BayeShield: A Conversational Anti-Phishing User Interface

ABSTRACT
In this paper we present BayeShield, a novel anti-phishing tool that uses a conversational approach to partner with users in determining whether a website is phishing when the website is suspicious but not blacklisted. We describe the iterative user-centered development of BayeShield's user interface, discussing its evolution and the design principles we followed. In an empirical evaluation, BayeShield performed better than Firefox 2.0 in preventing participants from entering information on phishing sites. In a second portion of the study, we evaluate BayeShield's usability and obtain positive results including high user satisfaction ratings, and a high-level of engagement as demonstrated by perceived duration of tasks being lower than actual durations. In addition, we learned user characteristics that affect the likelihood users will enter information on phishing websites.

Author Keywords
Phishing, anti-phishing, HCI, conversational, Cognitive Load Theory.

ACM Classification Keywords

INTRODUCTION
Phishing is an attack in which users are fooled into entering personal information into a “spoof” website instead of the intended legitimate website. In this paper we present BayeShield, an advanced anti-phishing User Interface (UI) designed to capture at-risk users' attention and convince them of the danger. BayeShield is the first anti-phishing UI designed specifically to serve as the front-end for highly accurate Information Retrieval-(IR)-based phishing detection tools that go beyond blacklist-based phishing detection. The problem with current blacklist-based phishing tools is that phishing sites, on average, are not blacklisted for up to 9.3 hours[9]. In on-going work, we have observed that an IR-based approach can boost detection to 97% of phishing attacks but must allow for a small number of false positives (legitimate websites classified as phishing attacks). This provides a challenge as false positives could reduce user confidence in anti-phishing warnings[6]. However, we contrast this rate to our observed detection rate of the Firefox 2.0 anti-phishing tool of ~62%[11]. To address this, we designed the BayeShield Analyzer, a conversational UI that walks users through the decision-making process experts use to distinguish between phishing attacks and legitimate websites. The ultimate goal of the Analyzer is that, through repeated exposure, users learn to identify phishing attacks even without an anti-phishing tool.

In order to validate BayeShield's design, we present the results of a user study. In the study, we compared BayeShield's ability to prevent users from falling for phishing attacks with Mozilla Firefox 2.0's (FF2.0) anti-phishing tool and found that BayeShield outperformed FF2.0. We also evaluated BayeShield's usability and user engagement based on reported satisfaction and perceived vs. actual task duration. In addition, we analyzed correlations between participant behavior during our study and information on participants' demographics, attitudes towards computers, and computer literacy, which provide a better understanding of the user characteristics that affect user behavior when dealing with phishing sites.

The next section of our paper explores existing anti-phishing literature and relates it to our approach. This is followed by an examination of BayeShield's development with an emphasis on cultivating user trust and improving security software usability by incorporating Cognitive Load Theory (CLT). Next, we present a description of our methodology and then conclude with results of our study and a discussion of those results.

RELATED WORK
Anti-Phishing User Interfaces
Dynamic Security Skins, by Dhamija and Tygar[4], Passpet, by Yee and Sitaker[18], and WebWallet, by Wu et al, [17] are all anti-phishing UIs that establish a set of sites into which the user can safely enter personal information. These tools suffer when shared amongst multiple users on a single
Bayesian Classifier detected 90% of phishing attacks. 97% of phishing sites, classifier trained to identify phishing with even slightly system. Our group has found that combining a Bayesian UI acts as a front-end to an advanced phishing detection Markup Language for graphical overlays), and HTML. The combination of Javascript, XUL (Mozilla's eXtensible SYSTEM DESIGN help the user when they are at-risk instead of beforehand. Our group hope to educate users but we hope to warn and users about phishing and URLs. Both Sheng et al. and Dhamija et al. have found that SiteKey and, by extension, other pre-selected image authentication schemes failed to prevent phishing attacks[12].

The most commonly used anti-phishing tools are those incorporated in Internet Explorer 7 (IE7) and Mozilla Firefox 2.0 (FF2.0). In IE7, if a user browses to a blacklisted website, IE7 redirects them to a specially designed webpage. Firefox covers the screen in a grey overlay and displays a specialized pop-up warning telling the user they are on a site associated with identity theft. Both IE7 and FF2.0 rely on blacklists which limits their ability to detect new phishing attacks and research has shown these tools miss many phishing attacks[9][19].

Evaluating User Behavior on Phishing Websites
In 2006, studies by Wu et al. and Dhamija et al. found that users fail to notice traditional security indicators[3][16]. Wu et al. claim convincing website appearance trumped all warnings tested. Studies have found that users fall for phishing attacks even after being warned and sophisticated attacks fool up to 90% of users[3][12][16]. Wu et al. asked participants to protect their boss's personal information while responding to his/her email[16]. We decided against this scenario after people in pilot studies informed us they treat a second party's information differently from their own. In “Why Phishing Works,” Dhamija et al. asked participants to distinguish between legitimate and “spoof” websites. Even when primed, users made incorrect decisions up to 90% of the time[3].

In 2008, Egelman et al. published a study evaluating the anti-phishing tools bundled with FF2.0 and IE7. 97% of their participants clicked on a link in a phishing email but only 21% of users fell for phishing attacks when warned with active indicators (an indicator which prevents users from proceeding before acknowledging it). FF2.0's active indicator outperformed IE7's[6].

Finally, in 2008, Sheng et al. created a game to educate users about phishing and URLs[13]. Both Sheng et al. and our group hope to educate users but we hope to warn and help the user when they are at-risk instead of beforehand.

SYSTEM DESIGN
BayeShield is a Firefox 2.0+ extension implemented in a combination of Javascript, XUL (Mozilla's eXtensible Markup Language for graphical overlays), and HTML. The UI acts as a front-end to an advanced phishing detection system. Our group has found that combining a Bayesian classifier trained to identify phishing with even slightly effective blacklisting tools, BayeShield correctly detects 97% of phishing sites,1 far exceeding the performance of solely blacklist-based approaches [11].

Bayesian classifiers rely on calculating the probability a website is phishing and there is a small chance BayeShield will flag a legitimate site as phishing (~3%). These incorrect warnings are known as false positives. These false positives are necessary in order to reach acceptable levels of phishing attack detection. As a result, BayeShield's UI has two aims:

- To warn users when they are on a website that is highly likely to be phishing.
- To partner with users to distinguish between legitimate sites and phishing sites.

The UI consists of three components:

- A browser-based toolbar,
- A custom warning consisting of a pop-up with warnings about identity theft and an overlay,
- The “BayeShield Analyzer” which helps users to decide whether to enter information on a website.

Designing BayeShield
We followed an iterative design process in which we alternated between brainstorming, prototyping and seeking feedback via pilot studies and “quick-and-dirty” evaluation sessions.

Our first attempt at a User Interface illustrates the difficulty inherent in the problem. Initially, we crafted a warning mechanism that allowed users to continue browsing until entering information while still providing indicators they were on a suspicious website. After BayeShield detected an attack, the browser toolbar grew to three times its original size, forcing the participant's region of interest to track down the screen. When they began to type, a message appeared on the toolbar with a button to activate BayeShield's Analyzer. Justification for these early measures came from research in Cognitive Psychology suggesting that unique transient aspects of a scene on the periphery of user's visual focus, or changes in the field of focus, will capture their attention[8]. No users who tried this toolbar noticed the increase in size. This provides further confirmation of the failure of “passive” security indicators.

To address the shortcomings of the first iteration, we introduced an overlay and pop-up combination to block users' progress. We also developed a decision-tree mapping ways in which a user could reach a phishing site2 and the mental process security experts would apply to recognize the attack. We formulated the process as series of questions to which the user could easily respond. We assigned values to the user's responses based on our group's analysis of the

---

1Without the addition of a blacklist, BayeShield's Naive Bayesian Classifier detected 90% of phishing attacks.

2Email is the predominant phishing vector but it's easy to imagine spear phishing attacks utilizing social networks, IM clients or blog comments as potential vectors.
likelihood the response is indicative of a phishing attack. We used these values to raise and lower a “threat meter,” meant to visually convey the threat posed by the website.

From this tree, we constructed a low-fidelity notecard prototype (Figure 1) of the Analyzer and asked people to answer questions with our prototype to identify missing links in our tree and poorly phrased questions. This led to several improvements in the progression of the questions and generally improved the information architecture.

The next iteration centered on aesthetics and translating the low-fidelity prototype into high-fidelity. We used Photoshop CS3, Javascript, XHTML and CSS to build the software’s UI and translated it into XUL and Javascript for FF2.0. We aimed for a professional aesthetic that promoted trust and a sense of credibility [7]. The logo appears in the upper left of the header and the brand slightly lower to create a step effect leading the user to the body of the page. The layout is placed on this diagonal which leads finally to the navigation options at the bottom. Icons and thumbnails combine image/text features to convey messages without alienating users unfamiliar with the icon symbolism or text terminology. The icons and thumbnails are glossy with a glassy effect to lend a contemporary feel to the design.

Designing for Security with Cognitive Load Theory
Security software often asks users to consider complex ideas in order to make correct decisions. We incorporated Cognitive Load Theory (CLT) into BayeShield’s design to simplify the decision-making process for the user. CLT assumes a limited working memory capacity when dealing with novel or new information but unlimited long term memory holding cognitive schemas of varying degrees of complexity. These schemata reduce working memory load because the mind deals with them as a single element[14]. CLT is concerned with designing instructions to facilitate schema construction and enable novice users to evolve into expert users[15]. The following goals are aligned with our attempts to incorporate CLT.

Make Security the User’s Goal
We aimed to make security part of the task as opposed to an obstacle preventing progress. In order to achieve this, our pop-up+overlay warning was designed to convey imminent danger and make it clear we would help them understand if it was safe to proceed.

Reduce Cognitive Load
We sought to make the UI simple by using low-text density. Text presented in short chunks with non-technical wording is used in web-based learning to facilitate comprehension[1]. We used CSS stylings to emphasize keywords and allowed users to click on keywords to learn more. We displayed information according to inverted pyramid construction[1].

Studies from Cognitive/Educational Psychology support our language decisions[7][11]. Our main concern was to avoid straining beginners’ cognitive load with unfamiliar ideas or terminology. It is easy to set such goals but more difficult to implement them. Would all users understand what a “white-list” is? Should we include definitions directly in the text? How simply can we phrase a question without compromising its intent? Answering these questions resulted in a large amount of spirited debate and required obtaining feedback from users.

Educate the User
To encourage learning, we provide motivational language that engages the user rather than passively parroting information. Active engagement in the decision-making process increases the likelihood of schema development. We hope to instill, through repeated engagement, the decision-making process into user’s ordinary web experience.

We now describe each portion of BayeShield in detail.

BayeShield Toolbar
The BayeShield toolbar (Figure 2) is positioned below the address bar in Firefox and consists of the following: name-branding, add/remove buttons allowing the user to add or remove sites from BayeShield’s local list of safe sites and the domain of the current page. It is intended to provide visual clues to the user that BayeShield is installed and functioning and habituates them to BayeShield prior to exposure to warnings, building trust and familiarity.

BayeShield Warning: Pop-up+Overlay
As mentioned in Related Work, the pop-up+overlay used by Firefox 2.0 outperforms IE7’s security indicators[5]. Our group independently reached similar conclusions through pilot studies. When BayeShield determines there is a high probability a user has reached a phishing site, it displays both a custom pop-up and an overlay (Figure 3). We have included additional innovations to capture a user’s attention

Similar to a newspaper article which outlines the entire story first and attends to details later.
We distort the suspected phishing attack website with a grey, semi-opaque overlay covering all fields on the page. The overlay displays the phrase “WARNING THIS WEBSITE MAY BE FRAUDULENT” in large, red lettering.

The pop-up is displayed just below the address bar and extends over the browser's chrome and onto the pane displaying the website, arrows visually link the areas and associating the URL with the website. The pop-up contains only two sentences: the first warning the user that this site may be attempting to steal their personal information and the second offering to help them decide if it is safe to proceed by using the BayeShield Analyzer. The user is presented with two buttons: “Cancel” and “Open Analyzer”.

We do not allow the user to click “Cancel” until four seconds have elapsed. We selected four seconds as it is similar to the amount of time Firefox pauses before allowing users to install extensions. Forcing the user to wait and examine the warning underscores its importance. We keep users informed by displaying a countdown in the “Cancel” button.

We have skinned the pop-up to have a professional and reassuring appearance as well as a look-and-feel consistent with the BayeShield Analyzer. We emphasize the importance of appearance in this case. Research has found that users consider appearance above other indicators [3]. Looking “more convincing” than the phishing attack is essential. This is in line with Li and Helenius' recommendations from their usability evaluation of phishing UIs [10]. We draw attention to the contrast between our warning and that of FF2.0+ which has a more generic appearance with no branding.

BayeShield Analyzer

The BayeShield Analyzer adopts a conversational approach to helping the user distinguish between phishing attacks and legitimate sites. The Analyzer asks the user a series of questions tailored based on previous responses, and then presents a judgment page indicating whether the website is “Safe” or “Not Safe” and providing a summary of their responses as well as advice on how to proceed.

Screen 1: Warning

When the user chooses to open the Analyzer, they are presented immediately with “Warning!” in large, red letters. The screen concisely describes why BayeShield believe the website to be suspicious. We list three common criteria:

- The website appears to be asking for personal or financial information,
- The website resembles known identity theft sites,
- The website is not in a list of safe sites.

Notice that we avoid jargon and use understandable terms. When the user clicks on any of the highlighted words, help messages scroll open to give additional details. If the website is blacklisted, we alter the wording of the above criteria and provide information about the blacklist on which it appears.

Screen 2: Instructions

The next page instructs the user on interacting with the Analyzer. They are informed that the Analyzer will ask a series of questions and, based on their responses, a meter at the right of the screen will raise and lower. They are told the higher the meter goes, the more suspicious the website is.

Screen 3: Personal Information?

An overview of the current question is displayed in red at the top of the page (Figure 4). Below it is a re-statement of the question, reinforcing what is being asked. In this case, “Is the website asking for any of the following?” They are then given four categories of information: Online account, Financial info, Identification and Personal Info. All four are represented graphically and examples of the type of information in each category are listed next to the graphical representations. For instance, Identification (in the lower left quadrant) contains a picture of an ID card as well as the following examples: Social Security #, Driver's License and Passport. If the page asks for any of these, the user can check the box in that quadrant. Checking a box causes the meter to rise. This page compactly represents a large amount of information and is a good example of our use of CLT.

Screen 4: How did you get to the website in question?

In this screen (Figure 5), the user is asked how they reached the current page. This is the beginning of making a determination about the safety or danger of a specific page and from here on, follow-up questions vary depending on the answers to previous questions. We account for many possible ways the user could have reached the page to help...
Table 1: One progression of questions through the BayeShield Analyzer (questions abbreviated for the sake of space)

<table>
<thead>
<tr>
<th>Question</th>
<th>User Choices and Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>How did you arrive at the website?</td>
<td>From email</td>
</tr>
<tr>
<td>Do you recognize the company/person?</td>
<td>Yes</td>
</tr>
<tr>
<td>This email was:</td>
<td>A reply-</td>
</tr>
<tr>
<td>Did the email convey urgency?</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Expected-</td>
</tr>
<tr>
<td></td>
<td>Unexpected-</td>
</tr>
</tbody>
</table>

distinguish between false positives and actual attacks. The meter rises and, in a few cases, lowers in concert with answers. Next to each answer is an arrow symbol indicating what direction the meter will move if that choice is selected. Table 1 presents an example of how a user might advance through a typical series of questions.

**Summary Page**
Having answered a series of questions, the user is presented with a page either headed with the phrase “Safe” or “Not Safe” depending on the height of the meter. Their answers are summarized and presented in a tabular format so that they can review the progression and are not forced to remember it. In addition, users receive advice on how to proceed. The vast majority of the time, the user will be informed it is highly likely the site is not safe and advised to close the site and contact the company via another means.

**USER STUDY METHODOLOGY**
We conducted a user study with several aims in mind:

- To study the categories of sites into which users will submit personal information in a realistic setting.
- To compare BayeShield's effectiveness at preventing users from falling for phishing attacks to Firefox 2.0.
- To evaluate BayeShield's ability to help users distinguish between phishing and legitimate sites.
- To learn about how user characteristics are related to their behavior with respect to phishing attacks and anti-phishing software.

We recruited 20 participants. Participants completed an online survey evaluating their attitudes towards computers and computer literacy. An in-lab session of two blocks of tasks followed. After the in-lab session participants completed a user satisfaction questionnaire.

**Environment**
The study was conducted in one session approximately forty five minutes in length between Sept. 1st-9th 2008. The session occurred in an on-campus laboratory. Participants all used a Dell Dimension C521 running Windows Vista Home Premium with a Samsung SyncMaster 920NW monitor and the standard Dell keyboard and mouse. Participants used Mozilla Thunderbird 2.0.0.16 and Firefox 2.0.16.

**Block 1 Overview**
This block was a between-subjects design using matched pairs, dividing the participants into groups of ten. Each group was assigned to one of two treatments: FF2.0 or BayeShield. Pairs were determined from the online survey taken prior to the in-lab session. The goals of Block 1 were to determine: if BayeShield prevented phishing attacks, if BayeShield was more effective than FF2.0 at preventing phishing attacks and on what sites users enter personal information after clicking on links in emails.

**Matched Pairs**
We matched pairs based on survey responses. We treated each participant as a vector based on responses. To generate pairs we took the first ten participants, repeatedly grouped them into pairs, and calculated the vector distance between the pairs. The pairing that minimized the sum of the total distance was selected. We repeated the procedure for the next ten participants to obtained ten pairs. We did this to ensure a fair comparison by having the two groups of participants be as equivalent to each other as possible while preserving random assignment.

**Technical Details**
We locally hosted nine legitimate login pages from well-known institutions and four phishing attacks against: Wachovia Bank, Amazon, PayPal, Inc. and Bank of...
America. We added invisible divs to phishing pages instructing BayeShield to classify them as phishing. Although hosted locally, all pages appeared indistinguishable from the actual sites. In order to test FF2.0, we hosted the phishing attacks at domains in the FF2.0 blacklist so FF2.0's warning would appear. Finally, we short-circuited the login form on each page to re-direct the user to a page asking them to continue.

We crafted a total of 44 emails. We altered the headers so they appeared to be addressed to “Sam Smith.” Four emails were copies of phishing emails. The links in these emails pointed to our attacks. Nine emails had links to the other login pages. Two emails were spam messages. The remainder were personally addressed to Sam. Some of these contained links to sites with no login page. The content of each email was based on email our group received. The legitimate email was randomly divided into five tasks. Four of the tasks contained one phishing email and the fifth contained an email leading to a legitimate site (Orkut) on which BayeShield falsely warned users it was suspicious. In order to preserve realism, we set up the phishing attacks so that FF2.0 “missed” one of them since it has a much lower detection rate than BayeShield. In each task, the order in which the emails appeared was randomly determined but kept static across all participants. Each participant was exposed one of five conditions (two participants, per condition, per treatment) that altered the order in which the tasks were presented (based on a partially-balanced Latin Square).

The user selects “From email” and the meter moves higher compared with Figure 4.

Task Description and Data Collection
We asked the users to progress through each task, pretending they were Sam Smith, by reading each email and informing us whether they would “reply,” “delete,” or “click on the link.” They were instructed not to write the reply, that they could delete the email and told to click any link they believed Sam would follow. Each participant was provided with a sheet of Sam's personal information. We recorded each site the user chose to enter Sam's information into, whether the user noticed the warning, and if they used BayeShield (when applicable).

Block 2 Overview
Block 2 focuses on BayeShield's ability to prevent users from falling for phishing attacks. All twenty participants were asked to complete four tasks presented in one of four conditions (arranged according to a balanced Latin Square design). Two of the tasks were phishing attacks. In the other cases BayeShield incorrectly warns the user about a legitimate website. Here is a brief overview of each task:

- **Email**: The participant clicks on a link in an email which warns them their account will be disabled in 24 hours if they do not log-in. BayeShield warns them of the phishing attack.
- **Copy/Paste**: The participant pastes a URL containing a misspelling (www.bank0famerica.com) into the browser and BayeShield warns them the site may be phishing.
- **Bookmark**: The participant selects a bookmark to their stock broker and BayeShield warns them. The site asks for a large amount of personal information including their Social Security Number (SSN).
- **Brochure**: The participant picks up a brochure after visiting a state park. The brochure contains a URL, they type it to donate to the park and BayeShield warns them the legitimate site is suspicious.

Data Collection
For each task we observed whether the user followed the “expected” path through the Analyzer, whether BayeShield correctly classified the site as “safe” or “not safe” and whether the participant agreed. We were particularly interested in agreement as it is related to their trust in BayeShield. We also kept track of the duration of each task and asked participants to estimate the durations of a task after they completing it.

Recruitment
We posted flyers at and near our University and advertised our study in class sessions. Potential participants were first directed to an online survey, all participants who completed the survey were selected to participate in the in-lab session.

Demographics
Participants completed a demographic questionnaire. Fourteen were female, six male. Our age range was 19 to 46 years of age with a median of 27.6 years. Our cohort was highly educated. All 20 of our participants completed some college and 8 of the 20 completed at least some graduate work. The majority of the cohort used FF2.0, followed by...
IE. They had used computers for an average of 15 years, used a computer for 5.5 hours a day and the Internet for 1.5.

RESULTS

Personal Information Submitted by Website Category
In block 1, participants had the option of entering Sam's personal information into thirteen sites. We classified the thirteen sites into the following categories (number of sites in each category in parentheses): banking/e-commerce (5), social networking (3), school/work related (2), news (1), fun/entertainment (1), insurance (1). 87.5% of participants entered information into official school-based websites, 75% into the insurance site, 45% into a social networking site. Only 24% of participants entered information into banking and e-commerce sites. While only four of these thirteen sites were phishing, the distribution provides and idea of the types of sites in which users are willing to enter personal information.

Block 1: BayeShield vs. Firefox 2.0
BayeShield prevented more participants from falling for phishing attacks than FF2.0. Sixty percent of BayeShield participants clicked on phishing links for a total of sixteen clicks. Of these participants, two fell for a total of four phishing attacks (one fell for three attacks, the other for one attack). BayeShield protected two-thirds of at-risk participants and convinced participants not to enter information 75% of the time after they clicked on links in phishing emails.

For FF2.0, 60% of participants clicked on links leading to phishing websites for a total of fourteen clicks. FF2.0 protected one third of at-risk participants (two fell for one attack, one for two attacks, and one for three attacks). In half the cases in which participants clicked on phishing links, they proceeded to enter data into phishing sites. Three FF2.0 participants fell for the Amazon phishing attack which FF2.0 failed to detect. This attack was the only one that showed a statistically significant difference for participants entering information into a phishing website between BayeShield and FF2.0 (p<.05, likelihood ratio chi-square test).

All BayeShield participants who clicked on the link leading to the Orkut false positive noticed BayeShield's warning but still entered Sam's information. Thus, BayeShield prevented more phishing attacks than FF2.0 without blocking any legitimate access.

Qualitative Observations
The BayeShield participants who fell for phishing attacks were special cases. Without prompting one volunteered, “I'm not a person who is concerned about online security so I'd do it [enter information] and deal with the consequences later.” The participant correctly interpreted BayeShield's warning and ignored it as a result of a personal philosophy. The other participant noticed BayeShield's warning but stated, “I wouldn’t know what to do with this.” This participant fell for three phishing attacks. The participant had not heard of phishing and was not aware of the dangers associated with it.

Participants spent more time examining BayeShield's warning, due in part to the required delay of four seconds before closing it. Participants rarely expressed surprise after examining BayeShield. In contrast, responses to FF2.0's warning included, “I've never seen this before,” and, “what's this?” This is likely due to its generic appearance.

Almost half of participants did not click on links, dismissing or deleting the phishing emails. When prompted for the reasons, participants did not name phishing but typically responded: “I think it’s spam,” “I get these all the time,” or “I am afraid it will hack my computer.”

Block 2: User Trust and Usability Evaluation
In block 2, after using BayeShield participants reached the correct verdict 80% of the time (even with a breakdown in our information architecture in the Brochure task). Participants agreed with BayeShield's verdict 85% of the time which corresponded with a high trustworthiness rating in the satisfaction questionnaires. All participants reached the correct determination of “not safe” in the Email task.

Email: Phishing attack
This task represents the most common user interaction with BayeShield and the results are encouraging. BayeShield correctly informed 19 participants it was “not safe” to enter their information (19 participants followed the expected path through BayeShield). All 20 participants correctly identified the email as “not safe.” Participants agreed with BayeShield 95% of the time.

Copy/Paste (Incorrect URL): Phishing attack
BayeShield correctly informed 17 participants it was “not safe” to enter their information. Fifteen participants decided it was “not safe.” Participants agreed with BayeShield 90% of the time and followed the expected path 80% of the time. Although users typically fail to notice URLs, when prompted by BayeShield, most participants recognized the URL contained an error or indicated they were uncertain it was the correct URL.

Brochure: Legitimate site
This task revealed an issue with our information architecture. We predicted most participants would select “from printed material” when asked how they arrived at the site but many participants selected “tying or copy/paste.” As a result, BayeShield correctly identified the site as “safe” only 65% of the time. Despite this, 16 participants correctly identified the site as safe. We were encouraged that participants agreed with BayeShield 80% of the time and will correct the mistake by adding an additional question to BayeShield.
Evaluation of User Satisfaction
Participants rated BayeShield highly. BayeShield scored well in the following categories of the satisfaction questionnaire\(^6\) (average in parentheses): trustworthiness (7.5), ease of use (8.5), clarity (8.2), language made sense (8), BayeShield kept the user informed (8.5), provided sufficient options (8), responsiveness (8.6), BayeShield had a clear purpose (8.75), confidence in BayeShield's accuracy (6.95). We present perceived vs. actual time sorted in the order the participant experienced the tasks to support our claim that BayeShield is highly usable (Figure 6). On average, each successive use of BayeShield took less time than the previous use. This was statistically significant (\(p<.05\), repeated measures analysis, Huynh-Feldt test). For all four tasks, the average perceived time was lower than actual elapsed time, indicating that participants were engaged in the task[2].

Qualitative Observations
While proceeding through the tasks, participants asked very few questions and did not seem hesitant. Several participants were impressed by the meter when they first saw it in operation. Two participants mentioned that we should hide the instructions screen after participants have used the software several times.

Supporting the claim that users were “in the flow” and engaged with the software is the fact that users read both instruction pages and often selected multiple options for each question before settling on one just to see how the meter would react. We had hoped for this response as we believe it will encourage users to more quickly differentiate between safe and risky behaviors based on the meter's rise and fall. Even with the time spent reading and exploring, users underestimated the amount of time they had spent.

One participant illustrated the inherent difficulty in protecting users, informing us, “this software works well but my own intuition is both faster and more secure.” They did not fall for an attack but displayed disturbingly risky behavior and relied on faulty logic to reach conclusions.

Participant Characteristics Affecting Behavior
We checked for correlations between all the information we collected about participant characteristics (demographic information, attitudes towards computers, familiarity with technical terms) and their performance in both blocks. We found many statistically significant correlations. In this paper, we present some of the most representative correlations. A full list may be found at http://www.bayeshield.com. We used Spearman’s rho to check for statistically significant correlations.

For block 1, we looked for correlations with entering information into phishing websites and with an additional “score” which we gave a value of -1 if participants decided not to enter information after clicking on a link, 0 if they did not click on a link, and 1 if they clicked on a link and entered information. Thus, the score made not clicking on a link neutral, gave a positive value to entering phishing information, and gave a negative value to heeding the warnings and not entering information after clicking on a link.

Block 1: BayeShield
For participants who used BayeShield in block 1, there were statistically significant correlations indicating that participants who are more familiar and comfortable with computers were less likely to fall for phishing attacks. For example, agreeing with the statement “Using a computer is very frustrating” was positively correlated with score for the Amazon phishing attack (.688, \(p<.05\)). Similarly, agreeing with “I feel comfortable using a computer” was negatively correlated with score on the PayPal attack (-.702, \(p<.05\)). We also found statistically significant correlations where participants who prefer using a computer to watching television were less likely to fall for phishing on the Wachovia (-.677, \(p<.05\)), Amazon (-.820, \(p<.01\)) and PayPal (-.677, \(p<.05\)) attacks based on score, with similar results for participants who prefer using a computer to reading (-.697, \(p<.05\)) for both the Wachovia and PayPal attacks. On a humorous, but potentially interesting note, some of the highest correlations we found were between familiarity with the term “flux capacitor”\(^7\), and score for

---

\(^6\)Participants answered on a scale from 0-9

\(^7\)http://en.wikipedia.org/wiki/Back_to_the_Future
Wachovia (-.677, p<.05), Amazon (-.818, p<.01), and PayPal (-.677, p<.05) attacks. It seems people familiar with 1980s science fiction movies pay more attention to warnings.

**Block 1: Firefox**

For participants using FF2.0 in block 1, there were similar statistically significant correlations indicating that participants who feel more comfortable with computers were less likely to fall for phishing attacks. For example, agreeing with the statement “I am tired of using a computer” was positively correlated with entering personal information on the Wachovia and PayPal attacks (.640, p<.05 for both). Likewise, familiarity with the term “Ethernet” was negatively correlated with entering personal information on the Amazon attack (the one that FF2.0 did not detect) (-.757, p<.05).

On the other hand, there were statistically significant correlations indicating that people more familiar with some computer terms were more likely to ignore FF2.0’s warnings. For example, familiarity with the term “USB device” was positively correlated with the score for the Bank of America and the PayPal attacks (.667 and .645 respectively, p<.05 for both). There were similar correlations for familiarity with “web browser” and “ISP”. This suggests that participants who are more computer literate and comfortable with computers are less likely to click on links to phishing websites, but if they do, they are more likely to ignore FF2.0’s warnings.

Other interesting findings related to personalities include a negative correlation between agreement with “I do things by myself without depending on others” and entering information on the Wachovia and PayPal attacks (-.787, p<.01 for both), and a similar negative correlation between agreement with “I examine unknown issues to try to understand them” and entering information on the Wachovia and PayPal attacks (-.640, p<.05 for both).

**Block 2**

For block 2, we looked for statistically significant correlations with whether BayeShield provided the correct verdict, participants agreed with BayeShield, and participants followed the correct path. Overall, we found that participants more comfortable with computers were more likely to perform better by following the correct path, reaching the right verdict and agreeing with it. For example, agreeing that it is more difficult to use a computer than to write was negatively correlated with all three measures for the email task (-.546, p<.05) and having BayeShield reach the correct verdict on the copy/paste task (-.608, p<.01). For the bookmark task, familiarity with computer terms such as Bluetooth (.694, p<.01), DSL (.456, p<.05), Ethernet (.455, p<.05) and IP address (.473, p<.05) was positively correlated with participants reaching the correct verdict.

We also found that participants who are more comfortable with computers spent less time using BayeShield. For example, agreeing with the statement “I feel comfortable using a computer” was negatively correlated with the amount of time participants interacted with BayeShield on the bookmark task (-.444, p<.05), while agreeing with “I think it takes a long time to accomplish tasks using a computer” proved an accurate prediction in having a positive correlation with the amount of time spent on the bookmark task (.586, p<.01). Similarly, agreeing with “I can learn more from books than a computer” was positively correlated with the time spent on the copy/paste task (.517, p<.05).

We also looked for correlations with the difference between actual and perceived times for block 2. Positive differences suggest participants are more engaged in tasks, while negative differences suggest participants do not have a flowing interaction with the technology[2]. The correlations suggest that participants who feel less comfortable using computers were more engaged in completing tasks with BayeShield in block 2. For example, agreeing with “I believe it is important for me to learn how to use a computer” was positively correlated with the difference for the email task (.536, p<.05). Agreeing with “working with a computer makes me feel nervous” was positively correlated with the difference for the bookmark task (.560, p<.05), while agreeing with “I feel comfortable working with a computer” was negatively correlated with the difference for the same task (-.511, p<.05).

**DISCUSSION**

Although BayeShield outperformed FF2.0, the only statistically significant difference was for the phishing website for which BayeShield provided a warning and FF2.0 did not. Given that BayeShield has a higher level of detection of phishing websites than FF2.0, our results suggest that BayeShield would outperform FF2.0 in real-world situations.

We found the email task from block 2 to be highly encouraging insofar as all participants identified the website as a phishing attack when presented with our warning. User trust deteriorates quickly when software looks unprofessional or communicates poorly. In order to be certain BayeShield was well designed, our task descriptions gave participants no a priori reason to trust the software. Instead, participants relied on their interactions with BayeShield. Even so, participants agreed with BayeShield 85% of the time across all four tasks in block 2 and BayeShield received high marks for appearance, communication and performance in the satisfaction questionnaire. Finally, only two participants entered information on a phishing site (on the copy/paste task) in block 2 after being warned that the site was not safe.

The statistically significant correlations we found provide...
valuable evidence that should guide future research efforts. It uncovered two challenges. The first is to keep users who feel comfortable using computers from ignoring warnings, which was a problem for FF2.0 but not for BayeShield. The second is to better help those who do not feel as comfortable in navigating security software. While BayeShield did a good job engaging participants who often do not feel comfortable using computers and have lower computer literacy, these participants did not make a correct determination on whether a site was phishing as often as others who are more experienced. In addition to this, the wealth of information we obtained on user characteristics that affect the likelihood of phishing should encourage future researchers to do the same and go beyond our efforts to better understand the factors that make it more likely for people to fall for phishing attacks.

FUTURE WORK
As future work, our group proposes a longitudinal study designed to capture whether or not users' abilities to distinguish between legitimate and phishing sites improves after repeated exposure to BayeShield. In addition, our group is interested in evaluating the interaction between trust and security software appearance. An additional line of research of interest to us is in developing educational security software for children aged 12-15 and older adults.

CONCLUSION
We have presented BayeShield, a novel anti-phishing tool, the process that led to its current design, and a study comparing it against FF2.0 and evaluating its usability. The results of our study suggest that BayeShield provides better protection against phishing attacks than FF2.0. While noting a few areas in which we can improve BayeShield's communication with the user, our results suggest that BayeShield is highly usable based on differences in perceived vs. actual time, and high marks in our satisfaction questionnaire.

ACKNOWLEDGMENTS

REFERENCES
11. Anonymized for review.