Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 654–661

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

# Malicious Domain Detection Using Machine Learning On Domain Name Features, Host-Based Features and Web-Based Features

Gopinath Palaniappan[a], Sangeetha S[b], Balaji Rajendran[a], Sanjay[a], Shubham Goyal[a], Bindhumadhava B S[a]

[a]*CDAC, Electronics city, Bangalore, India*
[b]*NIT, Tiruchirappalli, India*

**Abstract**

Internet has plenty of vulnerabilities which are exploited by cyber criminals to send spam, commit financial frauds, perform phishing, indulge in command & control, disseminate malware and other malicious activities. Many times these exploits are carried out through malicious domain names which are the vital part of an Internet resource URL. Few vulnerabilities in the Internet setup and its related administrative policies allows such malicious domain names to be registered with the DNS servers. Though blacklisting happens to be the simplest and quickest solution to identify such malicious domains, the technique cannot cope up with the speed at which the domain names are generated and registered, and hence we look forward for other effective means of identifying malicious domains. The researchers have been using features from DNS data and features from lexical analysis of domain names, but there exists a need to identify more related features and introduce machine-learning to meet challenges due to IP flux and domain flux.

In this paper, we have introduced usage of web-based features of domain names in addition to using blacklists, DNS data and lexical features to identify malicious domains. Using the features extracted from the domain names, we build a classifier model using the logistic regression classification algorithm and use that classifier to identify benign and malicious domains. Our experiment is based on active DNS analysis and we look forward to take this work for passive DNS analysis.



*Keywords:* malicious domain detection; DNS; cyber security; malware; phishing; spam; domain generation;


10.1016/j.procs.2020.04.071

## 1. Introduction

The World Wide Web has seen tremendous increase in the usage of Internet with exponentially growing number of varied software applications in a wide-spectrum of areas such as finance, e-commerce, social-networking, automation systems and others. This increase of traffic in the whole of Internet infrastructure leaves several loopholes consequently paving way for more and more cyber threats. In fact, the WWW has also become a platform for directly or indirectly supporting malicious activities happening in the cyber space. One such malicious activity happening at an alarming rate on the cyber space is the downloading or spreading of malware. The primary means of the propagation or spreading of malware is through URLs (Uniform Resource Locators) hosted by domains created with harmful intents referred to as malicious domains or else through compromised websites.

The websites may be compromised due to loopholes in the hosting infrastructure or the hosting software tier or compromised credentials of the hosting network or machines through some means. Such hacked or compromised or defaced websites are an attractive target for cybercriminals who take advantage of its reputation and misuse it to perform malicious activities or to deliver malware.

While the domains refer to the domain names listed at the Domain Name System (DNS), they may also be malicious in nature, disseminating malware, hosting phishing or spam or scam webpages, or facilitating command and control (C&C) communications [1]. The detection of such malicious domains has become important for ensuring network and infrastructural security and preserving privacy.

**Nomenclature**

| | |
|---|---|
| DNS | Domain Name System |
| URL | Uniform Resource Locator |
| DGA | Domain Generation Algorithms |
| C&C | Command & Control |
| ICANN | Internet Corporation for Assigned Names and Numbers |
| ccTLD | Country Code Top Level Domains |
| gTLD | Generic Top Level Domains |
| IETF | Internet Engineering Task Force |

### *1.1. Malicious Domains*

Malicious domains are one among the primary resources leveraging attacks over the Internet. These malicious domains used for constructing malicious URLs are a very common and serious threat to the security of cyberspace. They can be used to lure users into becoming victims when they visit its phishing, drive-by downloads, spam and other contents, which may result in compromise of user's privacy, or may incur financial loss or may result in a malware installation onto the user's machine.

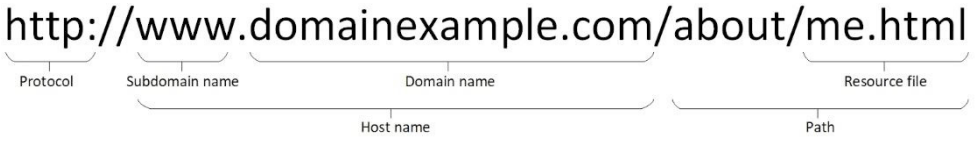


Fig. 1. Structure of a URL the domain name and subdomain name

As shown in Fig. 1, the domain name and the subdomain name are the important parts of a URL. The combination of domain name and the subdomain name together form the host name and represents the physical machine on which the resource is hosted.

### *1.2. Domain Name System (DNS)*

The Domain Name System (DNS) is one among the core protocol suites of the Internet, which takes the responsibility of directing requests for Internet resources to the absolute hosting machine and these Internet resources are identified by URLs which comprises of the domain names [4]. Domain Name System (DNS) is a decentralized, well-distributed, hierarchical naming system for resources connected to the Internet, which translates human readable and remembrable domain names into their respective IP (Internet Protocol) addresses, and sometimes the vice-versa too. In brief, DNS is a simple service for lookup and translation of URL into IP address and vice versa [3].

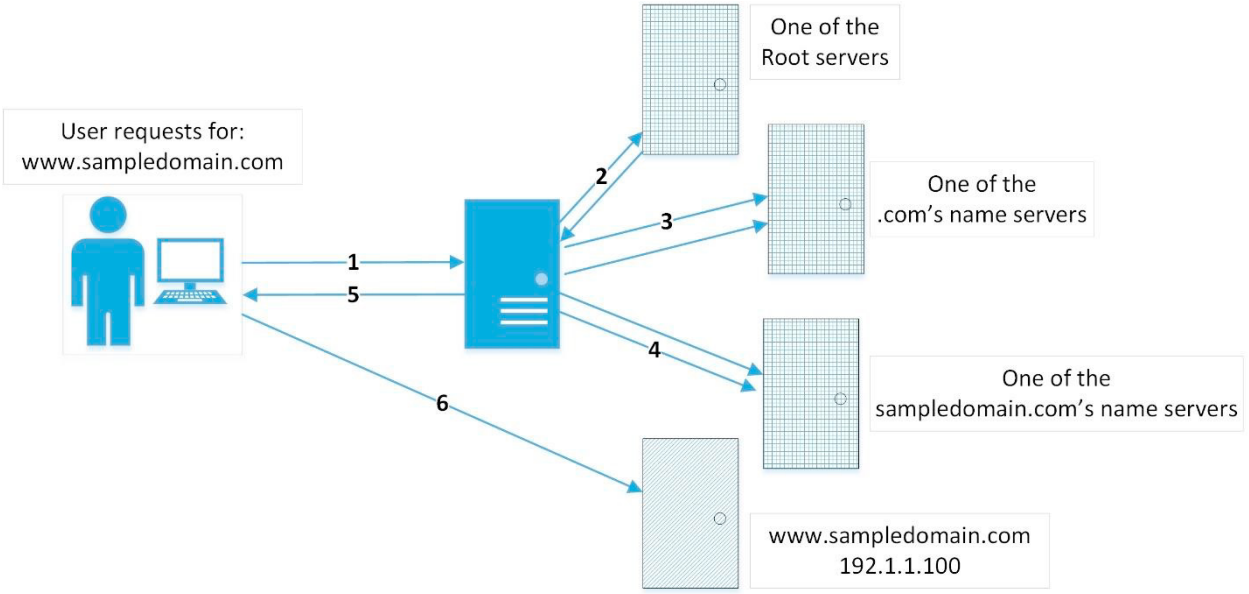


Fig. 2. Working of Domain Name System (DNS)

The domain data in the DNS is organized in the structure of an inverted tree with the root at the top represented by a "." (dot). The Internet Corporation for Assigned Names and Numbers (ICANN) is the International body given the authority of maintaining the root. Under root are the so-called Top Level Domains (TLDs) which includes country code top level domains (ccTLDs), unsponsored top level domains, generic top level domains (gTLDs) etc. The distributed DNS comprises of 13 root name servers named from A to M with replicated root zone spread across the Globe. The DNS data is divided into zones, where each zone is served by a set of authoritative name servers which hold the responsibility of providing authoritative answers for data present in their respective zones [5]. Another type of name server referred to Recursive Resolvers (RR) provide non-authoritative answers to clients. These RRs start at the root servers and follow zone delegations by processing the authoritative answers got at each level, until they reach the destined authoritative name server for the exact zone as represented in Fig. 2 above. These DNS data are very volatile, and unwanted records often sneak into this domain name system because of several technical and administrative loopholes, opening gateways for malicious activities.

### *1.3. Malicious Domain Detection*

The rapid growth of Internet across varied thematic areas and their related threats, keeps alive the challenge of malicious domain detection. Broadly malicious domain detection can be performed either on active DNS data or passive DNS data [11], where active data implies obtaining DNS data by purposefully sending DNS queries and recording the respective responses for analysis and detection, and passive data implies accessing DNS logs to get real DNS queries and responses, or else plant a sensor at the DNS servers to capture the queries and responses for

analysis and detection. The attackers keep switching their malicious system between different IP addresses using the domain names. In the recent times, a popular technique is of creating malware embedded with domain generation algorithm in order to generated domain names automatically and establish communication with their Command & Control systems hosted at those domains. Also, IP-Flux (Fast-Flux) and Domain-Flux are adding to the complexity of malicious domain detection [6].

The conventional approaches used by many for malicious domain detection include the usual blacklist of domain names [2], analysis of the network traffic [13], dissection of the webpage content [14], DNS traffic analysis [6], and analysis of salient lexical features [8]. Most of the work on content-based or non-content based malicious URL detection do not take the domain name and the DNS data into consideration for computing the maliciousness of the URL, and hence the results obtained lack precision. Hence effective mechanism for malicious domain detection will also help in improving the precision of malicious URL detection.

In this paper, we present a unique approach to detect malicious domains using a model trained by a machine learning algorithm using a combination of features of a domain name such as DNS data, lexical characteristics, and website reputation. We create separate data sets for benign and malicious domain names from various well-known and reliable sources and extract the above mentioned features from those domain names and feed them to logistic regression machine learning algorithm and generate a model. The generated model is then experimented with new list of domain names to classify them as benign or malicious.

## 2. Related Work

The domain names are closely related to its underlying infrastructure, the DNS infrastructure, and so analysis of these domain names needs to be done with data from the DNS. To do that, we have two approaches, the active approach where we request DNS data for a given domain name, while the other is a passive approach where we investigate the real-time data at the Recursive DNS servers or somewhere on the DNS infrastructure. In this paper, we have restricted ourselves to active DNS analysis. In this section of the paper we list few of the popular and effective works in the areas of both active and passive DNS analyses.

Leyla Bilge et al. (2014) proposed a passive DNS analysis service to detect malicious domains, EXPOSURE. EXPOSURE worked on real-time on 15 unique features categorized under one of time-based features, DNS answer-based features, TTL value-based features and domain name-based features [6]. They built a classifier using the J48 decision tree algorithm and reaped successful results, with their deployments running at several geographical locations across the globe.

Khulood Al Messabi et al. (2018) proposed a technique based on domain name features and DNS records to identify malicious domains [2]. They took efforts in identifying and preparing a list of malicious TLDs and inappropriate words, which if used in domain names adds to their maliciousness. They created a model using J48 decision tree classifier using 10-folds cross-validation.

Chunyu Han and Yongzheng Zhang (2017) proposed a technique based on domain names and their TTL (time-to-live) features to identify benign domain names [15]. They used naïve bayesian classifier algorithm. They focused only on determining benign-ness of a domain name, which in turn can be used in determining the maliciousness of a domain name.

Panpan Zhang et al. (2017) proposed an approach (DomainWatcher) to detect malicious domains based on local and global textual features [8]. They used lexical features, imitation features and bigram features of domain names to identify its maliciousness.

## 3. Proposed Approach

Our proposed approach is based on four features which we extract from a given domain name, (a) a blacklist of domain names and IP addresses collated from reliable and reputed resources, (b) DNS-based features extracted with support of various protocols which work on the DNS infrastructure, (c) web-based features, and finally (d) the lexical features.

### *3.1. Blacklist of domain names and IP addresses:*

We create and maintain two blacklists, one for malicious domain names and another for IP addresses used to host malicious domains. We populate our list of blacklist domains using data sets obtained from PhishTank, DNS-BH, and Reputation Blacklist (RBL) from ICANN. We also populate our list of blacklist IP addresses using data sets obtained from DNSBL (spamhaus). Though blacklisting happens to be a quick, simple solution to identify malicious domains, continuous increase in domain registrations, IP-Flux and Domain-Flux allowing frequent and dynamic change of IPs and domain names cause miserable failure of this straight-forward technique. Hence we consider this feature of occurrence of a domain name or its associated IP address in the blacklist as just one among several parameters to classify the domain name as malicious or benign.

### *3.2. DNS-based features:*

As discussed earlier, the Domain Name System (DNS) plays a vital role in the functioning and maintenance of the WWW space. For a given domain name, we find all its host IP addresses, the Autonomous System Number (ASN), and hence also the countries or the areas where it is hosted. Determining the IP addresses helps us to classify the type of IP address, the geo-location of the IP address, and the type of connectivity at the region. The ASN helps us determine the Border-Gateway in which the hosting is present.

Millions of domain names are registered every year. The domain name registrant provides information such as name, address, email, and administrative and technical contact numbers to the registrars and registries which are have the accreditation of ICANN. This information provided by the registrant to the registrars or registries for registration of a domain name is referred to as the WHOIS data. The registrar and registries started allowing public access to the WHOIS data of domain names as per certain policies and contracts defined by ICANN. In the recent times, the WHOIS is replaced with Registration Data Access Protocol (RDAP). RDAP was evolved by the Internet Engineering Task Force (IETF). RDAP has better standards for query format and data access when compared to WHOIS. The Table 1 below captures the DNS-based features of a domain name that can be obtained either by WHOIS or RDAP.

Table 1. DNS-based features of a domain name

| Attribute | Description |
|---|---|
| Autonomous System Number (ASN) | ASN number, ASN registry, ASN Country Code, ASN date |
| Registrant | Registrant organization and contact details |
| Registrar | Registrar name and contact details |
| Date | Creation date, Expiry date |
| IP addresses | All the IP addresses hosting the given domain name |
| PTR record | Reverse lookup. Given the IP determine the domain name it hosts. |

### *3.3. Web-based features:*

Conventionally, malicious domain detection is performed on the basis of DNS data and lexical features, here we determine several web-based features of the domain name too. The list of feature and their description are provided in Table 2 below. Several malicious domain detection techniques use DNS-based data analysis and lexical features analysis, whereas we have included web-based features in our procedure for detection of malicious domain because we are performing active DNS analysis. We have come up with a python-based implementation for extraction of web-based features for a given domain name. Here we also avail the services of Alexa, a popular, reliable, dynamic ranking service of websites. There exist rich libraries in Python facilitate in collecting such web-based features of domain names.

Table 2. Web-based features of a domain name

| Attribute | Description |
|---|---|
| Global and Country ranking | Alexa, a popular ranking service, provides reliable ranking to websites based on various parameters |
| Webpages | The number of webpages. More the number of webpages and lack of broken links ensures authenticity of the host |
| Time spent by visitor | A visitor spending relative high time in the website improves authenticity of the host |
| Web referrals | More the number of web referrals of a website in other Internet resources improves the credibility of the website |
| Web traffic | (a) The consistency in the number of visitors to the website adds to the trustworthiness of the website<br>(b) Number of page views and<br>(c) amount of traffic reaching it through search engines |
| Category | Identify the category or the subject matter to which the domain name belongs |
| Geo-location | Geo-location of the IP addresses on which the domain name is hosted |

### *3.4. Lexical features:*

A lexical or textual analysis the domain name itself can yield parameters which provide support in classifying a domain name as benign or malicious [2][7][12]. The parameters are listed in Table 3.

Table 3. Lexical features of a domain name

| Attribute | Description |
|---|---|
| Dots | Number of dots in the domain name should not be greater than 3 |
| Underscores and hyphens | Total count of underscores and hyphens in the domain name to be less than 4 |
| Digits | Count of number of digits in the domain name to be less than 4 |
| Illegitimate contents | We prepared a list of illegitimate words |

### *3.5. Building Model and domain classification:*

We prepared labelled list of around 20,000 domain names taken from PhishTank, DNS-BH, Alexa, and Reputation Blacklist (RBL) from ICANN. A part of the list comprising of about 10,000 domain names was used for building a model by training and the other part with the labels removed was used to evaluate the model and judge its effectiveness in classifying a domain name as benign or malicious. We used the logistic regression algorithm for creation of our model for classification of the domain name.

Logistic regression is one of the most scalable, probabilistic classifier algorithms. It stems from the sigmoid (logistic) function which basically maps an input to an output of values between 0 and 1. In logistic regression, the input function I is expressed as weighted sum of features as follows: $I = w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n = WX$, where $x_1$, $x_2$,…..,$x_n$ are the features of the feature vector and $w_1$, $w_2$,……., $w_n$ are the respective weight coefficients, as represented in Fig. 3. The weight coefficients of the logistic regression model are learned from the training data using maximum-likelihood estimation with the motive of predicting a positive sample as close to 1 as possible and predicting a negative sample as close to 0 as possible.


Fig. 3. Using the Logistic Regression Model

The probability of classifying an input is found by:

$$p(x) = 1 / (1 + e^{-WX}) \quad (1)$$

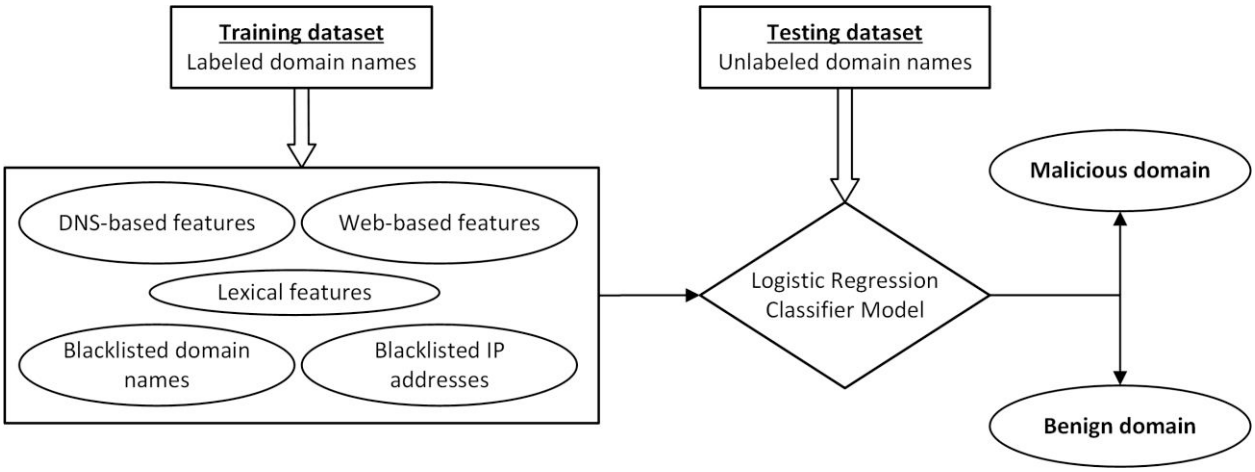


Fig. 4. The proposed approach for malicious domain detection

## 4. Experiment

A dataset of about 20000 labelled domain names were collected from various reliable sources. The domain names contained both benign and malicious domains. The malicious domain names were either spam, phishing, malware or defacement domains. The dataset was divided into two parts, each comprising of about 10000 domain names each, and in one of the parts the labels were removed and the data shuffled, while the other part remained as a labelled dataset. We took the labelled data set, extracted features such as blacklisted domain names, blacklisted IP addresses, DNS-based features, web-based features and lexical features as listed in the above section of proposed approach and trained the logistic regression classifier model using it to identify benign and malicious domain names. The trained logistic regression classifier is then fed with the unlabeled dataset to classify the domain names as benign or malicious as represented in Fig. 4, and listed in Table 4.

Table 4. Domain name classification

| | Two category classification | |
|---|---|---|
| Probability $\epsilon$ {0,1]} | 1 or positive classification | 0 or negative classification |
| Domain name | Malicious | Benign |

With our limited dataset we could achieve an accuracy of about 60% in classifying the unlabeled dataset as benign or malicious. We computed our accuracy in the following method:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

## 5. Conclusion and Future Work

This paper explored an active DNS analysis approach for classifying a domain name as benign or malicious by including the web-based features of the domain name in addition to the usually used lexical-based and DNS-based features. We extracted features of a domain name under DNS-based, web-based, blacklisting and lexical-based categories, and trained a logistic regression classifier and tested the classifier to classify unlabeled dataset of domain names and got an accuracy of about 60% in our attempt using a small dataset of about 10000 domain names. We look forward to improvise on this binary classifier and make it a multi-classifier to identify the type of maliciousness for a given domain name, such as spam, phishing, malware or defacement.

We plan our future work with larger datasets as follows: (a) we have already hosted a Recursive DNS (RDNS) and we plan to perform passive DNS analysis and feed its result to perform active DNS analysis, (b) improvise on the present implementation to use logistic regression to extract features in addition to using it to build the classifier, (c) use few more machine learning algorithms with improved features list and strive for better accuracies, and (d) work on detection of domain names generated by DGA (Domain Generation Algorithms) to prevent botnets [9][10].

## References

[1] Yury Zhauniarovich, Issa Khalil, Ting Yu, and Marc Dacier. 2018. A Survey on Malicious Domains Detection through DNS Data Analysis. ACM Computing Survev. 1, 1, Article 1 (May 2018), 35 pages.

[2] Khulood Al Messabi, Monther Aldwairi, Ayesha Al Yousif, Anoud Thoban, and Fatna Belqasmi. 2018. Malware Detection using DNS Records and Domain Name Features. In ICFNDS'18: International Conference on Future Networks and Distributed Systems, June 26–27, 2018, Amman, Jordan. ACM, New York, NY, USA, 7 pages.

[3] Tejaswini Yadav C.Y, Balaji Rajendran, and Rajani P. 2014. An Approach for Determining the Health of the DNS. International Journal of Computer Science and Mobile Computing, Vol.3 Issue.9, September- 2014, pg. 442-449.

[4] Adiwal, Sanjay & Rajendran, Balaji & Shetty, Pushparaj. 2018. Domain Name System (DNS) Security: Attacks Identification and Protection Methods. In SAM'18: International Conference on Security and Management, Las Vegas, USA.

[5] F. Weimer. 2005. Passive DNS Replication. In FIRST Conference on Computer Security Incident, 2005

[6] Leyla Bilge, Sevil Sen, Davide Balzarotti , Engin Kirda , Christopher Kruegel. 2014. Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains, ACM Transactions on Information and System Security (TISSEC), v.16 n.4, p.1-28, April 2014

[7] W. Wang and K. E. Shirley. 2015. Breaking bad: Detecting malicious domains using word segmentation. In IEEE Web 2.0 Security and Privacy Workshop (W2SP), 2015.

[8] P. Zhang, T. Liu, Y. Zhang, J. Ya, and J. Shi. 2017. "Domain Watcher: detecting malicious domains based on local and global textual features", in Proceedings of the International Conference on Computational Science, pp. 2408–2412, Zurich, Switzerland, June 2017.

[9] Yadav, Sandeep & Reddy, A. & Ranjan, Supranamaya. 2010. Detecting Algorithmically Generated Malicious Domain Names. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC. 48-61.

[10] C. Choudhary, R. Sivaguru, M. Pereira, B. Yu, A. C. Nascimento, M. De Cock. 2018. "Algorithmically generated domain detection and malware family classification", Proceedings of the Sixth International Symposium on Security in Computing and Communications, 2018.

[11] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, and D. Dagon. 2011. Detecting malware domains in the upper DNS hierarchy. In the Proceedings of 20th USENIX Security Symposium (USENIX Security '11), 2011.

[12] Shuang Hao, Matthew Thomas, Vern Paxson, Nick Feamster, Christian Kreibich, Chris Grier, and Scott Hollenbeck. 2013. Understanding the domain registration behavior of spammers. In Proceedings of the 2013 conference on Internet measurement conference (IMC '13). ACM, New York, NY, USA, 63-76.

[13] ANDYEN T, REITER M. 2008. Traffic aggregation for malware detection. In Conference on Detection of Intrusions and Malware& Vulnerability Assessment (DIMVA) (2008).

[14] Eshete, Birhanu & Villafiorita, Adolfo & Weldemariam, Komminist. 2013. BINSPECT: Holistic analysis and detection of malicious web pages. SecureComm. 2012.

[15] Han, Chunyu & Zhang, Yongzheng. (2017). CLEAN: An approach for detecting benign domain names based on passive DNS traffic, in Proceedings of 6th International Conference on Computer Science and Network Technology (ICCSNT) 343-346.