Does Domain Highlighting Help People Identify Phishing Sites?

Eric Lin, Saul Greenberg, Eileah Trotter, David Ma and John Aycock
Department of Computer Science
University of Calgary, Calgary, Alberta, Canada T2N 1N4
{linyc, saul.greenberg, davma, aycock}@ucalgary.ca, eileaht@gmail.com

ABSTRACT
Phishers are fraudsters that mimic legitimate websites to steal user's credential information and exploit that information for identity theft and other criminal activities. Various anti-phishing techniques attempt to mitigate such attacks. *Domain highlighting* is one such approach recently incorporated by several popular web browsers. The idea is simple: the domain name of an address is highlighted in the address bar, so that users can inspect it to determine a web site's legitimacy. Our research asks a basic question: how well does domain highlighting work? To answer this, we showed 22 participants 16 web pages typical of those targeted for phishing attacks, where participants had to determine the page's legitimacy. In the first round, they judged the page's legitimacy by whatever means they chose. In the second round, they were directed specifically to look at the address bar. We found that participants fell into 3 types in terms of how they determined the legitimacy of a web page; while domain highlighting was somewhat effective for one user type, it was much less effective for others. We conclude that domain highlighting, while providing some benefit, cannot be relied upon as the sole method to prevent phishing attacks.

Author Keywords
Phishing, domain highlighting, usable security.

ACM Classification Keywords
H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces. K.6.5 Security and protection.

General Terms. Security, Human Factors, Design.

INTRODUCTION
Phishing attacks are a method whereby fraudsters try to steal user's credentials – user names, passwords, bank accounts, credit card numbers, and so on – via technical means and by social engineering [6]. Typical phishing attacks involve a message (e.g., email) sent to a user, purporting to be from a legitimate entity like a bank; the message would try to convince the user to follow a URL to the phisher's web site. At this web site, the phisher visually mimics the real web site, and will save and later exploit the sensitive information sent to that fake web site by the user/victim.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

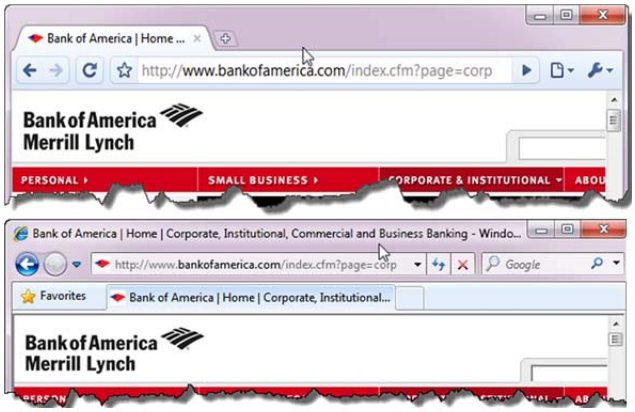

Figure 1. Domain name highlighting in Google Chrome (top) and Internet Explorer (bottom)

Of course, the web security landscape is not blind to phishing attacks, and consequently many methods try to thwart them. Some methods are technical ones that work under the covers, such as creation of blacklists that identify suspect phishing sites, and filtering of suspect sites by internet providers. Other methods try to help web users identify suspect sites, as discussed in the related work section. Yet in spite of these efforts to combat the threat of phishing, recent statistics paint a bleak picture of their overall effectiveness. The 2008 Gartner study [10] estimated that upwards of 5 million people in the U.S. alone were subject to financial losses due to phishing (non-financial losses also exist, but [10] did not consider these), even though safe-browsing features were used by 36% of the people surveyed. Furthermore, at least 9400 phishing attempts were identified in the month of August 2010 alone [12]. Thus phishing continues to occur, and users continue to fall victim despite efforts to deter attacks.

One problem is that a well-crafted phishing site will be visually identical to the original site. Indeed, the sole telltale sign of such a site may only be that its domain name is incorrect, something users can easily miss (e.g., [6]. The new user-centric anti-phishing method of *domain highlighting* [15] exploits this: it visually enhances the domain name portion of the web address in the browser's address bar. Figure 1 illustrates how domain highlighting is incorporated in the default interface of recent releases of Google Chrome (top) and Microsoft's Internet Explorer [15] (bottom). In both cases, the host domain is drawn in black, while the rest of the address is rendered in gray. The

only visible difference between the two browsers is that Explorer does not highlight the "www" prefix.

Domain highlighting is based on two assumptions which, if correct, would make it a critical anti-phishing measure:

1. Users can recognize legitimate domain names, likely from their knowledge of the web, their expectations of what a well-formed domain name should be, and their prior experiences visiting particular web sites.
2. Users will naturally attend to the address bar as part of their normal browsing practice, where they will quickly examine the highlighted domain name portion of the address to judge the legitimacy of the web site.

We ask in this paper if domain highlighting really meets its potential. Our objective is to provide insight into the effectiveness of domain highlighting in user identification of phishing sites. We first review related work that investigates phishing security as part of interface design. We then describe our study, its methodology, and results. We identify limitations and faults with domain highlighting and its assumptions, which imply possible directions for security developers who wish to improve the effectiveness of phishing deterrents. We also identify three types of users. One type consistently cited brand, content and previous experiences with the web page as major factors of trust; a second type scrutinized the address bar as part of their evaluation; the third type was a hybrid between these two.

RELATED WORK

Much prior work on anti-phishing security focuses on its technical aspects. Yet many technical solutions are often thwarted because they expect users to understand them (most cannot), or because phishers and other spammers used social engineering methods to fool users into taking actions they would not normally do. This is why researchers are now seriously focusing on the gap between technical approaches to security *vs.* user interface design (e.g., ACM SOUPS conference focuses on user aspects of security).

Within the context of phishing and the web, the majority of attention on user-oriented security methods has been on security toolbars and visual security indicators. In general, these methods situate visuals (such as icons) on the edges of the browsing window or in the address bar that display the security status of a particular web site (e.g., whether it is SSL). Where a phishing site is blacklisted, more prominent and active notifications may be used [3].

Yet why, despite the existence of such measures, do phishing attacks continue to be effective against users? Wu et al. [18] identified many drawbacks with existing security toolbars. First, people rarely noticed them because these were located in the peripheral area of a browser. Second, users did not necessarily care about the toolbar even if they did take notice. Third, the toolbar often mis-classified sites (e.g., legitimate sites as phishing sites), which led users to mistrust the toolbar. Other visual methods were similarly ineffective: subjects typically ignored visual lock icons located at the periphery of their attention, and continued their browsing even when confronted with a blatant pop-up warnings for fraudulent certificates [2,16].

The problem is that the intuition used to develop the above visual security cues assumes that what works for protecting the technically-knowledgeable specialist must, by extension, also work for the average consumer. This is not true. When researchers examined what people actually use to form their opinion of trust, they found that non-technical people judge security primarily from the visual content of the page itself and its relevance to their own task [9,6,4]. Seldom do they consider security icons and toolbars.

Other approaches move security indicators directly into the user's field of view, although this has to be carefully done as it could affect the usability of a browser. With enhanced sign-in indicators, first-time users of a legitimate site are asked to choose an image and/or enter some personal text during login. That image and text are then displayed on all subsequent logins. If they are missing or wrong, then he or she should suspect that they have been misdirected to a visually similar but fraudulent phishing site [13].

Domain highlighting is yet another approach added to the arsenal of security indicators. To our knowledge, the actual effectiveness of domain highlighting has not been examined in detail. While our study here is somewhat similar in nature to the above-mentioned studies, it differs from them in three significant ways. First, ours is the first formal study of domain highlighting. Second, our methodology masked our interest in domain highlighting (following fraud experiment recommendations by [8]), where participants only knew that we were conducting a security study of some kind. This allowed us to evaluate domain highlighting in the context of how users may behave outside of a lab setting. Third, during a second evaluation round, we did tell people to focus their attention on the address bar, but we did not tell them about domain names or how to identify phishing sites. This allowed us to evaluate how people used their own knowledge while removing problems associated with attention as found in prior studies of security icons.

DESCRIPTION OF THE EXPERIMENT

We ask: to what extent does domain highlighting aid in identifying phishing attacks? To answer this, our study evaluates the effectiveness of domain highlighting as an anti-phishing measure, where effectiveness is measured by whether a person makes appropriate judgments of the trustworthiness of legitimate and fraudulent web sites.

Our study was a controlled experiment. 22 participants were shown 16 web pages (8 legitimate and 8 phishing pages) in randomized order[1] across two phases. In phase 1, subjects were merely asked to classify how "safe" each page appeared to be. Phase 2 repeated the exercise, except this time they were told to focus on the address bar area.

[1] We also ran a pilot study involving 7 participants to help form and debug our study. The method and results obtained in the pilot are essentially identical to those found in this full study.

Focus of Participants

Participants were *not* told or instructed about domain name spoofing, nor domain highlighting, nor that we were specifically concerned with phishing attacks. Rather, we told them that they would be involved in a study on the visual security aids as generally found in Internet Explorer 8. They were consciously aware that they were being asked to judge a web site's trustworthiness. While this sensitized them to security, it was up to them to decide what informational cues were present and how to use them to evaluate a site's legitimacy. We did this because we wanted to evaluate the effects of domain highlighting in a semi-realistic context. Had we made participants specifically aware of domain highlighting, they would likely have knowledge atypical of everyday users, and they likely would have given that feature far more scrutiny than they would normally do while browsing. In real world contexts, security concerns are secondary goals in relation to the primary goal of completing a specific task [17,10]. Our emphasis on security shifted their primary goal to security. Consequently, we expect our participants to perform better than in a typical browsing scenario where judging security is at best a secondary goal. That is, if domain highlighting performs poorly when people are asked to judge a web site's security, it would certainly perform more poorly under more routine browsing conditions. Thus our study portrays a reasonable *upper bound* on the effectiveness of domain highlighting in standard browsing environments[2].

Participants

22 participants (17 male and 5 female), recruited from the University of Calgary, were compensated with $15 cash. Ages ranged from 19-41 (mean = 27.7, σ=6.2). They comprised 10 graduate engineering students, 2 computer science graduate students, 6 science or social science students, and four university employees or previous university graduates from non-technical fields. None were trained in security. Average weekly computer usage was ~47 hours (σ=24.0, median=49). All regularly banked online, and all were familiar with Internet Explorer. 18 shopped online regularly, 2 occasionally, and 2 not at all.

Method

Each ~45 minute study, conducted in an office, was comprised of an initial interview, 2 study phases, and an exit interview. Interviews were audio recorded.

The initial interview asked about computer experience, prior use of web sites similar to those in our study, and knowledge of security concepts. We wanted to see if any prior experience would affect user performance.

In phase 1, no particular instructions were given on how to evaluate the web sites' trustworthiness. In phase 2, we asked subjects to re-evaluate the web sites' trustworthiness where we directed them to focus on the address bar area (we did not tell them what to look for in that area). If subjects changed their ratings between these two phases, then we could attribute this to their using the extra information supplied by domain highlighting in the address bar. If there was no change, then we could assume that the participant either could not recognize or could not exploit the extra information supplied by domain highlighting.

In each phase, participants were shown a series of 16 web pages representing web sites (presented in randomized order across participants) and asked to gauge how "safe" a site was on a 5-point scale. In case subjects were unfamiliar with a site, we explained generally what each site was for (e.g., online banking, social networking, shopping). For our analysis, we collapse their ratings as unsafe (1-2), unsure (3) and safe (4-5)[3]. As they did this, we asked participants to describe their rationale for their ratings as well as what visual element(s) they were currently inspecting on the web page via a think-aloud process.

In the exit interview, we explained domain highlighting and how it could help people evaluate phishing attacks. Participants were asked if they knew about this feature, whether they had noticed it[4], and were asked for their thoughts on its potential effectiveness and what could be improved.

Materials

System. We used a standard computer, Windows XP and Internet Explorer 8. The only change was to point the web browser to a proxy server, which cached and presented the spoofed domain name / URL address in the participant's web browser for the fraudulent web pages. Legitimate pages were loaded by the proxy server directly from the real location and displayed without modifications.

Web sites. We used 16 different web pages representing 16 web sites. 8 were legitimate (Table 1a), while the other 8 were simulated phishing pages that we created along with a spoofed domain name / URL address (Table 1b). As common in 'better' phishing sites, each phishing page was a perfect replica of the original site. Thus the URL, including its domain name, was the only discernible visual difference between a fraudulent and its legitimate page.

For realism, the 16 web sites in Table 1 were selected according to one or both of the following criteria.

1. *Familiarity to participant.* We wanted participants to be either familiar with the displayed site or other sites like it. Thus we selected popular sites (as described in the rightmost columns of Table 1) within our geographic region, where those sites were also representative of online tasks our target group would likely encounter during routine web browsing. If participants were

[2] Identifying domain highlighting as a study goal and instructing people on typical phishing attacks would produce a 'true' upper bound, but one likely not reachable (on average) in the real world.

[3] A 5-point rating reasonably captures a person's judgement. For the analysis, however, we assume the collapsed 3-point scale is a reasonable approximation of the likelihood of that person actually using the website.

[4] While eye-tracking could tell us if people looked at the address bar (see [16]), self-disclosure reveals what people actually looked for.

unfamiliar with a particular site as shown by the web page they saw, we described (at a conceptual level) the kind of organization hosting that site, and then presented a sample scenario illustrating the purpose of that site.

2. *Frequently targeted market sector.* The Anti-Phishing Working Group [1] names 'payment services', 'financial', 'auction', and 'retail' as heavily targeted areas, while PhishTank [12] adds social networking. As described in the right columns, most pages in Table 1a/b are login pages to sites that ask for account credentials. If our fraudulent pages were actual phishing attacks, then people entering information into them would likely incur financial misappropriation, identity theft, or both.

Phishing Domain Names / URLs. Phishers use domain name obfuscation techniques to fool people into believing that the domain name / URL in the address bar is not fraudulent. For our fraudulent sites, we crafted URLs representative of typical obfuscation techniques (see Table 1b) as listed below. While we name these methods below, the community has no consistent names for most of them.

Similar-name attacks are addresses that are similar sounding to the legitimate address. For example, the fraudulent amazon.checkingoutbooksonline.ca appears similar to the legitimate www.amazon.ca, even though the second-level domain portion is quite different (checkingoutbooksonline *vs.* amazon).

IP-address attacks display a cryptic IP address as the domain rather than the real domain name, e.g., http://192.168.111.112/login *vs.* http://www.facebook.com/login.

Letter substitution attacks substitute one or more characters in the domain name with a character that is visually similar in appearance. These can be seen as crude homograph attacks [4]. An example is 'www.uca1gary.ca' rather than 'www.ucalgary.ca' (note the number '1' instead of the letter 'l'). The user may miss this in a quick glance.

Complex URL attacks are addresses that span the length of the address bar or that may contain nonsensical characters, making interpretation of the URL difficult.

Metrics

Our primary quantitative metric used to evaluate the effectiveness of domain name highlighting was derived by comparing user ratings (safe *vs.* unsafe) to web page legitimacy (legitimate *vs.* fraudulent). This led to the measures below, illustrated in Table 2 as a matrix.

Table 1a. Legitimate web pages (Row order matches top half of Figures 2 and 3)

URL (abbreviated here with '…')	**Company / Description / Information requested**
https://www1.royalbank.com/ cgi-bin/rbaccess/…	**Royal Bank of Canada.** Bank site's log-in page. Requests user account credentials.
http://clothing.shop.ebay.com/ i.html?_sacat=11450&_nkw…	**eBay.** Online auction for posted items; browsing, purchasing and paying for various items.
http://alumni.lib.ucalgary.ca:3048/ login?url=http://proquest.umi.com/...	**UC-library.** University library login page to access online library services. Requests user account credentials that could be used to access other university services.
https://www.google.com/ accounts/ServiceLogin?uilel=3&service=youtube…	**Youtube.** Video-sharing login page. Requests user account credentials.
https://auth.me.com/ authenticate?service=…	**Mobile Me.** Login page to syncronize user devices, e.g., iPod, iPhone, iPad, and laptop or desktop computer. Requests user account credentials.
http://websms.fido.page.ca/ 2way/	**Fido.** Phone company login page for text messaging services. Requests user account credentials.
https://canada.frenchconnection.com/ login.htm?returnUrl=/…	**French Connection.** Clothes store log-in page for online shopping. Requests user account credentials.
http://www.facebook.com/ r.php?invid=10000…	**Facebook (invitation).** Social network invitation sent by another person. Requests user account credentials.

Table 1b. Fraudulent web pages (Row order matches bottom half of Figures 2 and 3)

Phishing Type	**Spoofed URL** *long URLs abbreviated here with '…'*	**Original (non-spoofed) URL** *for comparison purposes*	**Supposed Company / Description**
IP address	http://192.168.111.112/ login	http://www.facebook.com/	**Facebook.** Social network login page. Requests user account credentials.
similar	http://www.easyweb.td-canadatrust.ca/	https://easywebcpo.td.com/ waw/idp/login.htm?...	**Canada Trust Bank.** Bank site's log-in page. Requests user account credentials.
similar	http://www.meebo.webmessenger.com/	http://www.meebo.com/	**Meebo.** Login page. Requests user account credentials for multiple social networking sites.
letter	http://www.uca1gary.ca/ registrar/payment	http://www.ucalgary.ca/ registrar/payment	**UC- Enrollment.** For fee payment. Describes how to pay tuition fees, but doesn't require any personal information.
letter	http://www.paypa1.ca	http://www.paypal.ca	**Paypal.** Login page for payment of goods via bank / credit cards. Requests account credentials.
complex	http://www.amazon.ca.checkingoutbook online.ca/Golden-Mean-Annabel-Lyon/…	http://www.amazon.ca/ Golden-Mean-Annabel-Lyon/	**Amazon.** Online store for browsing and purchasing books and other items.
similar	http://login.hotmailsecure.com	http://login.live.com/ login.srf?wa=wsignin1.0…	**Hotmail.** E-mail/messaging login page. Requests user account credentials.
complex	http://login.flickr.net/ config/login?.src=flickr…	https://login.yahoo.com/config/ login?.src=flickr	**Flickr.** Photo-sharing login page that requests user account credentials.

Table 2: Measurement matrix

		Subject decision	
		Site is safe	**Site is unsafe**
Site type	**Legitimate**	Correct: safe	Wrong: unsafe
	Fraudulent	Wrong: safe	Correct: unsafe

Correct decision is the percentage of participants who correctly classified (as an average) a page as legitimate or fraudulent. It is the sum of the two measures below.

- ***Correct***: ***safe*** occurs when one correctly rates a legitimate web page as safe.
- ***Correct: unsafe*** occurs when one correctly rates a fraudulent web page as unsafe.

Wrong decision is the percentage of participants who incorrectly classified (as an average) a page as legitimate or fraudulent. It is the sum of the two measures below, which are equivalent to Type I (false positive) and Type II (false negative) errors in hypothesis testing.

- ***Wrong: safe*** occurs when one incorrectly rates a fraudulent web page as safe. This is very critical factor, because this incorrect decision directly measures the vulnerability of participants to a phishing attack.
- ***Wrong: unsafe*** occurs when one incorrectly rates a legitimate page as unsafe. This means that the participant will not continue to use the page even though it was, in fact, safe to do so. While a (possibly serious) inconvenience, this error is not a security threat.

Focus

Our quantitative analyses were driven by three questions of interest.

Is there a difference in web page ratings in terms of correct decisions, wrong decisions, wrong:safe, and wrong:unsafe

1. regardless of whether or not subjects were told to pay attention to the URL address?
2. regardless of whether the page is actually legitimate or fraudulent?
3. regardless of the type of phishing method used (letter, IP-address, complex, and letter substitution attacks)?

Question 1 considers differences that happen before and after a person is told to pay attention to the URL, i.e., if they even notice and exploit domain highlighting more effectively after being told to look at that area of the page. Question 2 considers differences between measures when we compare rates on legitimate *vs.* fraudulent pages. Question 3 considers how well people can or cannot detect particular types of domain name attacks.

QUANTITATIVE ANALYSIS AND RESULTS

We use a combination of descriptive statistics plus hypothesis testing.

Hypothesis testing. Our null hypothesis is a rephrasing of Questions 1 and 2 above to state that no such difference exists between decisions over legitimate *vs.* fraudulent pages across the two phases. For hypothesis testing, we used a 2x2 ANOVA repeated measures: Phase (1 *vs.* 2) x Web Page Type (legitimate *vs.* fraudulent*)* with a threshold of $p < .05$ (with a Bonferroni correction of .0125 for post-hoc tests). We calculated frequency values (i.e., parametric proportion data) of correct/unsure/wrong decisions by averaging particular ratings across participants. As is common with proportional data in ANOVA that ranges between 0 and 1, we transformed them to arcsin values.

For correct and unsure decisions, a significant difference was found between the frequency of people's rating of Web Page Type ($F=10.24$, $p = .004$ for correct; $F=11.35$, $p = .003$ for unsure, $df=21$). That is, people rated fraudulent and legitimate web pages differently in terms of their correct / unsure frequency. There was no significant difference between the phases; the fact that they were made aware of the address bar in Phase 2 did not change these frequencies.

For incorrect decisions, an interaction effect existed for the frequency of people's ratings. Post-hoc tests revealed there was a significant difference between people's incorrect frequency ratings of legitimate *vs.* fraudulent sites in both phases ($t=-5.06$, $p=.00$ in Phase 1, $t=-3.61$, $p=.002$ in Phase 2), and between their ratings of fraudulent pages between Phases 1 and 2 ($t=3.96$, $p = .001$). However, there was no difference in people's ratings of legitimate sites between Phases 1 and 2 ($t=-.35$, $p=.73$).

The means describing these differences (and other descriptive statistics) are detailed and discussed below.

Correct *vs.* wrong decisions (Phase 1). Figure 2 is a visualization of each participant's decision data for each web page seen in Phase 1. Each column represents an individual participant, where columns are sorted by participants with the most correct results on the left. Each row represents a particular web page, with the 8 legitimate pages at the top and the 8 fraudulent pages at the bottom. For cross referencing URL and obfuscation type, row order is identical to rows in Tables 1 and 2. Each set of rows in the legitimate and fraudulent section is sorted by pages with the most correct score at the top. Green cells are correct decisions, red are incorrect, and yellow are those rated as unsure. Participant numbers are shown in the topmost row. Participant type (A, AB, or B) - explained in the qualitative results section - is shown on the bottom row.

For legitimate pages, participants rated 54% of them correctly (correct:safe), 15% incorrectly (wrong:unsafe), and were unsure of 31% of the pages. Figure 2, top, illustrates this by the moderate amount of green squares showing the correct ratings.

For fraudulent pages, participants rated 25% of them correctly (correct:unsafe), 57% incorrectly (wrong:safe), and were unsure of 18%. Figure 2,vividly reveals the high number of incorrect ratings shown in red at its bottom.

Correct vs. wrong decisions (Phase 2). Figure 3 is a visualization of how people changed their ratings of a page after they were told to attend to the address bar. White cells are those that were unchanged, or where the new rating was

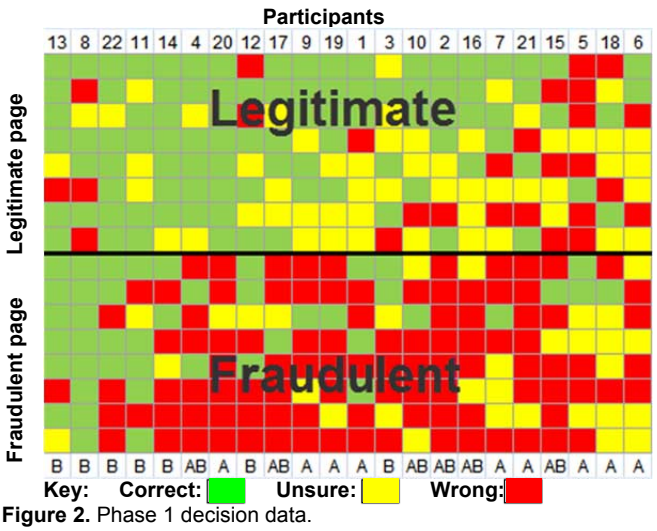

Figure 2. Phase 1 decision data.

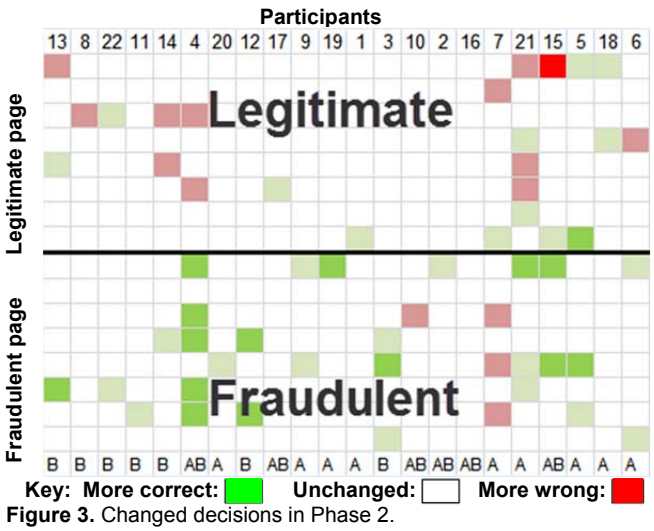

Figure 3. Changed decisions in Phase 2.

insufficient to move that page into a different safe/unsafe/unsure category. Pale green cells are page ratings that improved by one category (e.g., unsure to safe for a legitimate page, or unsure to unsafe for a fraudulent page). The dark green cells occurred when a person correctly changed their decision between safe and unsafe, while the (single) dark red cell occurred when a person incorrectly changed their decision between safe *vs*. unsafe.

For the legitimate pages, the overall percentages of ratings barely changed between Phases 1 and 2; as stated previously, no significant difference was found for legitimate pages between these phases. While changes were made, the number of incorrect re-ratings approximately equaled the correct re-ratings to produce no net gain. This can be seen visually by the balanced scattering of red and green cells in Figure 3, top.

For fraudulent pages, participants' correct ratings improved from 25 to 34% (correct:unsafe), with a decrease in incorrect ratings from 57 to 44% (wrong:safe). As mentioned, these differences were statistically significant. Unsure ratings increased from 18 to 23%. The overall gain in correct performance in re-rating these fraudulent pages can be seen visually by the predominance of light and dark green cells in Figure 3, bottom.

Discussion. Participants appeared reasonably accurate at identifying legitimate sites, where they wrongly rated only 15% of them as unsafe (the caveat is that they were still unsure of 31% of these pages). Yet their performance was effectively unchanged between the phases. We suspect that this good performance is due more to people's disposition to trust familiar-looking pages *vs.* their ability to recognize the domain name as legitimate. As we will see, this suspicion is verified in the qualitative analysis.

For fraudulent pages, participants fared poorly, where they correctly identified phishing pages only about 25% of the time. Fully 57% of these pages were incorrectly rated as safe (i.e., the wrong:safe measure), which would have put them at substantial risk of a successful phishing attack. Even when they were told to attend to the address bar, their performance only improved marginally (although it was statistically significant); fully 44% of the pages were still incorrectly rated (i.e., again as wrong:safe).

We interpret this as follows.

- Domain name highlighting is, at best, only somewhat effective at helping people identify fraudulent sites.
- Participants do not always attend to the address bar (and thus the domain name) to determine site legitimacy. If they had attended the address bar in Phase 1, we would not have found a difference between the two phases.
- However, when participants were told to attend to the address bar, they did do somewhat better – but not much better – in correctly identifying fraudulent sites.

Participant performance. As mentioned, the columns of Figures 2 and 3 are participants sorted by those with the most correct results on the left. Evident from these figures is that there is marked variation in participant performance. We will revisit this difference between participants in the qualitative analysis section, where we identify three user types and how they differ in what they look for.

Ratings by obfuscation method. While we did not use obfuscation method as a statistical factor in our analysis, we inspected our results for gross effects. The order of web pages within Tables 1a and 1b correspond with the row ordering in Figures 2 and 3. That is, both are ordered from those with the most correct ratings (top rows in each section) to the most incorrect. As evident in Figure 2, the majority of people frequently and incorrectly rated a fraudulent page as safe, regardless of the type of obfuscation method used. When people were asked to look at the address bar, people did improve the correctness of their ratings somewhat (but not that much) across all obfuscation methods. The exception is the IP-address attack (top row in the fraudulent section of Tab1e 1b and Figures 2 and 3). More people correctly identified the IP-address attack as a fraudulent page: 9 people in Phase 1, and 14 people in Phase 2 after being told to look at the address bar.

Discussion. Legitimate web sites rarely use an IP address as a domain name. If the domain name was displayed as an IP address, this should be an obvious cue suggesting a phishing attack. Our results indicate that this is the case, but only marginally.

- IP-address attacks are still an effective phishing method. Even when participants were asked to attend the address bar, only 64% of them were able to spot the attack. That is, people either did not notice or, more likely, they did not understand the implications of seeing an IP address in the domain name.
- The other three types of phishing attacks seem equally effective at fooling people, as there is little practical difference between them in terms of people's ability to spot them. This strongly suggests that people are equally bad at judging domain name legitimacy when slightly more sophisticated obfuscation methods are used.

QUALITATIVE DATA ANALYSIS AND RESULTS

Analysis method

We used a form of open coding [14] to analyse user comments and behaviours observed during the think-aloud phases and interviews. We first identified four overarching category types of information used to form trust judgments: (a) institutional brand; (b) content as presented in the main pane of the browser; (c) input information requested, and (d) information in the address bar and other security indicators. These will be described in more detail shortly.

Within these categories, we then grouped comments into two sub-categories that asked *what* specific information they scrutinized, and *how* they used that information to form their judgments. We then looked at differences between any of these categories as described by a participant between Phase 1 and Phase 2. Finally, during our analysis, we noted that participants did not use these categories equally. Consequently, we sorted participants with similar traits into a typology as described next.

Predisposed Perceptions: A Typology of Groups

Perhaps the most interesting behavioural pattern we identified concerned the differences between people in how and when they attended different elements of the browsing interface to judge a site's legitimacy. These differences suggest a typology, where people can be (loosely) categorized as clustering along points of a spectrum in terms of how they scrutinized a mix of information covering a web page's: (a) institutional brand; (b) content as presented in the main pane of the browser; (c) input information requested, and (d) information in the address bar and other security indicators. The three types of people are briefly described below and illustrated in Figure 4, Following sections detail how each type used this information.

Type A people focus solely on information present in the browser's content pane (9 participants). This confirms prior findings on how people judge web site credibility [9, 6, 4]. Information brand as identified in the content pane was

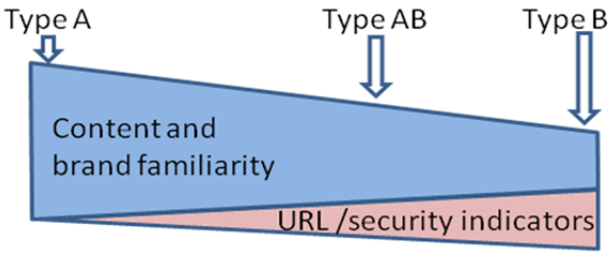

Figure 4. The typology spectrum.

critical. They did *not* consider information in the address bar. Nor did they consider other security indicators shown outside the content pane, verifying findings by [16,18].

Type B people consistently focused on the information found in the address bar, the information being requested, and other security indicators (7 participants). While they did use information visible in the content pane (e.g., to determine brand familiarity as with Type A people), it was not the sole source of their judgement.

Type AB people were a composite (6 participants). Like Type A, they relied primarily on the content pane and institutional brand. Like Type B, they sometimes considered information in the address bar and the kinds of information being requested. They were much more variable in the information they considered across pages.

Institutional Brand

A salient criterion people used for evaluating site security was the brand(s) depicted on the page. Both Type A and Type AB participants appeared predisposed to implicitly trust particular brands. The quintessential cases were financial institutions where: two people explicitly stated that banks *had* to be secure; and eight people used familiarity and/or personal experience with two banking sites (Canada Trust and Royal Bank) as their main or sole reasoning for trusting the sites. It is rather concerning that a good portion of our participants did not seem to even look at the URL, let alone correctly recognize one of them as a phishing site, for these high-risk banking pages. Other recognizable brands, such as Amazon, Google, the University of Calgary (UC), and Yahoo! were met with positive judgments, despite several of them being fraudulent. This suggests that for many users, the perception of safety and legitimacy of a site is tied heavily to their preconceived perception of the brand itself.

Type A and Type AB subjects were also often quick to identify symbols and logos associated with familiar brands. This included logos by other companies on a page, such as the Google logo on the YouTube page, the Yahoo! logo on the Flickr page, and the Apple logo on the MobileMe page. Any sign of these 'famous' and familiar big-brand logos seemed to instill these participants with a measure of (unwarranted) trust.

In contrast, the seven Type B participants showed a different pattern in their use of brand. Five mentioned brand for only two of the pages. Even when brand was mentioned, these participants never used brand as the sole category of

information to form their ratings. Most Type B comments involved only a passing mention of brand, typically to say whether they recognized the brand depicted or connected to the site.

Generally, participants were also inclined to distrust particular brands or types of sites. For example, a few participants mentioned privacy issues with Facebook, YouTube and Meebo as a reason to distrust the site. If anything, participants (especially Type A and AB) showed greater caution toward social networking sites, especially if they had little or no personal experience with that site, or had heard negative things about it.

Page Content

While it is natural to interpret page content as an indicator of trust, this is irrelevant in the detection of a phishing page that replicates the look and feel of the legitimate page that results in no discernible difference. What is alarming that Type A (and sometimes Type AB) instinctively relied on this page content to determine legitimacy. In contrast, Type B participants, for the most part, were not strongly influenced by page content in their evaluations of site legitimacy: they instead spoke mostly about the content of the address and status bars.

Type A and AB participants formed their judgment around the content information as present in the main content pane of the page. They were particularly sensitive to security and legitimacy indicators within the content pane, where their presence of these indicators positively influenced their perception of that page as trustworthy. These indicators fell into three groups: security and privacy related information, the content itself, and the type of information requested. Of course, a phisher can easily spoof all this information.

Security and privacy related information. Type A and Type AB subjects were often quick to identify symbols that associate a site as being verified as 'safe' by an external authority. An example is the appearance of the Verisign logo. They also pointed out internal indicators of security, such as a "security guarantee" logo on the Royal Bank page (9 of the 15 Type A and Type AB subjects mentioned this and also rated the site as trustworthy). The presence of a virus scan advertisement was interpreted by two participants as being a positive sign of site security, and a few mentioned the option to forget user login credentials ("Remember Me") as a sign that it would be safe to enter their login information. Indirect security-related content such as an 'updated' copyright date and links to contact/support information were also frequently reported by Type A, and some Type AB participants. In fact, Type A (and some Type AB) participants frequently spoke of the security and non-security-oriented vocabulary typical found on the edges of the content pane (copyright, security/privacy links, contact info link). In these instances, links to legal, privacy and security policies implied that the site had safeguards to ensure that disclosed information would be held in the sole possession of the particular site.

The content itself. Content amount, the kind of information, and its layout within a page were also important factors contributing to Type A and AB interpretations of a site's trustworthiness (see also [4]). One notable example involved customer reviews and book information in the amazon.com product page. The amount and layout of information on that page instilled a sense of confidence in its legitimacy (because it was information-rich and professionally laid out). The customer reviews on that page instilled a sense that others had visited the site before, and that others had purchased the depicted product.

Type of information requested.

All but one of our pages requests people to enter different kinds and amounts of information. Some pages request sensitive identification information, and/or financial information, and/or unusual or excess information that could seem disproportionate to that page type. Curiously, financial information requests did not seem to warrant greater caution by participants, and most participants did not even comment on it. Type AB people made the most comments about pages that seemed to request excessive or unusual information whereas Types A and B provided fewer comments; we have no explanation for this discrepancy.

Address Bar and Other Security Indicators

There was considerable disparity between our three participant types in whether they looked at the address bar and, if they did, their technical understanding of URLs and how they could be used to inform trust judgments.

Type A people focused almost exclusively on information present in the browser's content pane, and they rarely considered information in the address bar. Even when asked to look at the address bar, most Type A people clearly had no idea about 'unusual' URLs, nor what to do with the information contained within it. Comments included "this (URL) doesn't mean anything to me" or "this (URL) doesn't change anything for me". They mentioned they were uncertain of how to interpret '.net' or '.ca' suffixes, or the 'https' prefix. Some incorrectly considered the URL length as meaningful (i.e., perceived as 'too long' or 'too short'), while others thought that certain words in the URL had positive connotations, e.g., the appearance of the word 'secure', 'deliveryaddress', the brand name, and so on. Long URLs and those containing seemingly nonsensical argument characters (even if legitimate) prompted frequent comments about what the URL meant and why it was so complex. For IP attacks (e.g., top row, Table 1b), the three Type A people who noticed the IP numbers in the domain name mentioned their suspicions, but still rated that site as trustworthy: "the URL are just numbers".

Type A people tended to use brand identity and site type to over-ride any suspicions they may have had when seeing unusual URLs. To illustrate, one subject who was judging a fraudulent bank page commented "the highlighted [domain name] part is strange… [but] banks are secure" and then rated that page as trustworthy. Another subject spotted a

letter substitution attack (the '1' in 'paypa1'), yet also rated that page as trustworthy because of site familiarity.

Unlike Type A, Type AB paid more attention to URLs and to other security indicators, although not consistently. The content pane remained their primary source of information across all pages, but they did occasionally look at the URL on some pages. At least some of them seemed aware that the URL may contain security and other indicators that could be used to inform their trust judgements. For example, one said "https is more secure". Another recognized the "1" substitution attack as strong evidence of a fraudulent page. Yet Type AB people were often unsure of what to look for in the URL, and did not have a full understanding of phishing indicators. For example, 3 of the 6 Type AB people caught the "1" in the 'paypa1' domain name, but were confused as to how this should influence their ratings, i.e, they continued to trust that phishing site.

Unlike Type A and AB people, Type B participants paid heavy attention to the address bar. They knew of the security indicators built into the address bar area. They knew the URL was important to determine page legitimacy, and they consistently looked for clues in the address bar to appraise the legitimacy of the sites. For example, many of them consistently looked for 'https' and the lock icon in the URL. Some knew about the importance of the domain name, where they tried to match it with the page being displayed. For example, one person viewing the fraudulent Meebo page said "the URL is not meebo.com, and it looks phishy". While all this sounds like good news, Type B people still missed many fraudulent pages (see Figure 2). While one Type B person did correctly identify all fraudulent pages, four other Type B people missed 50% or more of the fraudulent pages.

Subjects with Type AB or Type B classifications demonstrated a marked improvement in successful identifications of phishing sites during the second phase. Contrary to Type A, these participants focused their attention more on the technical rather than superficial aspects of the address bar. Secure protocol indicators like 'https' or the lock icon were noted; the entire URL was examined, and (albeit with only a modest success rate), the obfuscation method was seen.

Perhaps surprisingly, when asked about it during the exit interview, most participants – including Type B – said they were not even aware that the domain name part of the URL was highlighted.

Correlating Type to Rating Correctness and Training

Between types and correctness of ratings. We did not statistically re-analyze our data based on these types, as our typology was determined post-hoc. However, a qualitative inspection suggests correlations between types and how correct their ratings are likely to be. Recall that the bottom row of Figures 2 and 3 indicates the participant type (A, AB, B) and that columns are sorted by participants with the most correct to least correct rating, from left to right. Not surprisingly, almost all Type B people are clustered on the left side, i.e., those with the most correct ratings. Types AB tended towards the middle, and Type A tended towards the right (least correct) side.

Between types to formal technical training. Some of our participants were trained as computer scientists and engineers. Most – but not all – of those fell into the Type B category. Similarly, most Type A people were in non-technical fields, with AB being a mix. However, this correlation was not perfect. For example, two of the nine Type A people were either computer scientists or engineers.

LIMITATIONS AND FUTURE STUDIES

Our methodology tries to balance the constraints of a laboratory study with real world validity, which introduces several limitations. Our results, while strongly indicative of real world behavior, are approximations. For example, by using real web sites, we could not easily control for people's familiarity particular sites and brands, nor with their familiarity (or expectations) of the actual URL address. Participants also knew that they were judging a web site's legitimacy, which is why we suggest this is a best case vs. average scenario. We left out other important factors that could have revealed additional nuances. One suggestion for future study could compare address bars with and without highlighting, thus providing a control comparison, i.e., to determine the marginal improvement of highlighting (if any). Another study could tease out obfuscation type as a statistical factor, i.e. to see if people do better at identifying particular obfuscation methods. A third study could incorporate user types (A, A/B, B) as a statistical factor.

CONCLUSIONS AND IMPLICATIONS

In spite of the above limitations, our study provides reasonable 'upper bounds' of how likely people are to detect phishing attacks based on domain name highlighting. That is, participants in Phase 1 were primed to judge security as their goal, and in Phase 2 they were specifically told to attend to information in the address bar. Our participants were also highly educated (although not in security), with several of them being technically knowledgeable.

What we found was that domain highlighting works, but nowhere near as well as we would like. Our participants still incorrectly rated about half to two-thirds of the fraudulent pages as safe, depending on whether they were told to attend to the address bar. For Type A users, domain name highlighting is rarely effective, as they judge a page almost solely by its content area (similar to [9, 6]). This is highly problematic, especially because we believe Type A behavior will characterize the general population, as most people are not technically sophisticated. And even Type AB and Type B people who do scrutinize address bars are at best 'hit and miss' at detecting domain name anomalies. To make matters worse, most people don't even bother to look at the address bar unless they are told to do so. Consequently, in everyday life, we expect people to be much worse at identifying phishing attacks via domain

name highlighting, as they lack technical knowledge and they will be focused on their normal web browsing activities rather than security.

Two questions worth asking are: should we abandon domain name highlighing, and are there ways to improve it? We speculate on three areas for future work.

Education. Our results showed several problems that could be associated with a lack of user knowledge. First, Type A and AB people mistakenly place great reliance on the information in the content area and the brand, which can be easily mimicked by phishers. Second, Type A (rarely) and Type AB (only sometimes) attend to the address bar, likely because they do not know of its importance. Third, even if they do look there, almost all participants – even Type B – frequently miss the spoofed domain name, likely because they are unaware of the common obfuscation methods used. Alternately, they use information in the URL to form their judgement that is simply not relevant. Education would likely help [10]. People need to know that they cannot rely on the content area, that they should scrutinize the domain name, and that they should be aware of common phishing methods. As one person said: "highlighting the domain name is fine for people who normally look at the URL, but we need to reach the general population… more knowledge would help." The challenge is to find effective ways to train (likely unmotivated) users without getting in the way of their primary task [10].

Attention. When people browse, they tend to focus on the content area as it is the reason why they are browsing. Security is a secondary concern [17,10], and consequently even the most careful (and knowledgeable) may fail to scrutinize the domain name if they are immersed in their task. This makes us question if there are methods that we can use to draw people's attention to the domain name, or if we can present that information another way. Solutions here may try to make the domain name even more obvious, or draw the person's focus onto the address bar (especially if the browser can detect anything suspicious about the site), or somehow place the domain name in a portion of the page where the person is more likely to attend to it.

URL Complexity. URLs can be quite complex, and it was clear that many of our participants were confused by this complexity. This could perhaps be resolved by either presenting the domain name by itself in a dedicated area separate from the rest of the URL, or by somehow reducing complexity of the URL by (perhaps) hiding extra information in URL address.

In summary, domain highlighting gives only marginal protection and cannot be relied upon as the sole means to identify a phishing site. Still, it is worth including as it comes at almost no cost in terms of interface clutter, browser performance, and interference with a user's task. Clearly, it must be used with other anti-phishing tools, where the combination of those tools may provide better protection (although this has not yet been shown to be true).

We believe that domain name highlighting can be made somewhat more effective by making the domain name even more obvious, by drawing the person's focus onto the address bar, by reducing URL complexity, and – most importantly – by educating people about the importance of the domain name in judging web sites, and what typical obfuscation methods are used by phishers.

Acknowldegewments. This research is partially supported by NSERC and ISSNet.

REFERENCES

1. APWG (2009) Phishing activity trends report, 4th quarter. http://www.antiphishing.org/
2. Dhamija, R., Tygar, J. D. and Hearst, M. (2006) Why phishing works. *Proc ACM CHI*, 581–590, ACM.
3. Egelman, S., Cranor, L., Hong, J. (2008) You've been warned: An empirical study of the effectiveness of web browser phishing warnings. *Proc. ACM CHI*, ACM.
4. Fogg, B. Marshall, J., Laraki, O. et. al. (2001) What makes Web sites credible?: A report on a large quantitative study. *Proc ACM CHI'2001,* 61-68, ACM.
5. Gabrilovich, E. and Gontmakher, A. (2002). The homograph attack. *CACM 45*(2).
6. Jagatic, T., Johnson, N., Jakobsson, M. and Menczer, F. (2007) Social phishing. *Comm. ACM* 50(10), 94-100
7. Jakobsson, M. (2007) The human factor in phishing. *Privacy and Security of Consumer Information '07*.
8. Jakobsson, M., Johnson, N. and Finn, P. (2008) Why and how to perform fraud experiments. *IEEE Security and Privacy*, 6(2):66–68.
9. Jakobsson, M., Tsow, A., Shah, A., Blevis, E. (2007) What instills trust? A qualitative study of phishing. *Proc Usable Security, LNCS, Springer-Verlag, 356-61*.
10. Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. and Hong, J. (2010) Teaching Johnny not to fall for phish. *ACM Trans. Internet Technol.* 10(2).
11. Litan, A. (2009) The war on phishing is far from over. Garnter, April. http://www.gartner.com
12. PhishTank (2010) Statistics about phishing activity and PhishTank usage. //phishtank.com/stats/2010/08/, Aug.
13. Schechter, S., Dhamija, R., Ozment, A. and Fischer, I. (2007) The Emperor's New Security Indicators. *IEEE Symposium on Security and Privacy*, 51-65.
14. Strauss, A. and Corbin, J. *Basics of Qualitative Research*. 2nd Edition. Sage Publications, 1998
15. Vaughan, C. (2008) Address Bar Improvements in Internet Explorer 8 Beta 1, *IEBlog*, 11 March.
16. Whalen, T. and Inkpen, K. M. (2005) Gathering evidence: use of visual security cues in web browsers. *Proc. Graphics Interface*, 137–144.
17. Whitten, A. and Tygar, J. D. (1999) Why Johnny can't encrypt: A usability evaluation of PGP 5.0. *Proc. USENIX Security Symposium*, USENIX Association.
18. Wu, M., Miller, R. C. and Garfinkel, S. L. (2006) Do security toolbars actually prevent phishing attacks? *Proc ACM CHI'2006,* 601–610, ACM.