for phishing pages. Phish Net [14] used approximate pattern matching algorithm to match URL components against blacklist entries.



Above techniques always tried to automatically manage blacklists and increase their accuracy although they are still insufficient and suffer from their increasing size and incorrect listing. Blacklists can be combined with other techniques that uses machine learning to classify malicious websites. One of the earliest classification systems for malicious websites was concerned with the detection of SPAM in blog posts. There are certain limitations for URL blacklisting it is ineffective for new malicious URLs and it takes time to analyze malicious URLs. Zhang et al. [22] proposed a more effective blacklisting approach, "predictive blacklists", which uses a relevance ranking algorithm to estimate the possibility that an IP address is malicious.

V. FEATURES as a URL CLASSIFIER

The popular features used for the URL classification fall into following categories McGrath et al. [5] studied phishing infrastructure and the framework of phishing URLs. They pointed out the importance of features such as the URL length, linked-to domains age, number of links in e-mails and the number of dots in the URL. Phish Def [1] used features that resist obfuscation and suggested used the AROW algorithm to achieve higher accuracy. Inconsistency detecting works by extracting features during the normal learning phase based on a specific model. In the testing phase the new feature values for the websites to be tested are checked against the training models representing the normal behavior. The features used include: the number of code executions, code length, number of bytes, shell codes and the difference in returned pages for different browsers and the number of redirections. The Prophiler by Canali [6] used HTML tag counts, percentage of the JavaScript code in the page, percentage of whitespace, entropy of the script, entropy of the strings declared, number of embed tags, presence of Meta refresh tags, the number of elements whose source is on an external domain and the number of characters in the page. While improving accuracy the Prophiler significantly increased the number of features. In addition to the increased overhead due to the statistically processing the page content, those techniques suffered from the inherent danger of having to access the malicious page and download the content before deciding it was malicious.

1) Lexical Features

URL stands for uniform resource locator or formerly the universal resource locator. URL and uniform resource identifier (URI) are comparable and used to identify any document retrieved over the WWW. The URL has distributed in three main parts: the protocol, hostname and path. Malicious URLs, esp. those for phishing attacks, usually have distinguishable patterns in their URL text. Lexical features are the

properties of the URL itself and do not include content of the page it points to. The URL properties include Domain token Count, Path token Count, Average domain token length, Average path token length, Longest domain token length, Longest path token length, Brand name presence, Length of hostname, Length of entire URL, Number of dots in URL, Top-level domain Integer, IP address presence Binary, Security sensitive word presence Binary and tokens in the path URL delimited by '/', '?', '+', '–', '%', '&', '.', '=', and '_'.

2) Host-based Features
Malicious Web sites may be hosted in less reputable hosting centers so it uses host-based features, on machines that are not conventional web hosts, or through disreputable registrars by using WHOIS properties. To an approximate degree, host based features can describe "where" malicious sites are hosted by using Geographic properties, "who" own them, and "how" they are managed by IP address properties and Domain name properties.
Host-based features are derived from the host properties such as the IP address, geographic properties, and domain name properties, DNS time to live (TTL), DNS A, DNS PTR and DNS MX records as well as WHOIS information and dates. Those features are very important and can help any classifier for detection process.

3) Special Features
Some features are simple to get a value for such as JS Enable/Disable, HTML Title tag content (<title></title>), 3-4-5 grams (n-grams) and Term Frequency and Inverse Document Frequency (TF-IDF). Term frequency is the number of times a term occurs in a document. The inverse document frequency is the logarithm of the number of documents divided by the number of documents containing the term and it measures the importance of a term. TF-IDF is commonly used in search engines, classification and data mining and finally 3-4-5 grams take longer to calculate than the other features.
Other features require significant computation time
Such as Anchors or bag-of-anchors which are extracted from all URLs in Anchor tags on the page being examined.

4) Link Popularity Features
One of the foremost necessary options utilized in classification of URLs is "link popularity", that is calculable by investigation the amount of incoming links from alternative websites. Link quality is often thought about as a name live of an address. Malicious sites tend to possess a little price of link quality, whereas several benign sites, particularly in style ones, tend to possess an oversized price of link quality. Each link quality of an address and link quality of the URL's domain are utilized.

5) Webpage Content Features
Recent development of the dynamic webpage technology has been exploited by hackers to inject malicious code in to sites through commerce and so activity exploits in webpage content. Therefore, applied math properties of client-side code within the online page are used as options to observe malicious sites. To extract webpage content options (CONTs), numbers of HTML tag count is considered, iframes, zero size iframes, lines, and hyperlinks within the webpage content. All the options in Table three area unit from the previous work [16].

6) DNS Features
The DNS options are a unit associated with the name of an address. It find that most spam is being sent from a few regions of IP address space, and that spammers appear to be using transient "bots" that send only a few pieces of email over very short periods of time showed that a major portion of spammers came from a comparatively little assortment of autonomous systems.

7) DNS Fluxiness Features
A freshly rising fast-flux service network (FFSN) establishes a proxy network to host extralegal online services with a really high convenience.

VI. MODELS as a URL CLASSIFIER

As described in previous section various features are use to encode URLs. This feature poses certain challenges for URL classification. There are various features which correlate with malicious URLs. It is the system in which statistical models are most prone to over fitting. This section, briefly review models applied for classification. Ma et al. [11] used Phish Tank dataset and validated their work using three machine learning models Naïve Bayes, SVM with an RBF kernel and regularized logistic regression. Later, Ma et al. [9] developed a light weight algorithm for website classification based on features while excluding page properties. It was designed as real-time, low-cost and fast alternatives for black listing.

Naive Bayes[11]: It is commonly used in spam filters. It is a probabilistic method that has a long history in information retrieval and text classification (Maron and Kuhns, 1960). It stores as its concept description the prior probability of each class, $P(Ci)$, and the conditional probability of each attribute value

given the class, $P(v\ jjCi)$. It estimates these quantities by counting in training data the frequency of occurrence of the classes and of the attribute values for each class. Then, assuming conditional independence of the attributes, it uses Baye's rule to compute the posterior probability of each class given an unknown instance, returning as its prediction the class with the highest such value:

$$C = \text{argmax}_{Ci} P(C_i)\prod_j P(v_j|C_i).$$

Support Vector Machine (SVM): To demonstrate the support-vector network method experiments in two parts are conducted. In first part artificial sets of patterns in the plane is constructed and experiment with second degree polynomial decision surfaces, and experiments with the real-life problem of digit recognition. Support Vector Machine (SVM). SVM could be a wide used machine learning methodology introduced by Vapnik et al. [2]. The machine conceptually implements the following concept input vectors are non-linearly mapped to a very high dimension feature space. This feature space constructs a linear decision surface. These Special properties of the decision surface ensure high generalization ability of the learning machine. The support-vector network combines 3 ideas: the solution technique from optimal hyper planes (that allows for an expansion of the solution vector on support vectors), the idea of convolution of the dot-product (that extends the solution surfaces from linear to non-linear), and the notion of soft margins (to allow for errors on the training set).

$$\sum_{i=1}^{n} \propto i - 1/2 \sum_{i,j=1}^{n} \propto i \propto jyiyjK(xi, xj)$$

Subject to

$$\sum_{i=1}^{n} \propto iyi = 0{,}0 \leq \propto i \leq \propto i \leq C, i = 1{,}2, \ldots, n$$

Where∝i and ∝j are a unit coefficients allotted to coaching samples xi and xj. K (xi, xj) may be a kernel operate wont to live similarity between the 2 samples. Once specifying the kernel operate, SVM computes the coefficients that maximize the margin of correct classification on the coaching set. C may be a regulation parameter used for tradeoffs between coaching error and margin, and coaching accuracy and model quality. Blog identification and splog detection by Kolari et al. [15] used the activity and comments generated by a blog post as the main classification feature. A key requirement of such systems is to identify blogs as they move slowly through the Web. While this ensures that only blogs are indexed. Splogs not only incur computational overheads but also reduce user satisfaction. It describes experimental results of blog identification using Support Vector Machines (SVM).

Logistic Regression: This is a simple parametric model for binary classification where examples are classified based on their distance from a hyper plane decision boundary As in linear regression, Aim of this model is to estimate the regression coefficients in a model, given a sample of (X, Y) pairs. In the case of logistic regression, the X's can be numerical or categorical, but Ys are generally coded as 0 (for those who do not have the event) or 1 (for those who have the event).

The simple logistic model is based on a linear relationship between the natural logarithm (ln) of the odds of an event and a numerical independent variable. The form of this relationship is as follows:

$$L = ln\ o = ln\ (\tfrac{p}{1-p}) = \beta_0 + \beta_1 X + \epsilon,$$

Where Y is binary and represent the event of interest (response), coded as 0/1 for failure/success,
p is the proportion of successes,
o is the odds of the event,
L is the ln(odds of event),
X is the independent variable,
β_0 *and* β_1 are the Y-intercept and the slope, respectively, and ϵ is the random error. Garera et al. [18] they used linear regression and compare millions of Google's toolbar URLs to identify phishing pages.

VII. CONCLUSION

A URLs classification approach is proposed, which identifies URL to be either malicious or benign based on different learning methods using various features like on Lexical, host-based, link popularity, DNS, DNS fluxiness, Web page Content, Network and some special feature. The ultimate aim of using various features is to identify the ones that improve the detection accuracy with minimum overhead.This approach is complementary to blacklisting which cannot predict maliciousness of previously unseen URLs. Also we have studied classification models like Naïve bays, support vector machine, logistic regression. It is a System in which statistical models are most prone to over fitting. An open issue is how to scale millions of URLs whose features evolve over time and how to handle them.

REFERENCES

[1] A. Le, A. Markopoulos and M. Faloutsos, "PhishDef: URL Names Say It All", In Proceedings of the 30th IEEE INFOCOM 2011 (Mini Conference), Shanghai, China, April 10-15, 2011.

[2] Cortes, C., and Vapnik, V. Support vector networks. Machine Learning (1995), 273–297.

[3] C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale Automatic Classification of Phishing Pages", In Proceedings of the 17th Annual Network and DistributedSystem Security Symposium (NDSS'10), San Diego, CA, Mar 2010.

[4] Iron Port Web Reputation: Protect and defend against URL-based threat. http://www.ironport.com.

[5] D. McGrath and M. Gupta, "Behind Phishing: An Examination of Phisher Modi Operandi", In Proceedings of The USENIX Workshop on Large-scale Exploits and Emergent Threats (LEET), San Fransicso, CA, Apr 2008.

[6] D. Canali, M. Cova, G. Vigna and C. Kruegel, "Prophiler: a Fast Filter for the Large-Scale Detection of Malicious Web Pages", In Proceedings of the 20th International World Wide Web Conference (WWW), Hyderabad, India, Mar 2011.

[7] DNS-BH. Malware prevention through domain blocking. http://www.malwaredomains.com.

[8] Fette, I., Sadeh, N., and Tomasic, A. Learning to detect phishing emails. In WWW: Proceedings of the international conference on World Wide Web (2007).

[9] J. Ma, L. Saul, S. Savage, and G. Voelker, "Identifying Suspicious URLs: An Application of Large-Scale Online Learning", In Proceedings of the International Conference.

[10]JWSPAMSPY. E-mail spam filter for Microsoft Windows. http://www.jwspamspy.net.

[11] J. Ma, L. Saul, S. Savage, and G. Voelker, "Beyond Blacklists: Learning to Detect Malicious Websites from Suspicious URLs", In Proceedings of the ACM SIGKDD Conference, Paris, France, Jun 2009.

[12] MCAFEE SITEADVISOR. Service for reporting the safety of web sites. http://www.siteadvisor.com/.

[13] Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. Detecting spam web pages through content analysis. In WWW: Proceedings of international conference on World Wide Web (2006).

[14] P. Prakash, M. Kumar, R. R. Kompella and M. Gupta, "PhishNet: Predictive Blacklisting to Detect Phishing Attacks", In Proceedings of INFOCOM '10, San Diego, California, Mar 2010.

[15] P. Kolari, T. Finin, and A. Joshi, "SVMs for the Blogosphere: Blog Identification and Splog Detection", In AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, Mar 2006.

[16] PHISHTANK. Free community site for anti-phishing service. http://www.phishtank.com/.

[17] Provos, N., Mavrommatis, P., Rajab, M. A., and Monrose, F. All your iFRAMEs point to us. In Security: Proceedings of the USENIX Security Symposium (2008).

[18] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A Framework for Detection and Measurement of Phishing Attacks", In Proceedings of the ACM Workshop on Rapid Malcode (WORM), Alexandria, VA, Nov 2007.

[19] WOT. Web of Trust community-based safe surfing tool. http://www.mywot.com/.

[20] TREND MICRO. Web reputation query-online system. http://reclassify.wrs.trendmicro.com/.

[21] Whittaker, C., Ryner, B., and Nazif, M. Large-scale automatic classification of phishing pages. In NDSS: Proceedings of the Symposium on Network and Distributed System Security (2010).

[22] Zhang, J., Porras, P., and Ullrich, J. Highly predictive blacklisting. In Security: Proceedings of the USENIX Security Symposium (2008).

[23] http://en.wikipedia.org/wiki/Malware (accessed on 20/06/2014)

[24] http://random.yahoo.com/bin/ryl)3 .