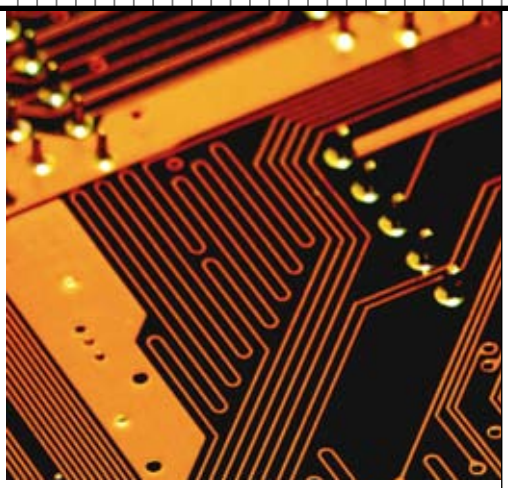


A Comparison of Tools for Detecting Fake Websites

Ahmed Abbasi, *University of Wisconsin-Milwaukee*

Hsinchun Chen, *University of Arizona*

As fake website developers become more innovative, so too must the tools used to protect Internet users. A proposed system combines a support vector machine classifier and a rich feature set derived from website text, linkage, and images to better detect fraudulent sites.

Fake websites are fictional, misrepresentative sites posing as legitimate providers of information, goods, or services used to garner illegitimate revenues by deceiving search engines or exploiting unsuspecting Internet users. Fraudsters have created several types of fake websites,[1] including web spam, concocted, and spoof sites, as Figure 1 shows.

Web spam sites attempt to deceive search engines to boost their rankings.[2] Leveraging link and content spamming methods, these websites engage in black hat search engine optimization in which highly ranked web spam sites are more visible and can be sold for a greater profit.[3,4] For instance, the cell phone spam domain in Figure 1a has an asking price of $350.

Concocted websites are deceptive sites attempting to appear as legitimate commercial entities. Figure 1b shows a concocted site for a counterfeit investment bank called "Troy Inc." The objective of such sites is failure-to-ship fraud; they collect unsuspecting users' money and disappear.[1] Concocted sites commonly pose as real escrow, financial, delivery, or retail companies.[5]

In contrast, spoof sites are imitations of real commercial sites, intended to deceive the authentic sites' customers.[6] The objective of spoofs is identity theft—capturing users' account information by having them log in to a fake site. Commonly spoofed sites include eBay, PayPal, and various banks.[7] Since considerable progress has been made on web spam detection,[2] we focus our attention on concocted and spoof sites.

Fake websites are often professional looking and difficult to identify as phony.[8] In response to increasing user awareness, fraudsters are also becoming more sophisticated,[9] while current security tools are unsuitable for handling fake websites' increasing complexity. Accordingly, a need has arisen for more refined fake website detection techniques.[6] Proposed tools have several shortcomings: Most are reactive lookup systems that rely solely on user-reported blacklists of fake URLs. Few systems use proactive classification techniques, and those that do utilize overly simplistic features and classification heuristics. Further, while developers have placed considerable focus on spoof site detection tools, concocted sites have received little attention despite their increasing prevalence.[5]

How effective existing tools would be at detecting concocted websites remains unclear. Since concocted sites do not simply mimic popular commercial websites, successfully identifying them requires more involved methods. To confront these challenges, we propose a support vector machine (SVM) classifier system for identifying fake websites. To further enhance performance, we combined the

 Published by the IEEE Computer Society

Figure 1. Example webpages for the three fake website categories: (a) Web spam site. (b) concocted site, and (c) spoof site.

Table 1. Summary of fake website detection tools.

Tool name	System type		Website type	Prior results (spoof sites)
	Classifier	Lookup		
CallingID	Domain registration information	Server-side blacklist	Spoof sites	Overall: 85.9% Spoof detection: 23.0%
Cloudmark	None	Server-side blacklist	Spoof sites	Overall: 83.9% Spoof detection: 45.0%
EarthLink toolbar	None	Server-side blacklist	Spoof sites	Overall: 90.5% Spoof detection: 68.5%
eBay Account Guard	Content similarity heuristics	Server-side blacklist	Spoof sites (primarily eBay and PayPal)	Overall: 83.2% Spoof detection: 40.0%
FirePhish	None	Server-side blacklist	Spoof sites	Overall: 89.2% Spoof detection: 61.5%
IE Phishing Filter	None	Client-side whitelist, server-side blacklist	Spoof sites	Overall: 92.0% Spoof detection: 71.5%
Netcraft	Domain registration information	Server-side blacklist	Concocted sites, spoof sites	Overall: 91.2% Spoof detection: 68.5%
Reasonable Anti-Phishing	Text and image feature similarity, stylistic feature correlation	Client-side whitelist	Spoof sites	N/A
Sitehound	None	Server-side blacklist downloaded by client	Concocted sites, spoof sites	N/A
SpoofGuard	Image hashes, password encryption, URL similarities, domain registration information	None	Concocted sites, spoof sites	Overall: 67.7% Spoof detection: 93.5%
TrustWatch	None	Server-side blacklist	Spoof sites	Overall: 85.1% Spoof detection: 46.5%

proposed classifier with a lookup mechanism to create a dynamic hybrid system.

FAKE WEBSITE DETECTION TOOLS

There are two types of fake website identification: *lookup systems* and *classifier systems*. Table 1 shows a summary of existing fake website detection tools. For each tool, the table lists the system type, applicable fake website categories, and prior results: overall accuracy (real and fake sites) and spoof site detection rates.[10] There has been no prior evaluation of concocted websites.

Lookup systems

Lookup systems use a client-server architecture in which the server side maintains a blacklist of known fake URLs,[10,11] and the client-side tool checks the blacklist and

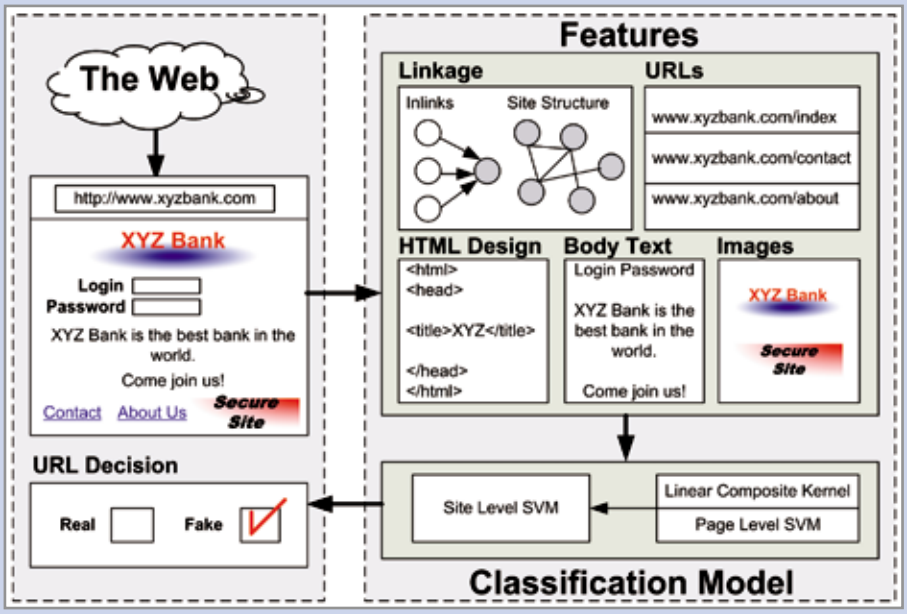


Figure 2. The AZProtect system uses a rich feature set and a kernel-based machine learning classifier.

provides a warning if a website poses a threat. Lookup systems employ collaborative sanctioning mechanisms similar to those in reputation ranking mechanisms.[12] Online communities of practice and system users provide information for the blacklists. Online communities such as the Anti-Phishing Working Group and the Artists Against 4-1-9 have developed databases of known concocted and spoof websites. Lookup systems also consider URLs directly reported or rated by system users.

Numerous lookup systems are available. Perhaps the most popular is Microsoft's IE Phishing Filter, which uses a client-side whitelist coupled with a server-side blacklist gathered from IE user reports and online databases. Similarly, Mozilla Firefox's FirePhish toolbar and the EarthLink toolbar also maintain a blacklist of spoof URLs. Firetrust's Sitehound system stores spoof and concocted site URLs taken from online sources such as the Artists Against 4-1-9.

An advantage of lookup systems is that they typically have high precision since they are less likely to consider authentic sites fake.[10] They are also easier to implement and computationally faster than most classifier systems; comparing URLs against a list of known fakes is fairly simple.

Lookup systems do, however, suffer from being susceptible to higher levels of false negatives—failing to identify fake websites. The blacklist is limited to a small number of online resources and lacks coverage. For example, the IE Phishing Filter and FirePhish tools only store URLs for spoof sites, making them inept against concocted sites. The performance of lookup systems might also vary based on the time of day and interval between report and evaluation time.[10] Blacklists are more likely to contain older fake websites than newer ones, which gives fraudsters a better opportunity of succeeding before being blacklisted. Five percent of spoof site recipients are defrauded in spite of the availability of a plethora of web browser integrated lookup systems.[7]

Classifier systems

Classifier systems are client-side tools that apply rule- or similarity-based heuristics to website content or domain registration information.[10,13] Developers have created a handful of classifier systems for fake website detection. SpoofGuard uses webpage features such as image hashes, password encryption checks, URL similarities, and domain registration information.[6] Netcraft's classifier relies on domain registration information such as the domain name, host name, host country, and registration date.[11] eBay's Account Guard tool compares the content of the URL of interest with legitimate eBay and PayPal sites.[10] Reasonable Anti-Phishing (formerly SiteWatcher) uses visual similarity assessment based on 40 body text, page style, and image features.[7] A page qualifies as a spoof if its similarity is above a certain threshold when compared to a client-side whitelist.

Classifier systems provide numerous benefits. They can offer better coverage for spoof and concocted sites than lookup systems.[5] Classifier systems are also proactive, capable of detecting fakes independent of blacklists. Consequently, classifier systems are not impacted by time of day and the interval between when a user visits a URL and the URL's first appearance in an online database.[10]

Nevertheless, classifier systems are not without their caveats. They can take longer to classify webpages than lookup systems. They are also more prone to false positives[10] (where positive refers to a legitimate website). Generalizability of classification models over time can be another issue, especially if the fake websites constantly evolve. For instance, the Escrow Fraud online database (http://escrow-fraud.com) has more than 250 unique templates for concocted sites with new ones added constantly. Effective classifier systems must employ a bevy of fraud cues and adapt and relearn to keep pace with the sophistication of fake websites.[7,9]

Table 2. AZProtect's feature set attributes.

Category	Feature	Description	IG weight
Body text	Word bigram "FREE HOSTING"	Fake websites are often hosted on websites that provide free hosting, such as Free Hostia.	1.000
	Word trigram "POWERED BY PHPBB"	Fake websites often use open source software packages (such as PHPBB) to generate website content.	1.000
	Word bigram "Member FDIC"	Legitimate websites usually contain information about their memberships with various government organizations, such as FDIC and BBB.	0.902
	Web unigram "español"	Many legitimate websites have multiple versions of their site in different languages.	0.534
Source code (HTML)	Links to careers/jobs webpage	Legitimate websites are more likely to place job postings on their website.	0.731
	Image preloading	This JavaScript code, which is used to preload images to decrease page loading times, rarely appears in fake websites.	0.688
URLs	URL token "HTTPS"	Fake websites rarely use the Secure Sockets Layer protocol.	0.933
	Percent of nonalphabetic characters in URL	Since fake websites are mass produced, they use random characters in URLs. This also allows new fake websites to easily circumvent lookup systems that rely on blacklists of exact URLs.	0.894
	Number of slashes "/" in URL	Spoof sites often piggyback off of legitimate websites or third-party hosts. The spoofs are buried deep on these websites' servers.	0.797
Links	Number of inlinks	Legitimate websites tend to have more websites point at them. Exceptions are some concocted websites that utilize link farms.	0.881
	Number of outlinks	Fake websites, particularly spoof sites, are generally partial replicas with only a handful of surface-level pages. As a result, they tend to contain fewer webpages (and less linkage).	0.817

Hybrid systems and dynamic classifiers

Hybrid systems combine classifier and lookup mechanisms. Such tools generally use simple content and domain registration information in unison with server-side blacklists.[11] The system blocks URLs on the blacklist, while the classifier evaluates others.

Examples of hybrid systems include Netcraft and eBay Account Guard. Dynamic hybrid systems using blacklists to update their classifiers could be highly effective against constantly changing fake website patterns. SpoofGuard does some updating; it stores image hashes for visited websites, allowing it to check for image duplication.[6] Nevertheless, work on dynamic classifiers for fake website detection has been limited.

PROPOSED APPROACH

AZProtect, our proposed classifier system, uses a rich feature set and a kernel-based machine learning classifier, as Figure 2 illustrates. The AZProtect system is capable of classifying concocted and spoof sites. Whereas existing systems only evaluate the current page's URL, the proposed system analyzes multiple webpages from the potentially fraudulent website for improved performance.

AZProtect utilizes a feature set containing nearly 6,000 attributes from five sources of information: body text, HTML design, images, linkage, and URLs. We collected these features by applying the information gain (IG) heuristic to a set of 500 training websites, encompassing concocted and spoof sites as well as 100 legitimate websites. We established these training websites approximately six months prior to the experimental testbed discussed here, selecting features with an IG value above a certain threshold (based on their occurrence in legitimate and fake websites).

The body text attributes consist of approximately 2,500 word-level (for example, "bank of", "bank of america") and character-level (for example, "pa", "pay") *n*-grams, while the HTML design features encompass more than 1,000 HTML tag *n*-grams (for example, "<html><head>"). The image features include pixel color frequencies arranged into 1,000 bins as well as 40 image structure attributes such as image height, width, file extension, and file size.

The feature set also includes 1,500 token- and character-level *n*-grams (unigrams, bigrams, and trigrams) derived from URLs and anchor text. We extracted the token-level *n*-grams (for example, "https," "org") by tokenizing URL

Represent each page a with the vectors:

$x_a = \{\text{Sim}_{ave}(a, b_1), \ldots, \text{Sim}_{ave}(a, b_p)\}$; $y_a = \{\text{Sim}_{max}(a, b_1), \ldots, \text{Sim}_{max}(a, b_p)\}$

Where:

$$\text{Sim}(a,k) = \frac{1}{2}\left(\left(1 - \frac{|1v_a - 1v_k|}{1v_a + 1v_k}\right) + \left(1 - \frac{|in_a - in_k|}{in_a + in_k}\right) + \left(1 - \frac{|out_a - out_k|}{out_a + out_k}\right)\right) + \frac{1}{2}\left(1 - \frac{1}{n}\sum_{i=1}^{n}\frac{|a_i - k_i|}{a_i + k_i}\right)$$

$$\text{Sim}_{ave}(a,b) = \frac{1}{m}\sum_{k=1}^{m}\text{Sim}(a,k)$$

$$\text{Sim}_{max}(a,b) = \underset{k \in \text{pages in site b}}{\text{agr max}}\ \text{Sim}(a,k)$$

For:

$b \in p$ web sites in the training set; $k \in m$ pages in site b; $a_1, \ldots a_n$ and $k_1, \ldots k_n$ are page a and k's feature vectors; $1v_a$, in_a, and out_a are the page level and number of in/out links for page a;

The similarity between two pages is defined as the inner product between their two vectors x_1, x_2, and y_1, y_2:

$$\text{Linear Composite Kernel: } K(x_1 + y_1, x_2 + y_2) = \frac{\langle x_1, x_2 \rangle}{\sqrt{\langle x_1, x_1 \rangle \langle x_2, x_2 \rangle}} + \frac{\langle y_1, y_2 \rangle}{\sqrt{\langle y_1, y_1 \rangle \langle y_2, y_2 \rangle}}$$

Figure 3. Linear composite kernel used by AZProtect's page-level classifier. The kernel function takes into account the content similarity and duplication tendencies of fake websites.

strings at the appearance of slashes, periods, and colons. Link and structure features included the total number of URL- and domain-level relative and absolute inlinks and outlinks for each webpage.[3] We also employed page-level frequency distribution for all inlink and outlink pages (the number of links for level 1 pages, level 2 pages, and so forth). We automatically derived inlink information using the Google search engine, as done in prior research.[14] Table 2 presents several attributes in the feature set, along with their IG weights on a 0-1 scale.

The classification model incorporates an SVM stack, composed of a page-level classifier and a site-level classifier. The page-level classifier uses a linear composite kernel, as Figure 3 shows. The kernel function is tailored to represent the content similarity and duplication tendencies of fake websites. It compares pages' feature vectors against training site pages and considers the average and maximum similarity for pattern and duplicate detection. The kernel also incorporates page linkage and structure information in each comparison, including inlinks/outlinks and page levels (the depth of pages based on the number of slashes in their URLs).

Given a website of interest, the kernel computes the similarity for each webpage a in that site against all webpages belonging to b, where b is part of the set of 500 real and fake websites in the training dataset. For a given webpage k in b, the similarity scores are on a 0-1 scale, with a score of 1 suggesting that a and k are identical. Scores are based on the occurrence of the aforementioned set of fraud cues in a and k, as well as the two pages' levels and number of inlinks and outlinks.

For each a, this results in a vector of similarity scores of length k (one vector for each b). For each vector, AZProtect computes the average and maximum similarity score. The average similarity score, $\text{Sim}_{ave}(a,b)$, is the average across all scores in the vector, while the maximum similarity, $\text{Sim}_{max}(a,b)$, is simply the highest similarity score in the vector. This results in a page-site similarity vector for each webpage. AZProtect computes the inner product between every two webpages' vectors to produce a kernel matrix that serves as the input into the page-level SVM classifier.

AZProtect then inputs the classifications from the page-level classifier into the site-level classifier. The site-level classifier uses three input attributes—the total number of pages classified, number classified as fake, and percentage of pages classified as fake—to make a decision regarding the website. The site-level classifier, which uses the classification results from multiple pages within a website, should allow better detection in situations in which a single fake page might not contain sufficient fraud cues.[5] Both SVM classifiers in the stack were trained on the set of 500 training websites.

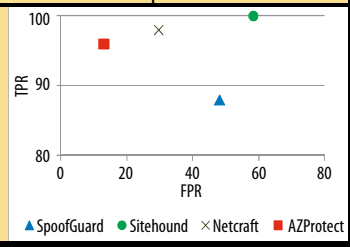
EXPERIMENTS AND RESULTS

Over a six-week period, we evaluated 350 concocted websites and 350 spoof sites taken from four online databases.[7,10] Concocted sites came from the Artists Against 4-1-9 (http://wiki.aa419.org) and Escrow Fraud (http://escrow-fraud.com) while the spoof sites came from PhishTank (www.phishtank.com) and the Anti-Phishing Working Group (www.antiphishing.org). We also evaluated 200 legitimate sites: 100 authentic websites to complement the 350 concocted sites and 100 legitimate websites commonly copied by the 350 spoofs. Overall, this resulted in two 450-website testbeds. The 500 sites used to train AZProtect did not overlap with the 900 sites in our testbeds.

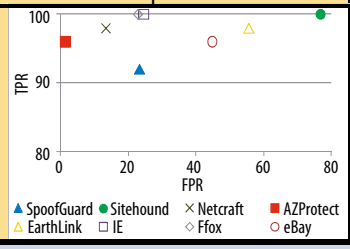
We evaluated fake sites between 9:00 a.m. and midnight. To assess the impact of evaluation time of day on performance, we collected several samples each hour. Because performance for certain lookup systems improves as the time interval increases,[10] we also evaluated the fake websites at different intervals between evaluation and report time in the online database. As a result, researchers collected a minimum of 10 evaluation samples for each hourly time interval between 0 and 24 hours (where all times were rounded to the nearest hour).

Table 3. Overall results for tool comparison on concocted and spoof site testbeds.

Concocted site test bed

Systems	All sites (n = 450)		Legit sites (n = 100)			Concocted sites (n = 350)		
	Acc.	ROC plots	F1	Prec.	Rec.	F1	Prec.	Rec.
SpoofGuard	59.77		49.30	34.24	88.00	66.66	93.78	51.71
Sitehound	54.45		49.38	32.79	**100.00**	58.59	**100.00**	41.43
Netcraft	76.44		64.90	48.52	98.00	82.28	99.19	70.29
AZProtect	**89.11**		**79.84**	**67.84**	97.00	**92.54**	99.02	**86.86**

Spoofed site test bed

Systems	All sites (n = 450)		Legit sites (n = 100)			Spoof sites (n = 350)		
	Acc.	ROC plots	F1	Prec.	Rec.	F1	Prec.	Rec.
SpoofGuard	80.00		67.15	52.87	92.00	85.62	98.53	76.57
Sitehound	40.22		42.64	27.10	**100.00**	37.58	**100.00**	23.14
Netcraft	89.11		80.00	67.58	98.00	92.52	99.67	86.57
AZProtect	**98.00**		**95.52**	**95.04**	96.00	**98.71**	99.42	**98.57**
EarthLink	56.23		49.87	33.45	98.00	61.15	99.36	44.29
IE	80.89		69.93	53.76	**100.00**	85.99	**100.00**	75.43
FirePhish	82.22		71.43	55.55	**100.00**	87.09	**100.00**	77.14
eBay	65.11		56.02	38.91	**100.00**	71.08	**100.00**	55.14

We evaluated the proposed AZProtect system's effectiveness in comparison with seven other state-of-the-art tools, some of which had performed well in prior testing[10,12] while others had not been evaluated. These included SpoofGuard, Netcraft, eBay Account Guard, IE Phishing Filter, FirePhish, EarthLink toolbar, and Sitehound. We compared only SpoofGuard, Netcraft, and Sitehound against AZProtect on the concocted site testbed, since the remaining tools do not effectively support concocted site detection. In contrast, we tested all eight tools on the spoof site testbed.

Since AZProtect examines multiple pages from the website of interest, we limited the maximum number of evaluated pages per site to 50 for computational reasons. AZProtect took an average of 2.9 seconds to evaluate a website, a number slightly higher than the 0.5- to 2.0-second times for other tools.[6,7] This includes the time necessary to collect webpages and images for a website, extract nearly 6,000 features from each collected webpage, and run the SVM classifier using the linear composite kernel.

Overall results

The evaluation metrics include overall accuracy, accuracy on legitimate sites, and accuracy on fake sites. The last is most important given the high cost of false negatives.[10] Table 3 shows the overall results on the two testbeds including the accuracy; receiver operating characteristic (ROC) plots showing true positive versus false positive; and the class-level precision, recall, and *f*-measures.

AZProtect had the best overall accuracy and class-level *f*-measures on both datasets. Based on the ROC plots, AZProtect also had the best ratio of true positives to false positives, indicated by its positioning in the top left corner on both plots. All p-values on pairwise t-tests were less than 0.0001 (n=450). Netcraft also performed well, but with 9 percent to 12 percent lower accuracy and *f*-measures. FirePhish, IE, and SpoofGuard fared decently on the spoof site testbed, while Sitehound performed poorly.

Impact of time of day and interval

Figures 4a and 4b show the results across times of day for various intervals between evaluation and report time on the 350 concocted sites. AZProtect had the best performance for interval between evaluation and report time and for evaluation time of day. Netcraft performed second best, followed by SpoofGuard and Sitehound. Netcraft's combination of classifier and lookup was beneficial; it detected many newly concocted sites by evaluating their domain registration information.

As expected, lookup systems such as Sitehound and Netcraft performed better as the interval between evaluation and report time increased. The two systems even outperformed AZProtect on longer intervals; however, AZProtect outperformed comparison techniques for all intervals less than 16 hours. Sitehound performed better when evaluating websites in the evening because that is when the tool's

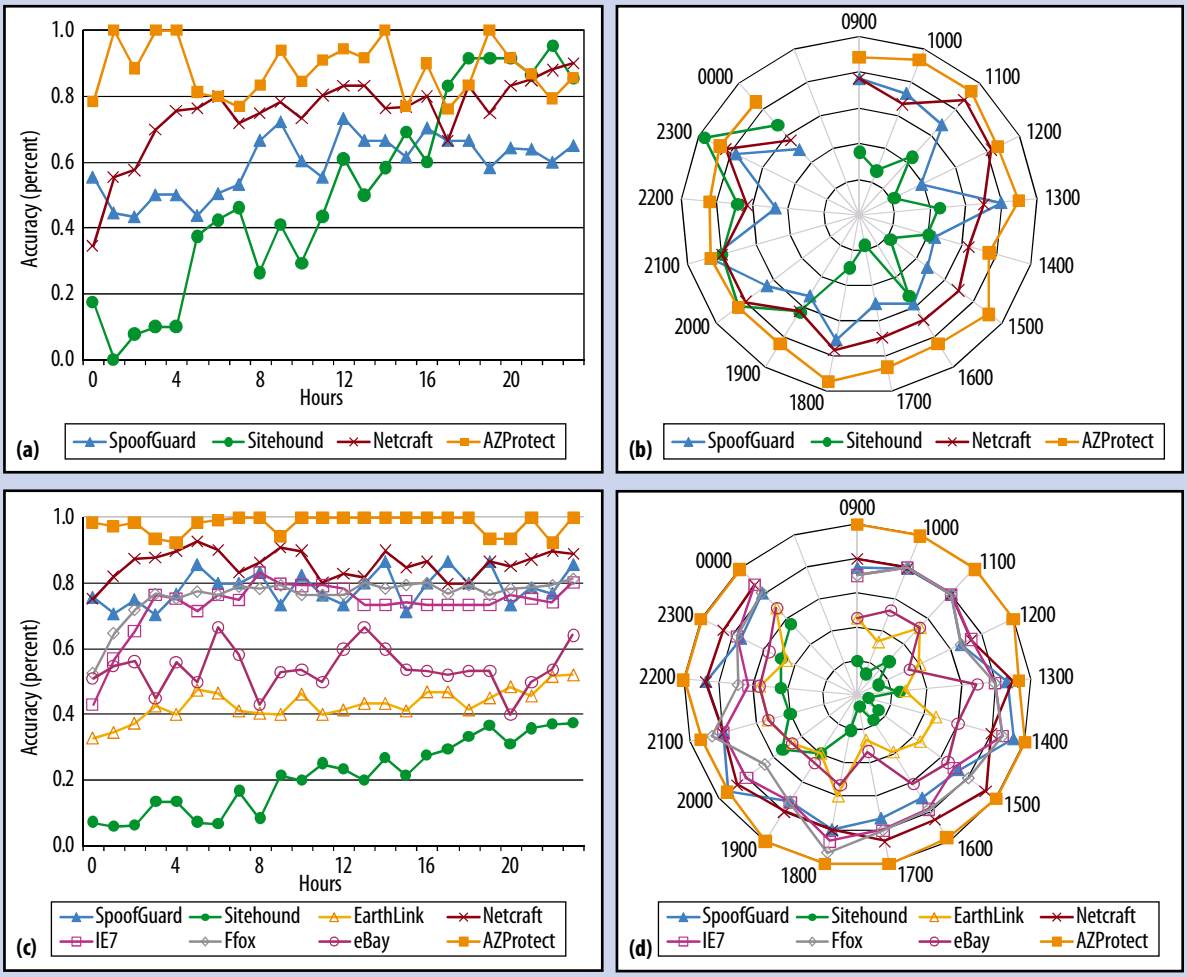


Figure 4. Impact of interval between evaluation and report time and time of day on accuracy for concocted and spoof site testbeds: (a) interval—concocted sites, (b) time of day—concocted sites, (c) interval—spoof sites, and (d) time of day—spoof sites.

server-side blacklist receives its daily updates, resulting in enhanced performance in subsequent hours.

Figures 4c and 4d show the spoof detection results. AZProtect again had the best performance, with more than 90 percent accuracy for all intervals and times of day. Netcraft, IE, FirePhish, and SpoofGuard also performed well for various times of day and intervals. Lookup systems such as IE and FirePhish only improved for time intervals up to four hours. Their accuracy leveled off to near 80 percent for longer time intervals because these tools update their blacklists more frequently.

EarthLink and Sitehound had detection rates under 50 percent for all time intervals. Sitehound once again performed better in the evening, while other tools performed consistently across times of day. The eBay tool performed well at identifying fake replicas of eBay and PayPal websites, which constitute a large portion of spoofs.[13] Interestingly, the results by time of day and interval for spoof sites were more stable than on the concocted sites, which tend to have greater content variability. In contrast, spoof sites usually replicate a handful of common sites.

Hybrid systems: Combining classifier and lookup methods

We assessed the effectiveness of combining the AZProtect classifier with a lookup mechanism on the same two sets of 350 fake website testbeds. The lookup component updated its blacklist every n hours, where n ranged from 1 to 24. We used the PhishTank and Artists Against 4-1-9 databases as blacklist sources. We compared three different systems: the standard AZProtect classifier, a hybrid classifier combining the classifier and lookup mechanism, and a hybrid classifier that combined the lookup mechanism with a dynamic classifier, which was updated every n hours with new blacklist URLs. The standard classifier ran on every URL. The two hybrid classifiers each compared URLs against the blacklist; the classifiers considered the URLs on the blacklist fake and evaluated the remainder.

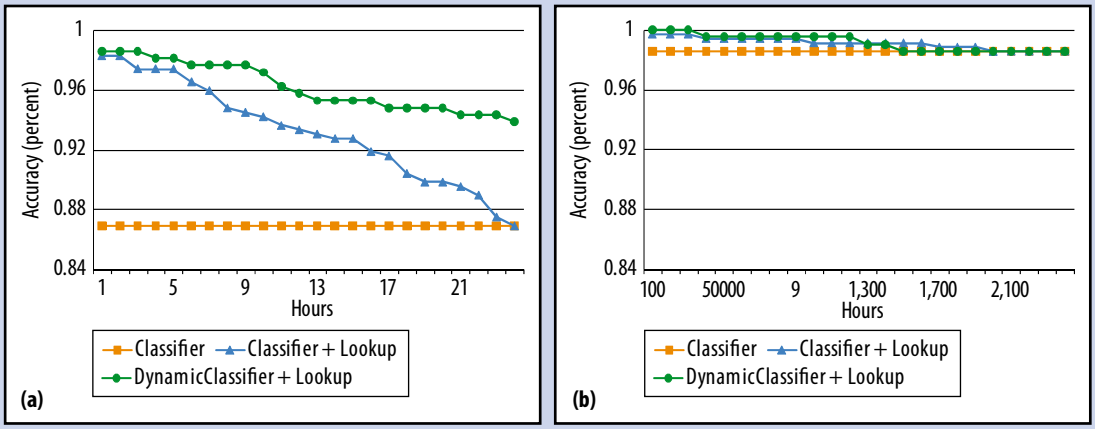


Figure 5. Impact of hybrid systems on fake website detection accuracy: (a) concocted sites and (b) spoof sites.

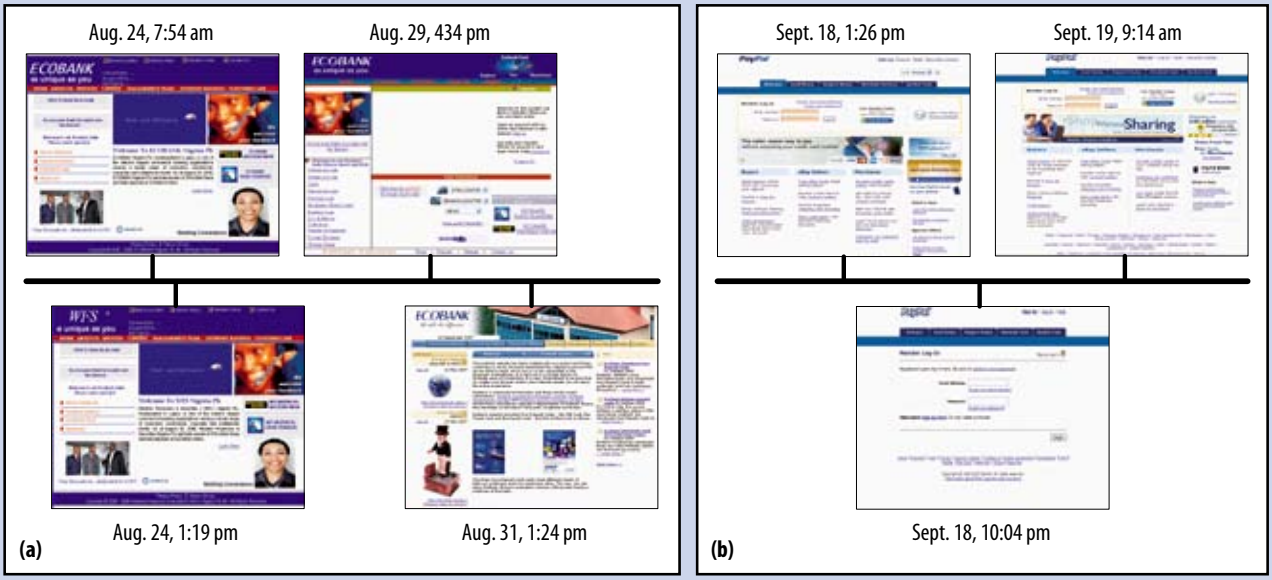


Figure 6. Fake website patterns over time: (a) concocted sites and (b) spoof sites.

Figure 5 shows the fake website detection percentage rates across the 24 values of n for the three systems. Both hybrid classifiers outperformed the standard classifier. As expected, using smaller time intervals between blacklist updates led to higher performance because the lookup mechanism was better able to identify recent fake sites. The use of a dynamic classifier further improved performance; however, the performance increase was more pronounced on the concocted sites. This is because concocted sites' patterns change over time, while spoof sites are more stagnant.

Figure 6a shows four concocted sites that appeared over a one-week period with an evolving template. The standard classifier could not identify these sites as fake due to new linkage, image, and text patterns not previously seen in the training data. The earlier two sites (Aug. 24) were almost identical, with the only difference being the company names. The third site (Aug. 29) had a somewhat different layout, while the fourth received a complete layout overhaul, though the body text was similar to its predecessors. Although the dynamic classifier also misclassified the first site, it identified the rest as fake (after update). In contrast, Figure 6b shows three spoofs of PayPal. Though slightly different, their page content was similar since they had to appear to be authentic PayPal sites. The standard and dynamic classifiers were able to identify the PayPal spoof site. Consequently, the dynamic classifier was more useful on the concocted website testbed.

A system containing a rich feature set and support vector machine classification model improves fake website detection performance. The proposed system outperformed comparison tools by a wide margin for concocted and spoof site detection. The results suggest that systems relying solely on lookup mechanisms, or classifier systems utilizing a small set of features, are ineffective in combating the myriad tactics employed by fraudsters. A good example is the Netcraft system, which only analyzes domain registration information; although it performed better than most, the system still could not detect between 15 and 30 percent of fake websites in our testbed.

Combining the classifier system with a lookup mechanism facilitated further performance enhancements, since the hybrid system benefited from periodic updates to the classification model.

In addition to providing improved detection accuracy in the short term, hybrid systems could offer an effective long-term solution. Further exploration of various forms of hybridization and different types of dynamic classification models is a potentially fruitful future endeavor. We intend to evaluate important usability issues related to fake website detection systems,[11,13] such as the tradeoffs associated with different interface design alternatives. In addition, we plan to assess various methods for decreasing AZProtect's average runtime without sacrificing the system's classification performance. C

References

1. C.E.H. Chua and J. Wareham, "Fighting Internet Auction Fraud: An Assessment and Proposal," *Computer*, Oct. 2004, pp. 31-37.
2. Z. Gyongyi and H. Garcia-Molina, "Spam: It's Not Just for Inboxes Anymore," *Computer*, Oct. 2005, pp. 28-34.
3. B. Wu and B.D. Davison, "Identifying Link Farm Spam Pages," *Proc. 14th Int'l Conf. World Wide Web*, ACM Press, 2005, pp. 820-829.
4. A. Ntoulas et al., "Detecting Spam Web Pages through Content Analysis," *Proc. 15th Int'l Conf. World Wide Web*, ACM Press, 2006, pp. 83-92.
5. A. Abbasi and H. Chen, "Detecting Fake Escrow Websites Using Rich Fraud Cues and Kernel-Based Methods," *Proc. Workshop on Information Technologies and Systems* (WITS), WITS, 2007, pp. 55-60.
6. N. Chou et al., "Client-Side Defense against Web-Based Identity Theft," *Proc. 11th Ann. Network and Distributed System Security Symposium*, Internet Society, 2004.
7. W. Liu et al., "An Antiphishing Strategy Based on Visual Similarity Assessment," *IEEE Internet Computing*, Mar./Apr. 2006, pp. 58-65.
8. I. MacInnes, D. Musgrave, and J. Laska, "Electronic Commerce Fraud: Towards an Understanding of the Phenomenon," *Proc. Hawaii International Conf. Systems Sciences* (HICSS), IEEE CS Press, 2005, pp. 181.1-181.11.
9. E. Levy, "Criminals Become Tech Savvy," *IEEE Security and Privacy*, Mar./Apr. 2004, pp. 65-68.
10. Y. Zhang et al., "Phinding Phish: Evaluating Anti-Phishing Tools," *Proc. 14th Ann. Network and Distributed System Security Symposium* (NDSS), Internet Society, 2007.
11. L. Li and M. Helenius, "Usability Evaluation of Anti-Phishing Toolbars," *J. Computer Virology*, vol. 3, no. 2, 2007, pp. 163-184.
12. P. Hariharan, F. Asgharpour, and L.J. Camp, "NetTrust—Recommendation System for Embedding Trust in a Virtual Realm," *Proc. ACM Conf. Recommender Systems*, ACM Press, 2007.
13. M. Wu, R.C. Miller, and S.L. Garfinkel, "Do Security Toolbars Actually Prevent Phishing Attacks?," *Proc. Conf. Human Factors in Computing Systems*, ACM Press, 2006, pp. 601-610.
14. M. Diligenti et al., "Focused Crawling Using Context Graphs," *Proc. 26th Conf. Very Large Databases*, Morgan Kaufmann, 2000, pp. 527-534.

Ahmed Abbasi *is an assistant professor of management information systems at the University of Wisconsin-Milwaukee. His research interests include online trust and security, computer-mediated communication, and text mining. Abbasi received a PhD in information systems from the University of Arizona. He is an IEEE member. Contact him at abbasi@uwm.edu.*

Hsinchun Chen *is the McClelland Professor of management information systems and director of the Artificial Intelligence Lab and Hoffman E-Commerce Lab in the Department of Management Information Systems at the University of Arizona. His research interests include digital libraries, knowledge discovery, and cybersecurity. Chen received a PhD in information systems from New York University. He is an IEEE and AAAS fellow. Contact him at hchen@eller.arizona.edu.*