Exploratory Experiments to Identify Fake Websites by using Features from the Network Stack

Jason Koepke, Siddharth Kaza
Department of Computer and Information Sciences
Towson University, Towson, MD 21252, USA
jkoepk1@students.towson.edu, skaza@towson.edu

Ahmed Abbasi
McIntire School of Commerce
University of Virginia, Charlottesville, VA 22902, USA
abbasi@comm.virginia.edu

***Abstract*—Users on the web are unknowingly becoming more susceptible to scams from cyber deviants and malicious websites. There has been much work in the identification of malicious websites using application layer features based on content (HTML, images, links, etc.) and a plethora of classification techniques. However, there has been little work on using features from the other layers in the Open Systems Interconnection (OSI) network stack. Capturing features from the transport and internet layers of the network stack based on responses to various Hypertext Transfer Protocol (HTTP) requests may allow for increased classification accuracy. In this paper, we use learning techniques (Winnow, Logit Regression, Naïve Bayes, J48, and Bayesian) utilizing these new features to identify fake pharmacy websites. The results show that using transport and Internet layer features yields an accuracy of 80% to 95% for detecting fake websites using standard machine learning algorithms. The results suggest that many organizations may be hosting multiple websites using shared code and hosting services to enable them to produce the maximum number of fraudulent websites.**

Keywords-fake websites, machine learning, web mining, cyber deviants, website signatures

I. INTRODUCTION

As long as the economy and national security remain dependent on the Internet, cyber-crime opportunities will continue to increase. Cybercrime has constantly been evolving since 1986 when the first well publicized security incident was reported on ARPANet. Malicious code, fraudulent websites, identity theft, and internet fraud have been growing in stature as the Internet becomes tightly woven into the fabric of people's lives [6]. As the number of reported computer crime cases grew so too did the associated number of victims and financial losses from thousands to hundreds of millions of dollars [6]. The work in this paper specifically focused on fraudulent pharmaceutical websites, however, these same methods and techniques could be applied to any type of fraudulent website and the results are generalizable to others kinds of cybercrime with similar data characteristics. If a consumer visited Google and searched for "pharmacy" there is a significant chance that many websites returned by search results would be fraudulent pharmacy websites [2]. Most users of the web may not have the ability to determine if a website is fraudulent just by looking at the website as most fraudulent websites will have a high degree of technical sophistication. Fraudulent pharmaceutical websites may commit failure-to-ship fraud and the type of product that consumers may receive when they use these online pharmacies is also suspect [3]. Recently, Google was the subject of a U.S. criminal investigation that alleged the search giant made millions of dollars by accepting advertisement revenue from online pharmacies that violate U.S. law [2]. This case also indicates that there might be a large number of fraudulent websites to financially scam Internet users.

Previous work on fraudulent website detection has focused on various content level features (like HTML text and images) from the application layer of the network packet. However, integrating information from the entire network stack can provide crucial input to aid in the detection of cyber deviants [1]. In this paper, we explore the use of features from the packet content at the transport and the internet layers (which we collectively refer to as 'network' layers henceforth). In particular we explore the following research questions:

- Does using the network stack information instead of just application level attributes provide a different perspective on the data?
- Can we just use the network stack information without application layer to identify fake websites?

The rest of this paper is organized as follows: in Section 2, we present pertinent literature related to this problem domain. Section 3 presents the research design, test bed, and the method used for data collection. In Section 4, we discuss the results and Section 5 concludes and presents future directions.

II. LITERATURE REVIEW

There has been much previous work in fake website detection. Most research in the area (with the exception of [4]) has only focused on information from the application layer in the network stack. Work using the remaining stack has been concentrated in Intrusion Prevention Systems (IPS) and Intrusion Detection Systems (IDS). IDS and IPS technologies use pattern matching algorithms on attributes including Transmission Control Protocol (TCP) fields, Internet Protocol (IP) addresses, TCP/User Datagram Protocol (UDP) port numbers, Internet Control Message Protocol (ICMP) type and code, and strings contained in the payload packet [7]. In this study, we use many of these attributes including IP address, Domain Name Service (DNS) server IP address, and the web server content responses to create a unique site signature. Another key feature in this study is using traceroute data that provides the path the packet traveled to reach the destination. These paths are not always unique but can be used to discern if two servers are hosted on the same network subnet [5]. This information is likely to provide better features to the classification algorithms and increase classification accuracies as compared to using content features alone. The aim here is to present the method for gathering features across all layers of

the network stack (rather than application layer alone) and using various classification algorithms to show the efficiency of these features.

III. RESEARCH DESIGN

This section presents the details on our test bed, the process of gathering network stack features, and the algorithms used in this study.

A. Test Bed

The test bed consisted of 300 websites from the medical and pharmaceutical industries; this included both legitimate and fake websites (Table I). The Uniform Resource Locator's (URL) for the fake pharmacy websites were collected from the National Association of Boards of Pharmacy (www.nabp.net) and LegitScript (www.legitscript.com). Immediately after we obtained the URLs for these websites, we used our collection mechanism to acquire the features for each of the URLs before the sites were taken down (as is common in fake websites).

TABLE I. SUMMARY OF DATASET

	Real	Fake	Total
Pharmacy Websites	150	150	300

B. Collection of Features from the Network Stack

We used several methods to gather features from the network layer. We focused on includes virtualized host information, DNS records (which includes IP addresses), traceroutes, and GET and POST responses. The subsections below explain these concepts.

1) Virtualized hosts

It is known that fraudsters reproduce the content of fake websites quickly to avoid detection. However, often they might use the same server architecture to host these sites as it is difficult to change architectures frequently. Many legitimate websites are hosted in a virtualized environment which concurrently runs multiple user websites and appears to be a stand-alone server when accessed from the Internet. Traditionally, web hosting by Internet Service Providers (ISP) has been done on a single web server that also hosts multiple websites. The second type of configuration used in web-hosting is virtualization, which consists of each web site being hosted on separate instances on a master server. Virtualization allows the hosting company to host many different web sites on a single server.

2) DNS Server information

DNS records are publicly accessible through the WHOIS database. The DNS records are going to provide a primary DNS IP address, a secondary DNS IP address, the date the site was created, and finally the date the was last updated.

3) Traceroute information

Tracerouting is a tool for measuring network transit time across an IP network that allows each user to see in detail which path the packet is taking. Further, this utility allows the end user to specifically trace the performance of each packet across each network hop (router) en route to its destination.

4) Valid/Invalid GET and POST responses

Valid and invalid HTTP GET and POST responses provide valuable information on the hosting server architecture. We gather data from four requests for each domain to gather the most common error and status messages provided by the destination web server (HTTP GET, HTTP POST, invalid HTTP GET, and invalid HTTP POST).

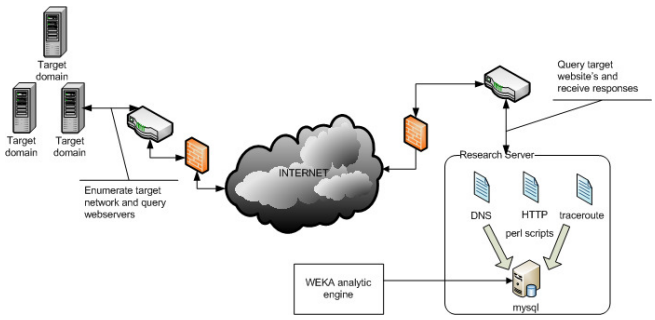

Figure 1 System Architecture to Gather Features

To gather the above features, three scripts were written in Perl to do data collection on the WHOIS, DNS, and HTTP services that run on the target web servers. The scripts were run on an Ubuntu 9.10 Linux system that was hosted in a virtualized environment. The scripts query WHOIS database servers on the Internet for Domain Registration information, the target HTTP web servers, and collect the traceroute data from each domain. All the results were stored in persistent storage for later analysis (Figure 1).

C. Classification algorithms compared

This study compared five different classifying algorithms that are commonly used in previous research in this problem domain (Table II lists these algorithms). Though other potential algorithms could have been used, since our focus in this study is on the efficacy of network layer features we picked a base set for the purpose.

TABLE II. CLASSIFICATION ALGORITHMS

Classification Algorithm	Description
Winnow	Machine learning from a linear classifier
Naïve Bayes	Probalistic classifier that applies Bayes theorem
Logit Regression	Prediction of an event occurring by using a logistic curve
J48	Statistical classifier that generates a decision tree
Bayesian	Probability that enables reasoning with uncertain statements

IV. EXPERIMENTAL RESULTS

In this section we will discuss the results from running five classification algorithms on the collected data set. The aim of the experiment was to correctly classify fake and legitimate websites using features from the network stack. WEKA [8]

was used for running all experiments and the best possible settings were used for all algorithms where applicable. WEKA is a collection of machine learning algorithms purposed for data mining tasks. To process the raw data in WEKA, it was necessary to convert all text to n-grams by applying strings to n-grams, setting no class, and converting numerical values to nominal values. These calculations were done using options within WEKA. Once these modifications to the data had occurred, WEKA was able to classify the data in the appropriate manner.

The results from running the classification algorithms are shown below (Table III). Overall, the Logit Regression algorithm was the most accurate in predicating the legitimacy of a website at a 93.8% accuracy.

TABLE III. EXPERIMENTAL RESULTS

Algorithm	Overall Accuracy	Legit			Fake		
		F-means	Precision	Recall	F-means	Precision	Recall
Winnow	82.9	76.6	80.7	72.9	89.1	87.1	91.3
Naive Bayes	91.2	88.3	88	88.6	94.1	94.3	94
Logit Reg.	**93.8**	91.7	93.2	90.4	96	95.3	96.7
J48	91.7	88.7	92.8	84.9	94.7	92.8	96.7
Bayesian	91.6	88.9	88.6	89.2	94.4	94.6	94.3

V. CONCLUSIONS

In the future, online services are going to be increasingly used by people for any need. The individuals using these services will unknowingly become more susceptible to scams from cyber deviants and malicious websites. This paper presented a new approach to capturing features from the network and internet layers of the network stack based on responses to various Hypertext Transfer Protocol (HTTP) requests which allowed for increased classification accuracy. In this paper, we used learning techniques that use network stack information (like responses to HTTP requests) to identify fraudulent websites. The experimental results show that using the network features yields an accuracy of 80% to 95%, depending on the algorithm used, for detecting fake websites using standard machine learning algorithms. The results suggest that many organizations may be hosting multiple websites using shared code and hosting services to enable them to produce the maximum number of fraudulent websites. In the future, it should be possible to use the features suggested in this work to aid in the detection of fake websites in other domains.

REFERENCES

[1] Abbasi, A. and Chen, H. "A Comparison of Tools for Detecting Fake Websites," IEEE Computer (42:10), 2009, pp. 78-86.Conclusions

[2] Catan, Thomas. Efrati, Amir. "Google Near DOJ Settlement Over Online Drug Ads - WSJ.com." Business News & Financial News - The Wall Street Journal - Wsj.com. Web. 16 May 2011. <http://online.wsj.com/article/SB10001424052748703730804576319572448399628.html>.

[3] Easton, G. "Clicking for Pills," British Medical Journal, 334(7583), 2007, pp. 14-15.

[4] Chun-Ying Huang, Shang-Pin Ma, Wei-Lin Yeh, Chia-Yi Lin, and Chien-Tsung Liu. "Mitigate Web Phishing Using Site Signatures", Proceedings of IEEE Region 10 Conference (TENCON-2010), Fukuoka, Japan, November 2010.

[5] Dall, Luca et al. "Exploring networks with traceroute-like probes : theory and simulations." Most (February 2008): 1-28.

[6] Hoar, S. B. 2005. "Trends in Cybercrime: The Darkside of theInternet," *Criminal Justice* (20:3), pp. 4-13.

[7] Patton, Samuel et al. "An Achilles ' Heel in Signature-Based IDS : Squealing False Positives in SNORT." Lecture Notes in Computer Science: 1-8.

[8] WEKA. The university of Waikato. http://www.cs.waikato.ac.nz/ml/weka/