Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection

Zuochao Dou, *Student Member, IEEE*, Issa Khalil, *Member, IEEE*, Abdallah Khreishah, *Member, IEEE*, Ala Al-Fuqaha, *Senior Member, IEEE*, and Mohsen Guizani, *Fellow, IEEE*

***Abstract*—Phishing is a form of cyber attack that leverages social engineering approaches and other sophisticated techniques to harvest personal information from users of websites. The average annual growth rate of the number of unique phishing websites detected by the Anti Phishing Working Group is 36.29% for the past six years and 97.36% for the past two years. In the wake of this rise, alleviating phishing attacks has received a growing interest from the cyber security community. Extensive research and development have been conducted to detect phishing attempts based on their unique content, network, and URL characteristics. Existing approaches differ significantly in terms of intuitions, data analysis methods, as well as evaluation methodologies. This warrants a careful systematization so that the advantages and limitations of each approach, as well as the applicability in different contexts, could be analyzed and contrasted in a rigorous and principled way. This paper presents a systematic study of phishing detection schemes, especially software based ones. Starting from the phishing detection taxonomy, we study evaluation datasets, detection features, detection techniques, and evaluation metrics. Finally, we provide insights that we believe will help guide the development of more effective and efficient phishing detection schemes.**

***Index Terms*—Phishing, Phishing website detection, software based methods.**

I. Introduction

PHISHING, one form of cyber-attacks, continues to be a growing concern not only to cyber security specialists but also to e-business users and owners. The severity of such cyber attack vector is continuously growing with the exponential increase in digital information generation and the increased reliance of people and business on cyber space. The Anti-Phishing Working Group (APWG) has seen rapid growth in the number of unique phishing websites detected from 2014 to 2016 [19]. The average annual growth rate is 97.36% and is expected to continue to grow. Estimates of annual direct financial loss to the U.S. economy caused by phishing activities range from $61 million to $3 billion [49].

To mitigate the increasing damage caused by phishing, a broad range of anti-phishing mechanisms have been proposed over the past two decades. These anti-phishing techniques can be categorized into three broad groups [12]: (1) Detective solutions (e.g., website filtering); (2) Preventive solutions (e.g., strong authentication [32]–[34], [43], [53], [54], [85]); and (3) Corrective solutions (e.g., Site takedown [57], [58]). In this paper, we focus on detective solutions. More specifically, we look at software-based phishing detection schemes that are specialized in identifying and classifying phishing websites. This class of approaches is arguably more important than other approaches because it helps in reducing human errors. Preventative and corrective solutions take a different approach, but if the user behind the keyboard has been successfully tricked by a phishing attempt, and willingly submitted sensitive information, then no firewall, encryption software, certificates, or authentication mechanism can help in preventing the attack from materializing [49]. Software-based phishing detection also delivers improved results compared to detection by user education (e.g., [60], [61], and [98]) because phishing attacks normally aim at exploiting human weaknesses [59]. For example, a study of phishing detection using user education [97] shows a 29% false negative rate (*FNR*) for the best performance, while the software based approaches that are surveyed by the same study have *FNR* in the range of 0.1% to 10%. For this reason, we focus our study on software based phishing detection systems, and the term “phishing detection" will refer only to this form of detection in the rest of the paper.

Although the research area of phishing detection and classification is relatively rich, there is a lack of systematic analysis of the requirements, the capabilities, and the shortcomings of the existing anti-phishing techniques. For example, websites that offer identification and classification of phishing as a service have been popular in recent years, however, those services leverage different evaluation datasets from various sources at different time periods to validate their outcomes. Albeit those schemes may have similar performance results (e.g., in terms of false positive rate, true positive rate, etc.), it is difficult to compare their performance because of the variation in the evaluation datasets employed. Consequently, a systematic assessment of the datasets used to validate phishing detection approaches is desired, as well as necessary, in

Manuscript received December 16, 2016; revised May 8, 2017; accepted August 16, 2017. Date of publication September 13, 2017; date of current version November 21, 2017. *(Corresponding author: Mohsen Guizani.)*

Z. Dou and A. Khreishah are with the Electrical and Computer Engineering Department, New Jersey Institute of Technology, Newark, NJ 07102-1982 USA (e-mail: zd36@njit.edu; abdallah@njit.edu).

I. Khalil is with the Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar (e-mail: ikhalil@hbku.edu.qa).

A. Al-Fuqaha is with the NEST Research Laboratory, College of Engineering and Applied Sciences Computer Science Department, Western Michigan University, Kalamazoo, MI 49008 USA (e-mail: ala.al-fuqaha@wmich.edu).

M. Guizani is with the Electrical and Computer Engineering Department, University of Idaho, Moscow, ID 83844-1023 USA (e-mail: mguizani@uidaho.edu).

Digital Object Identifier 10.1109/COMST.2017.2752087

order to provide a foundation for comprehensive comparisons among different phishing detection schemes, and ultimately, select the best in practice.

In this work, we complement the existing survey papers on phishing detection, including [49], [59], and [103], by providing a broad systematic analysis of software based anti-phishing approaches. Varshney *et al.* [103] focus on studying, analyzing, and classifying the most significant and novel detection techniques, and pointed out the advantages and disadvantages of each approach. On the other hand, we present a more comprehensive systematic review of phishing detection schemes, not only from the perspective of detection algorithms, but also from a broader perspective that covers other important aspects including the phishing detection life cycle, taxonomy of phishing detection schemes, evaluation datasets, detection features, and evaluation metrics and strategies. The work in [49] focuses more on the attack side of phishing. More specifically, it presents details about phishing attacks including the anatomy of such attacks, why people fall in phishing attacks and how bad phishing is. However, it only provides a high level analysis of the state-of-the-art phishing countermeasures. In order to provide a systematic review of the phishing detection research, we first present the necessary information about the phishing attacks by answering three questions: (1) What is phishing?, (2) How does phishing work? and (3) What is the current status of phishing? Then, we conduct systematic review of phishing detection schemes in a detailed and comprehensive manner. Finally, Khonji *et al.* [59] present a literature survey about anti-phishing solutions (e.g., user training, email filtering and website detection, etc.), including their classification, detection techniques and evaluation metrics. Compared to [59], we focus on the software based phishing website detection schemes, which are proved to be the most effective anti-phishing solutions and are not systematically studied in [59].

In a nutshell, the objective of this paper is to provide a systematic understanding of existing phishing detection studies and provide a comprehensive way to evaluate phishing detection approaches from different perspectives in order to guide future developments and validations of new or upgraded anti-phishing techniques.

We summarize our contributions in this work as follows:

- Compile a comprehensive profile of phishing through its various definitions, detailed ecosystem (i.e., in terms of phishing life cycle, actors involved and their operations, etc.), and the state-of-the-art phishing trends.
- Present a systematic review of the software based phishing detection schemes from different perspectives including the life cycle, taxonomy, evaluation datasets, detection features, detection techniques and evaluation metrics.
- Introduce a novel feature, Network Round Trip Time (*NRTT*), for efficient and real time detection of phishing attacks.
- Provide detailed takeaway lessons for researchers and practitioners in the area of phishing detection that we believe will help guide the development of effective phishing detection schemes.

TABLE I
MOST POPULAR DEFINITIONS OF PHISHING

	Definition
PhishTank [81]	Phishing is a fraudulent attempt, usually made through email, to steal your personal information.
APWG [44]	Phishing is a criminal mechanism employing both social engineering and technical subterfuge to steal consumers' personal identity data & financial account credentials.
Xiang et al. [118]	Phishing is a form of identity theft, in which criminals build replicas of target Web sites and lure unsuspecting victims to disclose their sensitive information like passwords, personal identification numbers (PINs), etc..
Whittaker et al. [108]	A phishing page is any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewer would only trust a true agent of the third party.
Khonji et al. [59]	Phishing is a type of computer attack that communicates socially engineered messages to humans via electronic communication channels in order to persuade them to perform certain actions for the attacker's benefit.
Ramesh et al. [90]	Phishing is a fraudulent act to acquire sensitive information from unsuspecting users by masking as a trustworthy entity in an electronic commerce.

The rest of the paper is organized as following: Section II describes the state-of-the-art phishing attacks, and presents the life cycle of phishing detection approaches. Section III introduces the taxonomy of phishing detection schemes with the corresponding literature review. Section IV presents a systematic review of software based phishing detection schemes from different perspectives: (1) phishing detection datasets; (2) phishing detection features; (3) phishing detection techniques; and (4) evaluation metrics. Section V provides detailed takeaway lessons for researchers and practitioners in the area of phishing detection. Section VI concludes the paper.

II. BACKGROUND

A. State-of-the-Art Phishing Attacks

In this section, we first present the various definitions of phishing, then we introduce some statistics about phishing between January 2010 and June 2016. Finally, we describe the phishing ecosystem.

1) What Is Phishing?: There is no consensus on how phishing should be defined. Different phishing definitions lead to different research directions and approaches (e.g., email filtering or website detection). It is important to clearly identify the target of any phishing detection approach to avoid confusion about its applicability in different scenarios. The target and scope of phishing detection approaches can be analyzed from the definition of phishing which has been adopted by such approaches. Therefore, presenting a background on the different definitions of phishing can help the readers understand the scope and the capabilities of different approaches. Table I summarizes the popular definitions of phishing. On one hand, the definitions of PhishTank [81], APWG [19], Xiang *et al.* [118],

TABLE II
TARGETS AND STRATEGIES OF PHISHING

	Target	Strategy
PhishTank [81]	Personal information	Social engineering
APWG [44]	Identity data, Financial account credentials	Social engineering
Xiang et al. [118]	Sensitive information	Not specified
Whittaker et al. [108]	Not specified	Not specified
Khonji et al. [59]	Not specified	Social engineering
Rameshe et al. [90]	Sensitive information	Not specified

TABLE III
OPERATIONS OF DIFFERENT PLAYERS INVOLVED IN PHISHING

	Basic Operations	Advanced Operations
A	1.Data collection 2.Website development 3.Email engineering	1. Evasion of anti-phishing techniques
B	1.Blacklist announcement	1.User behavior detection 2.Strong authentications
C	1.Human detection 2.Browser filter	1.Phishing detection toolbar
D	1.Policy enforcement	1.Brand monitoring
E	1.Phishing data analysis 2.Anti-Phishing solutions	1.Law enforcement 2.Government coalition
F	1.Employee training	1.Email filtering 2.Phishing detection software

A: Phisher; B: Web service provider; C: Web service subscriber; D: Web hosting provider; E: Anti-Phishing institute; F: Spear Phishing targets

Ramesh *et al.* [90] cover the majority of cases in which phishers aim at stealing sensitive personal information such as authentication credentials. Table II shows the comparison of those phishing definitions based on phishing target and phishing strategy. The most dominant phishing strategies are social engineering (e.g., through fraudulent emails) and technical subterfuge (e.g., malware infection). However, sophisticated techniques (e.g., pharming [52]) are also used to harvest users' personal information from the Internet. On the other hand, the definitions of Whittaker *et al.* [108] and Khonji *et al.* [59] do not limit the attacker's target (e.g., sensitive personal information). They describe the phishing strategy (e.g., phishing website or socially engineered messages) without stating a specific phishing target (e.g., only state the attackers' benefit). To sum up, the definition of Whittaker *et al*. [108] is the most general among those reviewed, while APWG [19] defines the most commonly used phishing attacks in a specific manner.

2) How Does Phishing Work?: In this section, we introduce the ecosystem of phishing in terms of phishing process, actors involved, their actions and interactions.

(i) Phishing Process: In a generic/traditional phishing scenario (i.e., mass-email phishing campaigns), an attacker hosts a fake website, and presents users of a Web service with convincing emails containing a link to the fake website. When a user of the Web service opens the link and enters her sensitive data, data is collected by the server hosting the fake website. As shown in Figure 1, Mihai and Giurea [78] suggest that a generic phishing process can be identified in five steps: (1) Reconnaissance: Phishers look for famous Web service brands with a broad customer base; (2) Weaponization: Phishers design the phishing websites and social engineer on email spam; (3) Distribution: Phishers deliver emails to the victims; (4) Exploitation: Phishers exploit weaknesses of humans to lure the victims into phishing traps via socially engineered emails. (5) Exfiltration: Phishers collect sensitive data from the phishing databases.

Unlike generic phishing attacks, spear phishing targets particular individuals or organizations. References [24], [86], and [104]. Spear phishing attacks typically extract sensitive data from their victims by attaching a type of malware to emails or in the phishing website. Industry statistics indicate that spear phishing attacks have a success rate of 19%, while the success rate of generic phishing attacks is less than 5% [86].

For the purpose of this paper, we will not consider email filtering (e.g., [21], [68], and [120]) as a phishing detection method. Our focus is on detection of website phishing for both generic and spear phishing attacks.

(ii) Phishing Actors: There are six actors involved in a typical phishing life cycle (see Figure 2), as defined in the following paragraphs:

- *Phisher:* Individuals or organizations that conduct phishing attacks in order to obtain a certain type of benefit, such as financial gain, identity hiding (e.g., refers to the situation in which phishers do not use the stolen identities directly, but rather sell them to interested criminals and cyber attackers.), fame and notoriety, etc. [59], [105].
- *Web service provider:* Companies that provide a certain type of service (e.g., email, social network, e-banking, on-line shopping, etc.) on the Internet (usually through a website).
- *Web service subscriber:* Customers who subscribe to Web services provided by the Web service provider. Subscribers are the potential targets of traditional phishing attacks.
- *Web hosting provider:* Companies that provide website hosting services to Web service companies.
- *Anti-phishing institutes:* Institutes that support those tackling the phishing menace and provide advice on anti-phishing controls and information on current trends [4].
- *Spear phishing targets:* Specific individuals or companies targeted by phishers.

Each actor involved in the phishing process has different actions and reactions (summarized in Table III). Phishers try to use sophisticated techniques to evade phishing detection approaches (e.g., DNS poisoning [11]). In addition, there is a growing trend in which phishers have decoupled the process of phishing website hosting from the process of sending phishing emails in order to evade the anti-phishing solutions (Han *et al.* [45]).

Web service providers usually announce blacklists of phishing websites and recommend users to use strong authentication schemes (e.g., [17], [32]–[34], and [55]). Additionally, Web service subscribers highly depend on browser filters (e.g., Google Safe Browser [50]) and other third party anti-phishing toolbars (e.g., Netcraft [3]) to detect and block phishing attempts.

The role of Web hosting providers is rather ambiguous in the phishing process. Reputable providers usually enforce strict "Terms of Use" and avail certain anti-phishing solutions

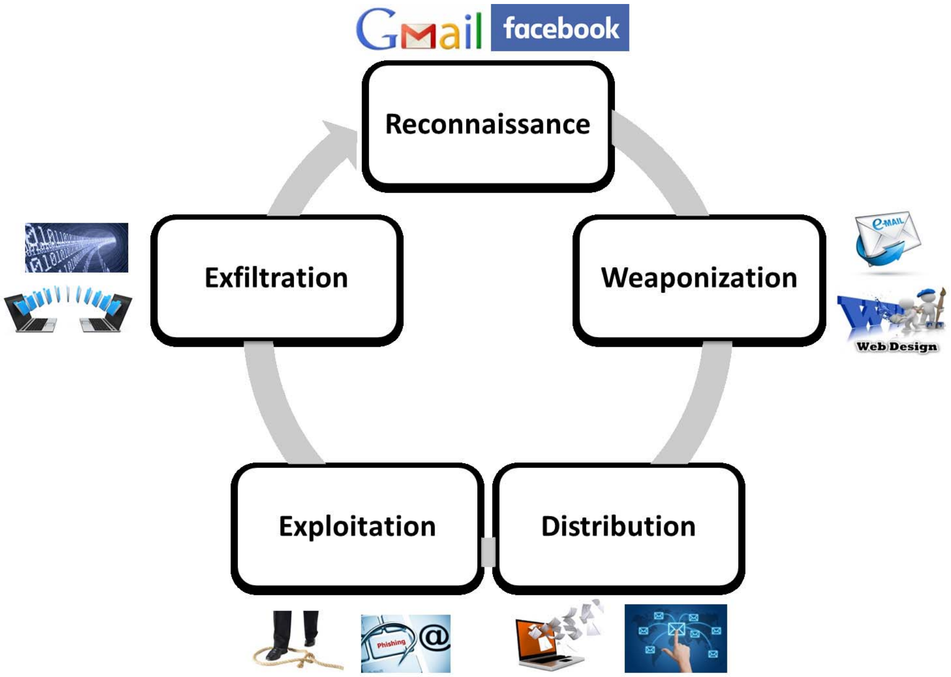


Fig. 1. Illustration of the phishing process.

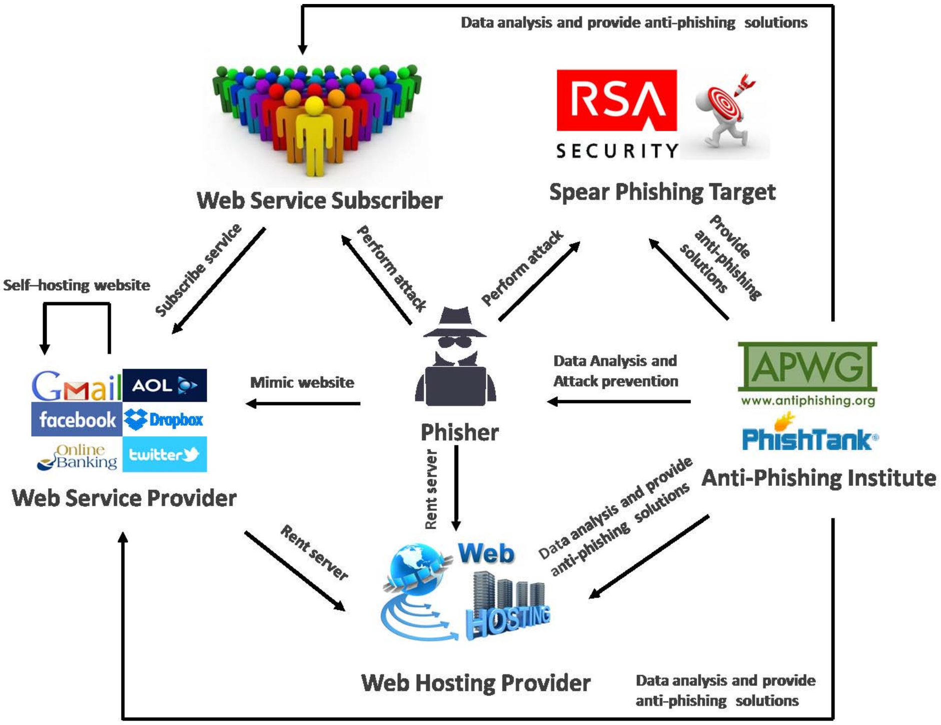


Fig. 2. Players involved in the phishing process.

(e.g., brand monitoring [12]). Due to financial constraints, many free-to-use Web hosting providers may not be able to afford deploying good anti-phishing security measures, which leaves their customers not only vulnerable, but even worse, attractive targets for phishing.

Anti-phishing institutes collect and analyze phishing data (e.g., suspicious websites reported by users) from various sources (e.g., users' reports via anti-phishing toolbars), and provide anti-phishing suggestions and solutions (e.g., up-to-date phishing website blacklist, phishing detection toolbars, etc.). In addition, they may also cooperate with government agencies such as public security and law enforcement to detect and prevent cyber attacks [4].

3) What Is the Current State of Phishing?: According to phishing activity trends reports published by APWG [19] from Jan. 2010 to Jun. 2016 (shown in Figure 3), the number of

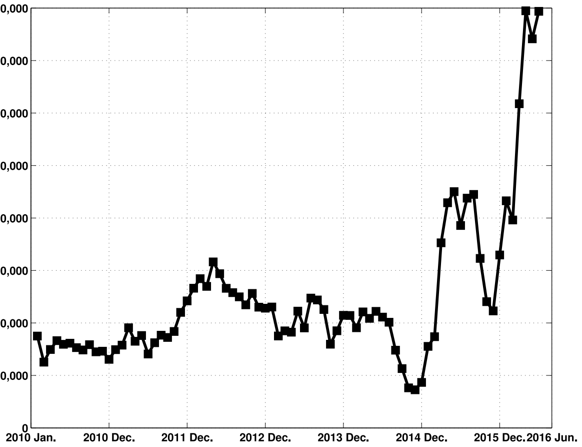


Fig. 3. The number of unique phishing sites per month from Jan. 2010 to Jun. 2016.

unique phishing websites established per month increased significantly since 2015 (i.e., the average number for 2016 is 2.93 times the average from prior years). It is clear that phishers profited from this type of cyber-attacks, which result in financial loss for both Web subscribers and business owners. Therefore, agile techniques to mitigate phishing will continue to be a pressing need.

Phishing attacks tend to employ advanced techniques to lure Web service users into their rogue websites. Using the database from Trend Micro Web reputation technology, Pajares [83] reports the number of phishing sites that use HTTPS connections increased significantly in 2014 compared to 2010 (shown in Figure 4). Attackers become more cautious and attentive when designing phishing websites to evade existing phishing detection methods [1]. Some phishing groups are capable and desire to perform more advanced phishing attacks. Avalanche (commonly known as the Avalanche Gang) is a criminal syndicate involved in phishing attacks [109]. In 2010, APWG reported that Avalanche was responsible for two-thirds of all phishing attacks in the second half of 2009, describing it as "one of the most sophisticated and damaging on the Internet" and "the world's most prolific phishing gang" [10]. It has been discovered that Avalanche uses different techniques to evade the anti-phishing mechanisms.

In addition, more and more sophisticated techniques are being used to implement phishing attacks. For example, the pharming attack, a refined version of phishing attacks, is designed to steal users' credentials by redirecting them to fraudulent websites using DNS-based techniques [41], [58]. Many computer security experts predict that the use of pharming attacks will continue to grow as more criminals embrace these techniques [52].

B. Life Cycle of Phishing Detection

As mentioned in Section I, we do not incorporate phishing detection approaches that rely on user education due to their poor performance. In addition, we do not cover phishing detection methods that perform email filtering because it is a different detection theme that warrants a separate comprehensive study on its own. we reemphasize here that our focus is on the area of software-based phishing detection which aims at detecting or blocking phishing websites.

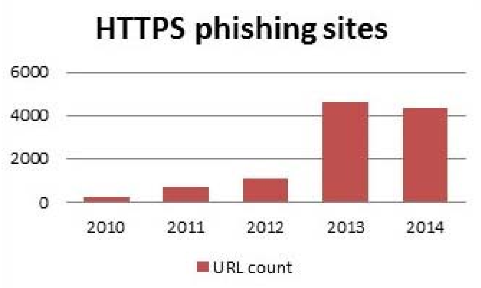


Fig. 4. The number of phishing sites that use HTTPS. Re-printed from [83].

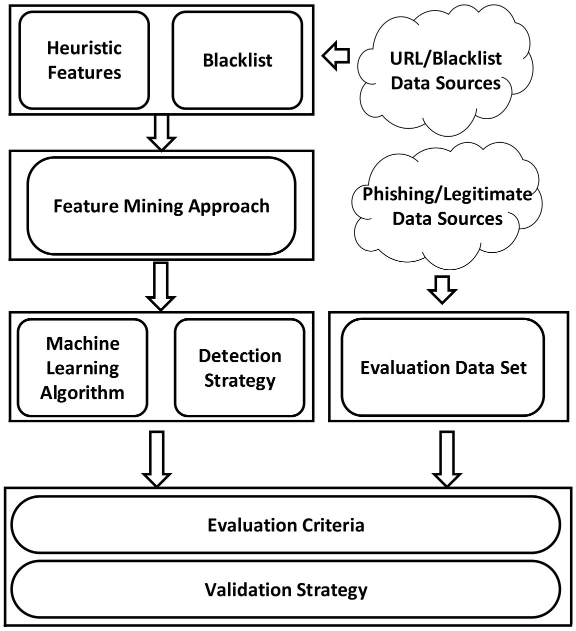


Fig. 5. Life cycle of typical phishing detection schemes.

The life-cycle of software-based phishing detection is illustrated in Figure 5. Starting from the initial inputs, the detection scheme extracts phishing detection features (or called heuristics, as detailed in Section IV-B) and/or blacklists from various sources (e.g., URL related information, trusted third party, WHOIS server, etc.) via different feature mining approaches (e.g., search engines, target identification algorithms, etc.). Then, it applies different data mining algorithms and/or proposes various detection strategies to the engineered features to achieve its objectives (e.g., identifying phishing links, blocking phishing websites, etc.). To evaluate the performance of phishing detection schemes, various evaluation datasets are collected from different sources (e.g., PhishTank, Yahoo directory, etc.). Finally, leveraging the collected datasets and following various validation strategies (e.g., cross validation), the proposed scheme is evaluated based on multiple metrics (e.g., False Positive Rate, False Negative Rate, etc.).

In the coming sections, following the life cycle of software-based phishing detection schemes, we present a comprehensive

TABLE IV
SUMMARY OF DIFFERENCES BETWEEN PHISHING DETECTION TOOLBARS AND ACADEMIC PHISHING DETECTION/CLASSIFICATION SCHEMES

	Public usage availability	**Data Mining Algorithms**	**Exact Heuristics**	**Blacklist**	**Scheme Details**	**Goal**
Publicly available Phishing detection toolbars	Yes	Very little	Very little	Yes	Little	**Detect and block phishing websites**
Academic phishing detection/classification schemes	Very little	Yes	Yes	Little	Yes	**Detect or classify phishing websites**

study of the phishing detection research from 5 different perspectives, namely, classification of phishing detection techniques, validation datasets, detection features, detection techniques and detection criteria.

III. PHISHING DETECTION SCHEMES: TAXONOMY AND THE CORRESPONDING LITERATURE REVIEW

In phishing literature, software-based phishing detection schemes are usually categorized into heuristic and blacklist based schemes [49], [59]. Heuristic-based approaches examine contents of the Web pages including: (1) surface level content (e.g., the URL); (2) textual content (e.g., terms or words that appear on a given Web page); (3) visual content (e.g., the layout, and the block regions etc.) [122]. These methods can detect phishing attacks as soon as they are launched but also introduce relatively high false positive rates (*FPR*). Blacklist-based approaches have a higher level of accuracy. However, they do not defend against zero-hour attacks [49], [115]. Combinations of heuristic and blacklist based approaches provide more robust and flexible defense against phishing attacks than either one on a standalone basis.

In this paper, we classify phishing detection approaches as either public phishing detection toolbars or academic phishing detection/classification schemes. Phishing detection toolbars use blacklists and/or selected heuristics to identify phishing websites. There is usually little information about what heuristics these toolbars use and how they are used. Academic phishing detection solutions are similar to phishing detection toolbars, but usually apply more complex technologies and are usually not available/feasible for public use. Most academic phishing classification schemes apply combinations of heuristics features into various data mining algorithms to enhance the classification accuracy. Table IV summarizes the differences between phishing detection toolbars and academic phishing detection/classification schemes. Note, the "scheme details" column in Table IV estimates the amount of publicly available details about detection schemes, such as detection methodology, data mining algorithms, and datasets.

Furthermore, based on the heuristic/blacklist classification, we further classify the academic phishing detection approaches into more specific and fine-grained sub-categories, namely, (1) heuristic: URL based methods; (2) heuristic: page content based methods; (3) heuristic: visual similarity based methods; (4) heuristic: other methods; (5) blacklist based methods; (6) hybrid methods. Details about each category are introduced in Section IV-B.

A. Public Phishing Detection Toolbars

Many freely available anti-phishing toolbars offer detection and blocking services against Internet phishing attacks. These toolbars typically come in the form of Web browser extensions (i.e., default extensions or third party extensions) that warn users about a suspicious phishing site after clicking on its URL.

Publicly available anti-phishing toolbars are either embedded in the browser as default extensions (e.g., Microsoft SmartScreen Filter [112]) or can be downloaded from third party websites (e.g., Netcraft [3]). They both display security warnings on screen when certain actions are triggered in the browser. These security warnings can be classified into two types [59]:

- *Passive warnings:* Passive warnings display various information (e.g., user ratings, site suggestions, etc.) about the website that is currently being visited but do not block the content of the website, as depicted in Figure 6.
- *Active warnings:* Active warnings display warning information about the website a user is trying to visit and block the content of the website, as depicted in Figure 7.

Many studies have shown that the majority of Web service users ignore security warnings provided by anti-phishing toolbars [31], [35], [116]. Furthermore, Egelman *et al.* [36] found that active warnings are much more effective than passive warnings (79% of participants paid attention to active warnings while only 13% participants paid attention to passive warnings). Table V summarizes the information gathered about the state-of-the-art anti-phishing toolbars. In the following paragraphs, we discuss the details of those toolbars:

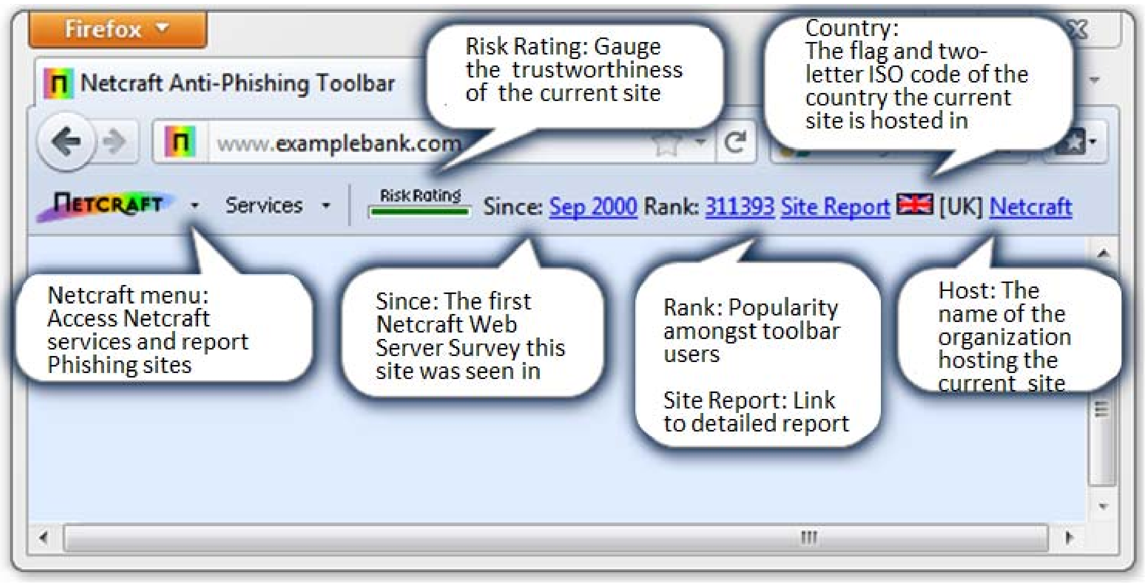
Google Safe Browsering: It uses a browser to check URLs against Google's constantly updated blacklist of unsafe Web resources (e.g., phishing websites) [50] and provides active warnings to the end users. According to Google Safe Browsing's website, for different platform and threat types, it examines pages against the safe browsing lists. It also issues reminders before users access risky links.

McAfee SiteAdvisor: This is a Web application that reports on the identity of websites by scanning them for potential malware and spam [111]. The detection result is decided according to a combination of heuristics and manual verification, such as the age and country of the domain registration, the number of links to other known-good sites, third-party cookies, and user reviews [30]. In addition, it provides passive warnings.

Netcraft Anti-Phishing Toolbar: Provides Internet security services including anti-fraud and anti-phishing services, application testing and PCI scanning [113]. According to its website, Netcraft's toolbar screens and identifies the deceiving contents in URLs. It also ensures that the navigational controls (e.g., toolbar and address bar) are activated in order to prevent pop-up windows (particularly for Firefox). In addition, it shows the geographic information of the hosting location of the sites and analyzes fraudulent URLs (e.g., the real

TABLE V
INFORMATION ABOUT SELECTED STATE-OF-THE-ART ANTI-PHISHING TOOLBARS

	Technology	Warning Type	Platform
Google Safe Browsing	BlacklistHeuristics(Suspected)	Active	Internet Explorer, Firefox and Chrome
McAfee SiteAdvisor	Heuristics, Blacklist(Manual verification)	Passive	Internet Explorer, Firefox and Chrome
Netcraft Anti-Phishing Toolbar	BlacklistHeuristics	Active	Internet Explorer, Firefox and Chrome
SpoofGuard	Heuristics	Passive	Internet Explorer
Microsoft SmartScreen Filter	Blacklist	Active	Internet Explorer
EarthLink Toolbar	Blacklist(User ratings and manual verification), Heuristics (Unknown)	Passive	Internet Explorer and Firefox
eBay Toolbar	BlacklistHeuristics	Passive	Internet Explorer
GeoTrust TrustWatch Toolbar	Blacklist Third-party reputation services and certificate authorities	Passive	Internet Explorer
WOT (Web of Trust)	Blacklist(User ratings, Third-party information)	Active	Internet Explorer, Firefox and Chrome


Firefox
Netcraft Anti-Phishing Toolbar
www.examplebank.com
NETCRAFT
Services
Risk Rating
Since: Sep 2000 Rank: 311393 Site Report [UK] Netcraft

Risk Rating: Gauge the trustworthiness of the current site

Country: The flag and two-letter ISO code of the country the current site is hosted in

Netcraft menu: Access Netcraft services and report Phishing sites

Since: The first Netcraft Web Server Survey this site was seen in

Rank: Popularity amongst toolbar users

Site Report: Link to detailed report

Host: The name of the organization hosting the current site

Fig. 6. Passive warnings from Netcraft anti-phishing toolbar. Reprinted from: http://toolbar.netcraft.com/.

citibank.com or barclays.co.uk sites have little possibility to be located in the former Soviet Union [3]).

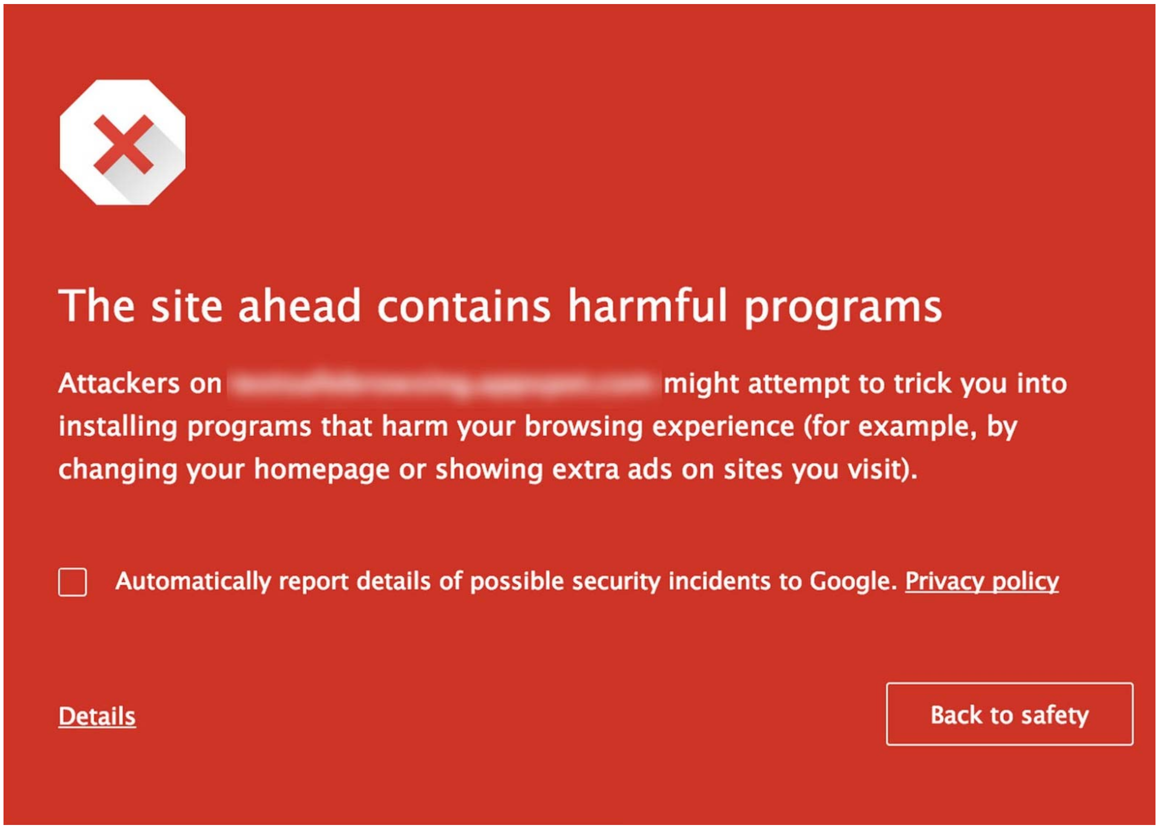
SpoofGuard: A heuristics-based anti-phishing toolbar developed for Internet Explorer with passive warnings. The heuristics used include (1) Domain name check: examines if the domain name for the attempted URL matches recent entries; (2) URL Check: checks if the username, the port number, as well as the domain name, are suspicious; (3) Email Check: determines whether the current URL directs to the browser via email; (4) Password Field Check: determines if the input fields of type “password" are located in the document; (5) Link Check: searches for risky links in the body of the document; (6) Image Check: analyzes the images of the new site vs. the previous sites; (7) Password Tracking: prevents the user from typing the same username and password for multiple sites [63].

Microsoft SmartScreen Filter: A blacklist-based phishing and malware filter implemented in several Microsoft browsers, including Internet Explorer and Microsoft Edge [112]. When browsing the site, SmartScreen helps monitor and identify the possibility of visiting a suspicious page. If so, it issues an active warning before next step is taken, as well as soliciting feedback from users. SmartScreen also maintains a list of reported phishing and software sites. It screens the list to check if a match is found. In that case, it issues a warning message while blocking the site for user’s safety. In addition, security checks are also performed when the user starts a download from the site. Moreover, SmartScreen compares the download to a list of existing downloads by other users. A warning is issued if it’s a brand new download.

EarthLink Toolbar: Helps to protect the user from on-line scams by displaying a security rating (i.e., passive warning) for all the websites the user visited previously. Additionally, it alerts the user if he tries to access a previously known fraudulent website. It appears to rely on a combination of heuristics, user ratings, and manual verification [30].

eBay Toolbar: Helps the buyers and sellers with real time alerts and keeps users safe from spoofing and fraudulent attacks by detecting fake sites via a combination of heuristics and blacklists through passive warnings [30].

GeoTrust TrustWatch Toolbar: Provides website verification service that alerts the users to potentially unsafe, or phishing Web sites based on the information of several third-party reputation services and certificate authorities via passive warnings [42]. TrustWatch notifies the users that the website has passed the verification scan based on a list of disreputable sites. It would also recommend additional caution when inputting



The site ahead contains harmful programs

Attackers on [redacted] might attempt to trick you into installing programs that harm your browsing experience (for example, by changing your homepage or showing extra ads on sites you visit).

□ Automatically report details of possible security incidents to Google. Privacy policy

Fig. 7. Active warning from Google Safe Browsering. Reprinted from: https://googleblog.blogspot.com/2015/03/protecting-people-across-web-with.html.

sensitive information to the website. Furthermore, it blocks the initial attempt when visiting potentially unsafe websites and warns users in case of a risk in revealing information to the site.

Web of Trust (WOT): A browser extension that tells the user which websites he can trust via active warnings [42]. It ensues the user's Internet safety from scams, malware, rogue Web stores and dangerous links based on community ratings and reviews.

B. Academic Phishing Detection/Classification Schemes

Unlike the public anti-phishing toolbars, which aim at providing real-time warnings about the legitimacy of visited websites, academic phishing detection and classification schemes normally focus on improving the detection accuracy and reducing the number of false alerts by employing sophisticated technologies and various machine learning algorithms. Table VI shows the time-based (from 2005 to 2016) development of 41 selected academic phishing detection/classification approaches. In order to choose the most representative studies, in this paper, we comply with the following criteria based on state-of-the-art literature:

- *Pioneering:* Research that introduces new ideas or methods to the literature.
- *Attention:* Research that receives more attentions in terms of the number of citations.
- *Completeness:* Research that presents their work following the entire life cycle of phishing detection in depth.

Based on the proposed criteria, all of the 41 selected works are introduced in the following sections and twelve representative studies are chosen as examples to illustrate the detailed detection methodology in each category. They are listed in Table VI and introduced below.

Visual similarity based methods: Chen *et al.* [27] describe a novel heuristic anti-phishing system that explicitly employs gestalt and decision theory concepts to model perceptual similarity. More specifically, they apply logistic regression algorithm to a set of normalized page content features. The proposed scheme can achieve 100% true positive rate and 0.74% false positive rate.

The most representative work in this category is done by Fu *et al.* [38]. They propose an effective phishing website detection approach via visual similarity assessment based on Earth Mover's Distance (EMD) [47]. The detection process contains two phases, namely, generating signature of Web pages and computing visual similarity score from EMD.

The Web page processing phase (i.e., generate the signature) contains three steps: (1) obtain the image of a Web page from its URL using Graphic Device Interface (GDI) API; (2) perform image normalization (the normalized image size is 100 x 100, and Lanczos algorithm [93] is used to resize the image); (3) transform the Web page image by a visual signature. The signature is comprised of the image color tuple using the [Alpha, Red, Green, and Blue] (ARGB) scheme and the centroid of its position in the image.

The second step is to compute the EMD between the visual similarity signatures of the two Web pages (legitimate site

TABLE VI
TIME-LINE BASED DEVELOPMENT OF PHISHING DETECTION SCHEMES FROM 2005 TO 2016

Research	Year	Represented(Y/N)
EMD based visual similarity for detection of phishing webpages [39]	2005	
Phishing web page detection [106]	2005	
Anomaly based web phishing page detection [84]	2006	
Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD) [38]	2006	Y
Cantina: a content-based approach to detecting phishing web sites [124]	2007	Y
A framework for detection and measurement of phishing attacks [40]	2007	Y
Anti-phishing based on automated individual white-list [23]	2008	
A phishing sites blacklist generator [96]	2008	
Visual similarity-based phishing detection [77]	2008	
B-apt: Bayesian anti-phishing toolbar [69]	2008	
Phishzoo: An automated web phishing detection approach based on profiling and fuzzy matching [15]	2009	
Fighting phishing with discriminative keypoint features [25]	2009	
Beyond blacklists: Learning to detect malicious web sites from suspicious URLs [72]	2009	Y
A hybrid phish detection approach by identity discovery and keywords retrieval [119]	2009	
Identifying suspicious URLs: An application of large-scale online learning [73]	2009	Y
Visual similarity-based phishing detection without victim site information [46]	2009	
Automatic detection of phishing target from phishing webpage [70]	2010	
Phishnet: predictive blacklisting to detect phishing attacks [119]	2010	Y
Large-scale automatic classification of phishing pages [108]	2010	Y
Lexical feature based phishing URL detection using online learning [22]	2010	
Detecting visually similar web pages: Application to phishing detection [26]	2010	
Intelligent phishing detection system for e-banking using fuzzy data mining [13]	2010	
Using domain top-page similarity feature in machine learning-based web phishing detection [94]	2010	
Textual and visual content based anti-phishing: A bayesian approach [122]	2011	
Cantina+: A feature-rich machine learning framework for detecting phishing web sites [118]	2011	Y
PhishDef: URL names say it all [66]	2011	
Design and evaluation of a real-time URL spam filtering service [102]	2011	Y
Antiphishing through phishing target discovery [107]	2012	
Using visual website similarity for phishing detection and reporting [76]	2012	
PhishAri: Automatic realtime phishing detection on twitter [16]	2012	
Phishing Website detection using latent Dirichlet allocation and AdaBoost [89]	2012	
Phishing detection plug-in toolbar using intelligent Fuzzy-classification mining techniques [14]	2013	
Phishstorm: Detecting phishing with streaming analytics [74]	2014	
An efficacious method for detecting phishing webpages through target domain identification [90]	2014	Y
An anti-phishing system employing diffused information [27]	2014	Y
Predicting phishing websites based on self-structuring neural network [80]	2014	
Examination of data, rule generation and detection of phishing URL using online logistic regression [37]	2014	
Feature extraction and classification phishing websites based on URL [20]	2015	
New rule-based phishing detection method [79]	2016	
Know Your Phish: Novel Techniques for Detecting Phishing Sites and their Targets [75]	2016	Y
PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder [101]	2016	

and phishing site). Firstly, the normalized Euclidean distance of the degraded ARGB colors and the centroids are computed. Then the two distances are added up with their corresponding weights (i.e., p and q, $p + q = 1$). The normalized feature distance between φ_i and φ_j is defined as:

$$d_{ij} = ND_{feature}\big(\varphi_i, \varphi_j\big) = p * ND_{color}\big(dc_i; dc_j\big) + q * ND_{centroid}\big(C_{dc_i}; C_{dc_j}\big)$$

where $\varphi_i =< dc_i, C_{dc_i} >$, $dc =< dA; dR; dG; dB >$ is the color tuple, and C_{dc} is the centroid value. Suppose we have signature $S_{s,a}$ and signature $S_{s,b}$, the EMD between $S_{s,a}$ and $S_{s,b}$ can be calculated as:

$$EMD\big\{S_{s,a}, S_{s,b}\big\} = \frac{\sum f_{ij} * d_{ij}}{\sum f_{ij}}$$

where f_{ij} is the flow matrix calculated through linear programming [47]. Note that if EMD=0, the two images are identical, if EMD=1, they are completely different.

Finally, the EMD-based visual similarity of two images is defined as:

$$VS\big\{S_{s,a}, S_{s,b}\big\} = 1 - \Big[EMD\big\{S_{s,a}, S_{s,b}\big\}\Big]^{\alpha}$$

where $\alpha \in (0, +\infty)$ is an amplification factor that limits the skewness of the visual similarity for the distributed in the (0,1) range.

Large-scale experiments with 10,281 suspected Web pages are carried out and the proposed scheme achieves 0.71% false positive rate and 89% true positive rate.

Similar works based on visual similarly include [15], [25], [26], [39], [46], [70], [76], [77], [94], [106], and [122].

Page content based methods: Zhang *et al.* [124] propose CANTINA, a novel content-based approach for detecting phishing Web sites based on the Term Frequency/Inverse Document Frequency (TF-IDF) information retrieval metric. In addition, using some heuristics, the false positive rate is reduced. Generally, CANTINA works as follows:

1) CANTINA calculates the TF-IDF scores of each term of the content in the given website.
2) CANTINA generates a lexical signature by taking the five terms with highest TF-IDF weights.
3) CANTINA sends the lexical signature to a search engine (i.e., in their case, Google Search).

4) If the domain name of the current website matches the domain name of the top N search results, it is considered to be a legitimate website. Otherwise, it is concluded to be a phishing site. Note that, the value of N affects the false positives.

CANTINA with TF-IDF alone results in a relatively high false positive rate. Therefore, several heuristics are used to reduce the false positive rate, including:

- Age of Domain: it examines the age of the domain name. If the page has been registered for more than 12 months, the heuristic returns $+1$ (i.e., legitimate), otherwise it returns -1 (phishing).
- Known Images: it examines whether a page contains inconsistent well-known logos.
- Suspicious URL: it examines if the URL contains an "@" or a "-" in the domain name.
- Suspicious Links: for each link in the webpage, it performs the above three URL checks.
- IP Address: it examines if the URL contains an IP address.
- Dots in URL: it examines the number of dots in the URL.
- Forms: it examines if a Web page contains any HTML text entry form requesting sensitive personal data (e.g., password).

In addition, CANTINA uses a simple forward linear model to make the decision:

$$S = f\left(\sum w_i * h_i\right)$$

where h_i is the result of each heuristic, w_i is the weight of each heuristic, and f is a simple threshold function.

$$f(x) = 1 \quad if \quad x > 0, \quad f(x) = -1 \quad if \quad x <= 0.$$

Here, 1 means legitimate site and -1 means a phishing site.

The proposed scheme could achieve 97% true positive rate while maintaining 1% false positive rate.

In 2011, Xiang *et al.* [118] extended the work of Zhang *et al.* [124] by proposing CANTINA+ which is claimed to be the most comprehensive feature-based approach in the literature. It exploits the HTML Document Object Model (DOM), search engines and third party services with machine learning techniques to detect phishing. It has been shown to achieve 0.4% false positive rate and over 92% true positive rate.

Similar works based on page content include [14], [89], and [119].

URL based methods: Garera *et al.* [40] claim that it is often possible to tell whether or not a URL belongs to a phishing attack without requiring any knowledge of the corresponding page data. By applying several selected features (i.e., page rank, domain name white list and URL based features) into logistic regression learning algorithm, the proposed scheme is efficient and has a high accuracy.

The most representative work in this category is done by Ma *et al.* [72]. They propose a phishing detection approach to automatically classify URLs based on different data mining algorithms across both lexical and host based URL features.

The lexical features selected in this method include the length of the hostname, the length of the entire URL, as well as the number of dots in the URL. In addition, the authors create a binary feature for each token in the hostname (delimited by ".") and in the path URL (strings delimited by "/", "?', ".", "=", "-" and "_"). The host-based features contain: (1) IP address properties (e.g., is the IP address in a blacklist?); (2) WHOIS properties (e.g., the date of registration, update, and expiration); (3) Domain name properties (e.g., the time-to-live (TTL) value for the DNS records associated with the hostname); (4) Geographic properties (e.g., the continent/country/city that the IP address belongs to).

All the features of the URL are encoded into high dimensional feature vectors and then different types of classifiers are applied to them. Here are some examples of the classifiers:

- *Naive Bayes:* Let x denote the feature vectors and $y \in \{0, 1\}$ denote the label of the website, with $y = 1$ for malicious and $y = 0$ for legitimate ones. $P(x|y)$ denotes the conditional probability of the feature vector given its label. Then, assuming that malicious and legitimate websites are equally probable, the posterior probability that the feature vector x belongs to a malicious URL is computed as:

$$P(y = 1|x) = \frac{P(x|y = 1)}{P(x|y = 1) + P(x|y = 0)}$$

  Finally, the right hand side of the equation is thresholded to predict the binary label of the feature vector x.
- *Support Vector Machine (SVM):* The decision using SVMs is expressed in terms of a kernel function $K(x, x')$ that computes the similarity between two feature vectors and non-negative coefficients α_i that indicate which training examples lie close to the decision boundary. SVMs classify new examples by computing their distance to the decision boundary:

$$h(x) = \sum_{1}^{n} \alpha_i (2y_i - 1) K(x_i, x)$$

  where $h(x)$ is the threshold to predict a binary label for the feature vector x.
- *Logistic Regression:* LR classification is based on the distance from a hyperplane decision boundary [71]. The decision function is $\sigma(z) = [1+e^{-z}]^{-1}$ that converts these distances into probabilities that feature vectors have positive or negative labels. The conditional probability that feature vector x has a label $y = 1$ is:

$$P(y = 1|x) = \sigma(wx + b)$$

  where w (i.e., the weight vector) and b (i.e., the scalar bias) are parameters computed based on the training data. Finally, the right hand side of the equation is thresholded to decide the label of the feature vector x.

The proposed scheme can achieve 0.1% false positive rate and 92.4% true positive rate.

Similar works based on URL related features include [20], [22], [37], [66], and [74].

Blacklist based methods: PhishNet [87] is a predictive blacklisting scheme to detect phishing attacks. Traditional blacklist approaches (i.e., exact match with the blacklisted entries) are easy for attackers to evade. Instead, PhishNet uses five heuristics (i.e., top-level domains, IP address, directory structure, query string, brand name) to compute simple combinations of blacklisted sites to discover new phishing sites. Also, it proposes an approximate matching algorithm to determine whether a given URL is a phishing site or not. PhishNet consists of two major components, namely, component I: predicting malicious URLs and component II: approximate matching.

The basic idea of component I is to combine different URL heuristics of known phishing URLs from a blacklist (i.e., PhishTank database) to generate new phishing URLs. These five URL heuristics include: (1) top-level domains (TLDs): by changing the TLDs of known blacklist entries, a list of new URLs can be obtained; (2) IP address: the predicted new phishing sites are obtained by enumerating all the combinations of the hostnames and pathnames of the known blacklisted websites with the same IP address; (3) directory structure: the idea is that two URLs sharing a common directory structure (e.g., www.abc.com/online/signin/paypal.htm and www.xyz.com/online/signin/ebay.htm) may have similar sets of file names. Therefore, the predicted new URLs are www.abc.com/online/signin/ebay.htm and www.xyz.com/online/signin/paypal.htm; (4) query string: starting from the observation that some URLs with the exact same directory structure differ only in query string (e.g., www.abc.com/online/signin/ebay?XYZ, and www.xyz.com/online/signin/paypal?ABC), two new URLs, www.abc.com/online/signin/ebay?ABC and www.xyz.com/online/signin/paypal?XYZ, are created; (5) brand name: the intuition here is that phishers often target multiple brand names using the same URL structure method. Therefore, the predicted URLs are obtained by changing the brand names embedded in the known phishing URLs.

After obtaining the whole set of the predicted URLs, PhishNet first performs a DNS lookup to filter out sites that cannot be resolved. Then it conducts a content similarity check (i.e., using an online tool at http://www.webconfs.com) between the known phishing URLs and the corresponding predicted URLs. The predicted URL is concluded to be a phishing site if the similarity score exceeds a certain threshold.

The second component performs an approximate match of a given URL to determine whether it is a phishing site or not. It first breaks the input URL into four different entities: IP address, hostname, directory structure and brand name. Then it assesses each entity by matching with the corresponding part of the known phishing URLs to generate an evaluation score. If the score is higher than a certain threshold, it is considered to be a phishing URL.

About 18,000 new phishing URLs are discovered from a set of 6,000 new blacklist entries. The proposed scheme achieves 3% false positive rate and 95% true positive rate.

Similar works based on blacklist/white-list include [23] and [96].

Hybrid methods: Whittaker *et al.* [108] use a logistic regression classifier to maintain Google's phishing blacklist automatically by examining the URL and the contents of a page. The proposed scheme correctly classifies more than 90% of phishing pages several weeks after training concludes. Marchal *et al.* [75] develop a phishing detection system that requires very little training data, which is language-independent, resilient to adaptive attacks and implemented entirely on client-side. The proposed target identification algorithm is faster than previous works and can help reduce false positives. The proposed scheme achieves 0.5% false positive rate and 99% true positive rate.

The most representative work in this category is Monarch [102], a real-time system that determines whether the submitted URL is spam or not. The authors deploy a real implementation to demonstrate its scalability, accuracy, and run time performance. Monarch consists of four components: (1) URL aggregation: it accepts URL submissions from a number of major email providers and Twitter's streaming API; (2) Feature collection: it visits a URL via Firefox Web browser to collect page content; (3) Feature extraction: it transforms the raw data generated from the feature collection component into a feature vector (e.g., transforming URLs into binary features and converting HTML content into a bag of words [110]). (4) Classification: feature vectors are applied to a proposed distributed logistic regression classifier for classification. The selected features in [102] are represented by a combination of URL based features, page content based features, whitelist and other features (e.g., routing data), including:

- Initial URL and Landing URL: domain tokens, path tokens, query parameters, number of sub-domains, length of domain, length of path, length of URL.
- Redirects: number of redirects, type of redirect.
- Sources and Frames: URL features for each embedded IFrame links and sources links.
- HTML Content: tokens of main HTML, frame HTML, and script content.
- Page Links: URL features for each link, number of links, ratio of internal domains to external domains.
- JavaScript Events: number of user prompts, tokens of prompts.
- Pop-up Windows: URL features for each window URL.
- Plugins: URL features for each plugin URL.
- HTTP Headers: tokens of all field names and values;
- DNS: IP of each host, mailserver domains and IPs, nameserver domains and IPs.
- Geolocation: country code, city code of each IP.
- Routing Data: ASN/BGP prefix for each IP encountered.
- Whitelist: a whitelist of known good domains.

Logistic Regression (LR) with L1-regularization is chosen as the classifier. To predict the class label ($y = -1$ means non-spam, $y = +1$ means spam) of a URL's feature vector x. We train a linear classifier characterized by weight vector w. Given a set of n labeled training points $(x_i; y_i)$, $i = 1{:}n$, the training process is to find w that minimizes the following objective function:

$$f(w) = \sum_{1}^{n} log\big(1 + exp\big[-y_i(x_i * w_i)\big]\big) + \lambda * ||w||_1$$

The first component is the log likelihood of the training data as a function of the weight vector. The second component is the regularization which adds a penalty to the objective function [71].

To perform the learning process over large-scale datasets in real time, the data is divided into *m* shares and then processed in a distributed manner (i.e., using Hadoop Spark [121]).

Monarch can achieve an overall accuracy of 91% with 0.87% false positives with a throughput of 638,000 URLs per day. Similar works that use hybrid features include [13], [16], [69], [79], [80], and [84].

Other methods: Ramesh *et al.* [90] present a phishing website detection approach based on the phishing target identification. After obtaining the target domain name, the proposed scheme performs third-party DNS look up for comparison to decide the legitimacy of the suspicious page. The proposed scheme achieves 0.32% false positive rate and 0.33% false negative rate.

Similar works based on phishing target identification include [101] and [107].

IV. Phishing Detection Schemes: A Systematic Study From Different Perspectives

In this section, we perform a systematic review of the software based phishing detection schemes from different perspectives including evaluation datasets, detection features, detection techniques and evaluation metrics.

A. Evaluation Datasets

The evaluation is tightly coupled with the ground truth datasets employed by the various approaches. Different approaches collect ground truth from different cyber intelligence sources. Such sources may employ different testing methodologies and target different types of phishing activities, and hence cover different phishing domains. That is, evaluation based on one dataset may differ from that based on another. Therefore, we argue that having a publicly available reference datasets is crucial for systematizing the evaluation of various approaches. Because it is an important step towards providing a benchmark to compare and contrast the efficiency of various approaches and it can help researchers to further advance the area in a more systematic way. The absence of reference sets combined with difficulties in sharing code, make it hard to repeat experiments for systematic comparison of effectiveness. In the following, we list the identifying features of the datasets used in the literature:

- *Dataset source:* Table VII lists the commonly used data sources of phishing websites and legitimate websites, together with the approaches that leverage each source. There is no common consensus on the quality of the different sources due to the lack of knowledge about the methodologies used in compiling and maintaining each source.
- *Dataset size:* the evaluation dataset size varies a lot among different approaches. Generally speaking, the larger the dataset, the more credible the results.
- *Dataset redundancy:* Datasets, especially those of phishing websites, usually contain repeated entries due to multiple submissions and overlap among different sources. However, little information is provided about datasets redundancy in the literature.
- *Dataset timeliness:* Phishing websites tend to have very short life time. Therefore, phishing blacklist providers usually update information in hourly, daily or weekly schedules. Even if two schemes use the same data source with the same dataset size, they may contain different phishing website information.
- *Ratio of legitimate to phishing websites:* the ratio of legitimate to phishing instances shows the extent to which the experiments represent a real world distribution ($\approx 100/1$).
- *Training set to testing set ratio:* the ratio of training to testing instances indicates the scalability of the approach.

TABLE VII
Sources of Datasets for Phishing and Legitimate Websites

Phishing websites		
Data source	**Description**	**Literature**
APWG [4]	The Anti Phishing Work Group maintains a blocked URL list in the form of an archive	[106] [13] [102]
PhishTank [81]	It contains downloadable databases which are available in multiple formats and updated hourly. It is easy to have fast and up to date phishing detection built into your application.	[124] [77] [70] [69] [15] [72] [46] [87] [13] [122] [118] [66] [102] [89] [107] [74] [90] [27] [80] [75] [37]
Millersmiles [8]	It has archive of spoof email and phishing scams from the last few years and currently stores all scam reports in the HoneyTrap database (i.e., 2025857 scams) which is now available for commercial license.	[84] [90]
SpamScatter [18]	It mines emails in real-time, follows the embedded link structure, and automatically clusters the destination web sites using image shingling to capture graphical similarity between rendered sites to identify and analyze scams	[72] [87]
MalwarePatrol [7]	It is an open source community for sharing malicious URLs.	[66]
SURBL [9]	It offer URL reputation data for professional users through faster updates and resulting fresher data.	[102]
CLEAN MX [5]	It provide real time public access query for malicious URLs	[94]
Other sources	E.g., webmail providers, google searching, etc.	[108] [73] [84] [20]
legitimate website		
Data source	**Description**	**Literature**
DMOZ [6]	Searchable people-reviewed web directory categorized by language, subject and location	[87] [72] [66] [74] [37]
Top Alexas [2]	The top 500 sites on the web	[46] [118] [89] [90] [27]
Yahoo [114]	Yahoo! Directory was a web directory which at one time rivaled DMOZ in size	[70] [72] [73] [87] [66] [107] [80]
Google	Keywords searching using Google	[38] [94] [122] [90] [20]
3Sharp [92]	3Sharp's list of legitimate URLs in their Gone Phishing report	[124] [69]
Other sources	E.g., user submission, third party providers etc.	[15] [108] [75]

In Section V, we use these aspects to perform a systematic and comprehensive evaluation of the various phishing detection approaches.

B. Phishing Detection Features

1) Most Commonly Used Phishing Detection Features: In this section, we summarize the most commonly used features by various phishing detection approaches. Even though the

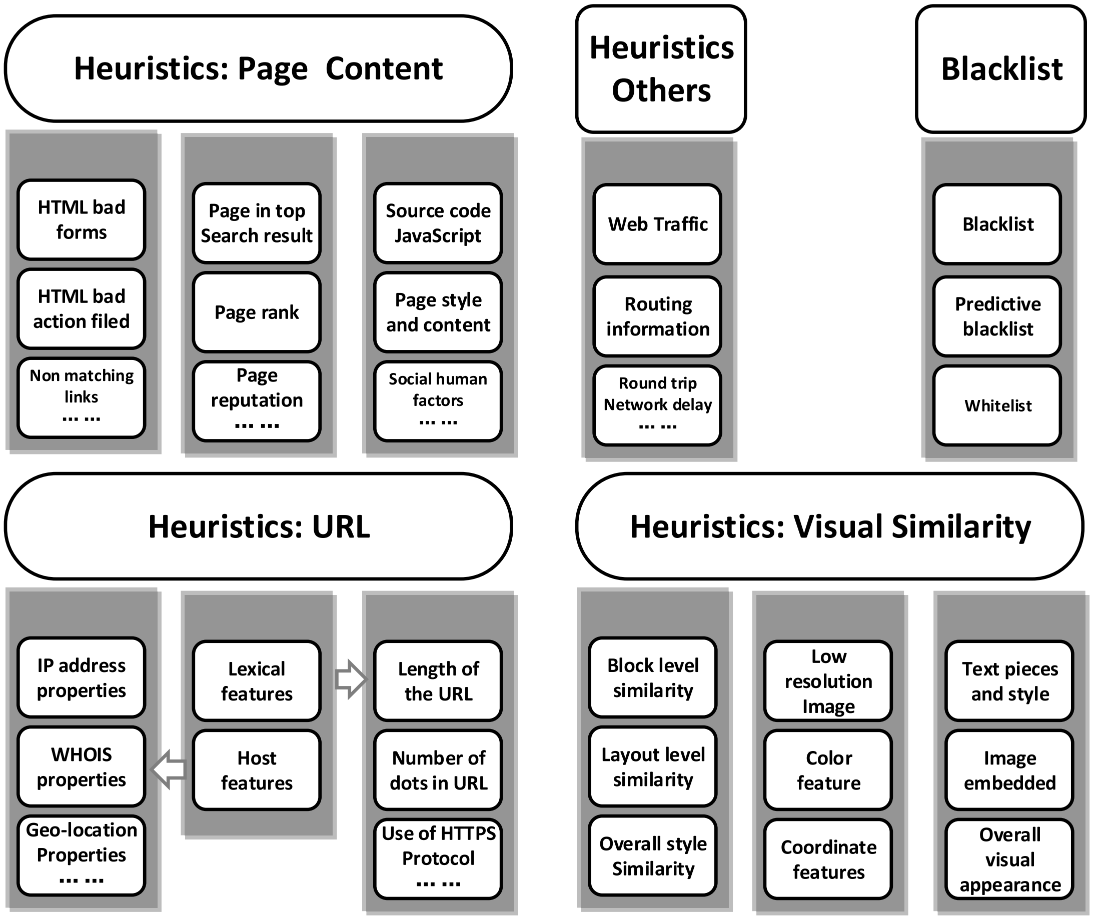


Fig. 8. Most commonly used phishing detection features.

listing of atomic features presented here is not exhaustive, it includes the popular features used in most of the state-of-the-art phishing detection approaches. Figure 8 summarizes the features.

(i) URL-based lexical features: URLs are rich of lexical features that have been widely used in various phishing detection approaches [22], [108], [118], including:

- *URL replaced with IP address:* Some phishing websites do not use host-names, but rather use IP address directly to locate the fake site. Such behavior is normally employed either to obfuscate the legitimate URL or simply to reduce cost.
- *URL Length*: Phishing websites usually have longer URLs compared to legitimate websites.
- *Number of dots and sub domains:* Phishing URLs often contain more "dots" and sub-domains compared to legitimate ones.
- *Number of re-directions:* Malicious URLs often have multiple URL redirects in order to evade detection by blacklists.
- *Use of HTTPS protocol:* Legitimate websites often use HTTPS protocol, while phishing sites usually do not.

(ii) URL-based host features:

- *WHOIS information:* WHOIS is a query and response protocol that is widely used for querying databases that store registration information about websites [51], [72].
  - *Registered information about domain names:* For most of the observed phishing sites, either the registration record is not available in WHOIS databases or the claimed identity is not accurate in the record.
  - *Age of Domain:* Many of the observed phishing websites have domains that are registered only a few days before phishing emails are sent out, that is, phishing domains are likely to be short lived.
- *Geographic information:* Geographical location is one of the most commonly used indicators in detecting phishing because phishing websites are likely to be hosted in locations different from those of legitimate websites [3]. For example, Netcraft [3] provides location information (i.e., IP-based country information) to help in identifying fraudulent URLs. For example, the real bankofamerica.com is unlikely to be hosted in Russia.
- *Domain name similarity:* A measure of the similarity between a potential phishing domain name and a target domain name. The similarity can be measured in many ways. For example, it can be measured based on the Edit Distance between the two domains [28]. The Edit Distance (a.k.a., Levenshtein distance) is the number of characters that need to be inserted or deleted in order to transform one domain into another. The smaller the number of insertions and deletions, the higher the similarity.

(iii) Website page content: The textual content of the phishing website can be used to determine the identity of the target website.

- *Page in top search result:* Xiang *et al.* [118] proposed a content-based approach to detect phishing websites based on the TF-IDF information with the help of Google search engine, as follows:
  - TF-IDF of each term on a suspected Web page is calculated.
  - Top 5 terms with highest TF-IDF values are selected.
  - The top 5 terms are submitted to a search engine and the domain names of the *n*-first returned entries are stored.
  - If the suspected domain name is found within the *n* returned results, the site is considered legitimate.
- *Page rank:* PageRank is a link analysis algorithm first used by Google, in which each document on the Web is assigned a numerical weight from 0 to 10, with 0 indicating “least popular" and 10 representing “most popular”.
- *Page style & contents:* Aburrous *et al.* [13] propose several page style and content heuristics including: spelling errors, copying website, using forms with submit button, using Pop-Up windows, and disabling Right-Click.
- *Bad forms:* Phishing attacks are usually accomplished through HTML forms. This feature checks if a page contains potentially harmful HTML forms.
- *Non-matching links:* Links on phishing sites are usually meaningless or contain URLs of the target legitimate sites. Therefore, this feature examines all the links in the HTML, and checks if the most frequent domain coincides with the page domain [118].

(iv) Website visual similarity:

- *Text, image and overall similarity:* Medvet *et al.* [77] present a technique to visually compare a suspected phishing page with the legitimate one via a set of visual features. These features include (i) each visible text section with its visual attributes, (ii) each visible image, and (iii) the overall visual look-and-feel (i.e., the larger composed image) of the Web page visible in the viewport (i.e., the part of the Web page that is visible in the browser window).
- *Dominant color and its centroid coordinate:* Fu *et al.* [38] first convert Web pages into low resolution images and then use the dominant color category and its corresponding centroid coordinate to represent the image signatures. Finally, they use Earth Mover’s Distance (EMD) to calculate the signature distances of the images of the Web pages (i.e., legitimate vs. phishing).

2) Network Round Trip Time (NRTT): A New Phishing Detection Feature: In this section, we propose a new phishing detection feature based on the Network Round Trip Time, dubbed as *NRTT*. *NRTT* has been introduced in [56] as a reliable and robust second Web authentication factor. *NRTT* simply captures the network round trip time that packets take in its journey from one Internet connected host to another and back to the original sender. We propose here a two-phase approach based on *NRTT* to detect phishing Web sites. In the first phase, the targeted victim site is identified through content analysis of the phishing website, URL characteristics, spam email content analysis, or combination of them. Victim website identification is well studied in literature (e.g., [70], [75], and [107]), and hence we capitalize on the state-of-the-art mechanisms to build our victim identification component. In the second phase, we compare the average *NRTT* values of both the phishing website and the targeted victim website from a certain vantage point. If the difference between the two *NRTT* averages is greater than a threshold, the link is highly likely a phishing link. The intuition here is that the phishing website is likely to be hosted on a server different from that of the victim website, and hence, they will experience different average *NRTT*.

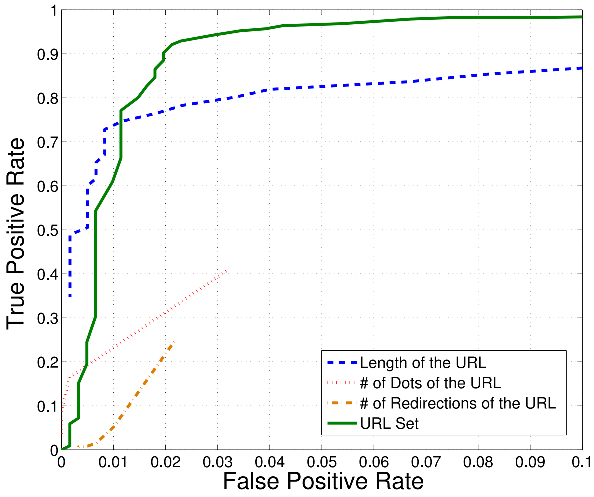


Fig. 9. ROC curves of *FPR-TPR* for different URL based features and URL sets.

To account for transient network instabilities, *NRTT* is measured by sending multiple packets (called profiling signals in [56]) from the vantage point to the target website, which acknowledges each packet back to the vantage point. The number of the profiling signals is optimized based on the trade-off between the accuracy of the measurements on one hand and the bandwidth and the delay overhead on the other hand. Based on the observation that *NRTT* follows a Gaussian distribution [56], [62], Khalil *et al.* [56] theoretically show that 27 profiling signals are sufficient to create reliable *NRTT*.

More importantly, *NRTT* should be robust enough to resist manipulation attacks and should cope with different kinds of network instabilities to avoid false positives. These *NRTT* challenges have been well studied and addressed in [56], where the authors define and address three types of network instabilities that may affect the authentication accuracy:

Instantaneous instabilities cause transient changes in *NRTT* and hence, they only affect a few of the profiling signals. This type of instability is addressed by outlier filtering based on median absolute deviation [67].

Long-term instabilities are instabilities that persist long enough to affect all or most of the of profiling signals yet not permanent. These instabilities are mainly caused by traffic congestion at the local network segment connecting to the network

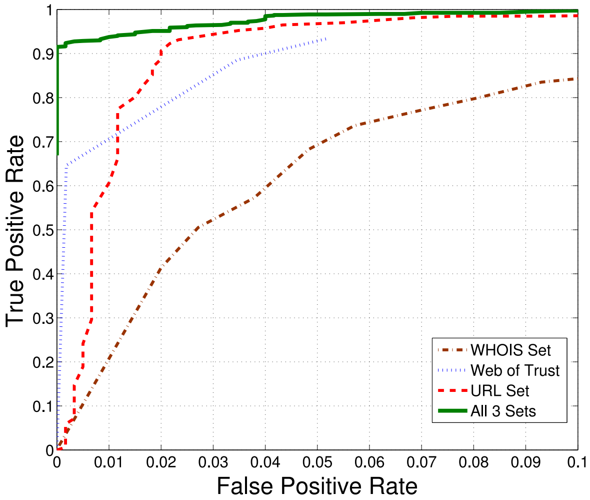


Fig. 10. ROC curves of *FPR-TPR* for different feature sets.

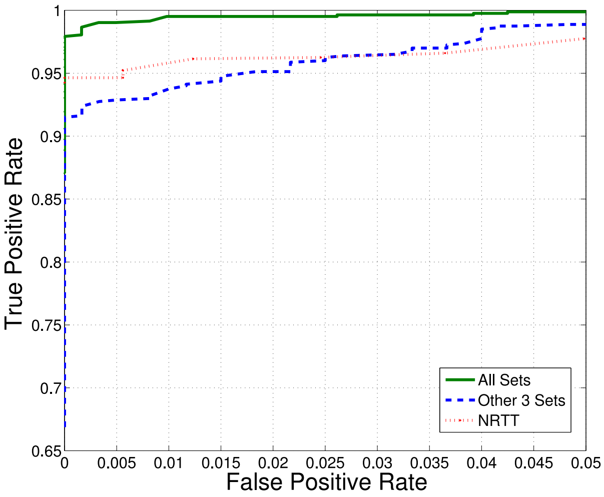


Fig. 11. ROC curves of *FPR-TPR* for selected feature sets, *NRTT* and the all feature sets.

backbone. This type of instability is addressed by measuring *NRTT* from different vantage points as detailed in [56].

Routing instabilities may result in permanent changes in network communications latency due to, for example, permanent network routing changes. It has been shown in many previous works [29], [64], [65], [91], [95], that only a small portion of the Internet is responsible for the vast majority of the routing instabilities and these routing changes exhibit a strong temporal periodicity, despite the growth of the Internet.

Leveraging *NRTT* for our problem, that is, distinguishing phishing from legitimate websites is much more practical than the application envisioned by Khalil *et al.* [56], that is, Web authentication: (i) Identifying phishing websites does not have the limitation and the concern of mobile clients, not only because Web servers are static but also because *NRTT* is only computed and compared on the fly for the two websites (the suspected phishing and the target website). (ii) No reference profiles are maintained and stored at the vantage point. (iii) The unsolved routing network instabilities mentioned above do not exist in our case. For Web authentication, the reference profile and the real time profiles are measured at different times. That is, it is possible that permanent route changes occur between measuring reference and real-time profiles, which may harshly affect the efficiency. On the other hand, *NRTT* of both the phishing and the legitimate website are measured at the same time in real time, and hence, permanent instabilities are not a concern. (iv) Long term instabilities are also not a concern in our problem. This is because local network congestion does not apply in the case of Web hosting servers compared to Web clients who may have poor network connections. Additionally, *NRTT* signals are sent at the same time for the phishing and legitimate websites, that is, the instabilities affect both and hence the difference between the two remain unchanged.

In order to demonstrate the effectiveness of *NRTT* as a phishing detection feature, we perform a set of experiments to evaluate the trade-off between True Positive Rate (*TPR*) and False Positive Rate (*FPR*) among different selected features and different feature sets (including *NRTT*). *TPR* measures the proportion of positives that are correctly identified (i.e., the percentage of phishing sites which are correctly identified):

$$TPR = \frac{\#\ \ of\ \ correctly\ \ detected\ \ phishing}{Total\ \ \#\ \ of\ \ phishing}$$

FPR measures the proportion of positives that are incorrectly identified (i.e., the percentage of legitimate sites which are wrongly identified as phishing sites):

$$FPR = \frac{\#\ \ of\ \ wrongly\ \ detected\ \ legitimate}{Total\ \ \#\ \ of\ \ legitimate}$$

Figure 9 shows the ROC (Receiver operating characteristic) curves of *FPR* vs. *TPR* for different URL based features including Length of the URL, Number of dots in the URL, Number of re-directions of the URL, and the URL set (i.e., combination of all of the 3 URL features plus the binary features: usage of HTTPS protocol and IP address in the URL). It shows that the URL set alone could achieve about 90% *TPR* with 2% *FPR*.

Figure 10 shows the ROC curves of *FPR* vs. *TPR* for different feature sets, including: WHOIS set, URL set, Web of trust score and the combination of all of the three sets "all 3 sets" (i.e., the combination of WHOIS set, URL set and Web of trust score). The WHOIS set contains two features, namely, the age of the domain and the existence of the registering information in WHOIS database. The Web of trust score is provided by SEO (search engine optimization) that collects all website ranking information based on Google, Bing, Yahoo, among others. The results show that the combination of all the selected feature sets can achieve about 93% *TPR* with 0.5% *FPR*.

Figure 11 shows the ROC curves of *FPR* vs. *TPR* for "other 3 sets" (i.e., the combination of WHOIS set, URL set and Web of trust score), *NRTT* and "all sets" (i.e., "other 3 sets" + *NRTT*). It clearly shows that, with the combination of all the features, the proposed scheme can achieve 99% *TPR* and 0.2% *FPR*.

The evaluation dataset contains 820 verified on-line phishing websites (redundancy reduced) collected from PhishTank

TABLE VIII
SUMMARY OF PHISHING DETECTION/CLASSIFICATION TECHNIQUES

	Literature
Support Vector Machine (SVM)	[84] [72] [73] [118] [94] [46] [66] [74] [79]
Earth Mover's Distance (EMD)	[39] [38]
Logistic Regression (LR)	[40] [72] [108] [118] [37] [27] [102]
Density-based spatial clustering of applications with noise (DBSCAN)	[70]
Term frequencyInverse Document Frequency (TF-IDF)	[124] [90] [118]
Bayesian-based	[69] [72] [122] [118] [37] [20]
Fuzzy-based	[15] [13] [14]
Confidence Weighted (CW)	[73] [22] [66]
Neural Network (NN)	[94] [80]
Random Forest (RF)	[94] [37] [74]
J48 Decision Tree	[94][37] [118]
AdaBoost	[94] [118]
Perceptron	[73] [66]
Adaptive Regularization of Weights (AROW)	[66]
latent Dirichlet allocation (LDA)	[89]
Sequential Minimal Optimization (SMO)	[20]
C4.5	[74]
Random Tree (RT)	[74]
Logistic Model Tree (LMT)	[74]
Jrip	[74]
Gradient Boosting	[75]

and 612 legitimate websites obtained from Alexa in September 2016.

C. Phishing Detection Techniques

In this section, we research the phishing detection/classification techniques and their supporting technologies (e.g., data mining algorithms, detection strategies, etc.). Table VIII summarizes the results. We can conclude that the Support Vector Machine (SVM), Logistic Regression (LR) and Bayesian-based classifiers are the most popular tools used to support phishing detection in the literature we covered.

We include feature mining algorithms in this category because they are increasingly used in recent detection techniques. For example, the TF-IDF algorithm (as introduced in Section III-B) is widely used in many page content based phishing detection schemes [90], [118], [124].

D. Evaluation Metrics

In this section, we summarize the most commonly used evaluation metrics in the literature.

- *Accuracy (ACC):* The number of correctly identified phishing and legitimate websites divided by the total number of websites.

$$ACC = \frac{\#\ \ of\ \ correctly\ \ identified\ \ sites}{Total\ \ \#\ \ of\ \ sites}$$

- *True Positive Rate (TPR):* The number of phishing sites correctly identified divided by the total number of phishing sites. It is also known as hit rate, recall (R), and sensitivity in different parts of the literature.
- *False Positive Rate (FPR):* The number of legitimate sites that are wrongly identified as phishing sites divided by the total number of legitimate sites. It is also known as false alarm rate and Type I error in some parts of the literature.
- *True Negative Rate (TNR):* The number of correctly identified legitimate websites divided by the total number of legitimate websites.

$$TNR = \frac{\#\ \ of\ \ correctly\ \ identified\ \ legitimate}{Total\ \ \#\ \ of\ \ legitimate}$$

- *False Negative Rate (FNR):* The number of phishing sites that incorrectly identified as legitimate sites divided by the number of phishing sites. It is also known as miss rate, Type II error in some parts of literature.

$$FNR = \frac{\#\ \ of\ \ phishing\ \ identified\ \ as\ \ legitimate}{Total\ \ \#\ \ of\ \ phishing}$$

- *Precision (P):* The rate of correctly detected phishing sites in relation to all sites that were detected as phishing.

$$P = \frac{\#\ \ of\ \ phishing\ \ correctly\ \ identified}{Total\ \ \#\ \ of\ \ sites\ \ detected\ \ as\ \ phishing}$$

- *F1 score:* The harmonic mean between precision P and recall R.

$$F1 = 2TPR/(2TPR + FPR + FNR)$$

Although two different phishing detection schemes may use the same evaluation metrics, we cannot simply compare the numerical results. We have to take other aspects into consideration when performing comparisons. For example, the use of different datasets may have high impact on the results, as discussed in Section IV-A.

Based on our investigations, two of the most commonly used evaluation metrics are *TPR* and *FPR*. Due to the relatively high percentage of legitimate websites compared to the phishing ones, the latter metric is considered to be of high importance for practicality and usability reasons. Even a very small *FPR* may result in a large number of wrongly identified legitimate websites as phishing and hence diminish any potential benefit of the approach.

V. INSIGHTS AND FUTURE RESEARCH DIRECTIONS

Most of the phishing detection related work focus on the detection quality but often overlook other important angles such as datasets, features, practicality and performance. In this section, we provide a comprehensive view of all the important aspects that we have discussed as part of studying and evaluating phishing detection schemes.

More importantly, this section provides detailed takeaway lessons for researchers and practitioners in the area of phishing detection. This section is structured following the four aspects of phishing detection that have been discussed in the survey, namely: datasets, detection features, detection schemes, and evaluation metrics. According to our study, every technique (e.g., machine learning algorithms) used in the literature has its own advantages and disadvantages, which has to be carefully selected in order to optimize the goals of the detection approach.

A. Dataset Selection

Datasets considerably affect the evaluation results. This evaluation criteria measures how easy it is to compile the datasets and to extract the necessary features from them. It also provides an understanding of the deployment environment. Different environments may call for different detection techniques. For example, in healthy environments where the vast majority of the websites are legitimate, schemes with low *FPR* are preferable over those with high *TPR* and relatively high *FPR*. In Section IV-A, we elaborate on the various important aspects relevant to datasets used in training and evaluating different phishing detection approaches.

One of the important challenges here is to compile a representative dataset that covers as much as possible of the behaviors of phishing attackers. It is relatively easy to collect phishing related datasets from a single organization. However, such datasets may only offer a limited local view of the threats. On the other hand, compiling datasets from multiple organizations could be challenging due to the potentially leakage of sensitive information. Therefore, such efforts are usually hampered by restrictive legal clauses including non-disclosure agreements and regulations against sensitive information leakage. Even though when such representative datasets may be successfully compiled by some entities, it is generally difficult to share the datasets with the bigger phishing detection community. This not only adds to the difficulty of systematically comparing different phishing detection approaches, but also considerably limits the potential benefits of the datasets. The takeaway lesson here is that it is not always true that the bigger the dataset, the better the detection outcome, but rather, the more representative the dataset, the more comprehensive the features collected and the better the detection performance.

The timeliness of phishing detection datasets is another challenging issue. As mentioned earlier, the cyber space is dynamic by nature, a characteristic that has been extensively exploited by attackers to evade detection. Attackers frequently change behavior and adjust their attack models to evade detection. The most obvious examples include Dynamic Generation Algorithms (DGAs) and Fast Flux Service Networks (FFSNs). DGAs (e.g., [100]) create phishing websites, among others, that are only accessible for short time periods. This makes the identification of a phishing email, for example, by a known suspicious link that it contains useless, because such link may continuously change, some times on daily bases. FFSNs (e.g., [48]) exploit short TTL values to frequently change the IP addresses assigned to a specific domain, mainly to evade IP-based blocking. Such behavior makes, for example, IP-based phishing identification inefficient. Dynamic behavior of attackers also complicates the process of dataset collection as it limits the validity of such datasets to only short time periods. For example, a dataset of phishing URLs that has been compiled 1 year ago, may not be valid now because most of the listed URLs may no longer be in service.

One last issue is related to the ground truth datasets. It has been shown [88], [99] that blacklists, which are usually used to compile ground truth datasets, have high false positive and false negative rate. Such impurities may negatively impact the detection accuracy of the underlying phishing detection approaches. Therefore, we recommend to always cross-check the ground truth in multiple blacklists, or use majority voting to decide the inclusion of entries in the ground truth dataset.

B. Feature Selection and Engineering

Different phishing detection systems use different combinations of detection features. In this survey, we have discussed the most important detection features used by different schemes. However, the literature lacks systematic evaluation of these features in terms of the availability of the detection features, the time it takes to mine the features, the complexity of extracting the features, and the robustness of the features. For example, a feature that can be easily manipulated by adversaries should not be used irrespective of its effectiveness in detecting current phishing attempts. This is intuitive as adversaries can simply manipulate the feature to avoid detection.

Feature extraction (a.k.a., feature engineering) is a challenging task that has a significant impact on the quality (accuracy and robustness) of the underlying phishing detection approaches. Well-crafted features lead to more successful detection approaches, while poor features may even ruin good detection algorithms. Typically, approaches look for features that maximize the detection accuracy while ignoring the robustness of the feature. Some features may have strong positive impact on the detection accuracy, however, they may be controlled by attackers and hence can be easily manipulated without affecting the attacker utility. For example, the URL-based lexical features (e.g., length of the URL, usage of HTTPS protocol, etc.) can be easily manipulated to evade detection. More specifically, it is no longer uncommon to see some phishing attackers increasingly use HTTPS, which makes such features less effective in identifying phishing attempts.

The most important takeaway here is that it is paramount for phishing detection approaches to carefully select the features that strike the right balance between detection accuracy and robustness in the face of potential manipulations. Our new *NRTT* feature is an example of a robust feature that not only provides excellent detection accuracy (as clearly shown by the experimental results), but also continuously adapt to changes in the underlying requirements. According to the work in [33], the design and measurement of *NRTT* aim at preventing attackers from being able to learn and mimic legitimate *NRTTs*, and hence the feature is robust. Additionally, the decision threshold specific to the feature is not fixed, but rather is adaptive and may change from the current measurement to the next one, depending on network conditions. More specifically, both the baseline *NRTT* and the *NRTT* of the claimed URL are measured and compared at the same time, and hence, congestion or other network conditions do not affect the outcome of the comparison.

However, it is quite challenging to evaluate the robustness of a feature in a systematic and measurable way. The importance of the problem has been recognized by many researchers in other domains as well (e.g., [82], [118], and [123]). However, to the best of our knowledge, none of the existing approaches provide a framework that can be used to quantitatively evaluate the robustness of features. Most of the approaches that recognize the problem only qualitatively discuss the robustness of some of the important features used in their approaches. Therefore, we believe that providing a framework that outlines qualitative and quantitative evaluation guidelines of the robustness of features is an open problem that calls for attention from the research community. Such framework has to consider both complexity of feature forging as well as its impact on attacker benefits. Such framework will be an important tool in the face of the ever evolving attack as it helps researchers and practitioners to design phishing detection techniques leveraging features that are both adaptive and hard to manipulate without considerably affecting attack utility.

One additional issue to consider while designing the detection features of an approach is the time it takes to mine the feature. Some features could be extremely useful in identifying and detecting phishing attempts, however, they may take a relatively long time to compute, such as page reputation and virtual appearance similarity. The use of such features may either result in user inconvenience due to service delays until computation completes, or may result in security risks in case the service is provided before computing the features.

C. Detection Schemes

As presented in Section IV-C, phishing detection systems use various data mining algorithms and detection approaches, each with its own advantages and disadvantages. Understanding the underlying data mining algorithms is important in evaluating the performance, the scalability and the robustness of phishing detection schemes.

When designing a phishing detection scheme, we recommend to follow the life cycle illustrated in Figure 5 in order to help fellow researchers obtain a comprehensive understanding of the proposed approach and to make it easy for future studies to conduct comparative evaluations. Specifically, a detection approach has to clearly state what it can and what it cannot do in terms of phishing detection and blocking, to avoid relying on the approach in scenarios where it may not be efficient. Additionally, details of dataset specifications in terms of content and volume that better support the approach should be clearly articulated and documented.

A very important lesson that we have learned is that, due to the dynamic nature of cyber attacks, the most reliable and efficient phishing detection approaches are those that can continuously adapt to cope with such dynamisms. References [108] and [118] are examples of such dynamic approaches. Additionally, the robustness of phishing detection approaches is tightly coupled with the robustness of the features used by the approach. Therefore, an approach may result in excellent detection accuracy at the time of design or in its early deployment, but fails miserably later due to either changes in the dataset or deliberate manipulation of the features utilized by the approach.

Two main categories have been considered in the underlying technologies (e.g., feature mining, classification, etc.) of phishing detection approaches. The first category of approaches utilizes human expertise to identify phishing attempts, which has been implemented using various heuristics. However, such approaches require man-in-the-loop, and hence are too slow. They fail to handle large scale datasets and cannot cope with high data rates, frequent dataset changes, or adaptive attack behaviors. Therefore, machine learning technologies, which utilize data-driven algorithms, were introduced to help automate the learning process. Different machine learning algorithms are used. Support Vector Machine (SVM), Logistic Regression (LR), and Bayesian-based classifiers, are among the mostly used algorithms in the literature.

Through our extensive investigation of the large body of phishing detection approaches, we learned that one size does not fit all and hence, it is extremely difficult to recommend one machine learning algorithm over another. Each machine learning algorithm has its own strengths and weaknesses, which has to be carefully considered to optimize the goals of the detection approach. For example, SVM is considered among the most robust and accurate classification algorithms [117]. However, it has the drawback of being computationally inefficient, and hence may not be appropriate for large scale datasets or high data rates. On the other hand, LR is one of the most widely used statistical models for binary data [12]. However, it performs poorly when nonlinear relationships exist between feature sets. Furthermore, even though Bayesian-based classifiers are easy to construct and can be readily applied to large scale datasets [103], they assume independent features, and hence are very restrictive.

Recent research efforts leverage Deep Learning (DL) algorithms to improve the performance of phishing detection schemes. DL allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction [102]. DL has been successfully applied in many research fields, such as speech recognition, visual object recognition, drug discovery and genomics. Therefore, we believe that DL could be a viable alternative to traditional machine learning algorithms (e.g., SVM, LR), especially when handling complex and large scale datasets.

Another important issue that we have identified through this survey is the absence of deep and systematic evaluation of the performance of phishing detection approaches. The vast majority of the approaches focus on evaluating and analyzing the detection accuracy, while they overlook the run-time performance of the approach. Some approaches may show acceptable performance during the design and test phases due to the relatively small size datasets used during these phases. However, real world datasets are usually more complex and much larger, which may cause such approaches to perform poorly in real world applications. Systematic performance analysis can provide important guidelines to evaluate the scalability of detection approaches, which in turn can help in improving performance by considering, for example, distributed platforms and parallel algorithms.

D. Evaluation Metrics

In addition to evaluating the quality of the detection scheme in terms of *FPR* and *TPR*, it is also imperative to have

TABLE IX
EVALUATION OF ACADEMIC PHISHING DETECTION SCHEMES

			[38]	[124]	[40]	[72]	[73]	[87]	[108]	[118]	[102]	[90]	[27]	[75]
Performance	*Accuracy*						99%					99.62%		
	True Positive Rate		89%	97%	88%	92.4%		95%	97%	99%	90.87%	99.67%	100%	>99%
	False Positive Rate		0.71%	1%	0.7%	0.1%		3%	<0.1%	0.4%	0.87%	0.32%	0.74%	0.5%
	True Negative Rate											99.5%		
	False Negative Rate													
	Precision											0.5%		
	Recall													
	F1 score													
Data Set	*Data set source*	*Phishing*		PhishTank	Google	PhishTank Spamscatter	Webmail provider	PhishTank Spamscatter	Gmail User submission	Phishtank	Google SURBL APWG PhishTank	Phishtank Others	Phishtank	PhishTank
		Legitimate	Google	3Sharp		DMOZ Yahoo	Yahoo	DMOZ Yahoo	Gmail submission	Alexa Yahoo 3Sharp		Alexa Google Others	Alexa	Intel Security
	Data set size	*Phishing*	9	100	Several million	20000	6000 7500	32000	16967	6943	500000	3374	1000	1553
		Legitimate	10272	100		15000	12500 to 14000	120000	74816740	2561	500000	1200	404	100000
	Data set redundancy									Reduced				
	Data set timeliness		2005	Nov.17th to 18th 2006	Aug.20th to 31st 2006	Aug.22nd to Sept.1st 2008	2009	2009	2009	2009	2011	2014	Jun. 4th to 22nd Jul. 13th to 25th 2012	2015 to 2016
	Legitimate sites to Phishing sites ratio		1141:1	1:1		4:3	2:1	15:4	4410:1	1:2.7	1:1	1:2.8	2.48:1	64.4:1
	Training set to testing set ratio		1:100			1:1			6:1	3:7	4:1			4:1
Feature	*Blacklist*				Y			Y						
	Heuristics: URL lexical			Y	Y	Y	Y	Y	Y	Y	Y			Y
	Heuristics: URL host			Y	Y	Y	Y	Y	Y	Y	Y			
	Heuristics: Page content			Y	Y				Y	Y	Y	Y	Y	Y
	Heuristics: Visual similarity		Y											
	Heuristics: Others										Y			
Techniques	*SVM*					Y	Y			Y				
	LR				Y	Y			Y	Y	Y		Y	
	Bayesian-based					Y				Y				
	EMD		Y											
	DBSCAN													
	TD-IDF			Y						Y		Y		
	Fuzzy-based													
	CW						Y							
	NN													
	RF													
	J48									Y				
	AdaBoost									Y				
	Perceptron						Y							
	AROW													
	LDA													
	SMO													
	C4.5													
	RT													
	LMT													
	Gradient Boosting													Y
	Jrip													

systematic evaluation of the effectiveness and the scalability of the underlying detection algorithms. Effectiveness metrics have been discussed in Section IV-D, however, we note that most of the work in the literature lacks evaluation of other performance aspects such as speed of detection, usability, and practical deployment, among others.

The majority of phishing detection approaches leverage machine learning concepts including clustering and classification techniques. Therefore, they adopt the evaluation metrics and strategies developed in this domain. However, as mentioned earlier, the cyber security domain is more challenging due to the adaptive nature of attackers. Therefore, the evaluation results during the design phase should be considered with caution, as they may not hold later. In other words, the design phase results are limited in time validity and scope, which calls for the phishing detection community to think about adaptive evaluation strategies that cope with the unique challenges in the cyber security domain. For example, the researchers could firstly classify the dataset into different categories (e.g., by type, time period, country, etc.), then perform the evaluation over every type of the dataset to obtain a more comprehensive and convincing results. Another important issue is the difficulty in providing comparative evaluations among different phishing detection techniques. This is mainly due to the lack of standard benchmarks, and the lack of reference datasets as a consequence of the dynamic nature of the attackers and the potential sensitivity of data, which restricts sharing.

Unfortunately, many of the above mentioned challenges across all the aspects continue to exist, and hence call for a collaborative effort among the research community to alleviate their negative impact on the effectiveness and coverage of phishing detection approaches.

Table IX shows the comparison results across the previous four evaluation dimensions. From the performance perspective, we can see that 11 out of 12 schemes focus on the evaluation of true positive rate, false positive rate or other equivalent evaluation metrics. This is mainly because of the fact that *TPR* determines the detection capability of the scheme while *FPR* represents its negative effects. Thus, they together provide the most valuable performance information about the quality of different approaches.

PhishTank is the most dominant source for phishing websites because it provides large quantity, up-to-date and verified phishing list for free. Yahoo and DMOZ were competitors to each other for providing legitimate websites information. However, Yahoo closed its service since the end of 2014 for some unknown reasons. Other favorite sources for legitimate websites include Alexa top sites and Google keywords searching. Although researchers try to use a larger number of datasets for more convincing evaluation results, few of them considered some fundamental aspects about the datasets. For example, the ratio of the number of legitimate websites to the number of phishing websites, which is about 100 to 1 in reality.

Blacklists are commonly used in the public phishing detection toolbars because they have the fastest response time. From Table IX, we can conclude that the most commonly used features (also with the best performance results) are URL based and page content based features. Also, studies that are incorporated with more features tend to have better performance results. The recent trend is to leverage the classifier itself to optimize the detection accuracy using a large number of various detection features.

VI. Conclusion

In this paper, we provide a systematic study of existing phishing detection works from different perspectives. We first describe the background knowledge about the phishing ecosystem and the state-of-the-art phishing statistics. Then we present a systematic review of the automatic phishing detection schemes. Specifically, we provide a taxonomy of the phishing detection schemes, discuss the datasets used in training and evaluating various detection approaches, discuss the features used by various detection schemes, discuss the underlying detection algorithms and the commonly used evaluation metrics. Finally, we provide recommendations that we believe will help guide the development of more effective phishing detection schemes and make it easy to compare and contrast various schemes.

References

[1] (2016). *Phishing Trends & Intelligence Report: Hacking the Human*. [Online]. Available: https://info.phishlabs.com/pti-report-download

[2] *The Alexa Top 500 Sites on the Web*. Accessed: Nov. 21, 2016. [Online]. Available: http://www.dmoz.org/

[3] *Anti-Phishing Extension: Netcraft*. Accessed: Dec. 5, 2016. [Online]. Available: http://toolbar.netcraft.com/

[4] *Anti-Phishing Working Group*. Accessed: Nov. 15, 2016. [Online]. Available: http:/www.antiphishing.org/

[5] *Clean MX Malicious URL List*. Accessed: Nov. 21, 2016. [Online]. Available: http://support.clean-mx.com/clean-mx/phishing.php?response=alive

[6] *DMOZ—The Directory of the Web*. Accessed: Nov. 21, 2016. [Online]. Available: http://www.dmoz.org/

[7] *Malwarepatrol*. Accessed: Oct. 31, 2016. [Online]. Available: https://www.malwarepatrol.net/open-source.shtml

[8] *Millersmiles Spoof Email and Phishing Scams List*. Accessed: Nov. 21, 2016. [Online]. Available: http://www.millersmiles.co.uk/scams.php

[9] *SURBL URL Reputation Data*. Accessed: Nov. 21, 2016. [Online]. Available: http://www.surbl.org/lists

[10] G. Aaron and R. Rasmussen, *Global Phishing Survey: Trends and Domain Name Use in 2h2009*, Anti-Phishing Working Group, Lexington, MA, USA, 2010.

[11] S. Abu-Nimeh and S. Nair, "Bypassing security toolbars and phishing filters via DNS poisoning," in *Proc. IEEE Glob. Telecommun. Conf. (GLOBECOM)*, New Orleans, LA, USA, 2008, pp. 1–6.

[12] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proc. ACM Anti Phishing Working Groups 2nd Annu. eCrime Researchers Summit*, Pittsburgh, PA, USA, 2007, pp. 60–69.

[13] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7913–7921, 2010.

[14] M. Aburrous and A. Khelifi, "Phishing detection plug-in toolbar using intelligent fuzzy-classification mining techniques," in *Proc. Int. Conf. Soft Comput. Softw. Eng.*, San Francisco, CA, USA, 2013.

[15] S. Afroz and R. Greenstadt, "Phishzoo: An automated Web phishing detection approach based on profiling and fuzzy matching," in *Proc. 5th IEEE Int. Conf. Semantic Comput. (ICSC)*, 2009.

[16] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on Twitter," in *Proc. IEEE eCrime Researchers Summit (eCrime)*, 2012, pp. 1–12.

[17] F. Aloul, S. Zahidi, and W. El-Hajj, "Two factor authentication using mobile phones," in *Proc. AICCSA*, Rabat, Morocco, 2009, pp. 641–644.

[18] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker, "Spamscatter: Characterizing Internet scam hosting infrastructure," in *Proc. Usenix Security*, Boston, MA, USA, 2007, pp. 1–14.

[19] *APWG Phishing Trends Reports*, Anti Phishing Working Group, 2016.

[20] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," in *Proc. IEEE Conf. Commun. Netw. Security (CNS)*, Florence, Italy, 2015, pp. 769–770.

[21] A. Bergholz *et al.*, "New filtering approaches for phishing email," *J. Comput. Security*, vol. 18. no. 1, pp. 7–35, 2010.

[22] A. Blum, B. Wardman, T. Solorio, and G. Warner, "Lexical feature based phishing URL detection using online learning," in *Proc. 3rd ACM Workshop Artif. Intell. Security*, Chicago, IL, USA, 2010, pp. 54–60.

[23] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in *Proc. 4th ACM Workshop Digit. Identity Manag.*, Alexandria, VA, USA, 2008, pp. 51–60.

[24] D. D. Caputo, S. L. Pfleeger, J. D. Freeman, and M. E. Johnson, "Going spear phishing: Exploring embedded training and awareness," *IEEE Security Privacy*, vol. 12, no. 1, pp. 28–38, Jan./Feb. 2014.

[25] K.-T. Chen, J.-Y. Chen, C.-R. Huang, and C.-S. Chen, "Fighting phishing with discriminative keypoint features," *IEEE Internet Comput.*, vol. 13, no. 3, pp. 56–63, May/Jun. 2009.

[26] T.-C. Chen, S. Dick, and J. Miller, "Detecting visually similar Web pages: Application to phishing detection," *ACM Trans. Internet Technol.*, vol. 10, no. 2, p. 5, 2010.

[27] T.-C. Chen, T. Stepan, S. Dick, and J. Miller, "An anti-phishing system employing diffused information," *ACM Trans. Inf. Syst. Security*, vol. 16, no. 4, p. 16, 2014.

[28] N. Chou, R. Ledesma, Y. Teraguchi, and J. C. Mitchell, "Client-side defense against Web-based identity theft," in *Proc. NDSS*, San Diego, CA, USA, 2004.

[29] G. Comarela, G. Gürsun, and M. Crovella, "Studying interdomain routing over long timescales," in *Proc. Conf. Internet Meas.*, Barcelona, Spain, 2013, pp. 227–234.

[30] L. F. Cranor, S. Egelman, J. I. Hong, and Y. Zhang, "Phinding phish: An evaluation of anti-phishing toolbars," in *Proc. NDSS*, San Diego, CA, USA, 2007.

[31] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Montreal, QC, Canada, 2006, pp. 581–590.

[32] Z. Dou, I. Khalil, and A. Khreishah, "CLAS: A novel communications latency based authentication scheme," *Security Commun. Netw.*, vol. 2017, 2017, Art. no. 4286903.

[33] Z. Dou, I. Khalil, and A. Khreishah, "A novel and robust authentication factor based on network communications latency," *IEEE Syst. J.*, to be published.

[34] Z. Dou, I. Khalil, A. Khreishah, and A. Al-Fuqaha, "Robust insider attacks countermeasure for Hadoop: Design and implementation," *IEEE Syst. J.*, to be published.

[35] J. S. Downs, M. B. Holbrook, and L. F. Cranor, "Decision strategies and susceptibility to phishing," in *Proc. 2nd Symp. Usable Privacy Security*, Pittsburgh, PA, USA, 2006, pp. 79–90.

[36] S. Egelman, L. F. Cranor, and J. Hong, "You've been warned: An empirical study of the effectiveness of Web browser phishing warnings," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Florence, Italy, 2008, pp. 1065–1074.

[37] M. N. Feroz and S. Mengel, "Examination of data, rule generation and detection of phishing URLs using online logistic regression," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Washington, DC, USA, 2014, pp. 241–250.

[38] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing Web pages with visual similarity assessment based on earth mover's distance (EMD)," *IEEE Trans. Depend. Secure Comput.*, vol. 3, no. 4, pp. 301–311, Oct./Dec. 2006.

[39] A. Y. Fu, L. Wenyin, and X. Deng, "EMD based visual similarity for detection of phishing webpages," in *Proc. Int. Workshop Web Doc. Anal.*, vol. 2005. 2005.

[40] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in *Proc. ACM Workshop Recurring Malcode*, Alexandria, VA, USA, 2007, pp. 1–8.

[41] S. Gastellier-Prevost, G. G. Granadillo, and M. Laurent, "A dual approach to detect pharming attacks at the client-side," in *Proc. 4th IFIP Int. Conf. New Technol. Mobility Security (NTMS)*, Paris, France, 2011, pp. 1–5.

[42] GeoTrust. *Geotrust TrustWatch Toolbar*. Accessed: Dec. 5, 2016. [Online]. Available: https://www.geotrust.com/comcasttoolbar/

[43] A. Gharaibeh *et al.*, "Smart cities: A survey on data management, security and enabling technologies," *IEEE Commun. Surveys Tuts.*, to be published.

[44] (2016). *Anti-Phishing Working Group*. [Online]. Available: http://www.antiphishing.org

[45] X. Han, N. Kheir, and D. Balzarotti, "Phisheye: Live monitoring of sandboxed phishing kits," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, Vienna, Austria, 2016, pp. 1402–1413.

[46] M. Hara, A. Yamada, and Y. Miyake, "Visual similarity-based phishing detection without victim site information," in *Proc. IEEE Symp. Comput. Intell. Cyber Security (CICS)*, Nashville, TN, USA, 2009, pp. 30–36.

[47] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *J. Math. Phys.*, vol. 20, nos. 1–4, pp. 224–230, 1941.

[48] T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling, "Measuring and detecting fast-flux service networks," in *Proc. 15th Netw. Distrib. Syst. Security Symp.*, 2008.

[49] J. Hong, "The state of phishing attacks," *Commun. ACM*, vol. 55, no. 1, pp. 74–81, 2012.

[50] Google Inc. *Google Safe Browsing*. Accessed: Dec. 5, 2016. [Online]. Available: https://developers.google.com/safe-browsing/

[51] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Commun. ACM*, vol. 50, no. 10, pp. 94–100, 2007.

[52] C. Karlof, U. Shankar, J. D. Tygar, and D. Wagner, "Dynamic pharming attacks and locked same-origin policies for Web browsers," in *Proc. 14th ACM Conf. Comput. Commun. Security*, Alexandria, VA, USA, 2007, pp. 58–71.

[53] I. Khalil and S. Bagchi, "Secos: Key management for scalable and energy efficient crypto on sensors," in *Proc. IEEE Depend. Syst. Netw.*, 2003.

[54] I. Khalil, S. Bagchi, and N. Shroff, "Analysis and evaluation of secos, a protocol for energy efficient and secure communication in sensor networks," *Ad Hoc Netw.*, vol. 5, no. 3, pp. 360–391, 2007.

[55] I. Khalil, Z. Dou, and A. Khreishah, "TPM-based authentication mechanism for apache hadoop," in *Proc. Int. Conf. Security Privacy Commun. Syst.*, Beijing, China, 2014, pp. 105–122.

[56] I. Khalil, Z. Dou, and A. Khreishah, "Your credentials are compromised, do not panic: You can be well protected," in *Proc. 11th ACM AsiaCCS*, Xi'an, China, 2016, pp. 925–930.

[57] I. Khalil, I. Hababeh, and A. Khreishah, "Secure inter cloud data migration," in *Proc. 7th Int. Conf. Inf. Commun. Syst. (ICICS)*, Irbid, Jordan, 2016, pp. 62–67.

[58] I. Khalil, T. Yu, and B. Guan, "Discovering malicious domains through passive DNS data graph analysis," in *Proc. 11th ACM Asia Conf. Comput. Commun. Security*, Xi'an, China, 2016, pp. 663–674.

[59] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2091–2121, 4th Quart., 2013.

[60] P. Kumaraguru *et al.*, "Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer," in *Proc. Anti Phishing Working Groups 2nd Annu. eCrime Researchers Summit*, Pittsburgh, PA, USA, 2007, pp. 70–81.

[61] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Teaching Johnny not to fall for phish," *ACM Trans. Internet Technol.*, vol. 10, no. 2, p. 7, 2010.

[62] M. Kwon *et al.*, "Use of network latency profiling and redundancy for cloud server selection," in *Proc. IEEE 7th Int. Conf. Cloud Comput.*, Anchorage, AK, USA, 2014, pp. 826–832.

[63] *Spoofguard*, Stanford Security Lab., Stanford, CA, USA, 2004. [Online]. Available: https://crypto.stanford.edu/SpoofGuard/

[64] C. Labovitz, G. R. Malan, and F. Jahanian, "Internet routing instability," *IEEE/ACM Trans. Netw.*, vol. 6, no. 5, pp. 515–528, Oct. 1998.

[65] M. Lad, J. H. Park, T. Refice, and L. Zhang, "A study of Internet routing stability using link weight," Dept. Comput. Sci., Univ. California at San Diego, San Diego, CA, USA, Tech. Rep., 2008.

[66] A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: Url names say it all," in *Proc. IEEE INFOCOM*, Shanghai, China, 2011, pp. 191–195.

[67] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, 2013.

[68] G. L'Huillier, A. Hevia, R. Weber, and S. Ríos, "Latent semantic analysis and keyword extraction for phishing classification," in *Proc. IEEE Int. Conf. Intell. Security Inf. (ISI)*, Vancouver, BC, Canada, 2010, pp. 129–131.

[69] P. Likarish, E. Jung, D. Dunbar, T. E. Hansen, and J. P. Hourcade, "B-APT: Bayesian anti-phishing toolbar," in *Proc. IEEE Int. Conf. Commun.*, Beijing, China, 2008, pp. 1745–1749.

[70] G. Liu, B. Qiu, and L. Wenyin, "Automatic detection of phishing target from phishing webpage," in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, Istanbul, Turkey, 2010, pp. 4153–4156.

[71] Z. Q. J. Lu, "The elements of statistical learning: Data mining, inference, and prediction," *J. Roy. Stat. Soc. A, Stat. Soc.*, vol. 173, no. 3, pp. 693–694, 2010.

[72] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious Web sites from suspicious URLs," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Paris, France, 2009, pp. 1245–1254.

[73] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious URLs: An application of large-scale online learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Montreal, QC, Canada, 2009, pp. 681–688.

[74] S. Marchal, J. François, R. State, and T. Engel, "PhishStorm: Detecting phishing with streaming analytics," *IEEE Trans. Netw. Service Manag.*, vol. 11, no. 4, pp. 458–471, Dec. 2014.

[75] S. Marchal, K. Saari, N. Singh, and N. Asokan, "Know your phish: Novel techniques for detecting phishing sites and their targets," in *Proc. IEEE 36th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Nara, Japan, 2016, pp. 323–333.

[76] M.-E. Maurer and D. Herzner, "Using visual website similarity for phishing detection and reporting," in *Proc. Extended Abstracts Human Factors Comput. Syst. CHI*, Austin, TX, USA, 2012, pp. 1625–1630.

[77] E. Medvet, E. Kirda, and C. Kruegel, "Visual-similarity-based phishing detection," in *Proc. 4th Int. Conf. Security Privacy Commun. Netw.*, Istanbul, Turkey, 2008, p. 22.

[78] I.-C. Mihai and L. Giurea, "Management of eLearning platforms security," in *Proc. Int. Sci. Conf. eLearn. Softw. Educ.*, vol. 1. 2016, pp. 422–427.

[79] M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," *Expert Syst. Appl.*, vol. 53, pp. 231–242, Jul. 2016.

[80] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Comput. Appl.*, vol. 25, no. 2, pp. 443–458, 2014.

[81] *PhishTank: An Anti-Phishing Site*, LLC OpenDNS, San Francisco, CA, USA, accessed: Dec. 5, 2016.

[82] A. Oprea, Z. Li, T.-F. Yen, S. H. Chin, and S. Alrwais, "Detection of early-stage enterprise infection by mining large-scale log data," in *Proc. 45th Annu. IEEE/IFIP Int. Conf. Depend. Syst. Netw.*, Rio de Janeiro, Brazil, Jun. 2015, pp. 45–56.

[83] P. Pajares. *Phishing Safety: Is HTTPS Enough?* [Online]. Available: http://blog.trendmicro.com/trendlabs-security-intelligence/phishing-safety-is-https-enough/

[84] Y. Pan and X. Ding, "Anomaly based Web phishing page detection," in *Proc. ACSAC*, vol. 6, 2006, pp. 381–392.

[85] R. K. Panta, S. Bagchi, and I. M. Khalil, "Efficient wireless reprogramming through reduced bandwidth usage and opportunistic sleeping," *Ad Hoc Netw.*, vol. 7, no. 1, pp. 42–62, 2009.

[86] B. Parmar, "Protecting against spear-phishing," *Comput. Fraud Security*, vol. 2012, no. 1, pp. 8–11, 2012.

[87] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "PhishNet: Predictive blacklisting to detect phishing attacks," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, 2010, pp. 1–5.

[88] A. Ramachandran, D. Dagon, and N. Feamster, "Can DNS-based blacklists keep up with bots," in *Proc. 3rd Conf. Email Anti Spam*, 2006.

[89] V. Ramanathan and H. Wechsler, "Phishing website detection using latent Dirichlet allocation and adaboost," in *Proc. IEEE Int. Conf. Intell. Security Informat. (ISI)*, Arlington, VA, USA, 2012, pp. 102–107.

[90] G. Ramesh, I. Krishnamurthi, and K. S. S. Kumar, "An efficacious method for detecting phishing webpages through target domain identification," *Decis. Support Syst.*, vol. 61, pp. 12–22, May 2014.

[91] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, "BGP routing stability of popular destinations," in *Proc. 2nd ACM SIGCOMM Workshop Internet Meas.*, Marseille, France, 2002, pp. 197–202.

[92] P. Robichaux and D. L. Ganger, "Gone phishing: Evaluating anti-phishing tools for windows," 3Sharp Project, Redmond, WA, USA, Tech. Rep., Sep. 2006.

[93] J. C. Russ and R. P. Woods, "The image processing handbook," *J. Comput. Assisted Tomograph.*, vol. 19, no. 6, pp. 979–981, 1995.

[94] N. Sanglerdsinlapachai and A. Rungsawang, "Using domain top-page similarity feature in machine learning-based Web phishing detection," in *Proc. 3rd Int. Conf. Knowl. Disc. Data Min. (WKDD)*, 2010, pp. 187–190.

[95] A. Shaikh, A. Varma, L. Kalampoukas, and R. Dube, "Routing stability in congested networks: Experimentation and analysis," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 30, no. 4, pp. 163–174, 2000.

[96] M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator," in *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl.*, Doha, Qatar, 2008, pp. 840–843.

[97] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Atlanta, GA, USA, 2010, pp. 373–382.

[98] S. Sheng *et al.*, "Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish," in *Proc. 3rd Symp. Usable Privacy Security*, Pittsburgh, PA, USA, 2007, pp. 88–99.

[99] S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based 'blacklists,'" in *Proc. 3rd Int. Conf. Malicious Unwanted Softw.*, Fairfax, VA, USA, Oct. 2008, pp. 57–64.

[100] A. K. Sood and S. Zeadally, "A taxonomy of domain-generation algorithms," *IEEE Security Privacy*, vol. 14, no. 4, pp. 46–53, Jul./Aug. 2016.

[101] C. L. Tan, K. L. Chiew, K. Wong, and S. N. Sze, "PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder," *Decis. Support Syst.*, vol. 88, pp. 18–27, Aug. 2016.

[102] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time URL spam filtering service," in *Proc. IEEE Symp. Security Privacy*, Berkeley, CA, USA, 2011, pp. 447–462.

[103] G. Varshney, M. Misra, and P. K. Atrey, "A survey and classification of Web phishing detection schemes," *Security Commun. Netw.*, vol. 9, no. 18, pp. 6266–6284, 2016.

[104] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao, "Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email," *IEEE Trans. Prof. Commun.*, vol. 55, no. 4, pp. 345–362, Dec. 2012.

[105] W. D. Yu, S. Nargundkar, and N. Tiruthani, "A phishing vulnerability analysis of Web based systems," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Marrakech, Morocco, 2008, pp. 326–331.

[106] L. Wenyin, G. Huang, L. Xiaoyue, X. Deng, and Z. Min, "Phishing Web page detection," in *Proc. 8th Int. Conf. Document Anal. Recognit. (ICDAR)*, Seoul, South Korea, 2005, pp. 560–564.

[107] L. Wenyin, G. Liu, B. Qiu, and X. Quan, "Antiphishing through phishing target discovery," *IEEE Internet Comput.*, vol. 16, no. 2, pp. 52–61, Mar./Apr. 2012.

[108] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in *Proc. NDSS*, vol. 10. San Diego, CA, USA, 2010.

[109] *Avalanche (Phishing Group)—Wikipedia, the Free Encyclopedia*, Wikipedia, San Francisco, CA, USA, 2016.

[110] *Bag-of-Words Model—Wikipedia, the Free Encyclopedia*, Wikipedia, San Francisco, CA, USA, 2016.

[111] *Mcafee Siteadvisor—Wikipedia, the Free Encyclopedia*, Wikipedia, San Francisco, CA, USA, 2016, accessed: Sep. 6, 2016.

[112] *Microsoft Smartscreen—Wikipedia, the Free Encyclopedia*, Wikipedia, San Francisco, CA, USA, 2016, accessed: Sep. 28, 2016.

[113] *Netcraft—Wikipedia, the Free Encyclopedia*, Wikipedia, San Francisco, CA, USA, 2016, accessed: Sep. 3, 2016.

[114] *Yahoo! Directory—Wikipedia, the Free Encyclopedia*, Wikipedia, San Francisco, CA, USA, 2016, accessed: Jun. 7, 2016.

[115] *Zero-Day (Computing)—Wikipedia, the Free Encyclopedia*, Wikipedia, San Francisco, CA, USA, 2016.

[116] M. Wu, R. C. Miller, and S. L. Garfinkel, "Do security toolbars actually prevent phishing attacks?" in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Montreal, QC, Canada, 2006, pp. 601–610.

[117] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[118] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing Web sites," *ACM Trans. Inf. Syst. Security*, vol. 4, no. 2, 2011, Art. no. 21.

[119] G. Xiang and J. I. Hong, "A hybrid phish detection approach by identity discovery and keywords retrieval," in *Proc. 18th Int. Conf. World Wide Web*, Madrid, Spain, 2009, pp. 571–580.

[120] J. Yearwood, M. Mammadov, and A. Banerjee, "Profiling phishing emails based on hyperlink information," in *Proc. Int. Conf. Adv. Soc. Netw. Anal. Min. (ASONAM)*, Odense, Denmark, 2010, pp. 120–127.

[121] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, “Spark: Cluster computing with working sets,” in *Proc. HotCloud*, Boston, MA, USA, 2010, p. 10.

[122] H. Zhang, G. Liu, T. W. S. Chow, and W. Liu, “Textual and visual content-based anti-phishing: A Bayesian approach,” *IEEE Trans. Neural Netw.*, vol. 22, no. 10, pp. 1532–1546, Oct. 2011.

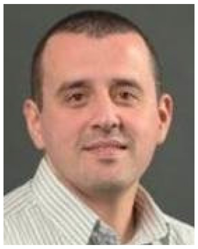
[123] J. Zhang, S. Saha, G. Gu, S.-J. Lee, and M. Mellia, “Systematic mining of associated server herds for malware campaign discovery,” in *Proc. 35th IEEE Int. Conf. Distrib. Comput. Syst.*, Columbus, OH, USA, 2015, pp. 630–641.

[124] Y. Zhang, J. I. Hong, and L. F. Cranor, “Cantina: A content-based approach to detecting phishing Web sites,” in *Proc. 16th Int. Conf. World Wide Web*, Banff, AB, Canada, 2007, pp. 639–648.

Zuochao Dou received the B.S. degree in electronics from the Beijing University of Technology in 2009, the M.S. degree from the University of Southern Denmark, concentrating on embedded control systems in 2011, and the M.S. degree from the University of Rochester majoring in communications and signal processing in 2013. He is currently pursuing the Ph.D. degree in cloud computing security and network security under the supervision of Dr. A. Khreishah and Dr. I. Khalil.

Issa Khalil (S’06–M’08) received the Ph.D. degree in computer engineering from Purdue University, USA, in 2007. He joined the College of Information Technology, United Arab Emirates University, where he served as an Associate Professor and the Department Head with the Information Security Department. In 2013, he joined the Cyber Security Group, Qatar Computing Research Institute, a member of Qatar Foundation, as a Senior Scientist, where he was recently promoted to a Principal Scientist. His research interests span the areas of wireless and wireline network security and privacy. He is especially interested in cloud security, malicious domain detection and takedown, and security data analytics. His novel technique to discover malicious domains following the guilt-by-association social principle attracts the attention of local media and stakeholders. He was a recipient of the CIT Outstanding Professor Award for outstanding performance in research, teaching, and service in 2011. He served as an Organizer, a Technical Program Committee Member, and a Reviewer for many international conferences and journals. He is a member of ACM and delivers invited talks and keynotes in many local and international forums.

Abdallah Khreishah received the B.S. (Hons.) degree from the Jordan University of Science and Technology in 2004, and the M.S. and Ph.D. degrees in electrical and computer engineering from Purdue University in 2010 and 2006, respectively. He was with NEESCOM. In 2012, he joined the Electrical and Computer Engineering Department, New Jersey Institute of Technology as an Assistant Professor and was promoted to an Associate Professor in 2017. His research spans the areas of wireless networks, visible-light communication, vehicular networks, congestion control, cloud and edge computing, and network security. His research projects are funded by the National Science Foundation, New Jersey Department of Transportation, and the UAE Research Foundation. He is currently serving as an Associate Editor for the *International Journal of Wireless Information Networks*. He served as the TPC Chair for WASA 2017, IEEE SNAMS 2014, IEEE SDS-2014, BDSN-2015, BSDN 2015, and IOTSMS-2105. He has also served on the TPC committee of IEEE Infocom 2017, IEEE Infocom 2016, IEEE PIMRC 2016, IEEE WCNC 2016, IEEE CCH 2016, IEEE PIMRC 2015, and ICCVE 2015. He is the Chair of the IEEE EMBS North Jersey Chapter.



Ala Al-Fuqaha (S’00–M’04–SM’09) received the M.S. degree in electrical and computer engineering from the University of Missouri-Columbia in 1999 and the Ph.D. degree in electrical and computer engineering from the University of Missouri-Kansas City in 2004. He is currently a Professor and the Director of NEST Research Laboratory, Computer Science Department, Western Michigan University. His research interests include wireless vehicular networks, cooperation and spectrum access etiquettes in cognitive radio networks, smart services in support of the Internet of Things, management and planning of software defined networks, and performance analysis and evaluation of high-speed computer and telecommunications networks. In 2014, he was a recipient of the Outstanding Researcher Award with the College of Engineering and Applied Sciences, Western Michigan University. He is currently serving on the Editorial Board for *Security and Communication Networks* (Wiley), *Wireless Communications and Mobile Computing* (Wiley), *EAI Transactions on Industrial Networks and Intelligent Systems*, and the *International Journal of Computing and Digital Systems*. He has served as a Technical Program Committee Member and a Reviewer of many international conferences and journals.

Mohsen Guizani (S’85–M’89–SM’99–F’09) received the B.S. (with Distinction) and M.S. degrees in electrical engineering, and the M.S. and Ph.D. degrees in computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, 1987, and 1990, respectively. He is currently a Professor and the ECE Department Chair with the University of Idaho, USA. He served as the Associate Vice President of Graduate Studies, Qatar University, the Chair of the Computer Science Department, Western Michigan University, and the Chair of the Computer Science Department, University of West Florida. He also served in academic positions with the University of Missouri-Kansas City, University of Colorado-Boulder, Syracuse University, and Kuwait University. His research interests include wireless communications and mobile computing, computer networks, mobile cloud computing, security, and smart grid. He currently serves on the Editorial Boards of several international technical journals and is the Founder and the Editor-in-Chief of *Wireless Communications and Mobile Computing* (Wiley). He has authored 9 books and over 450 publications in refereed journals and conferences. He guest edited a number of special issues in IEEE journals and magazines. He also served as a member, the Chair, and the General Chair of a number of international conferences. He was a recipient of teaching awards multiple times from different institutions as well as best research awards from three institutions. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker from 2003 to 2005. He is a Senior Member of ACM.