Detecting Visually Similar Web Pages: Application to Phishing Detection

TEH-CHUNG CHEN, SCOTT DICK, and JAMES MILLER
University of Alberta

We propose a novel approach for detecting visual similarity between two Web pages. The proposed approach applies Gestalt theory and considers a Web page as a single indivisible entity. The concept of supersignals, as a realization of Gestalt principles, supports our contention that Web pages must be treated as indivisible entities. We objectify, and directly compare, these indivisible supersignals using algorithmic complexity theory. We illustrate our approach by applying it to the problem of detecting phishing scams. Via a large-scale, real-world case study, we demonstrate that 1) our approach effectively detects similar Web pages; and 2) it accuractely distinguishes legitimate and phishing pages.

Categories and Subject Descriptors: H.3.5 [**Web-Based Services**]; H.5.3 [**Web-Based Interaction**]

General Terms: Security, Human Factors

Additional Key Words and Phrases: Algorithmic complexity theory, Gestalt theory, Web page similarity, anti-phishing technologies

ACM Reference Format:

Chen, T.-C., Dick, S., and Miller, J. 2010. Detecting visually similar Web pages: Application to phishing detection. ACM Trans. Intern. Tech. 10, 2, Article 5 (May 2010), 38 pages.
DOI = 10.1145/1754393.1754394 http://doi.acm.org/10.1145/1754393.1754394

1. INTRODUCTION

A fundamental idea is: are two items the same or different? In many situations, this binary decision has no absolute answer. Instead, the question must be evaluated in a probabilistic framework. Web pages fall into the category of entities where this question can be asked, and where the answer, and in fact the question, have no obvious unique definition. Alternatively, the answer can be recast onto a linear dimension that measures the similarity or difference between two

Author's address: S. Dick, Department of Electrical and Computer Engineering, University of Alberta, Canada; email: dick@ece.ualberta.ca.

DOI 10.1145/1754393.1754394 http://doi.acm.org/10.1145/1754393.1754394

Web pages. The construction of such a question, and the implementation of an approach to provide an answer to this question, are the principal themes of this article.

Web page similarity detection is widely used by many popular applications such as Web search engines, automated categorization systems, and phishing/spam filtering mechanisms. With precise similarity identification results, Web search engines and automated categorization systems can reduce their essential storage requirements. Phishing/spam mechanisms can prevent users from becoming the victims of malicious activities by filtering out suspicious Web pages with embedded similarity identification technology aimed at detecting malicious pages. An illustration (via a large-scale, real-world case study) of the effectiveness of similarity as a phishing-detection mechanism is the third principal theme of this article.

The goal of our research is to provide a robust way to evaluate the similarity of Web pages from the viewpoint of a human viewer. We start by considering human perception from a Gestalt viewpoint [Wertheimer 1944] as the theoretical foundation for our approach. Specifically, we follow the Gestalt viewpoint that images are interpreted in a holistic fashion rather than as a set of distinct features, which is common among other approaches. We augment our visual Gestalt viewpoint with the concept of supersignals [Dorner 1997], which provides an explanation of how humans use a holistic interpretation of visual input to drive rapid and frequent decision making. These concepts are expanded upon in Section 3. Finally, we show how these visual supersignals can be encoded (or compressed) into simple numerical values to facilitate automation of this decision-making process (i.e., similar or not). The supersignals are represented by an approximation of their algorithmic complexity description; and this description or the "distance" between two such descriptions is considered as an estimation of the perceived similarity of the two Web pages. We have undertaken four experiments to demonstrate that our concept is viable. All four of the experiments show that our new method is able to discriminate between similar and dissimilar Web pages.

Our proposed approach can be deployed in many different areas of applications. For example, an image search engine for visual object categorization [DSL Reports 2008]; detecting visual tricks used by spammers to fool Spam email filters [Wu et al. 2005]; and as an anti-phishing mechanism [Fu et al. 2006]. Clearly, however, the approach will require some tailoring to maximize its performance within any specific domain.

This article is organized as follow: Section 2 describes existing feature-based methods for detecting similarity between Web pages. In Section 3, we introduce the theoretical foundations for our proposed approach. In Section 4, we address the similarity metric, and discuss its implementation. In Section 5, we discuss how our similarity measure could be applied to detecting phishing scams. Our experiments reported in Section 6 demonstrate the efficacy of this approach, while we discuss related work in Section 7. We provide a summary and discussion of future work in Section 8.

2. SIMILARITY SIGNATURE

2.1 Feature-Based Similarity Measures

There are several approaches that utilize Web page components as features for detecting near-duplicate Web pages. Henzinger [2006] discusses and performs an evaluation of existing Web document identification methods. After comparing a variety of approaches, Broder et al. [1997]'s shingling method and Charikar [2002]'s random projection algorithm were found to represent the current state of the art in this domain. The shingling method is to use word sequences to detect the differences between documents. Charikar uses random projections of the words as a signature for their similarity identification method. In both algorithms, Web pages (in HTML format) are converted into a token sequence by specific rules to represent the "fingerprint" of the Web pages. Then these "fingerprints" are used as a signature to determine the similarity between two pages. The word sequences (random projections of the words) in the HTML Web page are the most important features within these two identification methods. A similar method, which uses the sequences of adjacent characters as the Web page signature, was developed by Manber [1994] and Heintze [1996]. Clearly, these methods are principally using the textual contents of a Web page as the main feature for the similarity comparison.

In the domain of Web page clustering or categorization, Web page elements such as Web page structure, text, and link structure are used as features for comparison. Haveliwala et al. [2002] explore text-based (with or without stemming), anchor-based, and text plus anchor–based approaches for describing features within a Web page. They extract these features and then seek to recombine them using a number of weighting approaches. The paper describes an extensive empirical exploration to find the best heuristics (based upon the empirical results) from the assembled list. In Cai et al. [2003], Web pages are compared based on their structure. A Web page is separated into "blocks" by extracting its layout (or tag) structure for further analysis. Only tags that directly impact the visual display of the page are considered in this work. The encoding of the relationships between these "blocks" of a Web page is considered as a description of the page. These descriptions are then compared as a measure of similarity between pages.

In Shen et al. [2004] a Web page classification algorithm is proposed based upon Web page textual summarization. Their approach is to extract the most relevant textual content from Web pages and then pass this information into a standard text classification algorithm.

In Dean and Henzinger [1999] and Hou and Zhang [2003] methods to cluster similar Web pages based on Web page hyperlinks are utilized. Web pages are classified as similar if they have similar hyperlink structures. In Wang and Kitsuregawa [2002], an algorithm is introduced to cluster Web pages by combining textual content and link analysis. The authors claim that this hybrid method can achieve superior identification performance than methods using either text or link analysis alone.

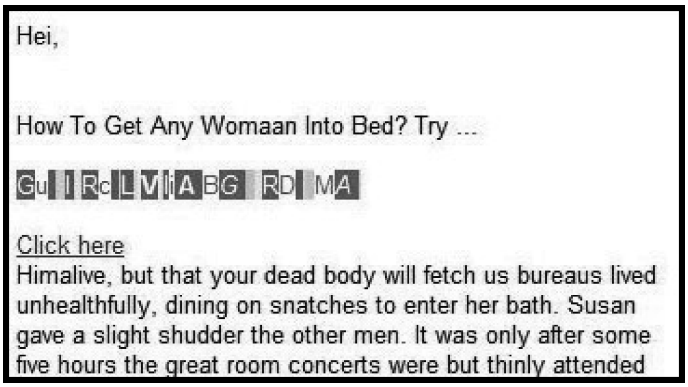

Fig. 1. Spam email with obfuscated characters made visible.

The outlined methods all transform Web pages elements into features to perform similarity identification or classification. These feature-based methods may identify textually related or structurally similar documents effectively for clustering applications, such as an automatic data categorization system. Nevertheless, these methods are often unable to recognize the similarity between two Web pages that, to a user, would appear essentially identical. This may be due to innocent differences in Web page implementations or deliberate countermeasures employed by a spammer or phisher. We will illustrate this idea in the following section.

2.2 What Can't We Count On for Visual Similarity Identification?

It is believed that traditional methods, based on Web page elements, cannot effectively identify visually similar Web pages in a manner congruent to user perceptions. Many identification methods that use features such as the textual content of a Web page can easily be evaded. Consider the Web-based email shown in Figure 1 (which passes the Thunderbird Spam filter) as an example. This malicious email defeats the spam detection mechanism, which is based on textual comparison, by arranging some meaningless character combinations in the Web email textual content. At the same time, they also create an illusion to unambiguously deliver their advertisement to the user (as shown in Figure 2). Spam detection methods based on textual content are easily foiled by this technique and its variations.

Similar countermeasures can be employed against other similarity-detection algorithms employing different feature sets. The specific feature set is irrelevant; Web page structure, hyperlinks, text, images, and their combinations have all been employed to generate feature sets. However, all of these features remain vulnerable to obfuscation attacks of one form or another. It is also entirely possible for obfuscation to occur by accident; the Web is now a rich media platform, and any given Web page can be encoded and presented using a great many alternative technologies. This implies that two pages that would be considered "identical" by users would exhibit vastly different "fingerprints" when feature-based techniques are employed.

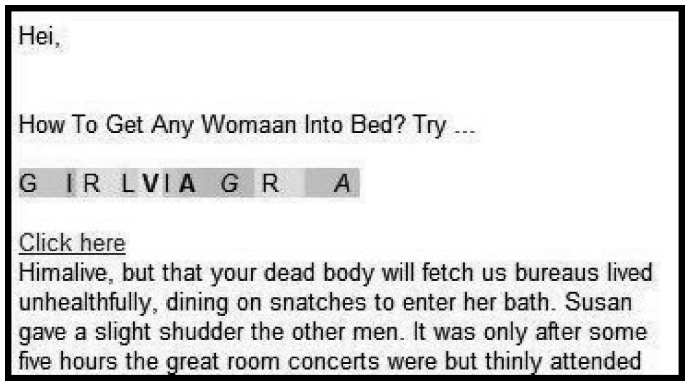

Fig. 2. Spam email as viewed by a user.

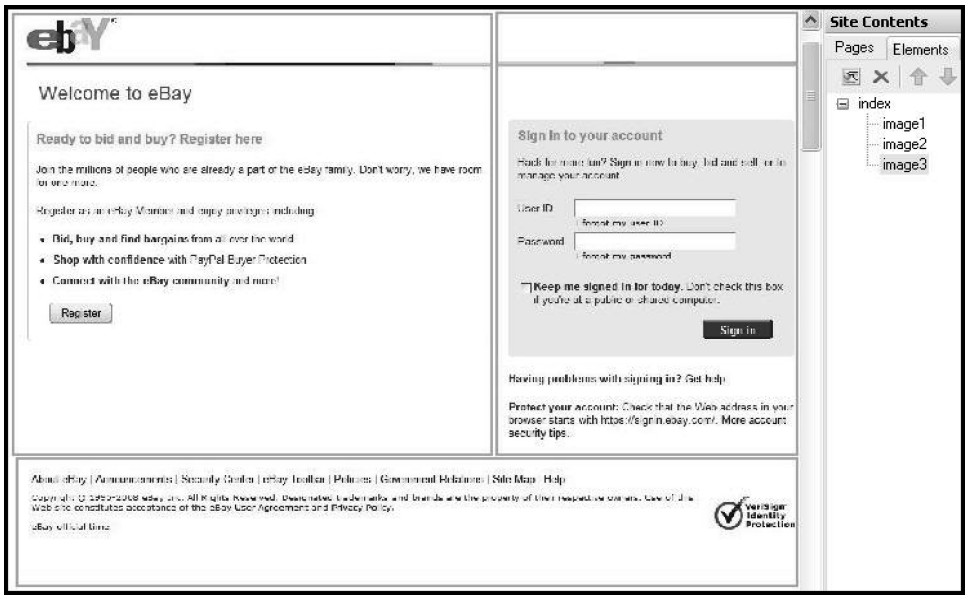

Fig. 3. Spoof Web page composed of three separate images.

As another example, consider the Web page in Figure 3. We built this Web page—which is visually similar to the legitimate eBay login Web page [eBay 2008]—in order to simulate a phishing attack. This Web page is composed of only three separate images. Consequently, the "fingerprint" of this phishing Web page—whether judged by the textual content, hyperlinks, and/or Web page structure—is totally different from the legitimate eBay Web page. This simple tactic causes similarity identification methods based on Web page element features [Heintze 1996; Broder et al. 1997; Charikar 2002; Haveliwala et al. 2002] to report that these Web pages are significantly different. Notice that this example also represents the case of unintentional obfuscation due to implementation differences. Perhaps a login page will not be presented as static

images, but a vanity or information page might well be (in order to precisely control the page's appearance). Web page elements become an unreliable clue when we are looking for a solution to the problem of visually similar Web page identification.

Therefore, for applications where we need to identify visually similar Web pages, such as phishing or spamming detection, we require an identification method that can identify the visual similarity of a Web document accurately, even in the presence of obfuscations. In considering this problem, we note that features based on page elements are inherently localized; that is, the information contained in a feature (in Shannon's sense) is concentrated in discrete, identifiable page elements. Any adversary seeking to evade a detector need only identify those features, and alter them. Likewise, any implementation decision that results in substantially different page elements will also confound a similarity comparison. Our objective is to design a heuristic method which can simulate the process of human visual perception and decision making in this situation.

3. THEORETICAL FOUNDATION

3.1 Gestalt Theory

Gestalt theory [Kalviainen 2007; Graham 2008] provides us with the theoretical basis for our similarity identification approach. One of its central ideas is that the whole of a perceived image is different from the sum of its parts acting in isolation [Gordon 2004].

"The fundamental formula of Gestalt theory might be expressed in this way: There are wholes, the behavior of which is not determined by that of their individual elements, but where the part-processes are themselves determined by the intrinsic nature of the whole. It is the hope of Gestalt theory to determine the nature of such wholes." [Wertheimer 1944].

For example, as shown in Figure 4, the left part of the figure is a set of simple shapes. The human perception of the left part is interpreted as "several simple shapes scattered all around." Nevertheless, when we rearrange those simple shapes in a certain way, suddenly the sum of those simple shapes becomes organized into to a recognizable illustration—a clown (as shown in the right part in Figure 4). Those simple shapes are endowed with a new interpretation when they are combined together. The parts are from the whole, but the whole changes the parts [Gordon 2004].

Gestalt visual psychology is based around a number of simple laws: figure/ground, proximity, closure, similarity, and continuation. With regard to our situation, the laws of proximity and similarity are the most important. In the proximity law, objects spatially near each other are grouped together. On the contrary, objects spatially apart are separated. Per Kepes [1944], words in a text are perceived as separate entities because of their spatial characteristics; letters in one word are close together, and not separated by a whitespace (ideographic languages obviously excepted). Different spacings, in turn, can change the entire meaning of a phrase. In general, the closer items are spatially or

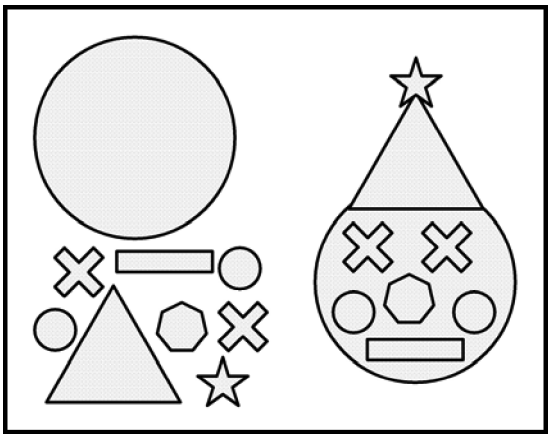

Fig. 4. The perception of parts and wholes.

temporally, the more likely they are to be considered part of an organized and unified group. In the similarity law, objects which are similar in shape, size, color, proximity, and direction are interpreted as part of a group. Even if the objects are spatially separated, we still tend to group them together. On the other hand, dissimilar objects can be separated even if they are spatially close.

The combined effect of these two laws illustrates why humans tend to perceive an object with several elements as a whole. Consequently, we argue that an individual does not interpret a Web page by examining individual Web page elements such as banners, colors, pictures, and icons. Instead, the structure, layout, and media design of the Web site creates a specific perception, which is more important than the individual components of the Web page.

3.2 Inattentional Blindness

Inattentional blindness (IB) provides another argument against the use of feature-based systems for similarity detection [Mack and Rock 1998a, 1998b]. IB can be summed up as the phenomenon of "looking without seeing." When IB happens, even though an individual's eyes are wide open and various objects are imaged on their retinas, individuals seem to perceive nothing. At the beginning stage of visual perception, the brain spontaneously and automatically performs Gestalt grouping operations to integrate all retinal inputs for further processing. However, only stimuli which capture attention at the later stages of processing are perceived. This has been experimentally demonstrated [Mack and Rock 1998a, 1998b]; these experiments show that objects' features or stimuli such as location, motion, frequency and color were not perceived about 75% of the time in an inattentional condition. This is a well-known psychological phenomenon.

Consequently, when users try to identify similar Web pages (that is, recognize pages that have been previously seen), they spontaneously ignore fine details and perceive only the whole image. If they are later asked to recall the details of the Web page (such as the exact color of icons, the exact text,

or other Web page components), a likely answer could be "I don't know" or "I didn't notice that" [Mack and Rock 1998a]. That is, the entire Web page is likely to leave a single impression on their memory. Empirical research into phishing scams supports this notion; users commonly do not observe the address bar, status bar, or security indicators displayed by modern Web browsers, and can be fooled up to 90% of the time by high-quality phishing sites [Dhamija and Tygar 2006].

3.3 Supersignals

Related to Gestalt psychology is the concept of "supersignals" [Dorner 1997], which seeks to explain how humans make rapid decisions when bombarded with a massive set of inputs. In this scenario, visual inputs tend to dominate, but other senses can play a role. We view the idea of supersignals as extending the basic Gestalt visual processing into a decision-making mechanism. In our situation, the Gestalt process transforms the visual representation of a Web page, producing a supersignal that acts as the input to the decision making process (is this page similar to, or the same as, a page that was previously viewed?). The content of these supersignals change with a number of factors—such as the individual's perceived familiarity with the situation. Supersignals can be thought of as trying to provide an explanation of an individual's behavior when they encounter a complex, but familiar, situation. The person can reduce the complexity of the situation by generating a supersignal that collapses a number of features into one impression, based upon their previous experience. Differences between the novice and the experienced driver are a good illustration of this idea. Novices need to direct attention to many variables and traffic situations at once. Driving is a highly complex business for them. Any input variables can cause unexpected circumstances that need more processing time for the novice to handle. Complexity causes trouble, stress, and anxiety, which make the situation more complicated to deal with. On the contrary, an experienced driver doesn't notice this complex situation as requiring the processing of so many independent variables. They are able to generate many supersignals (that is, they are able to integrate large numbers of inputs into a small number of holistic signals) to reduce the complexity. Once the complicated situation has been simplified, the driver can recognize similar situations from their prior experience, and apply an appropriate response from that experience. This recognition is also not an "optimized" comparison; decision theory tells us that people will select the first familiar situation that comes to mind and appears to fit the current circumstances; this will often take just a fraction of a second. People do not usually spend time reflecting on how closely the current situation matches the prior experience; it is simply "good enough" [Dorner 1997].

Similarity identification for humans is a decision-making process related to many complicated factors. To complete the analogy, an individual who is new to the Internet acts as an inexperienced driver. However, the vast majority of Internet users can be consider as "highly experienced drivers" and hence their similarity decision process only accepts a very small number (potentially only 1) of input(s), or supersignal(s), into their decision making process. In

order to design our similarity measure to be congruent to human perception, we will generate a "supersignal" that represents the whole of a Web page.

We view the construction of supersignals as a sampling or compression process. Cognitively, an individual is attempting to wade through a massive number of inputs and reduce it to a minimal, but sufficient, single representation to allow a decision to be successfully undertaken. This sampling approach is undoubtedly highly nonlinear, and is characterized by emphasizing and integrating seemly important aspects of the input; while ignoring or discarding the seemly unimportant aspects of the input. Hence, the process is unlikely to correspond to sampling in a traditional mathematical sense. In essence, an individual is seeking to reduce the input to a single irreducible form. Hence, we view the process as having parallels to approaches for defining Algorithmic Information theory [Chaitin 1987]; and hence, we seek to objectively measure the relationship (the basis of the decision) between supersignals within this framework. While, algorithmic information theory principally studies "complexity measures" on strings; clearly translating supersignals, or more accurately our representation of supersignals, into an appropriate form is straightforward. In addition, we seek to use these complexity measures as the basis of objectifying our supersignals representation. Within algorithmic information theory, Kolmogorov complexity [Li and Vitanyi 1997] can be used to provide a theoretical definition of an objective evaluation of a pseudo-irreducible form of a signal; and the numerical difference between two such objective approximations can be considered proportional to the actual difference between two arbitrary signals.

4. OBJECTIFICATION OF THE SIMILARITY METRIC

While it can be stated that Kolmogorov complexity is objective, this is clearly a theoretical position, as Kolmogorov complexity is incomputable[1] in anything apart from contrived situations. However, Cilibrasi and Vitanyi [2005] recently demonstrated that Kolmogorov complexity can be successfully approximated by current compression techniques. Kolmogorov complexity can be viewed as the ultimate compressor—producing for any arbitrary string (or file or image), a minimum description of that string, given some form of description language. Hence, practical compression approaches that compress arbitrary strings or files or images can be viewed as approximations to the optimal, however unattainable, compressor. Kolmogorov complexity can be viewed as the limiting case for compression technology. Specifically, Li et al. [2004] introduce the Normalized Information Distance (NID), which approximates Kolmogorov complexity within known limits. Further, they prove that NID is a valid metric within these limits. They claim that NID can "discover all similarities between two arbitrary entities; and represents object similarity according to the dominating shared features between two objects."

NID can be defined as follows: Let $K(x|y)$ refer to the Kolmogorov complexity, that is, the length of the shortest binary program, that accepts as input y and outputs x; and let $K(x)$ refers to the Kolmogorov complexity of x, that is, the length of the shortest binary program with no inputs that outputs x. The

[1]In a Turing sense.

value $max(K(y|x), K(x|y))$ can be considered as the length of the shortest binary program (with the reference universal prefix Turing machine) that with input y, computes x, and with input x, computes y. Given these definitions, NID can be defined as

$$NID(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}.$$

Further details about NID and its properties can be found in Li et al. [2004]. However, the fact that Kolmogorov complexity is incomputable implies again that NID can't be used directly. Hence, Li et al. [2004] and Cilibrasi and Vitanyi [2005] provide an approximation to this metric based upon real-world compression algorithms (denoted C) rather than the Kolmogorov complexity.

4.1 Normalized Compression Distance

Normalized Compression Distance (NCD) is described as a parameter-free distance metric that is believed to be able to uncover all similarities with a single metric [Li et al. 2004; Cilibrasi and Vitanyi 2005]. It is a practical metric approximating NID. It is computed from the lengths of compressed data files, images, strings, etc. using real-world compression algorithms. For an arbitrary compression algorithm C, NCD is given by:

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}.$$

Clearly, the relationship between the denominators of NID and NCD is straightforward; however, the relationship between the numerators is less so. It is shown that the numerator of NID can be rewritten as

$$\max\{K(x, y) - K(x), K(x, y) - K(y)\}.$$

Note that

$$K(x, y) = K(xy) = K(yx),$$

where xy and yx denote the concatenation of two signals. [Cilibrasi and Vitanyi 2005] argue that this numerator can be effectively approximated by

$$\min\{C(xy), C(yx)\} - \min\{C(x), C(y)\}.$$

If we now assume that the symmetry property holds for the compression algorithm C, we have [Cilibrasi and Vitanyi 2005]:

$$\min\{C(xy), C(yx)\} = C(xy).$$

Clearly, it is important to understand that NCD is an approximation of NID; and that the symmetry property may not hold for all real-world compression algorithms. In addition, practical compression algorithms may invalidate common properties found in theoretical measurement systems; for example, is monotonicity ($C(xy) \geq C(x)$) a guaranteed property of all block-coding compressors? Hence, this final approximation will require empirical verification within our context. (Note that much of the literature on applying the NCD uncritically treats it as a "universal similarity metric," [Delany and Bridge 2006; Cernian

et al. 2008; Feldt et al. 2008]. We find such a sweeping assertion to be dubious at best when modeling a phenomenon as complex as human perception.)

In our context, the NCD value is a nonnegative number representing the distance/difference between two images (approximations of supersignals) which in turn represent a Web page. Using the usual interpretation of similarity as an inverse of distance, we assert that the more similar two objects (images or Web pages) are, the smaller the NCD distance between them should be.

4.2 Compression Algorithms and Supersignals

The most commonly used data compression programs are gzip, bzip2, and PPM. Gzip is a Lempel-Ziv-type compressor with a 32-kilobyte window [Ziv and Lempel 1977]. Its reliability, speed, and simplicity make it the most popular compressor. Bzip2 is a fast compressor which uses the blocksorting algorithm [Burrows and Wheeler 1994]. It provides good compression and an expanded window of 900 kilobytes which has the ability to detect longer-range patterns. PPM (Prediction by Partial Matching) [Bell et al. 1984] is a compressor using a mix of statistical models arranged by trees, suffix trees or suffix arrays. It provides better performance but with the side effect of slower speed and heavy memory consumption.

Different data compression algorithms lead to different varieties of NCD. Some data compression programs use many complex schemes that involve stochastic modeling of the data at many levels simultaneously. While the NCD metric is in theory application neutral, in practice the choice of the compressor needs to be tailored to the application domain. For example, the "blocksort" virtual compressor (which also uses the blocksorting algorithm) is appropriate for frequency analysis, spectral analysis, and substring matching combined. Clearly no unique or optimal presentation exists for the construction of our pseudosupersignals—the input to the compression stage. This topic needs further research to find the most appropriate mechanism for this encoding; and a variety of empirical evaluations to confirm any such supposition.

Our long-term goals, research and hypothesis can be stated as follows: Can repeated research efforts and results into improving the representation of a supersignal (both the initial representation and the compression component) continually provide mechanisms which will defeat phishers' attempts to construct effective phishing Web sites? We view these phishing Web sites as a "moving target"; as researchers produce mechanisms to prevent successful phishing attacks; phishers will in turn produce new mechanisms to circumvent these defensive measures.

Our short-term goal, research and hypothesis (and the principal application in this paper) can be stated as: Can an initial representation of a supersignal (both the initial representation and the compression component) defeat phishers' current attempts to construct effective phishing Web sites? We will explore this hypothesis in the remainder of the paper by empirically evaluating our initial representation against current phishing sites found "in the wild." We have considered several possible representations of a supersignal (the DOM tree; the HTML code; the link structure of a page). We have chosen the rendered Web

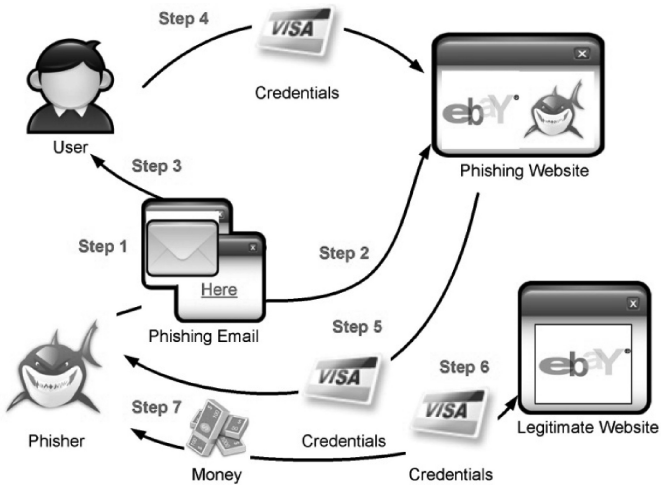

Fig. 5. The phishing seven-step process [Emigh 2005].

page as the initial representation because that—and not, for instance, the DOM tree or HTML code—is what the user *perceives* at a Web site. We therefore render the Web page and capture a high-resolution image representation thereof. And hence, subsequently two such images are the input to the compressor stage which produces the NCD value for the decision making process.

5. APPLICATION TO ANTI-PHISHING TECHNOLOGIES

5.1 Motivation

According to a report released by Gartner [McCall 2007], phishing scams cost $ 3.2 billion in losses, victimizing 3.6 million people in the United States alone in 2007. Phishing is a type of online identity theft in which sensitive information is obtained by misleading people to access a malicious Web page. While there are many different types of phishing attacks, the most common phishing scam seen today is a deceptive attack. Researchers usually conceptualize phishing scams as a seven-step process [Emigh 2005] as shown in Figure 5. The scam begins when a malicious payload is sent to a user (often an email asking the user to login to one of their accounts). The second stage is when the user attempts to navigate to the login page for that account—but is instead directed to an attack page. The third and fourth stages are when the user is prompted for their account's credentials—and provides them. In stage five, these credentials are transmitted to the phisher, who uses them to impersonate the user in stage six. Finally, in stage seven, the phisher is able to steal the assets held in the account. Phishing attack sites are designed to seem visually very similar to the legitimate site they are impersonating. As we have argued, user recognition of a Web site (or what they believe is the Web site) takes place in the blink of an eye. Users do not consciously reflect when they make this decision. We suggest that anti-phishing methods need to be developed to explicitly account for the user's perceptions and actions.

5.2 Existing Anti-Phishing Solutions

Email-filtering approaches are server-side techniques that try to interfere with stage 1 of the phishing scam. Email is a very common vector for delivering a malicious payload to a user—in this case, a message involving a "call to action" regarding one of the user's accounts, and a "helpful" link to the Web site. By preventing users from receiving these emails, these approaches attempt to completely deflect the phishing attack. They are closely related to anti-spam solutions, in that both rely on an analysis of the content of an email message. Therefore, they also suffer from the same weaknesses as anti-spam approaches. In particular, the image-based technique demonstrated in Figures 1 and 2 is a potent countermeasure against content analysis. Some well-known email filtering solutions include Thunderbird [Mozilla 2009] and PILFER [Fette et al. 2007].

Anti-phishing toolbars are the most popular and widely-deployed solutions to fight phishing Web sites. Most of the toolbars employ blacklists and whitelists. They determine the URL currently being viewed, and send it to the blacklist/whitelist (B/W) database for filtering. The result will be delivered back to the user with either an alert warning the user of a possible phishing scam, or an assurance that the site is legitimate. IE7 [Microsoft 2009], FireFox3 [Mozilla 2008], Spoofguard [Chou et al. 2004], and Netcraft [2009] are popular toolbars in wide usage. The performance of the toolbar depends on the B/W database (except for SpoofGuard, which uses heuristics). Unfortunately, the average life time of a phishing Web site is 3.4 days (APWG 2009). These short-lived Web sites can defeat these toolbars because the database is not updated fast enough to protect users from a brand-new phishing site. It simply takes time to detect a new phishing site. Another trick known as "DNS/URL redirection or domain forwarding" [Andresen et al. 1996] can fool the B/W databases by rapidly changing the DNS/URL IP address mapping in a dynamic DNS domain server. The "blocked" phishing Web site can be back to business in only one minute. Although other heuristic analysis methods such as domain registration lifetime checking are proposed and deployed in the toolbars, the phishers still can find a way to fool classifiers built on these heuristics. Human factors must also be acknowledged as another threat to the anti-phishing solutions. Some users input their credentials even when they receive warnings from the toolbars [Wu et al. 2006].

Mutual authentication is another common anti-phishing solution. By acquiring secret messages or preauthorized signals from the server (legitimate Web site) with a secure connection (usually SSL connection), the client (the user) can make sure they are browsing the legitimate Web site and their credentials can be safely transmitted. The most challenging part of this client and server-side method is human factors. First, users must choose to install the software on their system, and follow complicated instructions in setting it up. Research indicates that the majority of computer users will never change the default configuration of their software [MacKay 1991]. Secondly, once users have set the system up, false positives will reduce the alertness of users. Such false alarms are known to destroy user trust in any anti-phishing system [Dhamija and

Tygar 2006]. Thirdly, phishers still have social engineering tricks that can fool users into disabling the anti-phishing systems. For example, a letter entitled "Incompatibility notice for your anti-phishing system," if apparently sent from a trusted authority, can trick users into removing their anti-phishing system. Some well-known mutual authentication solutions include DSS [Dhamija and Tygar 2005], PassMark (now part of the RSA Identity Protection & Verification Suite) [RSA 2009], and YahooMail's sign-in seals (Yahoo 2009].

We believe that any effective anti-phishing solution must be robust against the counterattacks of determined, inventive adversaries. This means that it is not enough to find features that discriminate phishing Web pages from legitimate pages; those features must also be extremely difficult or impossible for the phisher to alter. Our approach is to examine the phishing scam and find a critical-path item in the scam that cannot be changed without severely weakening the effectiveness of the scam. By finding such critical characteristics of phishing Web sites, we can design a classifier that is robust against the phishers efforts to defeat it.

5.3 A Key Characteristic of Phishing Web Sites

Phishing Web sites usually look similar to a legitimate Web site, but not exactly the same. Designing a visually similar Web page is a crucial step in the phishing scam. On arrival at a Web site (the point between stages three and four of the scam), a user faces a choice: they must either choose to believe that the site is legitimate, or that it is a fake. This choice is not made after a period of reflection; instead, the user looks for the "supersignal" of a recognized, trusted Web site. This decision is clearly on the critical path for the phisher; if the user is suspicious at this point, they will probably not provide their account credentials, and the scam fails. Thus, the phisher must craft a page that closely imitates the legitimate page, causing the user to erroneously recognize the supersignal of the legitimate page. Once this decision is made, psychological studies [Dorner 1997] tell us that it is unlikely to be revisited until a significant amount of contrary evidence is observed. Therefore, a visually similar Web page becomes an inevitable element in the phishing scam—and thus a characteristic that can be considered as a fundamental component of a phishing attack. Although a phisher's goal is to make the phishing Web site as similar to the legitimate Web site as possible, there are still differences between them. This is mainly due to the frequent updates of the legitimate Web site. Pictures, advertisements, and new information are renewed by the legitimate Webmaster from time to time to keep the site fresh and interesting to users. Phishers, however, do not expend the effort on such renovations, leading to a difference. As long as the users believe they are browsing the legitimate Web site, this small difference is irrelevant to the phisher. Thus, by detecting visual similarity between an unknown page and a known legitimate page, we can recognize an attempt to trick the user. Because this attempted deception is on the critical path for the phisher, we believe that phishers will not be able to alter this attribute of their sites to avoid detection. This attribute seems robust against anything short of a wholesale revamping of the current phishing model.

A key point to note is that using phishing Web pages to test a similarity algorithm immediately gives us an excellent "ground truth" for our evaluations. The authors of phishing toolkits (who are now usually professional Internet criminals) go to great lengths to craft fraudulent pages that will fool human beings. This means that we can assume that captured phishing pages will be perceived as highly similar to the legitimate brand they imitate; to the point that human beings will confuse them even with significant financial consequences at stake. Furthermore, monitoring sites such as the PhishTank provide us with a corpus of phishing Web pages that essentially covers the entire domain of interest (current phishing scams); the entire purpose of the PhishTank is, after all, to provide "accurate and actionable" information to the anti-phishing community (Staff 2008). In contrast, corpora for Web clustering or search cannot possibly both cover the entire domain of "clustering Web pages" or "finding relevant Web pages" *and* provide a ground truth for that corpus. For instance, Henzinger [2006] created a corpus of 1.6 *billion* Web pages by using the Google Web crawler, in order to compare two existing similarity algorithms. However, as they freely acknowledge, it was impossible to create a ground truth for this dataset (i.e., what pairs of Web pages would *in fact* be perceived as highly similar). Haveliwala et al. [2002] attempt to use human-created Web directories (e.g., Yahoo! or the Open Directory Project) to create a ground truth, based on the relative position of two documents in the directory tree (the "familial distance"). The key assumption is that document similarity is monotonically related to this familial distance; however, as the authors acknowledge, this is not always so. Importantly, no evaluation or estimate of how frequently monotonicity fails is provided; thus, this evaluation metric cannot truly be said to be a ground truth. Our approach, by contrast, is a ground truth based on human perceptions. This does not mean (and we do not claim) that the NCD technique by itself is a complete anti-phishing solution.

Our choice of the NCD technique is intended to overcome the primary weakness of feature-based similarity comparisons: the ability of a phisher to easily craft a Web page that seems visually similar to the legitimate page (satisfying the critical-path constraint), but is not detected as such. The problem has some parallels to preventing message forgery in cryptographic systems; the mapping from the original to the encrypted message should be extremely difficult to reverse-engineer. The central characteristics for a successful encryption are confusion (a complex, nonmonotonic mapping from plaintext to ciphertext characters) and diffusion, which is the scattering of information across a message. However, phishing scams cannot be dealt with using cryptographic techniques, because the human user accepts many similar "messages" as being identical—whereas cryptographic techniques such as AES map a one-bit difference in the plaintext to a change in 50% of the bits in ciphertext. Such approaches are not congruent to human perceptions. We do believe, however, that the characteristic of diffusion is useful in preventing the phisher from reverse-engineering our similarity technique and finding an economical countermeasure. NCD, being based on compression techniques, naturally makes use of information (in Shannon's sense) that is diffused throughout the image of a Web page.

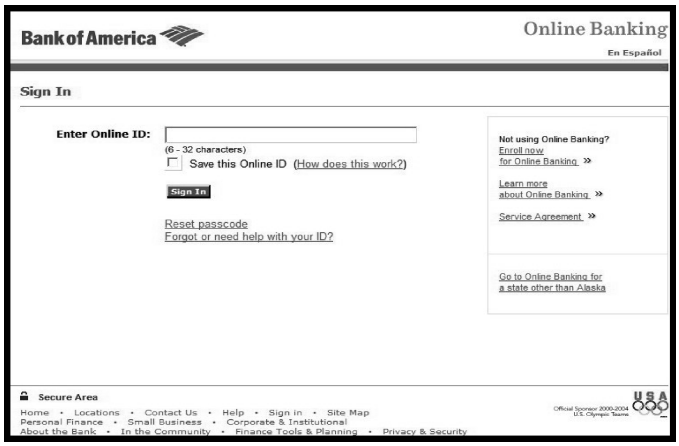

Fig. 6. Legitimate BOA Web page.

Fig. 7. Phishing BOA Web page.

Thus, we feel our NCD technique is likely to be highly resistant to obfuscation countermeasures.

As shown in Figures 6, 7, 8 there are slight differences between the legitimate and phishing Web sites targeting them. These real world Web pages were collected from the PhishTank through 20/05/08 to 22/05/08. We can clearly observe that the text, pictures, links, and Web structure in the phishing Web pages are not identical to the legitimate one. In the next section, we will report on experiments that demonstrate the utility of our similarity-based approach to detecting phishing scams. The phishing Web pages used in these experiments are actual phishing pages drawn from the PhishTank [PhishTank 2008], demonstrating that the method works in current real-world scenarios.

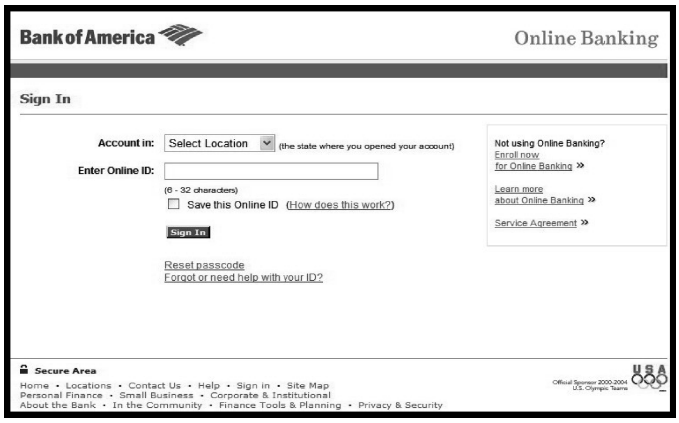

Fig. 8. Phishing BOA Web page.

6. EMPIRICAL EVALUATION

We conducted four experiments to validate our method for measuring similarity between Web pages. Due to the fact that all the samples are from the real world, this objective is combined with an assessment of our proposed similarity-based system for recognizing phishing Web pages. The first experiment, "12 Pairs" was designed to test if pairs of highly similar Web sites could be matched together while being distinguished from less-similar pages. The second experiment, "Clustering" was designed to test whether we could find a cluster of similar pages within a larger group of dissimilar pages (i.e., to see if any successes from the 12 Pairs experiment generalize to uneven distributions of similar and dissimilar Web pages). Thirdly, we perform a large-scale experiment specifically to test the ability of our similarity measure to distinguish between legitimate and phishing Web pages. This last experiment involved a control group of 120 comparisons between legitimate Web sites (legitimate versus legitimate), and 320 phishing Web pages targeting those sites (phishing site versus the legitimate they imitate). We analyze these results using both statistical tests for a difference in means, and using the ROC curve for a putative decision-threshold classifier. Finally, we conduct a small-scale test of possible obfuscations (based on image-processing techniques) to examine the robustness of our similarity technique.

In using the NCD, one key decision is *which* of the many compressors in the literature will be employed. There is very little guidance in the literature on this point; the most significant result available is that NCD is skewed by the size of the objects to be compared due to internal buffer limits in certain compression algorithms [Cebrian et al. 2005]. Moreover, our problem domain consists of rendered Web pages. Converting a fundamentally 2-dimensional object (a rendered Web page) into a one-dimensional string (i.e., treating an image as a row-major array) seems likely to destroy the spatial relationships in the image. However, among the very limited work in applying the NCD to images, there is virtually no exploration of two-dimensional compression techniques (e.g., the wavelet-based approach in JPEG images). Hescott and

Koulomzin [2007] used only black and white images; Li and Zhu [2006] and Lan and Harvey [2005] explored NCD for grayscale images, and Batista et al. [2005] used grayscale textures, all employing string-based compressors. Macedonas et al. [2008] created a new distance based on parsing the dictionary of a (one-dimensional) Lempel-Ziv-type compressor and applied it to color images. The only usage of a two-dimensional compressor in the NCD we could find was an image coregistration algorithm that compared JPEG and bzip2 [Bardera et al. 2006], and this also examined sets of monochrome images (the red, green and blue channels were separated).

The paucity of guidance on applying NCD to images means that we need to select our compressors by considering how browsers render Web pages from first principles. The Web page's source code (written in one or more markup and/or scripting languages) must be transformed into a visual representation by a fault-tolerant browser; indeed, browsers will render Web pages even if they contain substantial errors [Ofuonye et al. 2010], following the display preferences set by individual users. This means that even the same Web page might not be rendered to the same dimensions for two different users, and might be rendered quite differently by different browsers; this includes the size of the rendered image! Two-dimensional compressors *require* a rectangular image, and thus "concatenating" two images of differing sizes (an essential step in the NCD) makes no logical sense. (Note that we cannot just concatenate files; two-dimensional compression is appropriate only for image data, not the headers of a JPEG file.) A one-dimensional compressor, on the other hand, imposes no such requirement. Furthermore, the coregistration results of Bardera et al. [2006] for the bzip2 compressor indicate that one-dimensional compressors can be relatively robust against spatial shifts (the relative variation of the NCD values is quite small under translation). These reasons provide a sound rationale for the use of one-dimensional compressors with a row-major image representation in our experiments; while this means we are following existing practice in applying the NCD to images, previous work has not developed a sound foundation for these choices. We are also applying the NCD to RGB color images rather than monochrome images.

6.1 The Twelve-Pairs Experiment

The objective of this experiment is to see if we can group twelve legitimate Web pages and twelve phishing pages each targeting one of these pages together in pairs. Based on the argument that a legitimate page and a phishing page targeting it are highly similar to one another, this experiment also tests the validity of our proposed similarity metric. The expected result for this test is that each legitimate page and its single Phish will be paired together as the most similar to one another, for all twelve pairs.

6.1.1 *Design and Methodology.* We collected 12 different legitimate Web pages and 12 phishing Web pages targeting them as the samples in this experiment (Table I). We chose financial Web sites in Table I based on the frequency with which phishing sites attempt to imitate them. The Phish were captured in the PhishTank [PhishTank 2008]. In addition, one Italian and one Spanish

Table I. Samples List for 12 Pairs Test

Name of the Web Site	Collection Date (yy/mm/dd)
1. ArkValley	08/02/26
2. BancadiRoma(it)	08/02/26
3. Chase	08/02/26
4. CitiBank	08/02/26
5. FifthThird	08/02/26
6. ibercajadirecto(es)	08/02/26
7. LloydsTSB	08/02/26
8. RBC	08/02/26
9. USBank	08/02/26
10. Wachovia	08/02/27
11. WaMu	08/02/27
12. NatWest	08/02/28

Web site are added to the group to check if there was a language or regional dependency in our similarity metric. There are a total of 24 samples in this test, each of which is compared against all 23 other samples.

Note that, although NCD is theoretically a distance metric (and therefore commutative), in practice this would require a "perfect" compression algorithm, which does not exist. Thus, the NCD values we observe are not commutative. Therefore, all the NCD values shown below are the average value of both orderings of every two Web sites. Lower NCD values indicate greater similarity (i.e., we consider distance the inverse of similarity).

6.1.2 *Interpretation of Results.* Consider the Royal Bank of Canada (RBC) Web site as an example. The NCD values shown in Table II are the NCD of the remaining 23 samples against the legitimate RBC Web site (RBC-L). The "-L" in this table refers to the legitimate Web site of that brand, while "-P" denotes a phishing Web page targeting that brand. Most of the NCD values are between 1.01∼1.07. The values in Row 15 and 16 are different. An NCD of ∼0 in Row 15 (RBC-L against RBC-L) indicates that the algorithm properly finds that there is perfect similarity between a Web page and itself. NCD value 0.632, found in Row 16 (RBC-L against RBC-P) is far less than the values against the other 22 Web pages. According to this result, we can say RBC-L is most similar to RBC-P in this group of Web pages.

The same pattern holds true for all other Web sites; the most similar Web pages are always a phishing site against the legitimate site it targets. For example, the lowest NCD value for ArkValley-L is against ArkValley-P, and vice versa. As shown in Table III, all twelve legitimate Web sites are paired against the phishing sites targeting them. Again, the values reported for NCD are the average of both computations of NCD between two Web sites. To visualize these results, we have employed quartet trees [Strimmer and von Haeseler 1996] in Figure 9. We selected this technique because the quartet-puzzling algorithm is based on identifying locally optimal pairings of elements (in the maximum-likelihood sense), which is a good match to the data we wish to visualize. We interpret Figure 9 as indicating that the twelve pairs have been successfully grouped together, as the two members of a pair always share the same parent node (i.e., the branch length between them is minimal).

Table II. The NCD Values of RBC-L against Other 23 Web Sites

Name of the Web Site	NCD Value
1. ArkValley-L	1.059
2. ArkValley-P	1.047
3. BancadiRoma(it)-L	1.029
4. BancadiRoma(it)-P	1.031
5. Chase-L	1.041
6. Chase-P	1.055
7. CitiBank-L	1.039
8. CitiBank-P	1.052
9. FifthThird-L	1.037
10. FifthThird-P	1.051
11. ibercajadirecto(es)-L	1.025
12. ibercajadirecto(es)-P	1.025
13. LloydsTSB-L	1.059
14. LloydsTSB-P	1.057
15. **RBC-L**	**0.168**
16. **RBC-P**	**0.632**
17. USBank-L	1.024
18. USBank-P	1.040
19. Wachovia-L	1.073
20. Wachovia-P	1.073
21. WaMu-L	1.048
22. WaMu-P	1.050
23. NatWest-L	1.013
24. NatWest-P	1.013

Table III. The NCD Values for All 12 Pairs

Name of the Web Site	In Pairs	NCD Values
1. ArkValley-L against ArkValley-P	YES	0.558
2. BancadiRoma(it)-L against BancadiRoma(it)-P	YES	0.352
3. Chase-L against Chase-P	YES	0.748
4. CitiBank-L against CitiBank-P	YES	0.808
5. FifthThird-L against FifthThird-P	YES	0.890
6. ibercajadirecto(es)-L against ibercajadirecto(es)-P	YES	0.221
7. LloydsTSB-L against LloydsTSB-P	YES	0.284
8. RBC-L against RBC-P	YES	0.632
9. USBank-L against USBank-P	YES	0.834
10. Wachovia-L against Wachovia-P	YES	0.149
11. WaMu-L against WaMu-P	YES	0.218
12. NatWest-L against NatWest -P	YES	0.329

6.2 The Clustering Experiment

The objective for this experiment is to determine if the NCD similarity technique can detect a single "cluster" of highly similar Web pages within a larger group of Web pages. This experiment examines the performance of the NCD similarity technique when the groups of highly similar Web sites are not balanced in size. This is known as the imbalanced-dataset problem in machine learning (alternatively, the sample selection bias problem in statistical

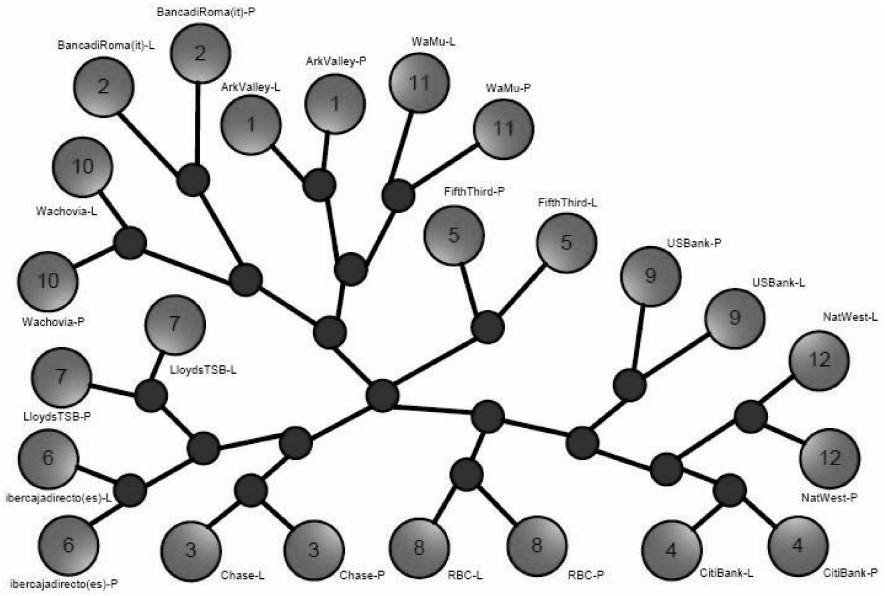

Fig. 9. Quartet tree visualization for 12 pairs experiment.

Table IV. Sample List for the Clustering Test

Name of the Web Site	Legitimate/Phishing	Collection Date (yy/mm/dd)
01-BOA-P	Phishing	08/05/09
02-BOA-P	Phishing	08/05/09
03-BOA-P	Phishing	08/05/09
04-BOA-P	Phishing	08/05/11
05-BOA-P	Phishing	08/05/13
06-BOA-L	Legitimate	08/05/09
07-Wachovia-L	Legitimate	08/05/11
08-LloydsTSB-L	Legitimate	08/05/09
09-AbbeyNational-L	Legitimate	08/05/10
10-NatWest-L	Legitimate	08/05/10

modeling), and can profoundly affect the performance of a model. The expected result in this experiment is that all of the highly similar Web pages will have a lower NCD value against one another than against the dissimilar pages, and that pages outside of the "cluster" will have higher NCD values against one another.

6.2.1 *Design and Methodology.* [45] We selected BOA (Bank of America) as the legitimate Web site and collected five phishing Web sites against it from the Phish Tank from 08/05/09 to 08/05/13. Then four other legitimate, and highly popular, financial Web sites were collected from 08/05/09 to 08/05/11. We selected financial Web sites because this should make the test more rigorous; conceptually, two financial Web sites should be at least somewhat more similar to each other than, say, an online auction site is to a bank Web site. These 10 samples are shown in Table IV.

Table V. NCD Values against BOA-L

Web Site	NCD Value
01-BOA-P	**0.802**
02-BOA-P	**0.741**
03-BOA-P	**0.743**
04-BOA-P	**0.704**
05-BOA-P	**0.663**
06-BOA-L	**0.183**
07-Wachovia-L	1.009
08-LloydsTSB-L	0.990
09-AbbeyNational-L	0.979
10-NatWest-L	0.966

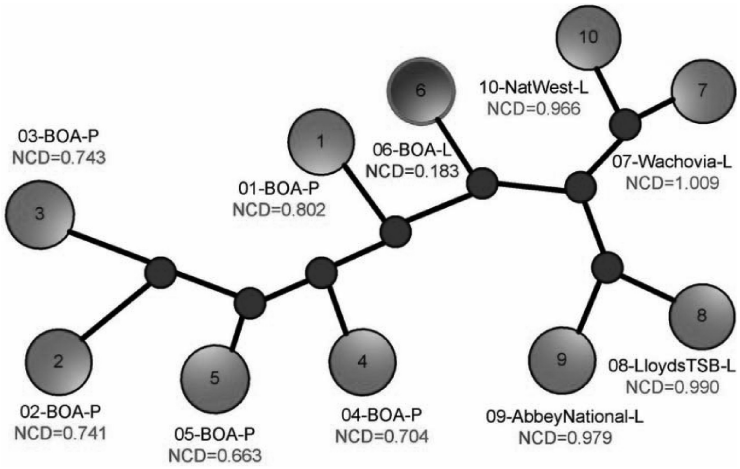

Fig. 10. Quartet tree visualization for clustering experiment.

6.2.2 *Interpretation of Results.* In Table V, we present the NCD of each Web page against the legitimate Bank of America Web page. Rows 1 to 5 correspond to phishing pages that target Bank of America, while rows 7 to 10 correspond to different, legitimate financial Web sites. There appears to be a considerable difference between the phishing Web pages targeting BOA and the other four financial Web sites. Therefore, we can say that the legitimate BOA Web site is more similar to the five phishing Web sites targeting it than the other four Web sites in our experiment, and these six Web pages should be clustered together. A quartet tree visualization of the NCD values is presented in Figure 10.

6.3 The Large-Scale Experiment

The objective of this experiment is to subject our proposed similarity-based anti-phishing technique to a realistic test. We collected 16 legitimate Web sites, and 20 phishing Web pages targeting each of these sites (total 320 phishing Web pages). Our initial experiments (12 Pairs and Clustering) have supported our assumption that a phishing page will be highly similar to the legitimate

Web site it targets. Furthermore, they also indicate that a legitimate page and the phish targeting it will be more similar to one another than to other legitimate Web sites. Thus, in this experiment, we will contrast two populations: one consists of the pairwise NCD between all of the legitimate sites, while the other consists of the NCD between a legitimate site and each of the phishing Web pages targeting it. The expected result of this experiment is that there should be a statistically significant difference in the means of the two populations, specifically with the mean of the latter group being lower.

6.3.1 *Design.* Our goal in this experiment is to examine how the NCD similarity technique would perform in a realistic, browser-level anti-phishing scenario. Assume that we have access to a whitelist of legitimate Web sites, which represent the likely targets of phishing scams. This is plausible because the number of brands that phishers attack is relatively constant (at less than 131 in January 2008), with a small number of brands (less than 15 in January 2008) accounting for over 80% of all phishing attacks [APWG 2008]. When we visit a Web site, we automatically execute an image capture, followed by a comparison (using the NCD similarity technique) against all Web sites in the whitelist. If there is a strong similarity to one of the whitelisted sites (i.e., the NCD is unusually low), we signal an alert. This experiment is designed to determine whether a population of phishing sites exhibits a statistically significant difference in the mean NCD against their target brand when compared to differences between different, legitimate Web sites. If such a difference exists, then the NCD similarity technique is viable in the anti-phishing scenario we have outlined.

Using the same criteria as in the previous experiments (frequency of phish targeting a legitimate site), we chose 16 legitimate Web sites (see Table VI). Evidence from the Anti-Phishing Working Group, an industry consortium dedicated to anti-phishing research, indicates that the great majority of Phishing scams (more than 80%) active in any one month target as few as 15 Web sites. We thus chose 16 as a reasonable number of legitimate sites to "protect." These 16 sites were also the most heavily phished during our collection period, allowing us to reach our goal of capturing 20 "live" phishing Web pages for each legitimate Web site. "Live" phish are phishing Web pages that have not yet been taken down from their host servers; we captured these phish by visiting them as soon as we observed them in the PhishTank [PhishTank 2008] or the Broadway PhishTracker [DSL Reports 2008] during the period May 10–July 10, 2008.

We designed two groups in this experiment (shown in Table VI). The samples for group one are the 16 legitimate Web sites. The samples for group two are the 320 phishing Web sites targeting the legitimate Web sites in group one.

6.3.2 *Methodology.* Two populations of NCD values are generated in this experiment, using the Blocksort compressor. Firstly, we compute all possible pairwise NCD values for the Web sites in group one (note that the NCD is, at this time, defined only for the comparison of two objects). This population represents the expected NCD between two different, legitimate sites, and is the

Table VI. Samples for the Large Scale Experiment

Group **One**	Group **Two**
01-ebay-L	01-ebay-P1 ∼ ebay-P20
02-PayPal-Home-L	02-PayPal-Home-P1 ∼ PayPal-Home-P20
03-PayPal-L	03-PayPal-P1 ∼ PayPal-P20
04-Halifax-L	04-Halifax-P1 ∼ Halifax-P20
05-NatWest-L	05-NatWest-P1 ∼ NatWest-P20
06-BOA-L	06-BOA-P1 ∼ BOA-P20
07-ebay(it)-L	07-ebay(it)-P1 ∼ ebay(it)-P20
08-Wachovia-L	08-Wachovia-P1 ∼ Wachovia-P20
09-LloydsTSB-L	09-LloydsTSB-P1 ∼ LloydsTSB-P20
10-RBS-L	10-RBS-P1 ∼ RBS-P20
11-AbbeyNational-L	11-AbbeyNational-P1∼ AbbeyNational-P20
12-PosteItaliane(it)-L	12- PosteItaliane(it)-P1 ∼ PosteItaliane(it)-P20
13-HSBC(uk)-L	13-HSBC(uk)-P1 ∼ HSBC(uk)-P20
14-Cartasi-L	14-Cartasi-P1 ∼ Cartasi-P20
15-WellsFargo-L	15-WellsFargo-P1 ∼ WellsFargo-P20
16-eppicard-L	16-eppicard-P1 ∼ eppicard-P20
Total samples count	
16	16*20 = **320**

control group for this experiment. The computation results in a 16*16 matrix; the diagonal values should be removed (NCD(x,x) $\sim = 0$), and the corresponding entries in the upper and lower triangle of the matrix are aggregated together, yielding $(16*16 - 16)/2 = 120$ elements in this population. We have used both the arithmetic mean and the maximum to aggregate each of the two corresponding NCD values; there is currently no guidance available on which would be the more effective choice. Our decision to use the pairwise differences between our legitimate sites seems likely to cause an overestimate of the false positive rate in our experiments; the 16 brands all have highly similar objectives, will likely share at least some common text, and the structural layout of the pages should be more similar than would a whitelist site and a random site. Given the importance the anti-phishing community places on minimizing false positives, this seems to be a reasonable approach. To form the second population, we computed the NCD between each phishing page and the legitimate page it targets. As we have captured 20 phishing pages for each legitimate page, this yields 320 elements in the group two populations (again, we compute the average and maximum of the two NCD values).

Our hypothesis is that the NCD values in group two are significantly less than group one. As we have more than 300 samples, we choose to employ the z-test for sample means to test this hypothesis. We repeat this experiment for both aggregating the NCD values by the arithmetic mean, and for aggregating by the maximum operation. Note that both the group one and group two populations are different in each of these repetitions, so the results are independent.

6.3.3 *Interpretation of Results.* The results of the z-tests are given in Table VII (aggregation by average) and in Table VIII (aggregation by maximum). In both cases, we reject the null hypothesis with $p < 0.05$, and so we

Table VII. Results from the Z-Test for Average

Two Sample for Means		
	Legitimate vs. Legitimate	Legitimate vs. Phishing
Mean	1.005	0.745
Known Variance	0.001	0.087
Observations	120	320
z	**15.577**	
P(Z ≤ z) one-tail	< 0.001	
z Critical one-tail	**1.645**	

Table VIII. Results from the Z-Test for Maximum

Two Sample for Means		
	Legitimate vs. Legitimate	Legitimate vs. Phishing
Mean	1.005	0.745
Known Variance	0.001	0.087
Observations	120	320
z	**15.578**	
P(Z ≤ z) one-tail	< 0.001	
z Critical one-tail	**1.645**	

conclude that our original hypothesis—that NCD values in group two are significantly less than group one—is supported. Furthermore, it appears that the choice of using arithmetic mean or maximum to aggregate corresponding NCD values does not influence this outcome. Thus, the NCD similarity technique is a viable anti-phishing strategy.

6.4 Effectiveness as an Anti-Phishing Classifier

The viability of our proposed approach has been evaluated in three experiments. 24 samples (12 legitimate, 12 phishing), 10 samples (5 phishing, 1 matching legitimate, and 4 other legitimates), and 440 samples (120 pairings of legitimate sites, 320 phishing) of real-world Web pages are used in these three experiments, respectively. In our small-scale "12-pairs" and "clustering" experiments, the NCD technique was 100% accurate in grouping a legitimate page with phish targeting that brand. In our large-scale test (which simulates a reasonable client-side anti-phishing scenario), a z-test reveals that the NCD between a phish and its target brand is significantly less than that between two different, legitimate sites.

The results presented in Tables VII and VIII show that our similarity metric is able to distinguish highly similar pages from dissimilar pages, which is the primary goal of this article. However, since our technique seems effective in detecting (in a statistical sense) phishing pages, it is only sensible to inquire what the performance of this technique would be if it were used specifically as a classifier. For this discussion, we assume that we are implementing a simple decision threshold over the NCD values, with no other features.

Anti-phishing researchers often evaluate their algorithms using two interrelated metrics: the true positive rate (true positives divided by the sum of true positives and false negatives) and the false positive rate (false positives divided by the sum of false positives and true negatives). (Note that these measures are

equivalent to sensitivity and (1–specificity) in the medical diagnostic testing literature.) However, the two measures are not considered equally important; false positives are extremely annoying to the average user, and even a fairly low false positive rate may well lead to the abandonment of the system [Dhamija and Tygar 2006; Florencio and Herley 2005; Yih et al. 2006]. In the machine learning literature, whenever such differential error costs are present, it is customary to analyze the performance of a classifier using the Receiver Operating Characteristic (ROC) curve [Provost et al. 1998]. The ROC curve is a plot of true positive rates against false positive rates, under a variety of "tradeoffs" between improving one metric or the other. The ROC curve allows one to visualize the capabilities of a classifier when the analyst is concerned with different costs (penalties) associated with false-negative errors versus false-positive errors, and to compare two different classifiers in the presence of differential error costs.

Our analysis in this section is a different presentation of the same results from Tables VII and VIII. Instead of a z-test, we subject the NCD values to a simple threshold-based decision rule: if the NCD value is less than the threshold, we judge the page in question to be a phish targeting one of our protected pages; if it is greater, we deem the page legitimate. We vary the threshold across an adequate range to produce false positive rates from 0% to roughly 100%. The spacing of thresholds is nonuniform so as to reduce the importance of interpolations in the ROC curves. (The distribution of NCD values for legitimate sites has a much smaller variance than that for phishing pages against their targets, meaning there could be a large jump in FP rates for uniformly spaced thresholds.)

We again compare aggregating NCD value pairs using the arithmetic mean or the maximum value. In addition, we also compare two different one-dimensional compression techniques: the Blocksort algorithm [Burrows and Wheeler 1994] and the LZMA algorithm [Salomon 2007; Pavlov 2009]. (See our discussion at the beginning of Section 6 for our rationale in choosing one-dimensional algorithms.) Plainly, these represent only two out of a great many possible choices of algorithms. At this time, we are not aware of any theoretical rationale for any given compression technique in this class to be more or less effective in computing the NCD value, and so we have simply picked two well-known techniques. These comparisons should thus be considered initial explorations of the impact of different compression techniques on an NCD-based decision threshold classifier.

As can be see in Figures 11–14, an NCD-based decision-threshold classifier would work extremely well on this dataset. The classifier would achieve a true positive rate of roughly 95% with a false positive rate of less than 1.7% for all combinations of compression algorithm and aggregation technique. At the "corner" of the curves, LZMA is slightly superior (TPR = 95.6%, FPR = 0.8%), but this difference is minuscule; it amounts to one less false positive at the true positive rate of 95.6%. This compares favorably with existing anti-phishing techniques (excluding blacklist-based approaches) such as CANTINA [Zhang et al. 2007] (TPR = 97%, FPR = 6%), or SpoofGuard [Chou et al. 2004; Cranor et al. 2007; Zhang et al. 2007] (TPR = 91%, FPR = 48%),

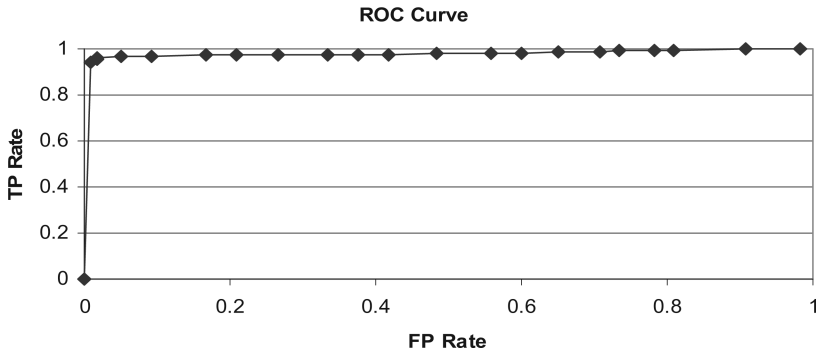

Fig. 11. ROC curve for Blocksort, aggregated by arithmetic mean.

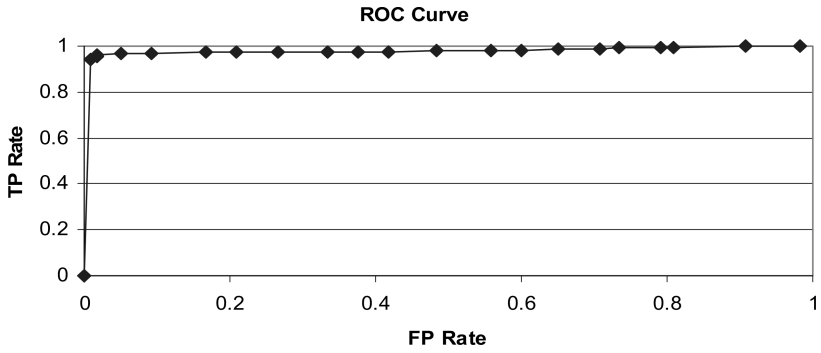

Fig. 12. ROC curve for Blocksort, aggregated by maximum value.

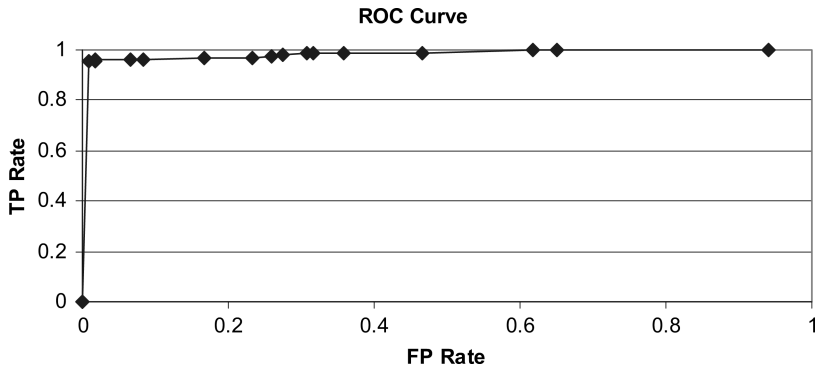

Fig. 13. ROC curve for LZMA, aggregated by arithmetic mean.

and is similar to the more recent hybrid approach in Xiang and Hong [2009] (TPR = 90.06%, FPR = 1.95%).

In Table IX, we compare our results at the "corner" of the curves with CANTINA, Spoof Guard, and the new hybrid method. We present five measures: the first is Cohen's Kappa statistic [Rourke et al. 2001], which measures

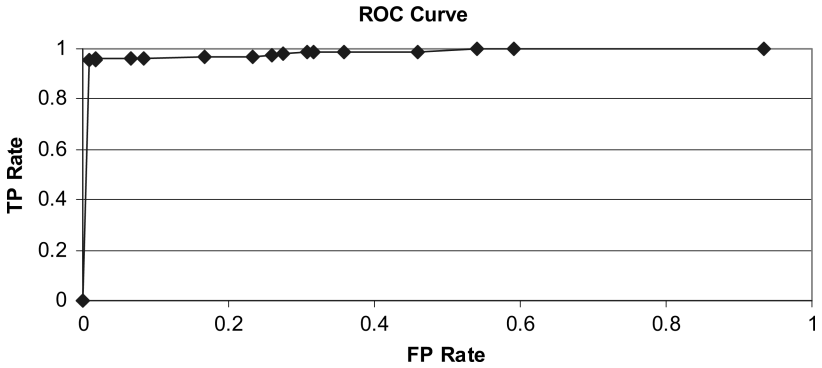

Fig. 14. ROC curve for LZMA, aggregated by maximum value.

Table IX. Comparison against Existing Anti-Phishing Solutions (without Blacklists)

	Kappa	MCC	Precision	Recall	F-Measure
Blocksort	0.9058	0.9082	0.9935	0.9531	0.9729
LZMA	0.9169	0.9193	0.9967	0.9563	0.9761
Hybrid	0.8345	0.8427	0.9904	0.9006	0.9434
CANTINA	0.91	0.9104	0.9417	0.97	0.9557
SpoofGuard	0.43	0.47	0.6547	0.91	0.7615

the chance-corrected agreement between the actual and predicted classification. Chance-correction means that the Kappa statistic explicitly accounts for classification biases due to uneven class distributions. Matthew's Correlation Coefficient (MCC) is equivalent to Pearson's correlation coefficient for binary data [Baldi et al. 2000]. Precision and Recall are normally presented together, as with TP and FP rates; precision represents the fraction of examples labeled positive that were in fact positive, while recall is again the true positive rate. Finally, the F-Measure is computed as 2*(precision*recall)/(precision + recall). (Technically, this is the F-1 measure.) As can be seen, our similarity technique with either compressor (there was no difference between averaging and maximum value) compares favorably with the three existing techniques on all five measures. Additionally, LZMA is slightly superior to Blocksort, but only by a small amount. These results also support our earlier finding that our similarity metric effectively discriminates between similar and dissimilar Web sites. They also indicate that our technique is at least potentially robust against different choices of compression algorithms. This is an important practical finding, as different algorithms may have radically different computational demands (in both time and space), rendering some ineffective on mobile Internet-enabled devices. LZMA, for instance, is a fast and memory-efficient algorithm [Salomon 2007; Pavlov 2009].

6.5 Robustness against Countermeasures

We have argued that our similarity technique could be robust against phishing countermeasures because it does not use localized features. In this section, we

will test this claim using obfuscations based on image processing techniques. This analysis will also indicate how our similarity technique will perform in general when a Web page is obfuscated, either deliberately or accidentally. However, the key characteristic of obfuscations employed by phishers is that they are *deliberately chosen* by these adversaries. The sheer variety of possible obfuscations that an adversary could employ to alter the phishing image has never been considered in the limited existing literature. Human experts readily recognize the difference between innocent similarity and fraud; crafting a computer-based technique to do so is another matter entirely.

Our threat model for these experiments assumes that the phisher will attempt to evade our similarity technique by manipulating the phishing page. Note, however, that our analysis of the phishing scam in Section 5.3 leads us to argue that changes that are noticeable to the human being lead to the failure of the scam. Thus, the changes the phisher makes must pass unnoticed. Specifically, while the phisher's goal is still to hijack a known brand by visually mimicking the page, they will also attempt to introduce discrepancies that are not visible to the human viewer but are significant to the system evaluating the similarity metric. Clearly, the phisher has a wide variety of options for pursuing this objective and it is impossible to visualize all of the possibilities. Therefore, in this section, we will seek to provide some initial explorations of this possibility. The results of these explorations seem to hold over a wide range of Web pages; however, in this section, the results will be given for a single (legitimate plus phish) pair, and only for the NCD average value, for the sake of brevity. The pair was randomly selected from the set of pairs with low NCD values before obfuscation; this allows us to observe how progressively increasing the distortion impacts the NCD values. In these trials, we utilize the LZMA compressor as it seems to perform slightly better than Blocksort in the anti-phishing classifier, and is known to be fast and memory-efficient.

We assume that the most likely attack possibilities are to change "small" details in the image; that is, in general, the phisher will work at the pixel level. Broadly, these types of changes can be characterized into two types:

- Nonstructural distortions (such as luminance changes, contrast changes, chromatic distortions, spatial shifts, etc.)
- Structural distortions (such as noise contamination, blurring, JPEG blocking, wavelet ringing, etc.)

Structural Distortions are unlikely to be effective attacks as they introduce "unnatural" characteristics into the image. Images representing Web pages are computer generated and tend to not suffer from structural distortions. By contrast, many nonstructural distortions do not degrade image structure or quality and hence are more difficult to detect. Hence in this section, we will first explore the potential impact of introducing such nonstructural distortions into phish as the detection of these types of distortions is not perfect. Nonstructural distortions based upon spatial shifts will not be actively explored as the compression techniques used in this paper are robust against spatial shifts which do not impact luminance, contrast, or chromatic

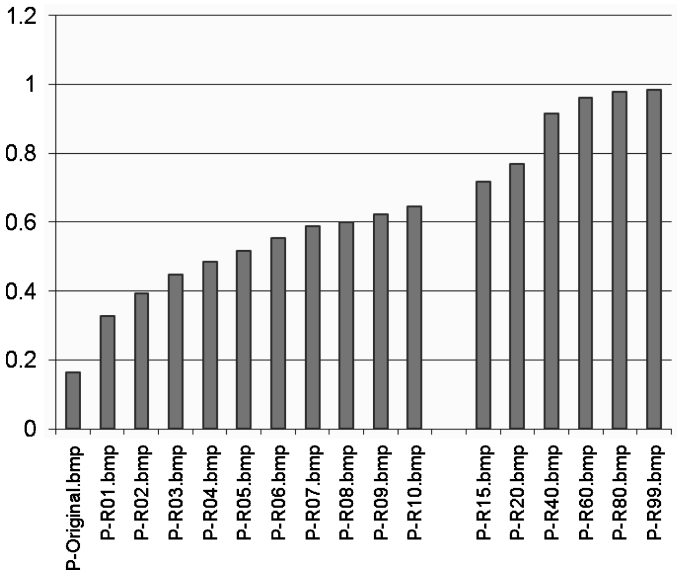

Fig. 15. The effects of local noise on NCD values.

characteristics. Unless otherwise stated, the legitimate image is left unaltered in these experiments.

6.5.1 *Nonstructural Distortions.* Our design for this experiment employs the decision-threshold classifier explored in Section 6.4. As this is only a preliminary exploration, we are only comparing a single legitimate page against one phish targeting it. We introduce controlled levels of nonstructural distortion into the image, and seek to determine what level of distortion results in an NCD value exceeding the "best" decision threshold (from Figures 11–14). We then have two judges visually compare the original and obfuscated phish. For this experiment, the nonstructural distortion is a replacement of a pixel value with one of its immediate neighbors (i.e., one of the eight pixels surrounding the chosen one in a 3×3 convolution mask). We control the level of distortion by varying the fraction of pixels chosen for replacement in the image, from 1% to 99%.

Figure 15 demonstrates a monotonic relationship between the level of the distortion and the NCD value. In the left-hand side of the figure, the NCD value rises steadily; the x-axis labels "P-Rxx" encode the level of distortion, which begins at 1%, and rises by one percentage point for each category on the left-hand portion of the plot. In the right-hand portion, the NCD value approaches and exceeds the decision threshold value. Using thresholds determined from Section 6.4, the phish is still correctly classified with 40% distortion, but becomes a false-negative error with 60% distortion. We then submit the legitimate and phish pages to our two judges, who unanimously agree that the phish no longer mimics the legitimate Web page at 60% distortion—*nor at 40%*. In Figure 16, we present the phish before and after 40% distortion; we believe it

Fig. 16. (a) Phish before 40% of the pixels have been changed. (b) Phish after 40% of the pixels have been changed.

is clear that the phisher would now have failed in their principal objective of visually mimicking the legitimate page.

6.5.2 *Structural Distortions.* This experiment explores the issue of detecting "pure" structural distortion. While most distortions have some structural impacts, especially at higher "noise" levels, several types of distortions are considered to be principally structural. Our experimental design is the same as in Section 6.5.1: a controlled level of distortion is introduced into one phish targeting one legitimate page, and we compare the resulting NCD values against the

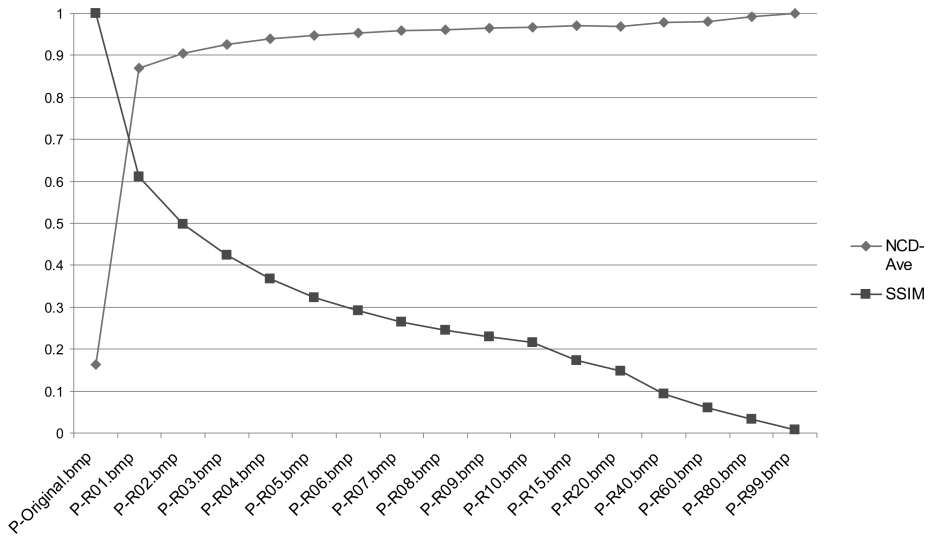

Fig. 17. Impact of Structural Noise on NCD and SSIM values.

decision threshold. The legitimate and phishing pages are then presented to our two judges for comparison. In this experiment, we are introducing random noise into the image; this is done by selecting a fraction of the pixels in an image, and randomizing the RGB color values for those pixels.

Figure 17 again shows a monotonic relationship between the distortion level and the NCD value. The decision-threshold classifier is expected to return a false-negative result for distortion levels above 3%, perhaps indicating that our similarity technique is more sensitive to this type of noise than to nonstructural noise. However, both judges agreed that 3% distortion made the legitimate and obfuscated phish pages distinct (see Figure 18). Again, the level of noise required to fool our similarity technique is great enough to be obvious to a human observer.

These results are consistent with the existing literature on the NCD metric, which generally shows it is quite robust against noise. Cebrián et al. [2007] studied how the NCD is affected by noise in the *symmetric channel* model, in which a random positive integer is added to individual bytes in a file, with the outcomes constrained to the legal domain for values in that file. For instance, genome data is limited to the characters {A, C, G, T}, while text bytes could be any ASCII character, and bytes in a MIDI file can be any value in [0,255]. Both theoretical analysis and empirical testing showed that NCD-based clustering degrades slowly with increasing levels of noise. Granados et al. [2008] examine a noise model in text corpora, in which some percentage of the words in the corpus are distorted by either replacing characters at random, or by replacing characters with asterisks. Six variations on this noise model were tested, and a cluster validity measure was computed from a dendrogram of the corpus. In this case, the NCD was very resilient when the most frequent words were distorted, but less so when words were randomly chosen. Finally, as noted earlier, the NCD seems to be resilient against translations [Bardera et al. 2006].

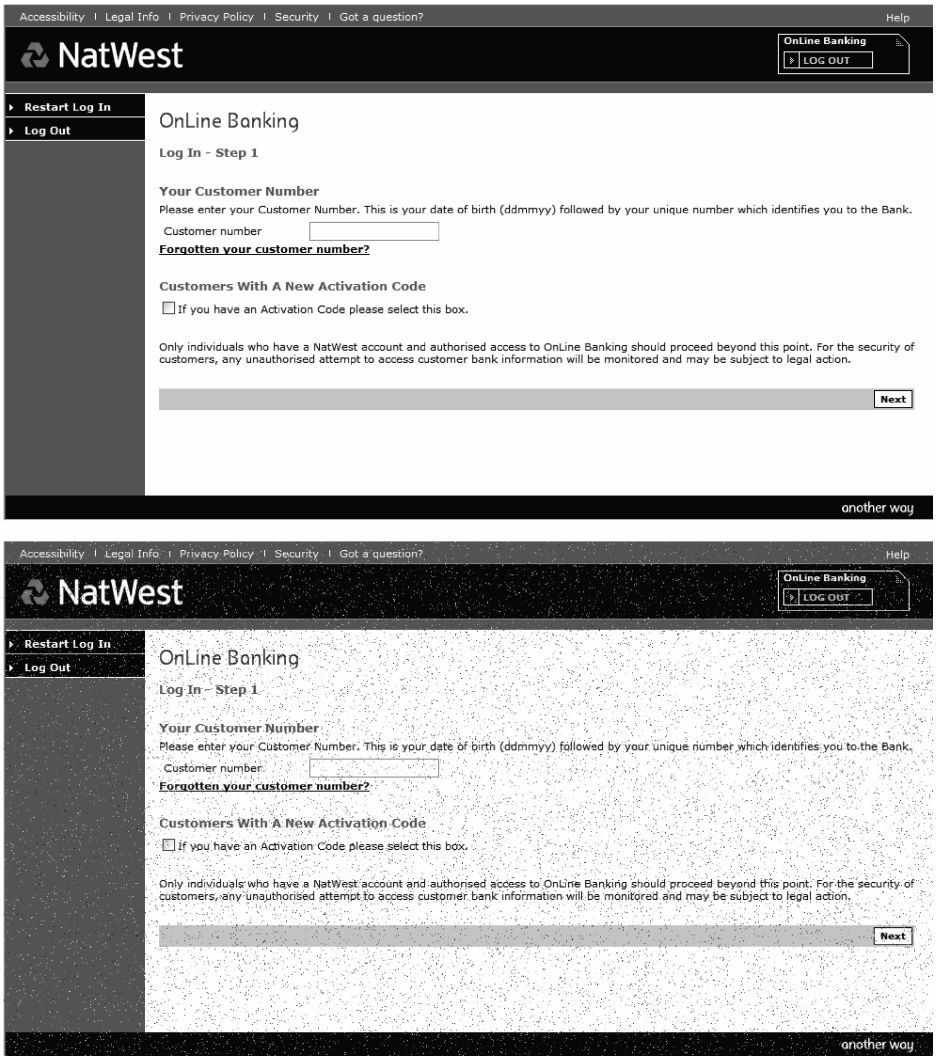

Fig. 18. Phish before and after 3% structural distortion.

In Figure 17, we compute and plot an additional metric, the Structural SIMilarity (SSIM) metric [Wang et al. 2004]. This is an "image quality" metric, which has been empirically shown to "match" human assessments of image quality in a number of experiments, [Sheikh et al. 2006]. We provide this metric to sketch a possible use of our similarity technique as a feature in a robust anti-phishing classifier. We believe this is the most appropriate usage of our similarity technique in a realistic anti-phishing scenario; the NCD feature is highly discriminative, and a well-designed anti-phishing system should incorporate other complementary features to help defeat phisher counter measures. In this experiment, the SSIM values also change dramatically as the distortion level rises.

We selected the SSIM metric because it is effective in detecting both structural and nonstructural distortions. Many nonstructural distortions do not degrade image structure or quality and hence are more difficult to detect. However, SSIM is known to be effective at the detection of global luminance shifts and contrast stretching [Wang et al. 2003]; in addition, a chromatic-variant of SSIM [Toet and Lucassen 2003] has been shown to be effective at detecting many chromatic, nonstructural, distortions. However, as both the legitimate and phishing pages can reasonably be considered montages of interrelated but independent images, it is unclear how successfully SSIM will be at detecting nonstructural distortions. Further empirical analysis will be required to determine if SSIM remains effective in this context.

SSIM, like NCD, is a full reference technique and as the number of brands being phished increases a risk exists that its discrimination performance will not scale. Under these circumstances, it may be necessary to adopt a no-reference image assessment approach [Sheikh et al. 2005; Venkatesh Babu et al. 2007; Brandao and Queluz 2008]. It should not be inferred from these experiments that the visual similarity metric is impervious to phisher countermeasures. It is believed that all phishing page detection techniques have limitations, and the visual similarity metric is likely to have limitations that the determined phisher can expose by undertaking some form of image manipulation on their phishing page. However, just as the phisher can manipulate their page, the authors can manipulate the similarity detection approach to counteract these manipulations. For example, extending the similarity approach to include image reconstruction [Quiney et al. 2006] and seam carving [Avidan and Shamir 2007] components represents interesting "defensive" possibilities.

7. RELATED WORK

In the anti-phishing literature, Fu et al. [2006] and Rosiello et al. [2007] are the most closely related research results to the experiments reported in this paper. Rosiello et al. [2007] analyze the similarity of Web pages by comparing HTML tags in the pages. By extracting and comparing regular subgraphs from the DOM tree representation they construct similarity metrics using Web page structure. Their approaches experienced significant false positive rates; for the identification of 200 phishing Web pages, the approaches experienced false positive rates of 16.90% and 30.29% for the isomorphic subtree identification algorithm and simple tags comparison approach, respectively. Moreover, phishers can avoid this mechanism by using a combination of images to create a phishing Web site that is visually recognized as the legitimate Web site (as we did with eBay in Figure 3). Due to the huge difference in Web page structure, this synthetic phishing Web page easily evades the similarity metric in Rosiello et al. [2007].

Another closely related approach is presented in Fu et al. [2006]. They first convert the Web page into low resolution images, and then extract features from this image (dominant color category and the corresponding coordinate). The Earth Mover's Distance approach (EMD) is then

employed to create a feature-based similarity metric. In their evaluation, they demonstrated 8 phishing Web pages (collected from the authors own email accounts) could be successfully identified within a collection of legitimate Web sites. This approach again appears vulnerable to obvious countermeasures.

8. CONCLUSIONS

Web page similarity detection is widely used by many popular applications. However, the existing methods cannot always successfully identify Web pages that humans would perceive as similar. We propose a robust way to evaluate the similarity of Web pages from the viewpoint of their reader. The concepts of Gestalt theory and supersignals provide us with a theoretical rationale for the conjecture that Web pages must be treated as indivisible entities (i.e., a whole) to be congruent to human perceptions. We use the domain of anti-phishing technology to derive test scenarios for our experiments, as visual similarity between a phishing page and its target is an essential part of the phishing scam. In a series of experiments, we have demonstrated that that we are able to consistently discriminate between similar and dissimilar Web pages. In a real-world case study, we showed that our approach is highly effective at detecting similar Web pages, in particular for anti-phishing applications. We hope to develop a novel anti-phishing system on this foundation; this will necessarily address issues including the sensitivity and specificity of the NCD technique; possible improvements through combinations with other sources of evidence (SSIM is only one possibility); and user-interface design.

In future work, we will undertake a more complete evaluation of the robustness of our NCD similarity technique against countermeasures phishers could employ in the future. While we think that it will be extremely difficult for phishers to evade the NCD similarity technique, we believe it is still important to seek empirical evidence to support this statement. We will attempt to quantify the robustness of the NCD similarity technique using a variation of mutation testing. A mutation will be defined as any operation that alters a Web page to avoid our similarity identification. These could include changes in image color, object coordinates (i.e., rearranging DOM elements), icon resolution, and textual contents (i.e., garbage text, out-of-context phrases, etc.). In mutation testing, the tester attempts to find a test suite that "kills" (i.e., detects) each mutation; in this approach multiple variants of a single program are developed, each containing one or more mutations. The effectiveness of a test suite is measured by the fraction of "mutant" programs killed. In this application, we will be attempting to quantify the effectiveness of a single test (NCD similarity) on a population of mutants. We will also seek an understanding of what mutations (if any) are able to evade this test, and hence of the limitations of the NCD similarity technique. We will also investigate the implementation issue of *which* compressor(s) are most effective for detecting similarity in rendered Web sites (and more generally, for clustering color images). As we noted, there is little to no guidance currently available on this issue.

REFERENCES

ANDRESEN, D., YANG, T., EGECIOGLU, O., IBARRA, O. H., AND SMITH, T. R. 1996. Scalability issues for high performance digital libraries on the World Wide Web. In *Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries.*

APWG. 2008. Phishing Attack Trends Report (Jan.). Anti-Phishing Working Group, http://www.antiphishing.org.

APWG. 2009. APWG. The Anti-Phishing Working Group, http://www.antiphishing.org.

AVIDAN, S. AND SHAMIR, A. 2007. Seam carving for content-aware image resizing. *ACM Trans. Graph. 26,* 3, 10, 1–9.

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. F., AND NIELSEN, H. 2000. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinform. Rev. 16,* 5, 412–424.

BARDERA, A., FEIXAS, M., BOADA, I., AND SBERT, M. 2006. Compression-based image registration. In *Proceedings of the IEEE International Symposium on Information Theory*.

BATISTA, L. V., MEIRA, M. M., AND CANALCANTI JR., N. L. 2005. Texture classification using local and global histogram equalization and the Lempel-Ziv-Welch algorithm. In *Proceedings of the 5th International Conference on Hybrid Intelligent Systems*.

BELL, T., CLEARY, J., AND WITTEN, I. 1984. Data compression using adaptive coding and partial string matching. *IEEE Trans Comm. 32,* 4, 396–402.

BRANDAO, T. AND M. P. QUELUZ. 2008. No-reference image quality assessment based on DCT domain statistics. *Signal Process. 88,* 4, 822–833.

BRODER, A. Z., GLASSMAN, S. C., MANASSE, M. S., AND ZWEIG, G. 1997. Syntactic clustering of the Web. *Comput. Netw. ISDN Syst. 29*, 8–13, 1157–1166.

BURROWS, M. AND WHEELER, D. J. 1994. A block-sorting loss less data compression algorithm. Tech. rep., Digital Systems Research Center.

CAI, D., YU, S., WEN, J. R., AND MA, W. Y. 2003. Extracting content structure for Web pages based on visual representation. In *Proceedings of the 5th Asian-Pacific Web Conference on Web Technologies and Applications*. Lecture Notes in Computer Science, vol. 2642, 406–417.

CEBRIAN, M., ALFONSECA, M., AND ORTEGA, A. 2005. Common pitfalls using the normalized compression distance: What to watch out for in a compressor. *Comm. Inform. Syst. 54*, 367–384.

CEBRIAN, M., ALFONSECA, M., AND ORTEGA, A. 2007. The normalized compression distance is resistant to noise. *IEEE Trans. Inform. Theory 53,* 5, 1895–1900.

CERNIAN, A., CARSTOIU, D., AND OLTEANU, A. 2008. Clustering heterogeneous Web data using clustering bv compression validity. In *Proceedings of the International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*.

CHAITIN, G. 1. 1987. *Algorithmic Information Theory.* Cambridge University Press.

CHARIKAR, M. S. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the ACM Symposium on Theory of Computing*.

CHOU, N., LEDESMA, R., TERAGUCHI, Y., BONEH, D., AND MITCHELL, J. C. 2004. Client-side defense against Web-based identity theft. In *Proceedings of the Annual Network and Distributed System Security Symposium.*

CILIBRASI, R. AND VITANYI, P. M. B. 2005. Clustering by compression. *IEEE Trans. Inform. Theory 51,* 4, 1523–545.

CRANOR, L., EGELMAN, S., HONG, J., AND ZHANG, Y. 2007. Phinding phish: Evaluating anti-phishing toolbars. In *Proceedings of the Annual Network and Distributed System Security Symposium.*

DEAN, J. AND HENZINGER, M. R. 1999. Finding related pages in the World Wide Web. *Comput. Netw. 31*, 11–16, 1467–1479.

DELANY, S. J. AND BRIDGE, D. 2006. Textual case-based reasoning for spam filtering: A comparison of feature-based and feature-free approaches. *Artif. Intell. Rev. 26*, 75–87.

DHAMIJA, R. AND TYGAR, J. D. 2006. Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

DHAMIJA. R. AND TYGAR, J. D. 2005. The battle against phishing: Dynamic security skins. In *Proceedings of the Symposium on Usable Privacy and Security.*

DORNER, D. 1997. *The Logic of Failure*. Metropolitan Books, Cambridge. MA.

DSLReports.com. 2008. Phish tracker. http://www.dslreports.com/phishtrack.

eBay. 2008. Weleome to eBay. https://signin.ebay.com/ws/eBayISAPl.dll?Signln&ru=http%3A %2F%2F.

EMIGH, A. 2005. Online identity theft: Phishing technology, chokepoints and countermeasures. Tech rep., Radix Labs.

FELDT, R., TORKAR, R., GORSCHEK, T., AND AFZAL, W. 2008. Searching for cognitively diverse tests: Towards universal test diversity metrics. In *Proceedings of the IEEE International Conference on Software Testing Verification and Validation Workshop*.

FETTE, I., SADEH, N., AND TOMASIC, A. 2007. Learning to detect phishing emails. In *Proceedings of the International World Wide Web Conference.*

FLORENCIO, D. AND HERLEY, C. 2005. Stopping a phishing attack, even when the victims ignore warnings. Tech. rep., Microsoft Research., Redmond, WA.

FU, A. Y., WENYIN, L., AND DENG, X. 2006. Detecting phishing Web pages with visual similarity assessment based on earth mover's distance (EMD). *IEEE Trans. Depend. Secure Comput. 3*, 4, 301–311.

GORDON, I. E. 2004. *Theories of Visual Perception, 3rd Ed*. Psychology Press, New York.

GRAHAM, L. 2008. Gestalt theory in interactive media design. *Human. Soc. Sci. 2*, 1, 3.1–3.12.

GRANADOS, A., CEBRIAN, M., CAMACHO, D., AND RODRIGUEZ, F. B. 2008. Evaluating the impact of information distortion on normalized compression distance. In *Proceedings of the 2nd International Castle Meeting on Coding Theory and Applications.*

HAVELIWALA, T. H., GIONIS, A., KLEIN, D., AND INDYK, P. 2002. Evaluating strategies for similarity search on the Web. In *Proceedings of the International World Wide Web Conference.*

HEINTZE, L. 1996. Scalable document fingerprinting. In *Proceedings of the USENIX Workshop on Electronic Commerce*.

HENZINGER, M. 2006. Finding near-duplicate Web pages: A large-scale evaluation of algorithms. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.*

HESCOTT, B. AND KOULOMZIN, D. 2007. On clustering images using compression. Tech. rep., Computer Science Department, Boston University.

HOU, 1. AND ZHANG, Y. 2003. Utilizing hyperlink transitivity to improve Web page clustering. In *Proceedings of the Australasian Database Conference.*

KALVIAINEN, M. 2007. The role of sign elements in holistic product meaning. In *Proceedings of the SeFun International Seminar on Design Semiotics in Use.*

KEPES, G. 1944. *Language of Vision*. Paul Theobald, Chicago, IL.

LAN, Y. AND HARVEY, R. 2005. Image classification using compression distance. In *Proceedings of the 2nd International Conference on Vision, Video and Graphics.*

LI, M. AND VITANYI, P. 1997. *An Introduction to Kolmogorov Complexity and its Applications,* 2nd Ed. Springer-Verlag, Berlin.

LI, M. AND ZHU, Y. 2006. Image classification via LZ78-based string kernel: A comparative study. In *Proceedings of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science, vol. 3918, 704–712.

LI, M., CHEN, X., LI, X., MA, B., AND VITANYI, P. M. B. 2004. The similarity metric. *IEEE Trans. Inform. Theory 50,* 12, 3250–3264.

MACEDONAS, A., BESIRIS, D., ECONOMOU, G., AND FOTOPOULOS, S. 2008. Dictionary based color image retrieval. *J. Vis. Comm. Image Rep. 19,* 464–470.

MACK, A. AND ROCK, I. 1998a. *lnattentional Blindness*. MIT Press.

MACK, A. AND ROCK, I. 1998b. Inattentional blindness: Perception without attention. In *Visual Attention*, R. D. Wright Ed., Oxford University Press, Oxford, UK, 55–76.

MACKAY, W. E. 1991. Triggers and barriers to customizing software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

MANBER, U. 1994. Finding similar files in a large file system. In *Proceedings of the USENIX Winter Technical Conference.*

MCCALL, I. 2007. Gartner survey shows phishing attacks escalated in 2007: More than $ 3 Billion lost to these attacks. Gartner, Inc., http://www.gartner.comit.pclgc.jsp?id=565l25.

MICROSOFT. 2009. Get Internet Explorer 7. http://www.microsoft.eom/windows/internet-explorer/ie7.

MOZILLA. 2008. FireFox Web Brower. http://www.mozilla.com.en-US firefox/.

MOZILLA. 2009. Thunderbird—Reclaim Your Inbox. http://www.mozilla.com/en-US/thunderbird.
NETCRAFT. 2009. Netcraft Anti-Phishing Toolbar. http: toolbar.netcraft.com.
OFUONYE, E., BEATTY, P., DICK, S., AND MILLER, J. 2010. Prevalence and classification of Web page defects. *Online Inform. Rev. 34*, 1, 160–174.
OPENDNS. 2008. PhishTank. Join the fight against phishing. http://www.phishtank.com/phish_archive.php.
PAVLOV, I. 2009. 7z Format. 7Zip, http://www.7-zip.org/.
PROVOST, F., FAWCETT, T., AND KOHAVI, R. 1998. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the International Conference on Machine Learning*.
QUINEY, H. M., NUGENT, K. A., AND PEELE, A. G. 2006. Iterative image reconstruction algorithms using wave-front intensity and phase variation. *Optics Lett. 30*, 13, 1638–1640.
ROSIELLO, A. P. E., KIRDA, E., KRUEGEL, C., AND FERRANDI, F. 2007. A layout-similaritv-based approach for detecting phishing pages. In *Proceedings of the IEEE International Conference on Security and Privacy in Communications Networks and the Workshops.*
ROURKE, L., ANDERSON, T., GARRISON, D. R., AND ARCHER, W. 2001. Methodological issues in the content analysis of computer conference transcripts. *Int. J. Artif. Intel. Educ. 12*, 8–22.
RSA. 2009. RSA Identity Protection and Veritication Suite. http://www.rsa.eom!node.aspx?id=30l7.
SALOMON, D. 2007. *Data Compression: The Complete Reference*. Springer-Verlag.
SHEIKH, H. R., BOVIK, A. C., AND CORMACK, L. K. 2005. No-reference quality assessment using natural scene statistics JPEG2000. *IEEE Trans. Image Process. 14,* 11, 1918–1927.
SHEIKH, H. R., SABIR, M. F., AND BOVIK, A. C. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process. 15*, 11, 3449–3451.
SHEN, D., CHEN, Z., YANG, Q., ZENG, H.-J., ZHANG, B., LU, Y., AND MA, W.-Y. 2004. Web-page classification through summarization. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.*
STRIMMER, K. AND VON HAESELER, A. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Molec. Biol. Evol. 13,* 7, 964–969.
TOET, A. AND LUCASSEN, M. P. 2003. A new universal colour image fidelity metric. *Displays 24*, 4–5, 197–207.
VENKATESH BABU, R., SURESH, S., AND PERKIS, A. 2007. No-reference JPEG-image quality assessment using GAP-RBF. *Signal Process. 87,* 6, 1493–1503.
WANG, Y. AND KITSUREGAWA, M. 2002. Evaluating contents-link coupled Web page clustering for Web search results. In *Proceedings of the International Conference on Information and Knowledge Management*.
WANG, Z., BOVIK, A. C., SHEIKH, H. R., AND SIMONCELLI, E. P. 2004. Image quality assessment: From error visibility to structural simihrity. *IEEE Trans. Image Process. 13,* 4, 600–612.
WANG, Z., SIMONCELLI, E. P., AND BOVIK, A. C. 2003. Translation insensitive image similarity for image quality assessment. In *Proceedings of the IEEE Asilomar Conference on Signals, Systems and Computers*.
WERTHEIMER, M. 1944. *Gestalt Theory*. Hayes Barton Press, New York.
WU, C.-T., CHENG, K.-T., ZHU, Q., AND WU, Y.-L. 2005. Using visual features for anti-spam filtering. In *Proceedings of the IEEE International Conference on Image Processing.*
WU, M., MILLER, R. C., AND GARFINKEL, S. L. 2006. Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*
XIANG. G. AND HONG, J. 2009. A hybrid phish detection approach by identity discovery and keywords retrieval. In *Proceedings of the International World-Wide Web Conference*.
YAHOO. 2009. Yahoo! Personalized Sign-In Seal. https://protect.login.yahoo.com.
YIH, W., 1. GOODMAN, J., AND HULTEN, G. 2006. Learning at low false positive rates. In *Proceedings of the 3rd Conference on Email and AntiSpam.*
ZHANG, Y., HONG, J. AND CRANOR, L. 2007. CANTINA: A content-based approach to detecting phishing Web sites. In *Proceedings of the International World-Wide Web Conference*.
ZIV. J. AND LEMPEL, A. 1977. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory 23*, 3, 337–343.

Received January 2009; revised June 2009, December 2009; accepted December 2009