Inf Technol Manag (2009) 10:83–101
DOI 10.1007/s10799-009-0059-0

A comparison of fraud cues and classification methods for fake escrow website detection

Ahmed Abbasi · Hsinchun Chen

Published online: 21 July 2009

Abstract The ability to automatically detect fraudulent escrow websites is important in order to alleviate online auction fraud. Despite research on related topics, such as web spam and spoof site detection, fake escrow website categorization has received little attention. The authentic appearance of fake escrow websites makes it difficult for Internet users to differentiate legitimate sites from phonies; making systems for detecting such websites an important endeavor. In this study we evaluated the effectiveness of various features and techniques for detecting fake escrow websites. Our analysis included a rich set of fraud cues extracted from web page text, image, and link information. We also compared several machine learning algorithms, including support vector machines, neural networks, decision trees, naïve bayes, and principal component analysis. Experiments were conducted to assess the proposed fraud cues and techniques on a test bed encompassing nearly 90,000 web pages derived from 410 legitimate and fake escrow websites. The combination of an extended feature set and a support vector machines ensemble classifier enabled accuracies over 90 and 96% for page and site level classification, respectively, when differentiating fake pages from real ones. Deeper analysis revealed that an extended set of fraud cues is necessary due to the broad spectrum of tactics employed by fraudsters. The study confirms the feasibility of using automated methods for detecting fake escrow websites. The results may also be useful for informing existing online escrow fraud resources and communities of practice about the plethora of fraud cues pervasive in fake websites.

Keywords Online escrow services · Internet fraud · Website classification · Fraud cues · Machine learning

1 Introduction

Electronic markets have seen unprecedented growth in recent years. Online auctions are a major category of electronic markets prone to Internet fraud stemming from asymmetric information [1]. The lack of physical contact and prior interaction makes such places susceptible to opportunistic member behavior [2]. While reputation systems attempt to alleviate some of the problems with electronic markets, these systems suffer from two problems: easy identity changes and reputation manipulation. Easy identity changes refer to the fact that online traders can create new identities, thereby refreshing their reputation [3]. Reputation manipulation enables individuals to inflate their own reputations, using multiple identities, or sabotage competitors' ranks [4]. Consequently, fraud and deception are highly prevalent in electronic markets, particularly online auctions, which account for 50% of internet fraud [5]. Approximately 40% of buyers in online auctions have reportedly had problems [6].

In light of the troubles associated with electronic marketplaces [7], many believe the solution is online escrow services. Online escrow services (OES) are intended to serve as trusted third parties protecting against Internet fraud [1]. OES play an integral role in the development of

A. Abbasi (✉)
Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA
e-mail: abbasi@uwm.edu

H. Chen
Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721, USA
e-mail: hchen@eller.arizona.edu

ÔÔinstitution-based trustÕÕ in online marketplaces [2]. Risk-averse online traders are especially likely to adopt OES [8]. The increased use of OES has inevitably brought about the rise of escrow fraud. Escrow fraud is a variant of the popular ÔÔfailure-to-shipÕÕ fraud; the seller creates a fake OES service coupled with an associated website, and disappears after collecting the buyerÕs money [5]. Such forms of internet fraud, involving fake escrow websites, are becoming increasingly prevalent. Online databases such as the Artists-Against-419 contain thousands of entries for fraudulent websites [9], with hundreds added daily. These fraudulent OES sites are often very professional looking and difÞcult to identify as fake by unsuspecting online traders [10, 11]. While there has been a recent effort to develop tools to combat spoof sites such as those used in phishing attacks [12], fake OES sites have received little attention despite their pervasiveness. There is therefore a need for automated categorization techniques capable of identifying fraudulent escrow websites. Fraudulent escrow site identiÞcation entails the use of cues from various information types [13, 14] including website content (i.e., body text), website design (i.e., HTML tags), URLs and anchor text, images, and website structure. The effective representation of the necessary information types introduces complexities which must be taken into consideration when developing an adequate automated approach to fake OES identiÞcation.

In this study, we evaluate the viability of automatic fake escrow website detection in order to improve online trust by thwarting web-based escrow service fraud. Our analysis evaluates a rich set of features (i.e., ÔÔfraud cuesÕÕ) for identifying fake escrow websites. These include stylistic features extracted from body, HTML, and URL and anchor text; image pixel features for identifying duplicate pictures, banners, and icons across fake escrow websites; and website structure and linkage based features. We also compare several machine learning algorithms that have been successfully applied to related document classiÞcation problems. Results from this research serve two important purposes. Firstly, the study assesses the feasibility of mechanisms for automated identiÞcation of fake websites that can help reduce the negative impact of online auction fraud stemming from fraudulent OES. Secondly, evaluating different features and techniques for fake OES categorization can provide insights into fraud patterns that can be used to help educate Internet users that use such trusted third party sites as a source of institution based trust.

2 Related work

According to Fraud.org [15], 15.6 million people (41% of online auction participants in the U.S.) have encountered Internet fraud [8]. The number of Internet complaints has increased in recent years, with the majority pertaining to online auctions [7]. Therefore, online trust instruments are highly important. Online feedback mechanisms and escrow services represent two vital sources for institution-based trust in electronic markets [2]. However, online feedback mechanisms such as reputation systems suffer from easy identity changes and reputation manipulation. Easy identity changes allow community members to build up a reputation, use it to deceive unsuspecting members, and start over under a new identity [16]. Reputation manipulation involves using additional (fake) identities to inßate ones reputation or threatening to post negative feedback against other traders [4, 16].

These problems have led to the increased popularity of trusted third parties such as OES [1, 8]. The use of OES involves a tradeoff between price premiums and enhanced transaction security. Online traders use escrow services as an insurance mechanism against Internet fraud. Therefore, the perceived effectiveness of OES plays a critical role in the amount of online trust [2]. Pavlou and Gefen [2] found that the seeming effectiveness of OES had a signiÞcant impact on buyersÕ trust in the community of sellers. While OES are intended to offer security against the lack of trader identity trust, ironically they themselves fall prey to similar concerns. It is often difÞcult for online traders to differentiate legitimate escrow sites from fraudulent ones, making them susceptible to escrow fraud; the unsuspecting use of a fake escrow website posing as a legitimate one [5]. In addition to monetary losses, OES fraud has social and psychological implications. Fraud in marketplaces results in a psychological contract violation as perceived by the defrauded trader [17]. Others refer to such infringements of consumer trust as breaches of the social contract [18]. This can result in a reluctance to engage in future online transactions, impacting the sustainability of electronic markets. Methods for reducing escrow fraud can mitigate these negative outcomes. There is hence a need for techniques capable of alleviating escrow fraud.

2.1 Escrow fraud

Many online resources have emerged in recent years for combating OES fraud. Communities of practice describing fraud victim experiences and best practices for online trading provide an invaluable knowledge base for online traders [5]. Online databases of known OES fraud sites feature URLs for these sites along with commonly used fraudulent website templates [9]. However, often the entries in these databases occur at the expense of fraud victims who report these sites after they have been scammed (i.e., these databases are reactive by nature). Furthermore, many online buyers and sellers lack awareness about Internet fraud and are even less cognizant of resources available to prevent it [8]. Figure 1 shows examples of fake escrow websites. Their professional appearance often makes it difficult for online buyers to identify such fraudulent websites [11], resulting in a need for methods capable of automatically identifying fake escrow websites. Effective automatic identification techniques could potentially be utilized in a pre-emptive fashion to alleviate OES fraud. The development of an automated approach requires considering the relevant features and techniques capable of extracting and utilizing fraud cues inherent in fake OES websites. These crucial elements are discussed below.

2.2 Escrow web page fraud cues

We're not aware of any prior research on automatic categorization of fraudulent escrow websites. However, there has been work on a related emerging website categorization problem: web spam categorization. Web Spam is the "injection of artificially created web pages into the web in order to influence the results from search engines, to drive traffic to certain pages for fun or profit" [19]. There are many commonalities between the features used for web spam categorization and those likely necessary for fake escrow website identification [20]. Analogous to fake escrow websites, web spam typically uses automatic content generation techniques to mass produce fake web pages [14]. Automated generation methods are employed due to the quick turnover of such content, with hundreds of fake escrow sites popping up monthly to replace ones already identified by online traders [9, 11]. Such use of machine generated pages results in many content similarities which may be discernable using statistical analysis of website and page level content [21]. Fake websites often duplicate content from previous fraudulent sites, thereby looking "templatic" [20]. Figure 2 shows 70 examples of fake escrow website templates. Hundreds of templates exist, with new ones generated daily.

Various Internet fraud watch organizations and prior web spam research have identified sets of features or fraud cues which may be applicable to fake escrow websites [5, 13, 14, 21]. It is important to note that while anecdotal evidence has been provided regarding various potential fraud cues inherent in fake OES websites, no formal evaluation has been conducted to assess the effectiveness of these cues for identification of fraudulent escrow sites. These include feature categories pertaining to the following website segments:

- Website content (i.e., body text)
- Website design (i.e., HTML)
- URL and anchor text
- Images
- Website linkage and structure

Figure 3 illustrates how fraud cues inherent in these website segments occur in fake OES sites. Relevant features include repetition of stylistic patterns in the page content (i.e., body text and URLs) and design (i.e., HTML) as well as duplication of images and icons.

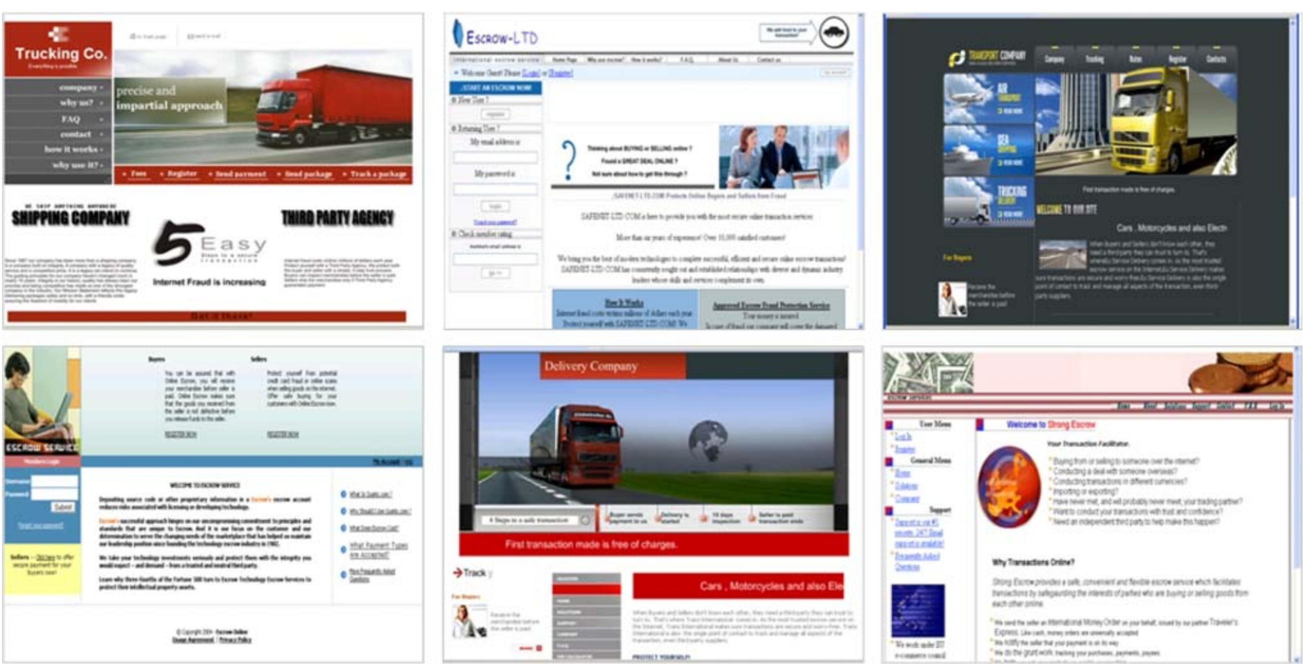

Fig. 1 Fake escrow web page examples

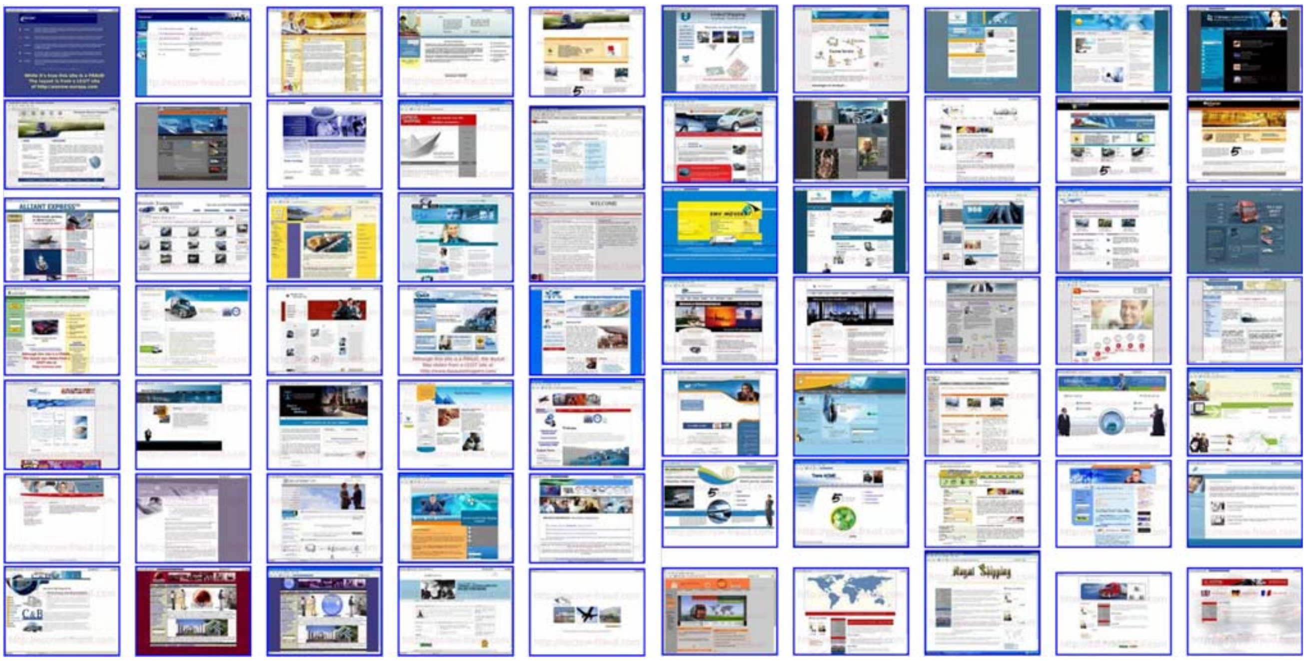

Fig. 2 Examples of fake escrow website templates

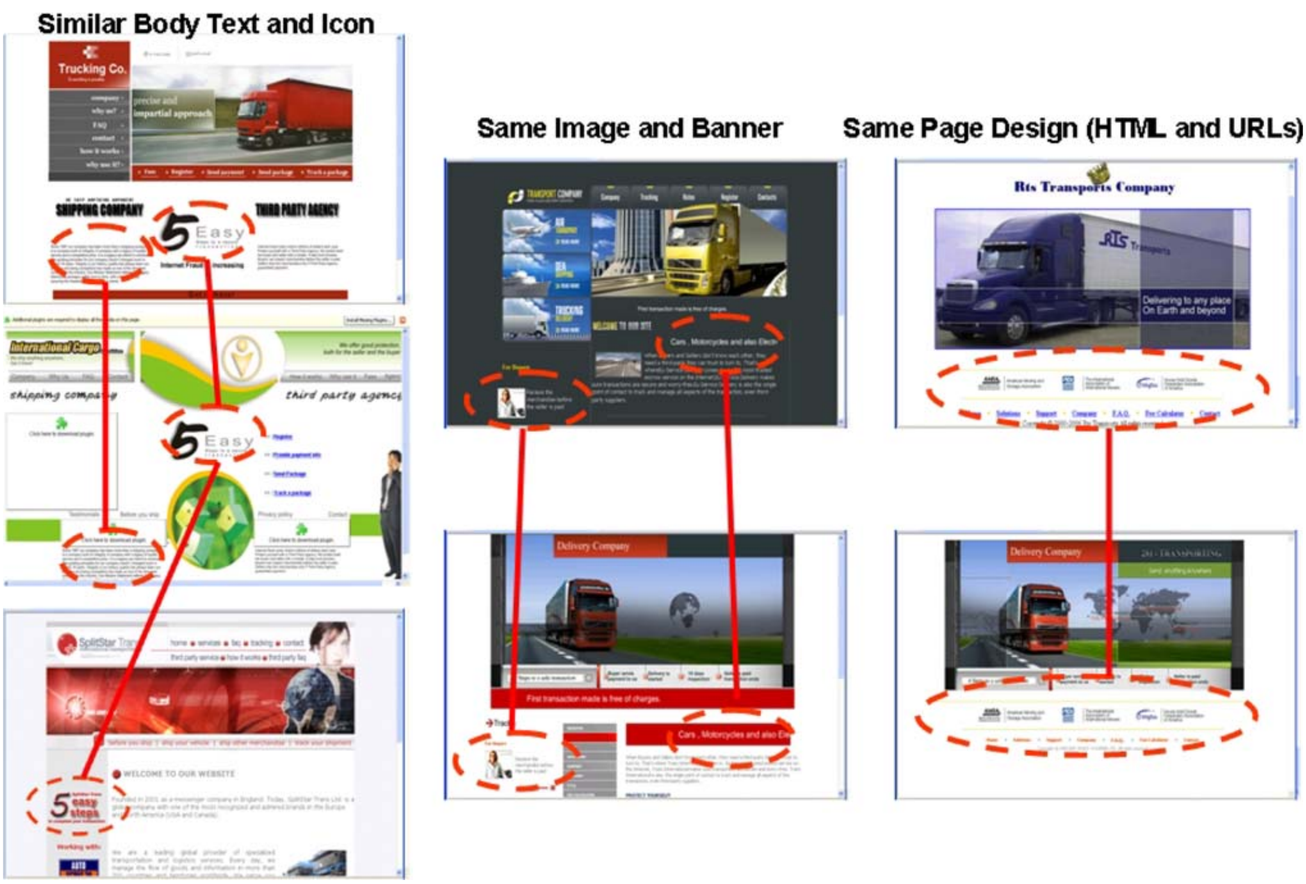

Fig. 3 Examples of fraud cues in fake escrow web pages

2.2.1 Body text style features

Fake OES sites occasionally contain misspellings and grammatical mistakes [22]. Such idiosyncrasies have been very informative in other stylistic categorization problems [23]. Additional text content features and style markers may also be useful. Ntoulas et al. [19] used various lexical measures including words per page, words per title, average word length, and word n-gram frequencies for categorizing web spam.

2.2.2 HTML features

HTML/source elements such as font types, sizes and colors have recently been used for stylometric categorization of authors [24]. The use of features taken from HTML source code, such as tag n-grams, is also useful for identifying web page design style similarities [14]. For instance, the websites shown in the third column in Fig. 2 have similar design which could potentially be detected using HTML source tag features.

2.2.3 URL and anchor text features

Certain text appearing in site URLs and anchor text can represent powerful escrow fraud cues [25]. URLs with dashes, digits, and more characters are often used by fake websites [21]. All websites engaging in online transactions should be secure (i.e., using ÔÔhttps); therefore fake OES sites may have URLs with ÔÔhttpÕÕ instead of ÔÔhttps.ÕÕ The number of slashes ÔÔ/ÕÕ (page levels) is another potential indicator since sites with deeper home pages are often suspicious. Website URL sufÞxes (e.g., ÔÔ.orgÕÕ ÔÔ.usÕÕ, ÔÔ.bizÕÕ) can provide fraud cues. For instance, escrow websites ending with ÔÔ.orgÕÕ are likely to be phony since this extension is typically used by non-proÞt organizations. Ntoulas et al. [19] randomly sampled over 105 million pages from the web and observed that 70% of ÔÔ.bizÕÕ and 35% of ÔÔ.usÕÕ pages selected were fake.

2.2.4 Image features

DifÞculties in indexing make multimedia web-content difÞcult to accurately collect and analyze. Consequently, many previous web mining and categorization studies have ignored multimedia content altogether [26]. While the use of image features may not reveal in depth patterns and tendencies, even simplistic image feature representations and categorization techniques can facilitate the identiÞcation of duplicate images [12]. This could be useful, given the pervasive nature of replicated photos, banners, and icons in escrow fraud sites, as depicted in the Þrst and second columns of Fig. 3.

2.2.5 Linkage features

Link information, coupled with text content, can dramatically improve web page categorization. In/back links and out links have been used effectively for categorizing web pages based on similar topics [26]. The context graphs method derives back links for each URL and uses these to construct a multilayer graph that provides a structural signature for different website types [27].

Most previous studies on web spam categorization have only adopted one or two of the aforementioned feature groups [19]. One important difference between fake escrow sites and web spam is that web spam is intended to deceive search engines [20] while fake escrow websites are designed to deceive online traders. In addition to site content (i.e., body text and URLs), fake OES sites must consider site design elements (e.g., HTML and images/banners) in order to aesthetically appeal to online buyers [11]. It is therefore unclear whether a single feature category (e.g., URL tokens) will be sufÞcient for automatically identifying fake escrow sites. The alternative is to incorporate multiple feature categories, however the use of rich feature sets comprised of text, link, and image features introduces representational complexities for the potential classiÞcation techniques employed.

2.3 Escrow web page categorization techniques

IdentiÞcation of fake escrow sites entails consideration of the various textual style and image elements described in the previous section. Several machine learning techniques have been used considerably for text and image categorization, including style classiÞcation. Methods such as support vector machines (SVM), neural networks, decision trees, and principal component analysis (PCA) have all been shown to be useful in related classiÞcation tasks [28, 29]. SVM is a popular classiÞcation technique that has been applied to topical categorization of web pages [30]. Grounded in Statistical Learning Theory [31], SVMÕs ability to learn from noisy data and its propensity to avoid over Þtting make it highly suitable for web mining applications. SVMÕs effectiveness for categorization of style makes it particularly suitable for fake OES detection [32, 33].

Principal component analysis (PCA) is another technique that has been used frequently for text and image processing. PCAÕs use of dimensionality reduction allows it to uncover important variation tendencies which are effective for text style and image categorization [28, 34Ð36]. Other relevant classiÞcation algorithms include the C4.5 decision tree [34, 37] and Winnow [38]. C4.5 uses the information gain heuristic to select attributes which provide the highest entropy reduction on the training data [39].

These features are used to build a decision tree model. Winnow is a variant of the multilayer perceptron neural network, that uses a weight update mechanism capable of handling large quantities of irrelevant or noisy attributes [38]. It has worked well for various text classiÞcation problems [40].

Based on Bayes Theorem [41], Naïve Bayes (NB) is a fairly simple probabilistic classiÞcation algorithm that uses strong independence assumptions regarding various features [42]. It assumes that the presence of any feature is entirely independent of the presence of any other feature(s), allowing it to build classiÞcation models in an efÞcient manner. However, this efÞciency often arises at the expense of classiÞcation performance.

SVM has outperformed other machine learning techniques, including decision trees and neural networks, in head-to-head comparisons on topic and style classiÞcation of online texts [29, 34]. It is therefore likely to perform well for fake escrow website detection as well. Nevertheless, it is difÞcult to surmise which classiÞcation algorithm would be best suited for identifying fake OES sites without performing a detailed comparison of the various methods.

2.4 Meta-learning strategies for escrow web page and website categorization

Based on the discussion presented in Sect. 2.2, it is likely that fake OES website detection requires the use of a rich heterogeneous set of fraud cues. A website contains many pages, and a page can contain many images, along with HTML, body text, URLs and anchor text, and site structure. The heterogeneous nature of fraud cues, as well as the many web pages encompassing a single website, presents an ideal opportunity for employing meta-learning strategies such as ensemble classiÞers and stacking.

2.4.1 Ensemble classifiers for page level classification

The use of multiple classiÞers (called an ÔÔensembleÕÕ) can allow complex information to be decomposed across a series of classiÞers [43]. Ensemble classiÞers are multiple classiÞers built using different techniques, training instances, or feature subsets [43]. Feature ensembles can be used to build multiple classiÞers with each classiÞer using a different feature category. Such a feature subset classiÞer approach has been effective for analysis of style and patterns. Stamatatos and Widmer [33] used an SVM ensemble for music performer recognition. They used multiple SVM classiÞers each trained using different feature subsets with each classiÞer acting as an ÔÔexpertÕÕ on its subset of features. Cherkauer [44] used a neural network ensemble for imagery analysis since the image recognition feature set was comprised of attributes with different properties (e.g., image pixel colors, object edge values, etc.). The imagery analysis ensemble involved the use of 32 neural networks trained on 8 different feature subsets.

Figure 4 shows an illustration of the feature based ensemble applied to an example escrow feature matrix. Each feature category uses a separate classiÞer, which allows handling of image features. The ensemble shown in Fig. 4 contains Þve classiÞers for body text, HTML, URL, link, and image features, respectively. Feature classiÞer ensembles have their strengths and weaknesses. They allow each classiÞer to become an ÔÔexpertÕÕ on subset of features

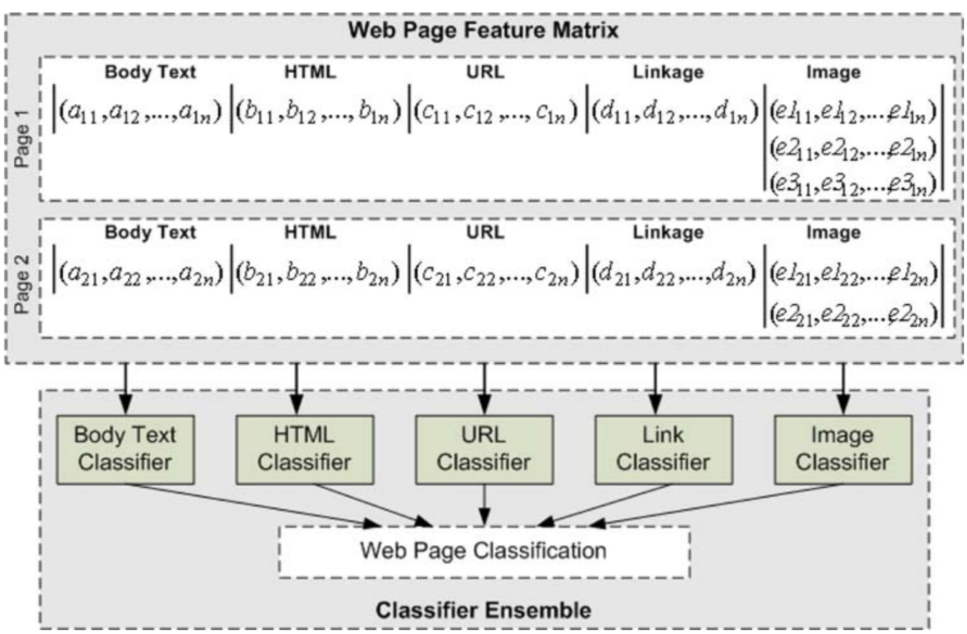

Fig. 4 Example classiÞer ensemble for web page feature matrix

[33]. However, feature set segmentation results in the loss of potentially informative feature interactions. For instance, information garnered by combining linkage and body text information would be lost using a feature ensemble. Hence, while ensemble classiÞers may facilitate the combination of rich sets of heterogeneous fraud cues, it is unclear whether they can improve OES website classiÞcation performance over individual classiÞers.

2.4.2 *Stacked classifiers for site level classification*

Prior research on website categorization has generally focused on page level classiÞcation. For instance, focused crawlers attempt to accurately categorize web pages based on their topical relevance, where relevant pages are collected [27]. Similarly, research on web spam categorization has also emphasized page level classiÞcation [13, 14, 19, 20, 45]. This is because these two research areas are related to the construction of improved search engines, which tend to operate at the web page level. In contrast, the end objective of fake OES detection is to classify websites as legitimate or fraudulent. Prior research has shown that treating an entire website as a single feature vector results in poor classiÞcation performance [46]. One effective approach has been to treat each website as a set of page feature vectors [47]. Hence, the site level classiÞcation task can incorporate information from underlying page level classiÞers [48]. Using this approach, a simplistic site level classiÞer could simply aggregate the page level feature vectorsÕ classiÞcation results and classify a website as real or fake if the percentage of its underlying real/fake page classiÞcations are above a certain threshold. Alternately, site level OES categorization could employ a stack, where the underlying page level classiÞcation results are used as input features into the top level classiÞer. Such a meta-learning approach is called stacking. It can be more effective than simple scoring/voting approaches since learning can occur within the top level classiÞer as well [49].

2.5 Site level classiÞcation parameters

Site level classiÞcation involves consideration of factors stemming from the quantity of web pages, and structure of these pages, within a website. Prior research has found that ÔÔwebsite pruning,ÕÕ i.e., selecting a subset of its pages, is essential to allow classiÞcation to occur in a computationally efÞcient manner [46, 47]. Pruning requires two parameters: the number of web pages to utilize, and the website region from which these pages should be sampled [47]. These parameters have important implications for fake OES website detection. Figure 5 represents a sample of web pages taken from two websites: a legitimate (left) and fake (right) OES. For each website, six pages from different levels are depicted, along with a description of each page. The legitimate website has hundreds of web pages, spanning 4-5 levels (a few level 0-2 pages are shown in Fig. 5). In contrast, the fake OES only has a dozen pages spanning two levels (many of which are shown in Fig. 5). It does not have any of the ÔÔdeeperÕÕ pages such as frequently asked questions and membership information pages. Selecting a few top level web pages per website may not effectively capture this important distinction. Similarly, the quantity of web pages employed per website requires balancing the tradeoffs between accuracy and computation time. For instance, in other website classiÞcation tasks, using the top 100Ð120 pages per site has been shown to be sufÞcient to accurately represent a websiteÕs content [47].

3 Research design

In this section we outline key research gaps as well as our research questions based on those gaps. We then describe

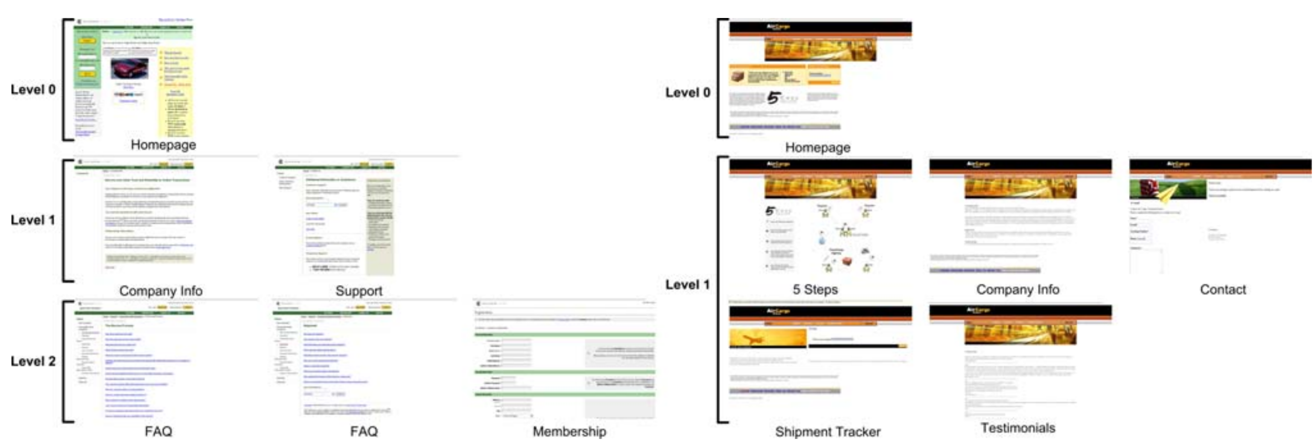

Fig. 5 Example showing two websitesÕ web pages from different levels

the features and techniques used for identifying fake OES websites.

3.1 Research gaps and questions

While there has been considerable work on detecting web spam to improve search engine performance, research on automated detection of fraudulent websites remains scarce. There is a need for resources that can facilitate identiÞcation of fake websites given the difÞculties people have in determining whether a particular site is legitimate [10]. However, given the lack of prior work, it is unclear what features (i.e., fraud cues) and techniques will be effective for automatic fake OES identiÞcation. Furthermore, fake website detection can be performed at the page or site level. Page level focuses on categorizing individual web pages, while site level is concerned with categorizing entire websites. While site level classiÞcation is likely more accurate (since there are more web pages available for evaluation), page level classiÞcation is faster (since less content needs to be evaluated). Although the objective is always site level classiÞcation, websites are browsed one page at a time. Hence, effective page level accuracy could expedite the discovery of fake websites over site level classiÞcation. Therefore, the interplay between page and site level classiÞcation has important implications for fake website detection toolbars, which must balance accuracy with computational efÞciency, while presenting end users with one page at a time. In order to attend to these research gaps, we seek to address the following research questions:

RQ1. Which feature categories are best at differentiating fake escrow web pages from real ones?

€ (Body text/HTML/URLs/Images/Linkage/All)

RQ2. Which technique is better suited at differentiating fake escrow web pages from real ones?

€ (SVM/Winnow/C4.5/NB/PCA classiÞers)

RQ3. Can the use of feature ensemble classiÞers improve page level performance over individual classiÞers?

RQ4. Which techniques are capable of providing the best site level classiÞcation performance?

€ (Page scores/Stacking)

RQ5. What impact will the number of pages per website and sampling regions have on site level classiÞcation performance (in terms of accuracy and computation time)?

3.2 Comparison fraud cues for fake OES website detection

The feature set utilized is comprised of fraud cues from the Þve categories described in the literature review: body text, HTML, URL, image, and linkage. These are summarized in Table 1. Body text features used include lexical measures incorporated in previous web spam studies [19] and style markers described in prior style categorization research [24, 28, 29, 32]. HTML tag n-grams were used for

Table 1 Fake OES website identiÞcation feature set

Feature group	Category	Quantity	Description/Examples
Body text	Letter N-grams	<18,278	Count of letters (e.g., a, at, ath)
	Digit N-grams	<1,110	Count of digits (e.g., 1, 12, 123)
	Word length dist.	20	Frequency distribution of 1Ð20 letter words
	Special characters	21	Ocurrences f special char. (e.g., @#$%^)
	Function words	300	Frequency of function words (e.g., of, for, to)
	Punctuation	8	Occurrence of punctuation marks (e.g., !;:,.?)
	POS tag N-grams	Varies	Part-of-speech tags (e.g., NNP, NNP JJ)
	Bag-of-word N-grams	Varies	e.g., ÔÔtrustedÕÕ, ÔÔthird partyÕÕ, ÔÔtrusted thirdÕÕ
	Misspelled words	<5,513	e.g., ÔÔbeleiveÕÕ, ÔÔthougthÕÕ
HTML	HTML tag N-grams	Varies	e.g.<HTML> , <HTML> <BODY>
URL	Character N-grams	Varies	e.g., a, at, ath,/, _, :
	Token N-grams	Varies	e.g. ÔÔspeditionÕÕ, ÔÔescrowÕÕ, ÔÔtrustÕÕ, ÔÔonlin
Image	Pixel colors	10,000	Frequency bins for pixel color ranges
	Image structure	40	Image extensions, heights, widths, Þle sizes
Link/structure	Site and page linkage	10	Site and page level relative/absolute in/out links
	Page structure	31	Page level, in/out link levels distribution

representing page design style [14]. URL features included character and token level n-grams [19]. The image features comprised of frequencies for pixel colors [50]. Link and structure features included page and site level relative and absolute in/out links for each web page [45] along with the page level frequency distribution for all in/out link pages. Site level in-links were derived from the Google search engine, as done in prior research [27]. All n-gram features require feature selection, commonly using the information gain heuristic to govern selection [23]. Therefore, the quantities for these features are unknown apriori.

3.3 Comparison classification techniques for fake OES website detection

We incorporated the SVM, PCA, NB, C4.5 decision tree, and Winnow algorithms described in Sect. 2.3. These five algorithms were adopted since they have been used heavily for text, style, and image categorization [28, 29, 40], all of which are relevant to fake OES website identification.

4 Experimental design

We conducted experiments to evaluate the proposed extended feature set, page level classification techniques, feature-based ensemble, and site level classification methods. This section includes a description of the experimental design, including the website test bed, experimental setup, and evaluation metrics.

4.1 Test bed

We collected 350 fake OES and 60 real escrow websites over a 3 month period between 12/2006–2/2007. The test bed was skewed because the number of fake OES sites significantly exceeds the number of legitimate ones. The fake OES website URLs were taken from two online databases that post the HTTP addresses for verified fraudulent escrow sites: Escrow Fraud Prevention (http://escrow-fraud.com) and The Artists Against 4-1-9 (http://wiki.aa419.org). These sites allow defrauded traders to post URLs for fake escrow sites. The site owners require all complaints to be accompanied with evidence that fraud occurred. Evidence includes payment receipts to the escrow site, transcripts of email exchanges, copies of reports filed with the appropriate authorities, etc. Such verification is important to ensure that the sites added to the databases are indeed fraudulent.

Since fake OES sites are often shut down or abandoned after they have been used, these sites typically have a short life span (often less than a few days). In order to effectively collect these websites, we developed a web spider program that monitored the online databases and collected newly posted URLs daily. This was done in order to retrieve the content from these fake OES sites before they disappeared. The spider collected all files, including static and dynamic indexable files and images/icons/buttons. Table 2 below shows the summary statistics for our test bed.

4.2 Experimental setup

We ran 50 bootstrap instances for each experiment condition, in which 100 OES websites (30 real and 70 fake) were randomly selected for training, while another 100 OES websites (30 real and 70 fake) were used for testing, in each bootstrap instance. All web pages from these 200 sites were used for training and testing, respectively. The websites (or any of their web pages) appearing in the training set for a particular run did not appear in the test set for that bootstrap run. Such a 30–70 split between real and fake websites was used since it resulted in an approximately proportionate number of real and fake web pages, ensuring a more balanced training and testing set for the classifiers. On average, there were approximately 10,000 pages (5,000 real and 5,000 fake) in the training and testing sets each, per bootstrap run. We used the following performance metrics and receiver operating characteristic curves, each of which has been utilized in prior research [14, 19, 45]:

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}}$$

$$\text{Class level recall} = \frac{\text{Number of correctly classified class instances}}{\text{Total number of instances in class}}$$

$$\text{Class level precision} = \frac{\text{Number of correctly classified class instances}}{\text{Total number of instances classified as belonging to class}}$$

$$\text{Class level } F\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Table 2 Escrow website test bed

Category	Number of sites	Number of pages	Number of images	Average pages per site	Average images per site
Real OES sites	60	19,812	6,653	330.20	110.88
Fake OES sites	350	69,684	29,764	199.10	85.04
Total	410	89,496	36,417	218.28	88.82

Table 3 Average number of features per bootstrap run and examples of key features

Feature category	# of Features	Example features	Description
Body text	10,086	Word bigram “FREE HOSTING”	Fake OES are often hosted on websites that provide free hosting
		Word trigram “POWERED BY PHPBB”	Fake OES often use open source software packages to generate content
		Word bigram “Member FDIC”	Legitimate websites usually contain information about their memberships with various government organizations, such as the FDIC, BBB, etc.
		Word unigram “español”	Many legitimate websites have multiple versions of their site in different languages
HTML	2,465	Links to Careers/Jobs web page	Legitimate websites are more likely to place job postings on their website
		Image Preloading	This Javascript code, which is used to preload images to decrease page loading times, rarely appears in fake websites
URL	2,469	URL token “HTTPS”	Fake websites rarely use the secure sockets layer protocol
Image	10,040	Recurrence of certain images	Fake OES tend to reuse images of consumers, employees, and company assets
Link	41	Number of inlinks	Legitimate websites tend to have more websites point at them. Exceptions are some fake websites that utilize link farms
		Number of outlinks	Fake OES attempting to spoof legitimate escrow websites are generally partial replicas with only a handful of surface level pages. As a result, they tend to contain fewer web pages (and less linkage)
Total (All features)	25,101		

5 Evaluation: page level classification

We initially conducted page level classification experiments to assess the effectiveness of various feature sets and classification techniques. These experiments are discussed below.

5.1 Experiment 1: comparison of features and techniques for page level classification

The experimental design included six feature sets (body text, HTML, URL, image, link, and all) and five techniques (SVM, Winnow, C4.5, NB, and PCA). This resulted in 30 total experimental conditions. SVM was run using a linear kernel. PCA was run using the Kaiser-Guttman stopping rule which selects all eigenvectors with an eigenvalue greater than one [51]. The training and testing instances were projected to an n-dimensional space (where n is the number of selected eigenvectors). Each test instance was assigned to the class with the lower average distance between its training points and the test point (across the n-dimensions). Such an approach has worked well in prior text categorization and analysis studies [28].

Consistent with previous research, information gain was used to select all n-gram quantities [23]. Information gain was performed on the 100 training websites’ pages for each of the 50 bootstrap runs. The average number of features used for each category is shown below in Table 3, along with a few examples of the key features for each category. The image and link feature sets were static since they did not include attributes such as n-grams. The “All” features used the total features for the five categories. Since individual classifiers are not capable of incorporating multiple images per web page into the feature matrix, a single feature vector comprised of the average image features (across all images in that web page) was used in the “All” feature set, for each web page row in the feature matrix.

Table 4 shows the page level classification experimental results for the various feature and technique combinations. Consistent with prior website categorization research, the

Table 4 Overall page-level classiÞcation accuracy (%) across 50 bootstrap runs

Technique	Feature set					
	Body text	HTML	URL	Image	Link	All
PCA	52.36	50.94	49.04	62.34	65.56	66.04
NB	60.34	61.04	62.18	51.04	70.46	71.22
C4.5	75.60	75.56	75.64	69.98	71.90	77.90
Winnow	76.78	78.88	82.16	60.12	72.44	85.28
SVM	79.98	84.74	81.84	70.38	73.44	88.38

Bold values indicate techniques that attained the best results on that feature set

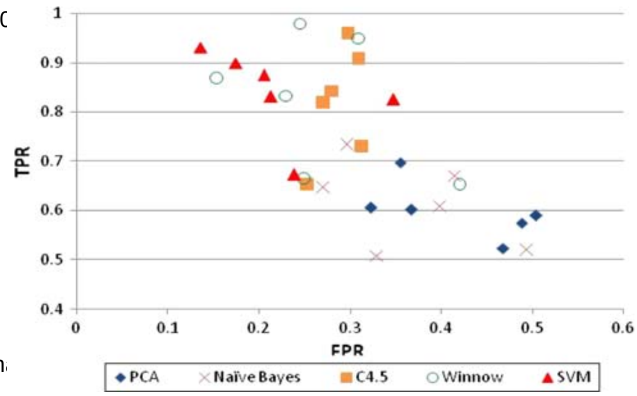

Fig. 6 Receiver operating characteristic plots for different techniques

body text, HTML, and URL features all performed well with accuracies near 80%. Although image features only had up to 70% accuracy, this is also quite promising. Our image feature set was fairly simplistic, only capable of matching duplicate or fairly identical images. The image performance suggests that image duplication is pervasive in fake OES sites. Link features were not as effective as we had anticipated. Further analysis revealed that many smaller legitimate OES sites have link patterns similar to those of fake OES websites (i.e., less external in-links). The use of all features (ÔÔAllÕÕ) outperformed individual feature categories, typically improving accuracy by at least 3Ð5%. This supports the notion that fraud cues in fake OES sites occur across feature types.

With respect to classiÞcation algorithms, SVM had the best performance, outperforming comparison algorithms on most feature sets. The best results were obtained using SVM in combination with all features, resulting in approximately 90% accuracy. However, Winnow had slightly better performance than SVM when using URL features. Winnow and C4.5 were competitive on many other feature sets as well. PCA and NB performed rather poorly, generally attaining 10Ð20% lower accuracy values than the other three methods.

Figure 6 shows the receiver operating characteristic (ROC) plots, showing the true positive and false positive rates for different techniques. Here, true positives refer to correctly classiÞed legitimate web pages. Each point signiÞes a particular feature/technique combination. Values closer to the top left corner signify better results, since they denote high true positive rates and low false positive rates. Looking at the Þgure, we can see that SVM, Winnow, and C4.5 had the best performance. Overall, the results suggest that SVM provides the most desirable combination of true and false positives.

Figure 7 shows the page level precision and recall for various classiÞcation methods on real and fake OES web sites, across feature sets. Looking at the results on real OES websites, we can see that the SVMÕs recall performance was similar to that of C4.5 (illustrated in 7a). Its precision on the fake OES sites was also comparable to other techniques. Hence, SVM did not do much better than C4.5 for classifying real OES websites, with best overall results of approximately 90% attained when using all features. However, SVM outperformed all four comparison algorithms on all feature sets except images, on the fake OES websites in terms of recall (Fig. 7b). When using all features, SVM was able to detect nearly 90% of fake OES pages. This was marginally better than Winnow and nearly 10Ð20% better than results attained using other comparison methods, including C4.5. Hence, SVM provided the best balance for real and fake OES categorization. While algorithms such as C4.5 and Winnow were competitive on one or the other, SVM had the least overall misclassiÞcations (sum of false positives and false negatives).

5.1.1 *Analysis of feature performance (RQ1)*

Pair wise t-tests were run on the 50 bootstrap instances to see which feature set had the best performance (Table 5). The tests were run on classiÞcation accuracy and class level F-measures. During the comparison, classiÞcation techniques were controlled. Given the large number of test conditions, a Bonferroni correction was performed in order to avoid spurious positive results. Only individual p-values less than 0.0033 were considered signiÞcant at alpha = 0.05. The use of all features signiÞcantly outperformed all Þve individual feature sets in terms of classiÞcation accuracy, and class level F-measures. The body text, HTML, URL, and link feature sets each signiÞcantly outperformed the image feature set. Interestingly, these four feature sets all provided comparable performance, with none of them signiÞcantly outperforming the other (with the exception of body textÕs F-measure on fake OES). This suggests that these individual feature sets provide important complementary discriminatory potential which can be exploited by incorporating them in unison. Therefore, the results lend validity to the notion that using a large set of rich heterogeneous fraud cues can be highly beneÞcial for automated fake website detection.

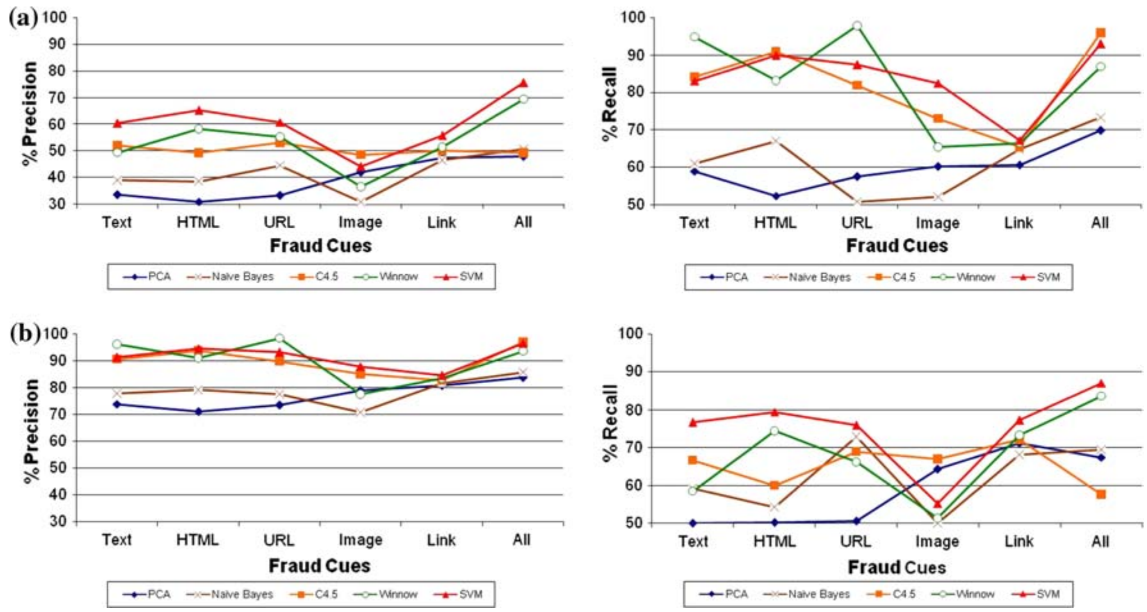

Fig. 7 **a** Page Level precision and recall on real OES websites for various features/techniques. **b** Page level precision and recall on fake OES websites for various features/techniques

Table 5 P-values for pair wise t-tests for Features (n = 250)

Techniques	Body text	HTML	URL	Image	Link
Accuracy					
All	<0.00001*	<0.00001*	0.00018*	<0.00001*	<0.00001*
Body text		0.18387	0.03269	0.00091*	0.47824
HTML			0.33842	0.00063*	0.34855
URL				0.00034*	0.25757
Image					0.00055*
F-measure real OES					
All	<0.00001*	<0.00001*	<0.00001*	<0.00001*	<0.00001*
Body text		0.01221	0.08578	<0.00001*	0.07584
HTML			0.24371	<0.00001*	0.03096
URL				0.00021*	0.06921
Image					0.00104*
F-measure fake OES					
All	<0.00001*	<0.00001*	<0.00001*	<0.00001*	<0.00001*
Body text		<0.00001*	<0.00001*	0.00401	0.02097
HTML			0.07747	0.00067*	0.16256
URL				0.00060*	0.25763
Image					<0.00001*

* Result significant at alpha = 0.05

We present two examples to illustrate why the extended feature set was able to garner improved performance. Figure 8 shows an example of a fake escrow website called ShipNanu (www.shipnanu.addr.com). The fraudulent website could not be categorized correctly using link features. This was because it had over 400 site level in-links (derived from Google) and numerous out-links. Furthermore, ShipNanu also had a large site map, with numerous inter-connected pages and images. Hence, the website's inter-link and intra-link features resembled those of legitimate escrow websites. However, the site shared content patterns with other fake OES sites, including similarities in body text (BT), image (IM), HTML (HS), and URL and anchor text (UA) attributes. Thus, while link features couldn't categorize this site, content features were able to.

Figure 9 shows the legitimate website Escrow.com along with two fake replicas. Since the replicas copied web pages directly from the original Escrow.com, text and

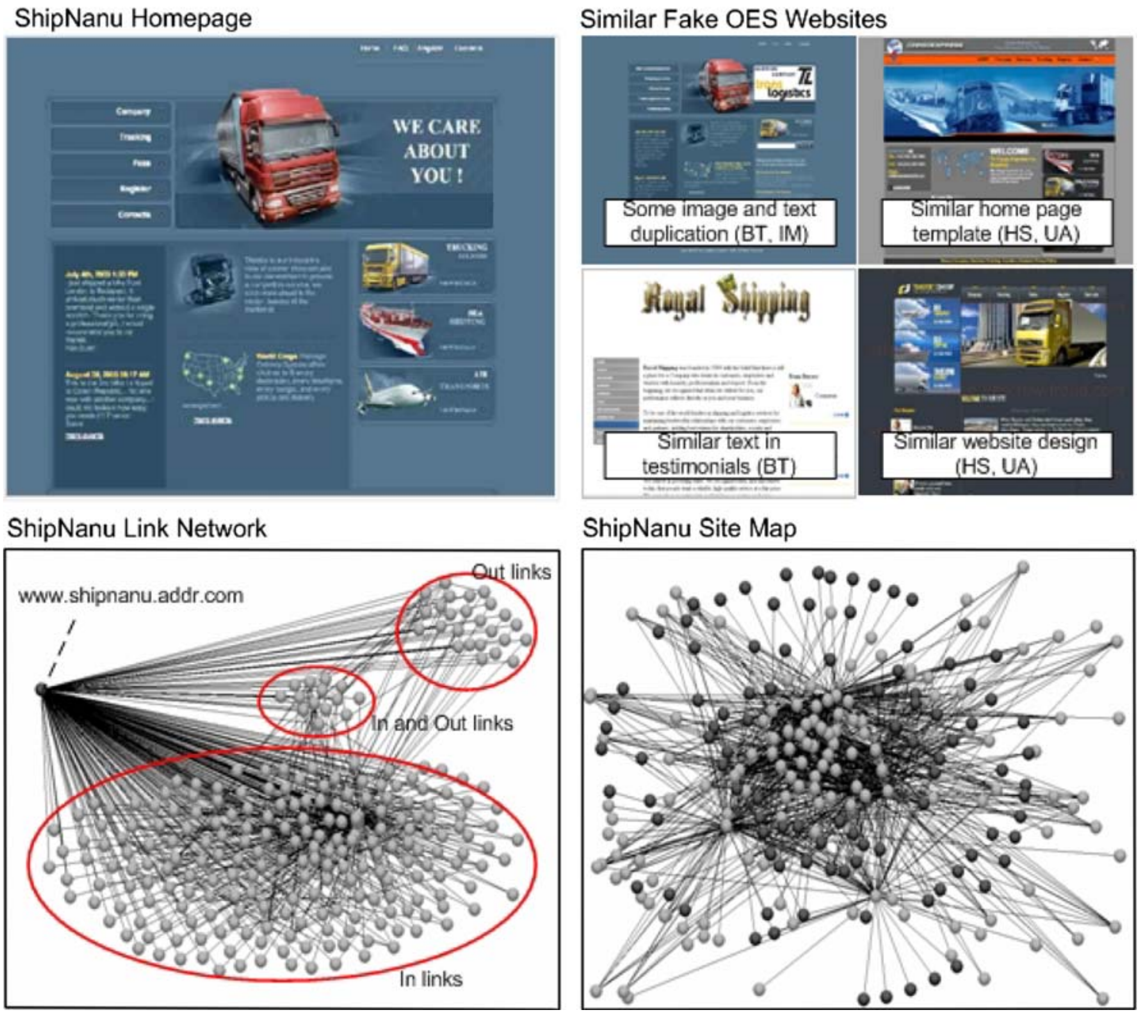

Fig. 8 Fake escrow website detected using content features but not linkage attributes

image content features were unable to identify web pages from the replicas as fake. However the replicas differed from the original in terms of link features (which allowed them to be detected as fraudulent). Replica #1 was a full replica with a similar site map but had only 3 site level in-links, a low number for legitimate sites. Replica #2 had higher in-links but was a partial copy devoid of a large portion of the originalÕs FAQ section, resulting in a less dense site map. The two examples presented in Figs. 8 and 9 exemplify how a rich holistic feature set incorporating a wide array of content and linkage based fraud cues can enhance the detection of fake OES websites.

5.1.2 Analysis of classification techniques (RQ2)

Pair wise *t*-tests were conducted on the 50 bootstrap runs to see which classiÞcation algorithm had the best performance (Table 6). During the comparison, feature sets were controlled. After applying the Bonferroni correction, *p*-values less than 0.005 were considered signiÞcant at alpha = 0.05. SVM signiÞcantly outperformed the four comparison classiÞcation algorithms. C4.5 and Winnow also performed well, with both outperforming PCA and NB. Although Winnow outperformed C4.5, the difference was not statistically signiÞcant. Overall, the results provide strong evidence that SVM is better suited for fake OES website classiÞcation than the comparison algorithms. This is consistent with related text and style classiÞcation research, where SVM has also performed well against other learning algorithms [29].

5.2 Experiment 2: comparing individual classiÞers against a feature ensemble

We compared the individual classiÞersÕ against feature ensemble classiÞers. All techniques were run using the entire (i.e., ÔÔAllÕÕ) feature set. The ensemble classiÞers encompassed 5 individual classiÞers, each trained on one

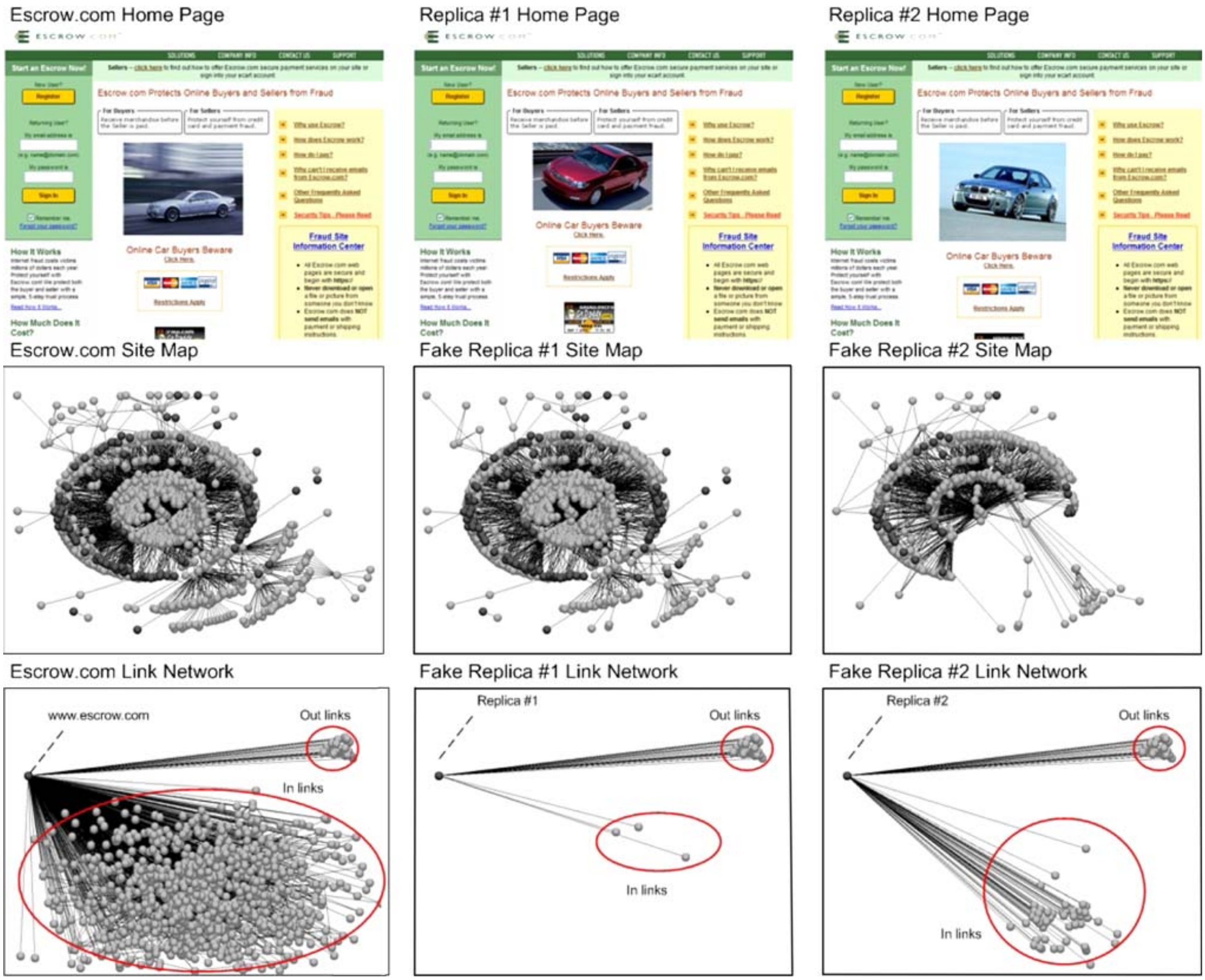

Fig. 9 Escrow website replicas detected using linkage features but not content attributes

of the individual feature categories (e.g., body text, HTML, URL, Image, and Link). For each of the 50 bootstrap runs, the ensemble and individual classiÞers were trained on the same 100 training websites. Table 7 shows the experimental results for each of the 5 individual and ensemble classiÞers, respectively. The table shows the overall accuracy and class level F-measure, precision, and recall. The ensemble classiÞers outperformed their individual classiÞer counterparts. Generally, the performance gain was consistently 1Ð2%. The exception was the Winnow classiÞer, where the performance gain was smaller.

5.2.1 *Analysis of individual classifiers and ensemble (RQ3)*

Pair wise t-tests were conducted on the 50 bootstrap runs to see whether the feature ensembles provided enhanced classiÞcation performance over the individual classiÞers. P-values less than 0.05 were considered signiÞcant. The t-test results are presented in Table 8. The ensemble classiÞers signiÞcantly outperformed their individual classiÞer counterparts in most instances. The feature ensembles had the added beneÞt of being able to consider each web page imageÕs feature vector in its entirety. As previously stated, the individual classiÞers were not capable of doing so. Instead, they could only consider the average of the aggregated image feature vectors for a particular web page. This allowed the ensembles to better preserve important image related information, thereby improving their discriminatory capabilities.

6 Evaluation: site level classiÞcation

Site level classiÞcation was performed on the 100 testing websites selected for each of the 50 bootstrap runs used in

Table 6 P-values for pair wise t-tests for techniques ($n = 300$)

Techniques	Winnow	C4.5	NB	PCA
Accuracy				
SVM	0.00047*	<0.00001*	<0.00001*	<0.00001*
Winnow		0.02218	<0.00001*	<0.00001*
C4.5			<0.00001*	<0.00001*
NB				0.00236*
F-measure real OES				
SVM	<0.00001*	<0.00001*	<0.00001*	<0.00001*
Winnow		0.04920	<0.00001*	<0.00001*
C4.5			<0.00001*	<0.00001*
NB				0.00619
F-measure fake OES				
SVM	<0.00001*	<0.00001*	<0.00001*	<0.00001*
Winnow		0.11460	<0.00001*	<0.00001*
C4.5			<0.00001*	<0.00001*
NB				0.00252

* Result significant at alpha = 0.05

the page level classification experiments. In addition to comparing different site level classification methods, we were also interested in assessing the impact of two important site level parameters: the maximum number of web pages used per test website (referred to as p) and the website region from which these p web pages were selected. As previously alluded to, these two parameters could have important implications for fake website classification performance, both in terms of accuracy/detection rates and computation times.

Five different values were used for p: 10, 25, 50, 100, and all pages. If a website had less than p pages in it, all pages from that website were used. Three different website regions were incorporated: top (select the first p pages), middle (select the middle p pages), and bottom (select the last p pages). Website regions were based on the web pages' directory structure (as described in section 2), where pages in deeper folders (i.e., ones with a greater number of slashes "/" in the URL) were considered to occur later than pages with fewer slashes "/". It is important to note that when using all pages, the region parameter became irrelevant (i.e., all-top, all-middle, and all-bottom would yield the same results).

Table 8 P-values for pair wise t-tests on number of pages per website ($n = 300$)

Comparison	Accuracy	F-measure real OES	F-measure fake OES
Individual vs. Ensemble	<0.00001*	<0.00001*	<0.00001*

* Result significant at alpha = 0.05

6.1 Experiment 3: comparison of stack and page scores methods

We compared the stack and page scores methods' performances for site level classification. Both methods used the SVM ensemble classifier's page-level classification results (using all features) as input. This particular page level classifier was used since it attained the best performance in experiment 2, with 89.60% classification accuracy. The page scores method categorizes a website as real or fake by comparing the percentage of its pages classified as fake against a pre-defined score threshold t. In other words, if $t = 0.4$, a website would be considered fake if more than 40% of its pages were classified as fake by the underlying SVM ensemble classifier. In order to identify an ideal threshold, we ran the page scores method for values of t ranging between 0.1 and 0.5, in increments of 0.05. For each t, we also ran the 5 different values of p for all three

Table 7 Performance results for individual and feature ensemble classifiers across 50 bootstrap runs

Technique	All websites	Real OES sites			Fake OES sites		
	Accuracy	F-Measure	Precision	Recall	F-Measure	Precision	Recall
PCA individual	66.04	55.21	45.68	69.77	72.65	83.26	64.44
NB individual	71.22	60.46	51.43	73.33	77.38	86.02	70.31
C4.5 individual	77.90	72.25	57.96	95.90	81.64	97.56	70.19
Winnow individual	85.28	77.98	70.73	86.90	88.94	93.78	84.59
SVM individual	88.38	82.79	74.55	93.07	91.24	96.67	86.39
PCA ensemble	68.42	57.96	48.25	72.57	74.71	85.00	66.64
NB ensemble	71.98	61.34	52.33	74.10	78.03	86.49	71.07
C4.5 ensemble	78.96	73.37	59.14	96.60	82.61	98.00	71.40
Winnow ensemble	85.32	78.06	70.80	86.97	88.98	93.81	84.63
SVM ensemble	89.60	84.39	76.76	93.70	92.20	97.02	87.84

Bold values indicate techniques that attained the best results for that particular metric

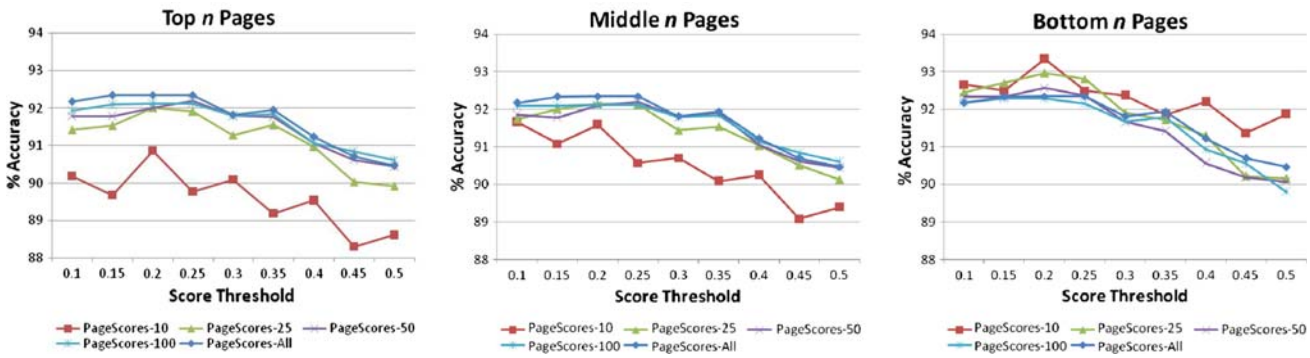

Fig. 10 Page scoresÕ accuracy using different thresholds, number of pages, and sampling regions

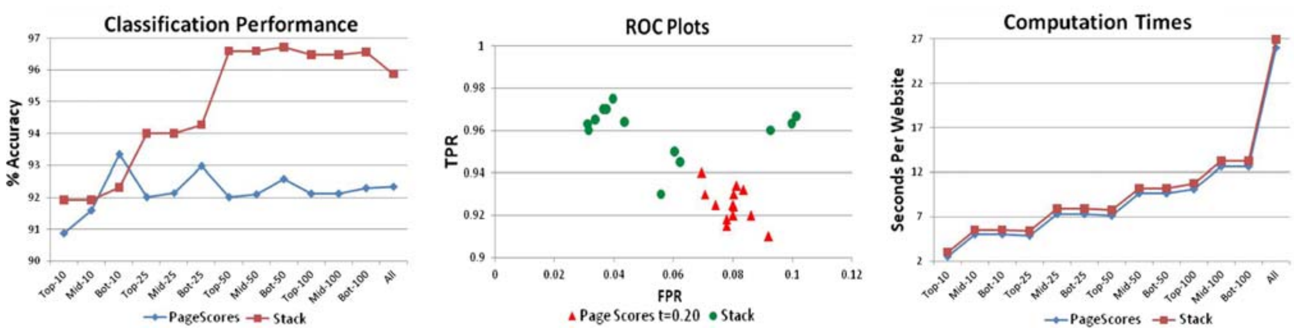

Fig. 11 Performance and computation times for different # of pages and sampling regions

website regions. The overall accuracy values using the page scores method are displayed in Fig. 10. Across various values of p, and for different sampling regions, the best results were attained when using a score threshold of 0.2. Interestingly, performance was improved when considering web pages from deeper within the websites. The best results were attained when using bottom pages, followed by middle and top. With respect to the maximum number of pages per website, using all pages tended to provide the best performance when compared against other values of p on the top and middle regions. However, the use of all pages was outperformed by $p = 10$, $p = 25$, and $p = 50$. Collectively, these results suggest that when using the page scores method, the bottom 10-50 pages in a website are the most effective indicators of whether a site is legitimate or fake. This is because these pages tended to be more accurately classiÞed by the SVM ensemble classiÞer.

Having identiÞed a suitable threshold for the page scores method, we compared its results against the stack classiÞer. The stack comprised of a top-level SVM classiÞer (with a linear kernel) that used the underlying SVM ensembleÕs page level classiÞcations as input features. The top-level SVM used two features for each website, the number of pages and the percentage classiÞed as fake (by the underlying SVM ensemble). For each of the 50 bootstrap runs, the top-level SVM was trained on the same 100 training websites used in the page level classiÞcation experiments (Experiments 1 and 2). Figure 11 shows the overall classiÞcation results for the stack, compared against the page scores method (with $t = 0.2$). The Þgure shows the overall accuracies (as a percentage) and computation times (in seconds per test website) for various values of p and different regions. The stack outperformed the page scores method by a wide margin across different website sampling regions for most values of p. For the stack, the best results were attained when using a maximum of 50 or 100 pages per website. As with the page scores method, the stack also performed better when using pages from the bottom region of the websites. However, this gain was only marginal for the stack.

Figure 11 also shows the average computation times for both techniques. These times encompass the time needed to collect websites (pages and images), extract all features from each page and image, perform page level classiÞcation using the SVM ensemble, and Þnally the time necessary to perform the site level classiÞcation. Looking at the computation times, it is apparent that the maximum number of pages used per website has the biggest impact on computation time. This is understandable since the feature extractor must be applied to every included page, and its images. The website sampling region also impacts computation times, with deeper regions taking somewhat longer. This is because sampling from deeper regions, even when selecting a small value for p, still requires knowing the entire site map for each website during the collection phase. In contrast, the difference in computation times between the page scores method and stack classifier is minor.

The computation times have important implications for determining the best site level classification technique and parameter settings. For example, when using the stack classifier, using all web pages from each site leads to an overall accuracy of approximately 96%, but with an average classification time of 27 s per website. On the other hand, using the top 10 pages per website takes 3 s on average, but results in an overall accuracy of approximately 92% (a 4% decrease). Assuming equal importance for computation time and accuracy, the ideal tradeoff probably lies somewhere in the middle.

Table 9 shows the experimental results for the three best parameter settings on the page scores method and stack classifier. The table shows the overall accuracy and class level F-measure, precision, and recall. In addition to providing enhanced overall accuracy, the stack classifier had better class level precision and recall on the real and fake OES sites. Hence, these classifiers were able to perform fake website detection with lower false positive and false negative rates than the page scores method.

6.1.1 *Analysis of page scores method and stack classifier (RQ4)*

A pair wise t-test was conducted in order to compare the performance of the stack classifier against the page scores method. Only the page scores with $t = 0.2$ was compared since it resulted in better performance than other score threshold values. During the comparison, the maximum number of pages per website and the sampling regions were controlled. The stack classifier significantly outperformed the page scores method (p-value < 0.00001, $n = 750$).

6.1.2 *Analysis of impact of number of page and sampling region on performance (RQ5)*

We conducted pair wise t-tests on the 50 bootstrap runs to see whether the maximum number of pages per website impacted classification performance (Table 10). During the comparison, sampling regions and classification techniques were controlled. After applying the Bonferroni correction, p-values less than 0.005 were considered to be significant at alpha $= 0.05$. Based on the t-test p-values, the best results were attained when using $p = 50$ pages. The use of $p = 10$ or $p = 25$ pages was significantly outperformed by $p = 50$, $p = 100$, and $p =$ all pages. Furthermore, $p = 50$ pages significantly outperformed $p =$ all pages and had better performance than $p = 100$ pages, but the difference was not significant.

Table 11 shows the pair wise t-test results comparing the classification performance for different sampling regions. During the comparison, the maximum number of pages per website and the classification techniques were controlled. P-values less than 0.05 were considered statistically significant. Based on the p-values, sampling region significantly impacted site level classification performance.

Table 10 P-values for pair wise t-tests on number of pages per website ($n = 300$)

# of Pages	25	50	100	All
Accuracy				
10	<0.00001*	<0.00001*	<0.00001*	<0.00001*
25		<0.00001*	<0.00001*	<0.00001*
50			0.02077	0.00049*
100				0.00678
F-measure–real OES				
10	<0.00001*	<0.00001*	<0.00001*	<0.00001*
25		<0.00001*	<0.00001*	<0.00001*
50			0.05621	<0.00001*
100				0.03438
F-measure–fake OES				
10	<0.00001*	<0.00001*	<0.00001*	<0.00001*
25		<0.00001*	<0.00001*	<0.00001*
50			0.01993*	0.00049*
100				0.00491*

* Result significant at alpha $= 0.05$

Table 9 Performance results for different combinations of number of pages and sampling regions across 50 bootstrap runs

Technique	All websites	Real OES sites			Fake OES sites		
	Accuracy	F-Measure	Precision	Recall	F-Measure	Precision	Recall
Page scores Bot-10	93.34	89.51	84.88	94.67	95.12	97.60	92.78
Page scores Bot-25	92.97	88.83	84.75	93.33	94.87	97.01	92.81
Page scores Bot-50	92.57	88.20	84.14	92.67	94.58	96.71	92.53
Stack Bot-50	96.72	94.65	92.71	96.67	97.63	98.54	96.74
Stack Mid-50	96.60	94.44	92.63	96.33	97.55	98.40	96.72
Stack Top-50	96.60	94.43	92.90	96.00	97.55	98.26	96.86

Bold values indicate number of page/sampling region setting that attained the best results for that particular metric

Table 11 P-values for pair wise t-tests on sampling regions (n = 400)

Sampling region	Accuracy		F-measureÐreal OES		F-measureÐfake OES	
	Middle	Bottom	Middle	Bottom	Middle	Bottom
Top	0.01471*	<0.00001*	0.02362*	<0.00001*	0.03924*	<0.00001*
Middle	Ð	<0.00001*	Ð	<0.00001*	Ð	0.00064*

* Result signiÞcant at alpha = 0.05

The use of bottom pages signiÞcantly outperformed using top and middle pages. Similarly, the use of middle pages also outperformed top level pages.

Based on the experimental results, the ideal parameter setting appears to be the use of 50 pages per website, sampled from the bottom region. However, taking computation time into consideration, using the top 50 pages provides slightly worse performance with a considerably lower computation time. Utilizing the top 50 pages per site with the stack classiÞer leads to 96.6% accuracy with an average computation time of 7.7 s per website. It is important to note that a major factor impacting computation times is the size of the feature set employed. When developing a fake website detection system, the size of the feature set would also need to be considered, in addition to the quantity of web pages and sampling region. The end decision would likely depend on whether the system is being run behind the scenes (in which case classiÞcation performance might be the more important factor), or in real-time, with end users awaiting the results (in which case shorter computation times are crucial).

7 Conclusions

In this study, we evaluated the effectiveness of automated approaches for fake OES website identiÞcation. To the best of our knowledge, this is the Þrst study to attempt to differentiate legitimate escrow websites from fraudulent ones, using automated classiÞcation procedures. Our study involved evaluation of various features and techniques for page and site level categorization of fraudulent escrow sites. The results indicated that the use of the proposed extended set of fraud cues coupled with an SVM ensemble classiÞer was capable of effectively identifying fake OES sites with 90% accuracy for page level classiÞcation. Combining the SVM feature ensemble with a top-level SVM classÞer (i.e., stacking) enabled over 96% site level classiÞcation performance. In addition to assessing the feasibility of automated fake OES detection, our analysis revealed several key Þndings. We observed that fake OES ÔÔfraud cuesÕÕ are inherent in body text, HTML, URL, link, and image features and that the use of a rich set of attributes is important for attaining a high level of detection accuracy. By using all Þve feature categories, our ability to detect phony escrow web pages typically improved by 2Ð10% over the use of individual feature groups.

The Þndings from this research have important implications (and potential future research directions) for developers of web security information systems as well as Internet users engaging in online transactions. Knowledge of key ÔÔfraud cuesÕÕ discovered in this study can be disseminated across the growing number of online resources and communities of practice that have emerged pertaining to Internet fraud. Additionally, we believe that fake OES detection systems can also be embedded into these online resource sites as an authentication mechanism, allowing Internet users to type in an escrow site URL in order to verify its legitimacy before transacting. Such a proactive authentication system would greatly complement the existing online databases which primarily offer retrospective support due to their reliance on individuals reporting fraudulent websites. Another important future direction could be the development of a browser plug-in to help protect against fake escrow websites by alerting Internet users. Current security toolbars only provide phishing Þlters for identifying spoof websites. In our future research, we intend to explore the effectiveness of an Internet browser toolbar for detecting fake escrow websites, based on the approach evaluated in this study.

References

1. Hu X, Lin Z, Whinston AB, Zhang H (2004) Hope or hype: on the viability of escrow services as trusted third parties in online auction environments. Inf Syst Res 15(3):236Ð249
2. Pavlou PA, Gefen D (2004) Building effective online marketplaces with institution-based trust. Inf Syst Res 15(1):37Ð59
3. Ba S, Whinston AB, Zhang H (2003) Building trust in online auction markets through an economic incentive mechanism. Decis Support Syst 35(3):273Ð286
4. Josang A, Ismail R, Boyd C (2007) A survey of trust and reputation systems for online service provision. Decis Support Syst 43(2):618Ð644
5. Chua CEH, Wareham J (2004) Fighting internet auction fraud: an assessment and proposal. IEEE Computer, pp. 31Ð37
6. Selis P, Ramasastry A, Wright CS (2001) Bidder beware: toward a fraud-free marketplaceÐbest practices for the online auction industry. Annual LCT Conference
7. IFCC (2003) IFCC internet fraud report: January 1, 2002Ð December 31, 2002, The National White Collar Crime Center

8. Antony S, Lin Z, Xu B (2001) Determinants of online escrow service adoption: an experimental study. In: Proceedings of the 11th workshop on information technology and systems (WITS Ô01) pp. 71Ð76
9. Airoldi E, Malin B (2004) Data mining challenges for electronic safety: the case of fraudulent intent detection in E-Mails. In: Proceedings of the workshop on privacy and security aspects of data mining
10. MacInnes I, Damani M, Laska J (2005) Electronic commerce fraud: towards an understanding of the phenomenon. In: Proceedings of the Hawaii international conference on systems sciences
11. Sullivan B (2002) Fake escrow site scam widens: auction winners sometimes lose $40,000. MSNBC, Dec 17 2002
12. Chou N, Ledesma R, Teraguchi Y, Boneh D, Mitchell JC (2004) Client-side defense against web-based identity theft. In: Proceedings of the network and distributed system security symposium, San Diego
13. Kolari P, Finin T, Joshi A (2006) SVMs for the blogosphere: blog identiÞcation and splog detection. In: AAAI spring symposium on computational approaches to analysing weblogs
14. Urvoy T, Lavergne T, Filoche P (2006) Tracking web spam with hidden style similarity. In: Proceedings of the 2nd international workshop on adversarial information retrieval on the web (AIRWeb)
15. Fraud.org ÔÔFraud Alert, 2001. http://www.fraud.org/news/newsset.htm
16. Dellarocas C (2003) The digitization of word of mouth: promise and challenges of online feedback mechanisms. Manage Sci 49(10):1407Ð1424
17. Pavlou PA, Gefen D (2005) Psychological contract violation in online marketplaces: antecedents, consequences, and moderating role. Inf Syst Res 16(4):372Ð399
18. Malhotra NK, Kim SS, Agarwal J (2004) Internet usersÕ information privacy concern (IUIPC): the construct, the scale, and a causal model. Inf Syst Res 15(4):336Ð355
19. Ntoulas A, Najork M, Manasse M, Fetterly D (2006) Detecting spam web pages through content analysis. In: Proceedings of the international world wide web conference (WWW Ô06), pp. 83Ð92
20. Gyongyi Z, Garcia-Molina H (2005) Spam: itÕs not just for inboxes anymore. IEEE Comput 38(10):28Ð34
21. Fetterly D, Manasse M, Najork M (2004) Spam, damn spam, and statistics. In: Proceedings of the seventh international workshop on the web and databases
22. Steiner I, Steiner D (2002) Online escrow fraud hits ebay members. AuctionBytes.com, 421
23. Koppel M, Schler J (2003) Exploiting stylistic idiosyncrasies for authorship attribution. In: Proceedings of IJCAIÕ03 workshop on computational approaches to style analysis and synthesis, Acapulco, Mexico
24. Abbasi A, Chen H (2005) IdentiÞcation and comparison of extremist-group web forum messages using authorship analysis. IEEE Intell Syst 20(5):67Ð75
25. Salvetti F, Nicolov N (2006) Weblog classiÞcation for fast splog Þltering: a URL language model segmentation approach. In: Proceedings of the human language technology conference, pp. 137Ð140
26. Menczer F, Pant G, Srinivasan ME (2004) Topical web crawlers: evaluating adaptive algorithms. ACM Trans Internet Technol 4(4):378Ð419
27. Diligenti M, Coetzee FM, Lawrence S, Giles CL, Gori M (2000) Focused crawling using context graphs. In: Proceedings of the 26th conference on very large databases, Cairo, Egypt, pp. 527Ð534
28. Abbasi A, Chen H (2008) Writeprints: a stylometric approach to identity-level identiÞcation and similarity detection in cyberspace. ACM Trans Inf Syst 26(2):7
29. Zheng R, Qin Y, Huang Z, Chen H (2006) A framework for authorship analysis of online messages: writing-style features and techniques. J Am Soc Inf Sci Technol 57(3):378Ð393
30. Joachims T, Cristianini N, Shawe-Taylor J (2001) Composite kernels for hypertext categorisation. In: Proceedings of the 18th international conference on machine learning, pp. 250Ð257
31. Vapnik V (1999) The nature of statistical learning theory. Springer, Berlin
32. Li J, Zheng R, Chen H (2006) From Þngerprint to writeprint. Commun ACM 49(4):76Ð82
33. Stamatatos E, Widmer G (2002) Music performer recognition using an ensemble of simple classiÞers. In: Proceedings of the 15th European conference on artiÞcial intelligence
34. Abbasi A, Chen H (2008) CyberGate: a design framework and system for text analysis of computer- mediated communication. MIS Q 32(4):811Ð837
35. Baayen RH, Halteren Hv, Neijt A, Tweedie F (2002) An experiment in authorship attribution. In: Proceedings of the 6th international conference on the statistical analysis of textual data
36. Binongo JNG, Smith MWA (1999) The application of principal component analysis to stylometry. Lit Linguist Comput 14(4):445Ð466
37. Apte C, Damerau F, Weiss SM (1994) Automated learning of decision rules for text categorization. ACM Trans Inf Syst 12(3):233Ð251
38. Littlestone N (1988) Learning quickly when irrelevant attributes are abound: a new linear threshold algorithm. Mach Learn 2: 285Ð318
39. Quinlan R (1986) Induction of decision trees. Mach Learn 1(1):81Ð106
40. Koppel M, Argamon S, Shimoni AR (2002) Automatically categorizing written texts by author gender. Lit Linguist Comput 17(4):401Ð412
41. Bayes T (1958) Studies in the history of probability and statistics: XI. Thomas bayesÕ essay towards solving a problem in the doctrine of chances. Biometrika 45:293Ð295
42. Yang Y, Slattery S, Ghani R (2002) A study of approaches to hypertext categorization. J Intell Inf Syst 18(2Ð3):219Ð241
43. Dietterich TG (2000) Ensemble methods in machine learning. In: Proceedings of the Þrst international workshop on multiple classiÞer systems, pp. 1Ð15
44. Cherkauer KJ (1996) Human expert-level performance on a scientiÞc image analysis task by a system using combined artiÞcial neural networks. In: Chan P (ed) Working notes of the AAAI workshop on integrating multiple learned models, pp. 15Ð21
45. Wu B, Davison BD (2006) Detecting semantic cloaking on the web. In: Proceedings of the world wide web conference (WWW Ô06), pp. 819Ð828
46. Kriegel H, Schubert M (2004) ClassiÞcation of websites as sets of feature vectors. In: Proceedings of the international conference on databases and applications, pp. 127Ð132
47. Ester M, Kriegel H, Schubert M (2002) Web site mining: a new way to spot competitors, customers, and suppliers in the world wide web. In: Proceedings of the 8th ACM SIGKDD, pp. 249Ð258
48. Kwon O, Lee J (2003) Text categorization based on k-nearest neighbor approach for web site classiÞcation. Inf Process Manage 39(1):25Ð44
49. Dzerosi S, Zenko B (2004) Is combining classiÞers with stacking better than selecting the best one? Mach Learn 54(3):255Ð273
50. Baldwin RG (2005) Image pixel analysis using Java. Online Press, Austin
51. Jackson D (1993) Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches. Ecology 74(8):2204Ð2214