Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings

Justin Petelka
The Information School
University of Washington
Seattle, WA, USA
jpetelka@uw.edu

Yixin Zou
School of Information
University of Michigan
Ann Arbor, MI, USA
yixinz@umich.edu

Florian Schaub
School of Information
University of Michigan
Ann Arbor, MI, USA
fschaub@umich.edu

ABSTRACT

Phishing emails often disguise a link's actual URL. Thus, common anti-phishing advice is to check a link's URL before clicking, but email clients do not support this well. Automated phishing detection enables email clients to warn users that an email is suspicious, but current warnings are often not specific. We evaluated the effects on phishing susceptibility of (1) moving phishing warnings close to the suspicious link in the email, (2) displaying the warning on hover interactions with the link, and (3) forcing attention to the warning by deactivating the original link, forcing users to click the URL in the warning. We assessed the effectiveness of such link-focused phishing warning designs in a between-subjects online experiment (n=701). We found that link-focused phishing warnings reduced phishing click-through rates compared to email banner warnings; forced attention warnings were most effective. We discuss the implications of our findings for phishing warning design.

CCS CONCEPTS

• **Security and privacy** → **Usability in security and privacy**; **Intrusion/anomaly detection and malware mitigation**; • **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Phishing; warning design; usability; security; privacy.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300748

ACM Reference Format:
Justin Petelka, Yixin Zou, and Florian Schaub. 2019. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3290605.3300748

1 INTRODUCTION

Phishing attacks typically involve an unsolicited email, text message, or telephone call purportedly from a legitimate organization requesting personal, financial, and/or login information [40]. The Anti Phishing Working Group reported half a million unique phishing campaigns in the second half of 2017 [6]. The FBI reported 25,344 victims of phishing in 2017, resulting in the loss of 30 million U.S. dollars [40]. Email is the most common medium for phishing attacks [61]. Phishing emails often contain links which, once clicked, take the victim to a legitimate-looking website, where victims are asked to input login, personal, or financial information. This information is then harvested by the attackers to gain unauthorized access to personal accounts, sometimes as the first step in a more sustained attack against an organization [53]. Phishing attacks vary with respect to target specialization (e.g., spear-phishing, leading to business email compromise) [27, 61] and attack vector (e.g., malicious attachment or phishing link). "Fire-and-forget" phishing attacks are the most common [53], where attackers send a large number of phishing emails with the objective of tricking a small percentage of users to click on a phishing link and visit the phishing website.

Efforts to combat phishing include (1) training users to identify phish [32, 49, 57], (2) automated identification of phishing emails [62, 64], domains, and websites [21, 59], and (3) providing warnings to aid users in spotting suspected phishing emails [55] or websites [18]. Common advice given in phishing training places the onus of discovering phishing links on the user, urging them to "hover to discover" [48] suspicious links and Uniform Resource Locators (URLs). Automated phishing detection approaches analyze link and URL features to determine whether an email or webpage is a phish [21, 39, 59, 62, 64], allowing certain phishing emails to be removed before users encounters them. However, despite high accuracy [39], phishing algorithms are probabilistic and produce false positives (i.e., removing legitimate emails), especially when deployed at scale. Phishing warnings commonly augment automated detection to manage detection uncertainty, and limit the impact of false positives by allowing users to override the system [17].

Phishing and other online security warning research has largely focused on browser-based warnings and indicators [2, 18, 19, 19, 50, 63]. *Browser warnings* appear when a user is attempting to load a suspicious website. However, if the email is the attack vector, a browser warning appears after a user has already decided to visit the link, at which point it may be difficult to overcome anchoring bias and loss aversion [1]. A potential solution is to warn users before they click an email's suspicious link, i.e., within the email client. Some current email clients use *banner warnings* (see Figure 2), which notify users that an email might be suspect. However, such banner warnings typically do not explain what specifically is suspicious. The lack of specific information in such banner warnings places the burden of locating suspicious cues (e.g., regarding the link) on the user. This increases the risk that the warning is ignored or misunderstood.

We propose and evaluate three warning design features for supporting users in more effectively assessing phishing risks and avoiding phishing websites. The first feature is (1) *warning placement.* In particular, we evaluated the impact of placing warnings near a suspicious link compared to banner or browser warnings. Similar to Volkamer et al.'s work [55], we developed *link-focused warnings*, which clearly display the underlying URL of a suspicious link, thus making it easier for users to notice discrepancies between where they expect the link to go and its actual destination. The second feature is (2) *forced attention.* We force the user's attention to the warning content [11] by deactivating the suspicious link in the email body and forcing the user to click the unmasked URL in the warning if they want to proceed. Forced attention allows users to safely hover over a link to identify it without the risk of clicking, while adding a small cognitive burden to the risky option [9]. The third feature is (3) *warning activation.* We hypothesize that dynamically displaying a warning only when the user hovers over a link would increase warning adherence compared to a static warning.

We assessed the effectiveness of these warning design features in a between-subjects controlled experiment (n=701) conducted on Amazon Mechanical Turk. Our evaluation metric for warning effectiveness was click-through rate (CTR). Our experiment had seven conditions: four link-focused warnings factored by forced attention (yes vs. no) and warning activation (activated on-hover vs. statically displayed), an email banner warning (based on Gmail's banner warning), a browser warning (based on Chrome's browser warning), and a no-warning condition (control). Participants were asked to test whether links worked in a set of emails presented in an online email client.

Our research contributes a new controlled method for evaluating phishing warning effectiveness, empirical evidence of different warning features' influence on CTR, and guidelines for phishing warning design. Our key findings are:

- Link-focused phishing warnings significantly reduced click-through rate compared to email banner warnings and no warning. This shows that warning placement has a significant impact on phishing warning adherence.
- Forced attention, i.e., requiring users to click the actual URL displayed in the warning, resulted in significantly lower click-through rates compared to other link-focused phishing warnings. This suggests that forcing interaction with the warning improves phishing warning adherence.
- Whether link-focused warnings were static or appeared only when hovering over the link had little effect on click-through rate. This suggests that warning placement and forced attention are more important for phishing warning effectiveness than the method of activation.

2 RELATED WORK

Phishing counter measures in HCI research fall into three major categories [32]: training users to detect and avoid phishing emails, silent identification and elimination of phishing emails and domains, and warning users about phishing emails and domains.

Phishing Training and Behavior

People become victims of phishing attacks because of inattention or not understanding browser-based cues regarding a website's authenticity [16]. In contrast, people who check URLs are less likely to click on a phishing link [17].

The importance of user perception and understanding for avoiding phish has been incorporated into anti-phishing training approaches. The Anti-Phishing Phil game [49] trained users to identify phishing URLs. PhishGuru [32] sent fake phishing mails to an organization and provided anti-phishing training once a person has clicked on one of those phishing links. The training emphasized phishing email cues, such as a potentially fraudulent sender or non-matching URLs (links that do not match the status bar). More recently, Wash and Cooper conducted a simulated phishing campaign similar to PhishGuru and found that stories are more effective when heard from peers, but facts and advice were more effective when heard from professionals [57]; four of six points covered in their training materials related to inspecting URLs by hovering over links. Lastdrager et al. evaluated the effectiveness of anti-phishing training for children [34]. When introducing phish identification, the first clue given to participants was "how to find a URL from a hyperlink and how to assess where a URL leads to."

While identifying a link's URL is common anti-phishing advice, current email clients do not support this task well. For instance, users are told to hover over links to check their URLs, but are not supported in doing so safely. Users should be able to evaluate a suspect link without being placed at risk of clicking a link. When the link is suspect, email phishing warnings should be *link-focused*, i.e., appear near the suspect link, clearly display the link's URL, and the suspect link should be unclickable so users can safely evaluate the URL.

Phishing Detection

Another approach to thwart phishing attacks is to automatically detect phishing websites or emails. Common strategies include blacklists of confirmed phishing website URLs [37, 45] and automated probabilistic detection. Such phishing detection algorithms analyze URL, website or email features to determine the probability that a website or email is a phish. Methods for phishing website detection include term frequency-inverse document frequency (TF-IDF) [64], semantic data models [58], and machine learning [38, 39, 59, 62]. Similarly, phishing email detection analyzes features such as message content, email structure, sender information, target identification, and links in the email [8, 14, 21, 62, 64]. While these approaches achieve high accuracy (e.g., CANTINA+ had a best case false positive rate of 1.35% [62]), problems still arise at scale. Considering an estimated 269 billion emails are sent each day [28], and one in 2,000 emails are phish [51], even such high accuracy in phishing email classification would still produce over 1.8 million false positives every day. Therefore, a common approach is to augment phishing detection with phishing warnings.

Security Warnings

Security warnings have been a particular focus of usable privacy and security research [24], including website security indicators [2, 19, 19, 47, 50], software installation dialogues [12], mobile app permissions [3, 26, 43, 44], privacy notices [15, 29, 46, 52], and phishing warnings [18, 55]. Security warnings and indicators aim to make users aware of a potential hazard and help them take informed actions. However, repeated exposure to warnings can result in habituation [30]. Proposals to mitigate habituation include dynamic [13, 18] or polymorphic warnings [4], or adding warning attractors to capture users' attention [9, 10, 12]. Other work has shown that certain security indicators and warnings may go unnoticed, such as the HTTPS icon in a browser's address bar; they may be misunderstood, or they may not be heeded [16, 47]. SSL certificate warnings also face misunderstandings [11] and lack of adherence [2, 19, 20]. Different variations in warning design have been shown to improve (or reduce) a warning's effectiveness [2, 4, 13, 18, 19, 19, 50].

Regarding phishing warnings, Egelman et al. found that active phishing warnings, which interrupt the user's process, are more effective than static indicators [18]. Akhawe and Felt evaluated the click-through rate of malware and phishing warnings using browser telemetry in Google Chrome and Mozilla Firefox, and found click-through rate differed by browser, release channel, and operating system [2]. Volkamer et al. found that drawing attention to a pruned URL (only the domain, e.g. www.chase.com) [54] and tooltip warnings appearing when hovering over a link [55] significantly helped people identify phishing URLs. They further used a time delay (i.e. a time limit before the link is clickable) as an inhibitive attractor. However, Volkamer et al. only evaluated their warning against a non-warning control, which makes it difficult to ascribe observed effects to specific warning design features. Instead, we conducted a between-subjects experiment to understand and isolate the effects of different warning features on click-through rate: warning placement (link, email banner, browser warning), warning activation (hover, static), and forced attention (yes, no). With our study, we validate Volkamer et al.'s finding that link-focused warnings are more effective than no warning. Additionally, we provide empirical effectiveness comparisons across different types of email client warnings.

3 LINK-FOCUSED PHISHING WARNING DESIGN

Security and phishing warnings have been studied extensively. We contribute to this body of work by assessing in a controlled manner what warning features effectively support people in applying prominent anti-phishing advice, namely checking a link's URL before clicking on it, and thus avoid phishing links. Building on prior work and warning design principles, we studied the effects of warning placement, forced attention, and warning activation in a controlled experiment. In order to be able to isolate the effects of specific warning features, we designed a link-focused phishing warning that allowed for controlled variation of variables rather than comparing existing warnings as-is.

Our link-focused phishing warning design, as shown in Figure 1, appears near a suspicious link, similar to Volkamer et al.'s proposal [55]. We implemented forced attention similar to Bravo-Lillo et al. [12] by deactivating the suspicious link in the email body and forcing the user to click the unmasked URL in the warning if they want to proceed. This allows users to safely hover over a link without the risk of clicking, while still allowing them to proceed. We further studied the effect of a warning's activation, since active warnings have been shown to improve warning effectiveness compared to static warnings [18]. Our active warnings displayed on-hover [55], while the static warnings were displayed statically with the email. Next, we elaborate on our design rationale for link-focused warnings with respect to warning placement, warning activation, warning content, and the use of forced attention.

Warning Placement

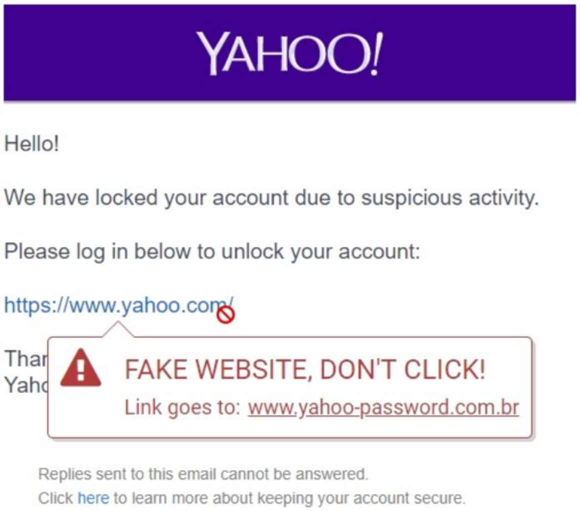
Wogalter et al. suggest that to be effective, a warning should be placed close in space and time to the hazard it guards against [60]. For victims of email-based phishing attacks, the hazard begins at the phishing email. A *browser*-based phishing warning (see Figure 3) appears after the user has already decided to click on an email link, at which point the browser warning must overcome the user's anchoring bias and loss aversion [1] to prevent them from proceeding to the phishing site. Phishing warnings in the email client can potentially prevent users from clicking a suspicious link in the first place. Current phishing warnings in email clients, such as Gmail's *banner* warning (see Figure 2), appear at the top of an email. Banner warnings often warn users to be cautious and state general anti-phishing advice. However, for phishing emails in which a link is the hazard, these warnings may (1) appear too far from the specific phishing link and (2) provide little support for helping users identify the specific hazard. Therefore, similar to Volkamer et al. [55], we propose placing *link-focused phishing warnings* inside the email, in close proximity to the suspected phishing link (see Figure 1).

Warning Activation

Prior research indicates that active warnings (those that interrupt the user's process) are more effective than static warnings [18]. This approach also aligns well with typical phishing training [32, 48], recommending that one should hover over links and check the URL before clicking. There are two options for realizing an active link-focused phishing warning: when the user hovers over a link [55] or clicks it. Since the warning's objective is to prevent people from clicking the phishing link at all, our warning appears when a user hovers over a suspicious link — more specifically, the warning is triggered by the `onmouseenter` event and remains visible until the cursor leaves the link or warning area and a 250 ms delay has elapsed.

Warning Content

Typically, when hovering over a link, the link's URL is displayed in the browser's status bar, which is decoupled from the suspicious link and may be ignored by the user. Lin et al. found that domain highlighting helps some people identify phishing websites [36]. Volkamer et al. found that displaying a pruned URL (i.e., only the domain without path or parameters, e.g., www.chase.com) significantly helped people identify phishing URLs [54].



Hello!

We have locked your account due to suspicious activity.

Please log in below to unlock your account:

https://www.yahoo.com/

Thar
Yahc

FAKE WEBSITE, DON'T CLICK!

Link goes to: www.yahoo-password.com.br

Replies sent to this email cannot be answered.
Click here to learn more about keeping your account secure.

Figure 1: Our link-focused phishing warning with forced attention. The warnings appears when a user hovers over a phishing link. The warning uses sparse text and draws attention to the link's actual URL in order to facilitate its inspection. We further force a user's attention to the warning content by deactivating the link in the email when the warning appears (and mouse cursor turns into a "no" symbol), forcing the user to click the URL displayed in the warning if they choose to proceed.

Therefore, we display a link's unmasked and pruned URL in our warning. This reduces the cognitive burden of having to look away from the link in order to evaluate it, and makes it easier for users to recognize when the URL differs from the domain they expect. The URL shown in the warning is also clickable, allowing the user to proceed to the website if they so choose with minimal extra effort. We also added a clear indication that the URL displayed in the warning is the link's actual destination (see Figure 1). Following hazard warning guidelines [60], the warning further includes a red triangle with an exclamation symbol and a warning heading.

For the heading text of our link-focused warnings, we considered the heading of Chrome's SafeBrowsing warning ("Deceptive site ahead;" see Figure 3) and a variant in plainer language ("fake" instead of "deceptive"), as it might be more succinct and legible, which is important for effective warning design [7, 60]. We conducted a between-subjects pre-study with Amazon MTurk (n=207) with two groups to determine the word choice for the warning heading. Participants were paid $0.30 for a 5-minute survey. Participants saw an image of an email that contained the warning (either with “fake” or “deceptive”). We then asked participants to describe in an open-response question what the warning text meant; we scored answers correct if they described the hazard as the link or the website that it leads to (evaluation), and mentioned phish or stolen credentials as opposed to spam (accuracy). Then we asked participants 2 seven-point Likert scale questions about how likely they would be to heed the warning, and how they would rate the effectiveness of the warning.

We found no significant differences between our groups, so we evaluated each of these four categories by the mean response. By this measure, our participants responded favorably to the word “fake” across all four measurements. Participants more frequently identified the hazard as the link or the link’s website (85%) and described the consequences of clicking the link correctly (77.5%). Participants also rated “fake” as being effective (5.25 out of 7) and would be less likely to proceed past the warning (1.55 out of 7). Thus, we used “Fake Website” as the warning heading and further added a call to action to clarify that one should not proceed. The final warning heading for the link-focused warning was “Fake Website, Don’t Click!”

Forced Attention

Hovering over suspicious links to inspect URLs to see warnings has two weaknesses: (1) it increases the risk of accidentally clicking the link, and (2) it is easy for users to miss or ignore the link-focused warning. Bravo-Lillo et al. proposed inhibiting warning attractors that temporarily prevent the user from taking a risky action until a set amount of time has passed or the user has taken a specific action [10, 12]. Volkamer et al. [55] used a time delay before the original link in the email becomes clickable. We use an action-based inhibitor instead. More specifically, we deactivate the link in the original email, but include the pruned URL as a clickable link in the warning. Thus, the user cannot just click through and ignore the warning, but has to click the pruned URL in the warning if they want to proceed. The user can still take the risky option without having to wait, but needs to overcome a small cognitive burden [9]. This approach also guards the user when evaluating the potential hazard [60], since a user can check a link’s URL by hovering without being at risk of clicking it.

4 STUDY DESIGN

We conducted a between-subjects online experiment to evaluate the effectiveness of three different warning features (forced attention, activation, placement) with our warning design. Participants were asked to log into an online email client and evaluate ten emails, three of which contained a phishing link. Specifically, we used a 2x2 design to isolate effects of warning activation and forced attention for link-focused warnings. We then compared the link-focused warnings to a banner and a browser warning (for which warning activation and forced attention would not be meaningfully comparable), as well as a no-warning control condition, resulting in seven conditions in total. We first describe the conditions in more detail, before outlining the study protocol, our online email client, the emails and phishing URLs used, and our analysis methods.

Link-focused Warning Conditions

The four link-focused warning conditions (shown in Figure 1) only varied in whether the warning was activated by hovering over the link or was statically displayed (*warning activation: hover vs. static*); and whether forced attention was used (*forced attention: yes vs. no*). In the non-forced attention conditions, both the original link in the email and the link in the warning were clickable; in the forced attention conditions, only the link in the warning was clickable.

To study differences in user behavior when confronted with different warning placements, we further created a comparable email banner warning and browser warning. Both warnings were modelled after existing warnings (Google’s Safe Browsing[1] warnings for Gmail and Chrome, respectively). However, to isolate the effects of our independent variables (i.e., placement, forced attention, activation), the style and content of the browser and banner warnings were adjusted to be consistent with our link-focused warnings.

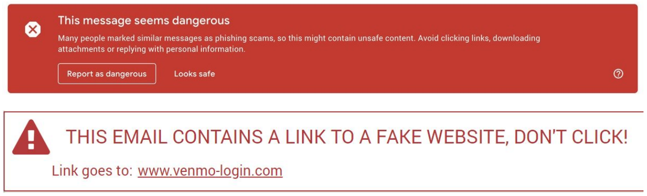
Figure 2 and Figure 3 show our adaptations in comparison to the original Gmail banner warning and Chrome’s browser warning, respectively. For the email banner warning, we adjusted the warning heading from the link-focused warning so the email was the focus of the warning (as opposed to the website) and kept the unmasked URL in the warning even though Gmail’s banner warning does not highlight suspicious URLs. Otherwise, we kept the warning content consistent with the link-focused warning. The browser warning appears after clicking on an email’s phishing link, so we used Chrome’s warning heading but replaced “deceptive” with “fake” to be consistent with our link-focused warning. We further reduced the amount of text in the browser warning to achieve consistency with the other warning conditions.

To obtain a baseline of participants’ unaided phishing detection ability and behavior, participants in the control condition saw no phishing warning, but could still see a link’s URL in the browser’s status bar.

Study Protocol

We conducted our study via Amazon Mechanical Turk. Our study used deception to avoid priming participants about

[1] https://safebrowsing.google.com/



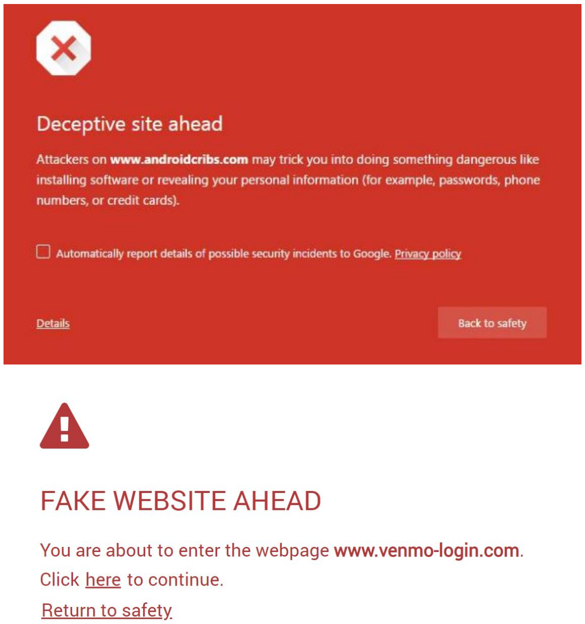
This message seems dangerous

Many people marked similar messages as phishing scams, so this might contain unsafe content. Avoid clicking links, downloading attachments or replying with personal information.

THIS EMAIL CONTAINS A LINK TO A FAKE WEBSITE, DON'T CLICK!

Link goes to: www.venmo-login.com

Figure 2: Gmail's (*top*) and our adapted (*bottom*) email banner phishing warnings. Both warnings appear above the email body.



Deceptive site ahead

Attackers on **www.androidcribs.com** may trick you into doing something dangerous like installing software or revealing your personal information (for example, passwords, phone numbers, or credit cards).

☐ Automatically report details of possible security incidents to Google. Privacy policy

FAKE WEBSITE AHEAD

You are about to enter the webpage **www.venmo-login.com**. Click here to continue.

Return to safety

Figure 3: Chrome's (*top*) and our adapted (*bottom*) browser phishing warnings. Both warnings appear in the browser when a person is about to visit a phishing website.

phishing risks. We described the goal of the task as evaluating whether links work in emails in order to help automated detection of dead hyperlinks.

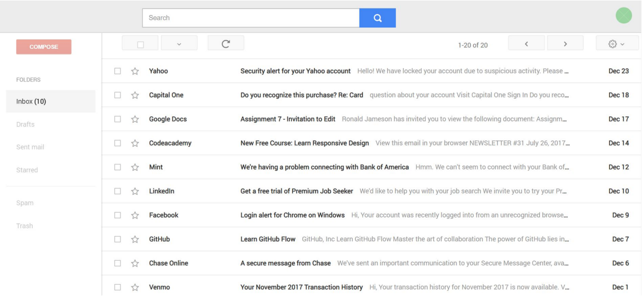
Upon accepting our HIT, participants were randomly assigned to one of the seven conditions. Participants were told that they would be given access to an email inbox containing ten emails drawn from a large dataset of real emails, and asked to provide information about each email in our survey. To further increase ecological validity, participants were then given a unique username and password and a link to log into 'their' inbox. We implemented an instrumented online email client, based on Gmail's aesthetics, in which participants then interacted with emails.

Participants were asked to log into the email client with their credentials, go through each email in their inbox, and answer questions about each email's contents in our Qualtrics survey. For each individual email, participants were asked (a) What is the title of the email?, (b) How many hyperlinks are in the email?, and (c) Are all the hyperlinks in the email working? These questions were designed to incentivize participants to click on links without explicitly instructing them to do so or priming their behavior. Participants have the freedom to click, hover, or use some other means of subjective evaluation.

Each inbox had the same ten emails in randomized order. Three of these emails contained phishing links. When participants opened any of the three phishing emails, they would encounter the phishing warning corresponding to their condition (i.e., banner and static link-focused warnings would be immediately visible; hover-activated link-focused warnings would load if the user hovered over the link; the browser warning would appear after the participant clicked the phishing link; no warning in the control). We recorded participants' click and hover interaction with the links in the inbox and emails.

After a participant finished evaluating all ten emails, we asked follow-up questions about their cybersecurity knowledge, past data theft experience, their frequency of using the vendors used as phishing targets, their impressions of our warnings, and demographic questions. Cybersecurity knowledge and past data theft experiences were assessed using questions from Pew Research Center surveys [41, 42]. Phishing vendor usage was measured on a 6-point scale for each company ranging from "less than once a month" to "daily", with an option for "I'm not sure". We assessed participants' impression of the warnings by asking them to recall if they had seen the warning, how useful the warning was in identifying the link, how annoying the warning was, and how much the warning affected their perception of the link. We also tested their comprehension of the warning by asking them to indicate which parts of the email seemed suspicious, with options including "the link shown in the email text," "the sender," and "the subject line." We further asked for optional feedback on how to improve the warning in a free text box before demographic questions.

At the end of the study, we debriefed participants that the true intention of the study was the assessment of phishing warnings and that their hover and click interactions with the email client and warnings had been recorded. They were provided with a completion code to copy-and-paste back into MTurk. Participants were compensated with $5.00 for work that was expected to take 20-25 minutes. Our study was approved by our institutional review board (IRB). The



Yahoo	Security alert for your Yahoo account	Hello! We have locked your account due to suspicious activity. Please ...	Dec 23
Capital One	Do you recognize this purchase? Re: Card	question about your account Visit Capital One Sign In Do you reco...	Dec 18
Google Docs	Assignment 7 - Invitation to Edit	Ronald Jameson has invited you to view the following document: Assignm...	Dec 17
Codeacademy	New Free Course: Learn Responsive Design	View this email in your browser NEWSLETTER #31 July 26, 2017...	Dec 14
Mint	We're having a problem connecting with Bank of America	Hmm. We can't seem to connect with your Bank of...	Dec 12
LinkedIn	Get a free trial of Premium Job Seeker	We'd like to help you with your job search We invite you to try your Pr...	Dec 10
Facebook	Login alert for Chrome on Windows	Hi, Your account was recently logged into from an unrecognized browse...	Dec 9
GitHub	Learn GitHub Flow	GitHub, Inc Learn GitHub Flow Master the art of collaboration The power of GitHub lies in...	Dec 7
Chase Online	A secure message from Chase	We've sent an important communication to your Secure Message Center, ava...	Dec 6
Venmo	Your November 2017 Transaction History	Hi, Your transaction history for November 2017 is now available. V...	Dec 1

Figure 4: Our online email client, modeled after Gmail's interface.

study materials, online email client code, and analysis code are publicly available.[2]

Online Email Client

Participants interacted with an online email client modeled after Gmail's interface, shown in Figure 4. Participants could log in and interact with emails, but additional buttons and features (e.g., "Compose") were visibly disabled. The email client did not have a name or logo to reduce brand effects.

While participants interacted with our email client, we recorded specific events: click and hover interactions, when a warning was rendered, and other metadata (e.g., link URL, their IP address). Events were detected on the client side with Javascript and sent back to our web server using AJAX requests. We also recorded website requests on the server side, both for pages of the email client, as well as the phishing domains used.

Consistent with prior work [2, 19, 31, 55], we used click-through rate (CTR), or "adherence rate," [19] as the metric to measure warning effectiveness: the number of people who clicked on the link, divided by the number of people who saw the link. The lower the CTR on phishing links, the more effective the warning is. We also recorded whether participants' hovered over links and for how long, which provides an indication of how long they engaged with a warning or URL displayed in the browser's status bar. We estimated that the threshold after which the browser's status bar is displayed is 250 ms. Since hover interactions less than 250 ms would not display a link's URL, we only recorded hover events longer than 250 ms. We further logged when a warning was rendered to assess potential changes in behavior after encountering multiple warnings. Using these event records, we split a participant's data into segments separated by the first, second or third warning displayed.

Email Selection and Phishing URLs

All participants saw the same ten emails in random order. These emails were modelled after real emails the authors had received. For benign (i.e., non-phish) emails, links that may have been difficult to see were removed to facilitate the participants' primary task, but were otherwise unaltered. Each email contained between two and seven unique links.

Three of the emails were modified to contain a link to a phishing domain controlled by the authors. We modeled our phishing domains and URLs based on actual phishing emails we received from our institution's IT department and trends in phishing reports [5, 25, 53]. The vendors (Chase, Venmo, Yahoo) were chosen because they are industries and/or organizations that are frequently targeted by phishing campaigns [5, 33]. The three domain names all contained a word or words related to the organization or industry (i.e., "chase" and "banking"). As top-level domains we used .com, .br, and .us. Each of these phishing emails contained a call to action such as "Click Here." We replaced the underlying URL of the call-to-action links with a phishing URL, thus replicating a common phishing attack in which an otherwise authentic looking email contains a non-matching URL [5, 21, 27, 62, 64]. Each phishing email had either two or three unique links, with the remaining unchanged links pointing to ancillary information such as privacy policies.

The three phishing domains pointed to our study server, ensuring participant safety. Both our email client and the phish links were configured for HTTPS, since APWG reports an increase in phishing websites using certificates [6]. When the server received a request for one of the phishing domains, the request was logged, and the server merely redirected the participant to the legitimate website.

Data Cleaning

After data collection, we found it necessary to filter out certain responses. We removed participants from the same IP address who began the survey within two hours of each other. Removing these participants ensures participants did not have prior knowledge of our study's deception. We further removed participants who did not appear in our server logs, and participants who answered factual questions about the emails incorrectly. Finally, we removed participants whose server records showed unread emails. After removing 56 incomplete and invalid responses, our final sample consisted of 701 participants.

All 701 participants opened all 10 emails and had click and/or hover interactions within every email. Furthermore, all participants responded with an approximation of the correct title of each email in our survey. This suggests that our participants attempted our task in earnest.

[2] https://github.com/spilab-umich/phishing-warning-experiment

We further analyzed the performance of websites under our control (email client and phishing domains). When a user clicks a link to one of our websites (i.e. the inbox, each email, and our phish domains), we should receive a click event from the client and a server record that the request was processed. Our results show that on average 8.02% of server records per condition did not have a corresponding client event log. This is likely due to interactions that we did not capture, such as the back and refresh buttons, or participant behavior. In an optional free-response question in our survey, several participants indicated they copied the phishing links and opened them in a separate window, browser or private browsing mode. Other participants mentioned our websites triggered their 3rd party security software. Note that this has no effect on the reliability of our phishing click-through rate analysis, because requests for the phishing domains were logged server-side. However, benign click-through rates and hover times may be under-reported, though equally across conditions. We discuss the implications of this on our findings where applicable.

Analysis Method

To examine how warning designs and personal characteristics affected participants' behavior, we conducted mixed-effect logistic regressions on click and hover actions (outcome: yes vs. no), and mixed-effect linear regressions on hover time (a continuous variable with 250 ms as the minimum value). For logistic regression results, we report the odds-ratio of the predictor as the effect size and its corresponding p-value. For linear regressions, the dependent variable (i.e., hover time) was normalized to maximize the model's accuracy.

Our mixed-effect regression models consisted of fixed-effect and random-effect variables. Fixed-effect variables included (1) warning-related variables: placement (link-focused, banner, or browser), activation (whether the email was activated by hover/click or loaded with the email), and forced attention (whether the phish link was unclickable or not); and (2) participant-related variables: cybersecurity knowledge, past data theft experience, vendor usage, and demographic information (gender, education, age, and occupation). Random-effect variables measure the random effects resulting from the differences between individual links and emails when fixed-effect variables are under control. We applied our models to both phishing links and benign links.

We further conducted qualitative analysis on participants' open-ended responses regarding (1) how the warning affected their perception of the link and behavior (Perception), and (2) how they would improve the warning (Improvement). Of 701 participants, we received 528 valid responses for Perception, and 89 for Improvement. We used thematic coding and affinity diagramming [35] to develop a codebook for analyzing these responses. Two researchers independently coded 20% of the responses, reconciled results and refined the codebook until reaching a high inter-rater reliability (Cohen's κ=83.3 for Perception; κ=73.3 for Improvement). One researcher then used the updated codebook to code all responses.

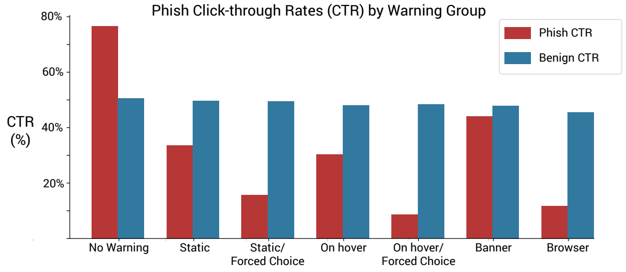


Figure 5: Click through rates for our warning designs for phishing links (*red*) and benign links (*blue*).

5 RESULTS

Our results show that placement influences click-through rate, as link-focused warnings were significantly more effective than banner warnings; browser warnings performed better than link-focused warnings, but this is likely due to the study design. Within link-focused warnings, we found that forced attention (i.e. restricting access to the link) was effective at reducing CTR. Warning activation had no significant impact on CTR.

Participant Demographics and Profile

Of our 701 participants, 395 responded male (56.34%), 301 female (42.94%), 3 non-binary/third gender (0.43%), and 2 preferred not to answer (0.29%). Our participants were 20 to 71 years old (mean: 34.38 years; median: 31 years). Each condition had 98 to 103 participants (mean: 100). Compared to the results from the Pew 2017 Cybersecurity Knowledge survey [41], our participants answered questions correctly more often (69.13% vs. 42.54%). This corroborates prior research that MTurk participants have better cybersecurity knowledge [22]. Our participants' aggregated responses regarding prior data theft experiences were within 6 percentage points of responses to Pew's Americans and Cybersecurity survey [42], suggesting our sample population had similar data theft experiences as the general population.

As shown in Figure 5, the different warning designs had an effect on whether a participant clicked a phishing link. Notably, participants in the control condition clicked on a greater percentage of phishing links than benign links. We attribute this difference to the call-to-action nature of the phishing links, which without a phishing warning, were effective at baiting participants to click. We would expect such call-to-action links to be clicked at a greater rate than peripheral links. Indeed, call-to-action links in other benign emails also had a high CTR (57–63%) across all conditions, suggesting participants generally clicked on links that called to be clicked. It is further notable that the average benign links CTR across all conditions was 48.5%, even though the task asked participants to check whether links worked. This is likely because not all benign links in an email (e.g. embedded images, fine print) are as obvious as the call-to-action links. It could also be that participants evaluated whether a link "worked" through other means than clicking; our intention in crafting the task is to avoid overly incentivizing participants to click the links.

Our logistic regression analysis showed that a participant's click-through rate for benign links was a significant predictor of their click actions on phish links (OR=4.34, p<.001). This suggests that participants who click more on benign links are more likely to also click on phishing links. Furthermore, participants who gained a higher score in the cybersecurity quiz were significantly less likely to click on phishing links (OR=0.89, p<.05), and were significantly more likely to hover over (OR=1.95, p<.001) and click on benign links (OR=1.35, p<.001). This suggests participants with higher cybersecurity knowledge were more cautious when interacting with these links, and could better distinguish phish from benign links. Moreover, we found that participants who used the vendor more often previously were more likely to click on phishing links (OR=1.05, p<.05), suggesting the existence of brand effects.

Warnings Reduced Phishing CTR

Setting the control condition (no warning) as the baseline for *placement* in the mixed-effect logistic regression model, we see that the presence of a warning led participants to be more cautious about clicking phishing links. Participants in the four link-focused warning groups (OR=.002, p<.001), the banner warning group (OR=.04, p<.01), and the browser warning group ($OR = .73 * 10^{-4}$, p<.001) were significantly less likely to click phishing links compared to participants who did not see a warning.

Next, we compared the click-through rates (CTR) for the four link-focused warning designs to current warning designs, namely browser and banner warnings, and the control. The most effective warning design with the lowest CTR was on-hover/forced attention (8.67%), followed by the browser (11.78%), static/forced attention (15.64%), on-hover (30.36%), static (33.65%), banner (44.00%), and finally the control group with no warning (76.67%). A Kruskal-Wallis test confirmed differences in phishing CTR were significant among different conditions ($\chi^2(6)$=180, p<.001, ϵ^2=.26). Post-hoc analysis (Dunn comparison test) showed that the CTR for the best-performing link-based warning (on-hover/forced attention) was significantly lower than the banner warning (p<.001), but does not differ significantly from the browser warning. This means link-focused warnings are more effective than banner warnings, but inconclusive when compared to the browser warning. Conversely, a one-way ANOVA showed that the CTR for benign links was consistent across conditions ($F(6, 694)$=.59, p=.74). This indicates that warning designs played a factor in determining whether a participant clicked a phishing link but not a benign link.

To determine if the difference in click and hover behaviors came from warning design variants and participant characteristics exclusively (and not random effects from links or emails, such as certain links or emails being more or less believable than others), we included individual links and emails as random-effect variables in our regression models. We found that the variances for the random effects for both links and emails were between .00 and .04, suggesting that participants' propensity to click or hover over links did not vary between emails or links. These results suggest that our emails and links were equally believable.

Placement: Link-focused Better than Banner

Using link-focused as the baseline for placement in the logistic regression model, we found that, compared to link-focused warnings, banner warnings led to higher phishing CTR (OR=3.26, p<.001), and higher hover rate for both benign (OR=1.60, p<.001) and phishing links (OR=6.65, p<.001). Participants who encountered the banner warning also spent more time hovering over phishing links (b=.18, p<.01). This suggests that participants noticed the banner warning, hovered over all the links to search for the suspect link, hovered longer over phishing links, yet, clicked them anyways. Due to likely under-reported hover times (see Section 4), we believe our results to be less pronounced than reality.

Compared to link-focused groups, participants in the browser warning group were significantly less likely to click phishing links (OR=.45, p<.001). This suggests that overall, browser warnings were the most effective at preventing participants from reaching phishing websites. However, the effectiveness of browser warnings may be an artifact of our study design. Since our task for participants was to check whether links worked, reaching a full-screen browser warning after clicking an email link may have been sufficient for them to conclude the link was suspicious with little incentive to proceed further. Furthermore, Figure 5 shows that the on-hover/forced attention condition had the lowest CTR on phishing links. This suggests that placement of the warning cannot be the single predictor of whether or not a participant will click on a phishing link, and the effect of forced attention and hover activation might add nuances to this interaction.

Forced Attention Is Effective

Our logistic regression analysis showed that within link-focused warnings, participants in forced attention conditions were significantly less likely to click phishing links (OR=.33, p<.001). A Mann-Whitney U test on forced attention confirms the significant difference between forced attention and non-forced attention groups regarding phishing CTR ($W=2.5 * 10^4$, p<.001, r=.27). This suggests that within link-focused warnings, forced attention is an effective approach for reducing phishing CTR.

Warning Activation: No Significant Effect

We further examined the effect of hover warnings (warnings that interrupt the user) and static warnings (warnings that display simultaneously with the email) on phishing link CTR. We found that activating a link-focused warning on hover had little to no effect on clicking phishing links (n.s.). This is also confirmed by a Mann-Whitney U test (n.s.). This indicates that prior work suggesting active warnings are more effective than passive warnings [18] might not apply to the kind of in-email, link-focused warnings we tested.

The only significant effect of warning activation we observed is that participants who encountered an active link-focused warning were significantly more likely to hover over benign links (OR=1.50, p<.001), and also spent more time hovering over benign links (b=.12, p<.001). Possibly when participants view a warning that appears on hover, they then might hover for longer over other links to see if a warning appears for them. Due to likely under-reported hover times (see Section 4), a more pronounced difference may exist.

Multiple Warnings Affect Benign Link Interactions

In addition to click-through rates, we examined how being exposed to multiple warnings would affect participants' interaction with phishing and benign links. We grouped participants' interactions into four groups: before a user sees a warning (0 warnings seen), after the user sees the first warning (1 warning seen), etc. For most conditions, phishing link click-through and hover rates decreased as participants saw additional warnings, as shown in Figure 6 and 7. This suggests that participants may have become habituated to the warnings. However, we also noticed that the hover rate over benign links increased as participants saw more warnings (see Figure 7). This suggests participants interacted with the phishing links less but hovered over benign links more as they saw more warnings.

Our regression models confirmed that, as participants saw more warnings, they were more likely to hover over benign links (OR=1.14, p<.001), but also spent less time hovering over both phishing (b=−.25, p<.001) and benign links (b=−.05, p<.001).

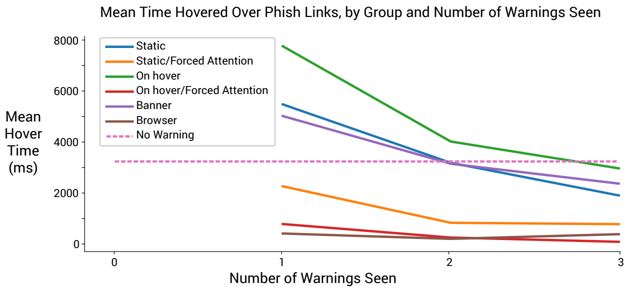


Figure 6: Mean time hovered over phishing links.

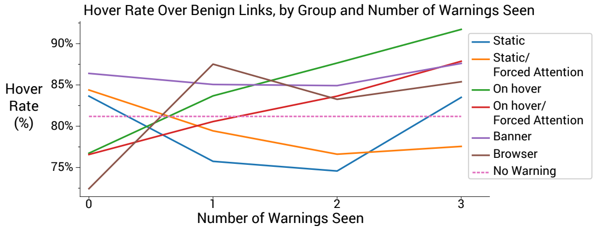


Figure 7: Hover rate over non-phishing (benign) links.

Particularly, as the number of warnings seen increased participants in the link-focused groups were less likely to click (OR=.87, p<.05) and hover over (OR=.68, p<.01) phishing links. They also spent less time hovering over phishing links (b=−.31, p<.001). This suggests that participants began to rely on our link-focused warnings to detect phishing warnings, which successfully helped reduce the cognitive load of seeking and examining phishing links, as reflected by the decreased time.

Warnings Aided Hazard Identification

Qualitative analysis of open-ended responses revealed that warnings aided participants in identifying phishing links more easily, and sometimes triggered protective behaviors. The majority of participants who answered the Perception question (325, 62%) reported that the warning helped them notice suspect links (e.g., "It made me think the link was bad"). Among them, 128 participants further reported perceptual or behavioral reactions to the warning. 105 participants explicitly said, that as a result of seeing the warning, they did not click suspect links. Several mentioned engaging in secure practices such as double-checking links (16), checking the email sender (3), or comparing the link to the URL (29). 4 participants mentioned opening a suspicious link in a separate browser, window, or tab during study; another 3 participants mentioned they would do so hypothetically. Additionally, 3 participants mentioned that our links triggered their previously installed 3rd party anti-virus software.

Based on participants' responses (multiple choice) regarding which parts of the email seemed suspicious, participants' ability to identify the hazard varied among conditions. Responses were categorized as correct (identified the link only), partially correct (identified the link and something else), and incorrect (did not identify the link). The two on hover, link-focused warning conditions had the most participants' correctly identifying only the link as the hazard (on hover: 61%, on hover/forced attention: 58%). A Kruskal-Wallis test revealed a significant difference among groups ($\chi^2(6)$=14.43, p<.05, ϵ^2=.02). However, post-hoc analysis (Mann-Whitney U) showed no significant pairwise differences.

Conversely, participants' perceptions of the different warning designs did not vary substantially among groups. A one-way ANOVA revealed no significant differences among groups for a warning's intrusiveness (n.s.), helpfulness (n.s.), or whether it changed their perception of the link (n.s.).

On a scale from 1 to 5, participants rated all warnings as similar with regard to intrusiveness (ranging from 1.36 (banner) to 1.5 (on hover/no forced attention)), helpfulness (from 4.0 (banner) to 4.42 (tatic/forced attention)), and perceptual change (from 3.54 (browser) to 4.0 (static/forced attention)).

While most participants had an accurate understanding of the warnings, 2 participants (<0.5%) confused the browser phishing warning with a Transport Layer Security (TLS) warning, reporting that the warning pertained to an expired security certificate. This aligns with prior studies which found that people tend to confuse browser-based phishing and malware warnings with TLS/SSL warnings [18, 19, 50]. Such confusions did not occur for link-focused and banner warnings, which might be an indication that placing phishing warnings in the email client helps users distinguish them from (potentially benign) TLS/SSL warnings.

Need for More Information

The qualitative analysis further revealed multiple suggestions for improving current phishing warning designs. Of 23 participants who reported clicking a phishing link, 5 noted they clicked the link because they were curious about where the link went; another 6 reported clicking despite being alert to the link's danger, e.g., *"It made me skeptical of the content of the link, or what the destination truly was. Yet, I still clicked it anyways."* Similarly, of the 89 valid responses for the Improvement question, 17 participants expressed a desire for more information, including an explanation of why the link/email was suspicious (10), what the consequences were of clicking (4), and what further actions should be taken (3). Prior work suggests that people will attend to an email message in their inbox if they are curious about its content [56]. We believe it is reasonable to extend this to phishing emails, phishing links, and phishing websites. While current warnings might include a "Learn More" button, those often point to general security advice, whereas providing more information about the specific suspected phishing link might also be important to satisfy users' curiosity.

6 DISCUSSION

Our results suggest that placing the warning near phishing links helps reduce click-through rate when compared to no warning and similar banner warnings. We also found that forced attention resulted in significantly lower click-through rates for link-focused warnings, though activation (on hover) appears to have had little effect on click-through rates.

Next, we first discuss potential limitations of our study before discussing our findings' implications in more detail.

Limitations

Our study has potential limitations. In our controlled experiment, we gave participants the artificial task to "Check emails for valid links," which differs from how people use email in their daily lives. We carefully crafted this task to get participants to engage with the emails in a meaningful way, without revealing the study's focus on phishing. More specifically, we gave participants credentials to a realistic online email client and asked them to evaluate whether links are "working," which was deliberately vague, in order to not over-incentivize clicking on links. Additionally, participants checked these email links on their own computers, further enhancing the ecological validity of perceived security risks. In fact, only 12 participants (1.7%) reported clicking a phishing link because of the task instructions; whereas 105 (15%) explicitly stated they hesitated to click because of the shown warning; 410 (58%) did not click any phishing links. This suggests participants were not unreasonably incentivized to click all links.

A further limitation is that our experiment did not assess the effects of false positives, i.e., erroneously displaying a warning for a benign link. In particular, participants may overly rely on the warning and thus refrain from clicking benign links when an erroneous warning is displayed; while a concern, common security advice encourages caution and suggests that people should validate an email request through other channels when unsure [48]. Another concern is that variance in phishing detection accuracy over time, i.e., repeatedly seeing false positive warnings, may affect the warning's credibility or lead to warning habituation. Our experiment provided new insights on the effectiveness of phishing warning features (placement, forced attention, activation). Further studies such as longitudinal field studies are necessary to assess the impact of false positives on those warning features and the interplay between warning design and phishing detection accuracy, in particular over time.

MTurk participants have been shown to be more security-aware than the general population [22], which was corroborated by our participants' cybersecurity knowledge. This means our participants may have clicked less frequently on links than the general population might. However, we still observed significant differences in CTR, especially among the control and treatment groups. This suggests that warnings which are effective for this population may also hold value for the general population.

Finally, phishing takes many different forms, from spear-phishing and whaling to business email compromise (BEC) to malicious attachments [51]. We focused on link-based phishing attacks. We acknowledge that phishing links are not always straightforward, as attackers can use URL shorteners to further obfuscate a link's URL. However, even in those cases, placing warnings at a suspicious link, unmasking a link's URL by showing it in the warning, and using forced attention to make it slightly harder for people to ignore the warning may still be useful in protecting people from falling for phish.

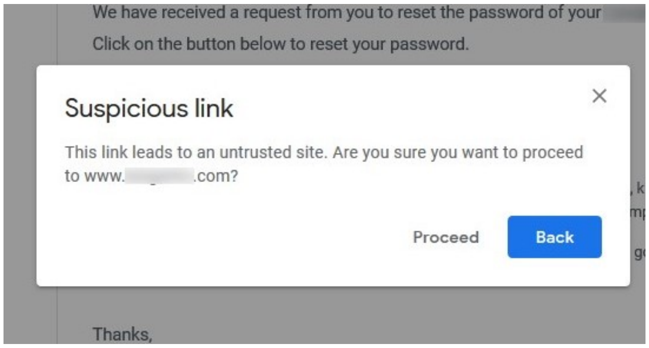
Put Your Warning Where The Link Is

When the link is the hazard, we found that placing the phishing warning near the link reduces phishing click-through rates. While the browser warning performed significantly better than our group of link-focused warnings as a whole, this may be a study design artifact, due to low incentives for participants to continue through a warning after having already clicked the link in the email.

However, our link-focused warnings performed significantly better than the banner warning group, with lower phish CTR, lower phish and benign hover rates, as well as shorter phish hover time. Our interpretation is that with a banner warning, users search for the suspicious link by hovering over multiple links, hover for a longer time over a phish link, yet might still click the phishing link.

Based on our results, we propose that if a link has been flagged as suspicious by phishing detection, email clients should place the warnings directly at the respective link and unmask the link's URL in order to help users apply common anti-phishing advice, namely checking the URL before clicking a link. Encouragingly, major email vendors are starting to adopt such approaches. While our study was in progress, Gmail introduced a new phishing warning, shown in Figure 8, which appears after clicking a suspected phishing link and includes the suspicious link's unmasked and pruned URL [54]. Our study provides empirical evidence for the positive impact of such link-focused warnings inside email clients and the use of forced attention on phishing warning adherence.

While this may be effective for phishing emails where the link is suspicious, the most effective method for calling attention to suspect senders or attachments may be different. Contextual phishing warnings that adapt in placement and content to highlight the suspected phishing hazard could be a worthwhile approach for help users more effectively identify phishing in all its forms.


We have received a request from you to reset the password of your [redacted]

Click on the button below to reset your password.

Suspicious link

This link leads to an untrusted site. Are you sure you want to proceed to www.[redacted].com?

Thanks,

Figure 8: Gmail's current phishing warning [23] appears after a user clicks but in the email client.

Implement Forced Attention

Our forced attention design deactivated the original phishing link in the email and provided a path to the phishing website through our warning, a type of inhibitive attractor [12]. Among the link-focused warnings, forced attention was significantly more effective at reducing phishing CTR. We recommend that link-focused warnings should apply forced attention to better protect people from phishing websites. We also propose that future work investigate the use of inhibitive attractors for other interactive phishing hazards, such as attachments.

Curiosity Phished Users

Prior work suggests that curiosity influences a user's attention towards email [56]. This is corroborated by our qualitative analysis, with several participants reporting they clicked phishing links out of curiosity despite risk awareness, or stating they wanted to learn more about the suspicious link. However, this curiosity is seemingly at odds with prior warning research, which suggests users do not often click "learn more" in browser warnings [2]. A more nuanced interpretation might be that people do not want to "learn more" about web security/phishing in general, but are curious about why an email in their inbox or a given link was flagged as suspicious. Addressing curiosity about the specific security incident could further help reduce phishing victims. For instance, a link-focused warning could provide an unclickable preview image of the linked website. Such a preview could highlight cues, drawn from phishing training research, on why the particular website is suspicious. This information could be coupled with the warning rather than hiding it behind a "learn more" button.

7 CONCLUSION

Despite extensive research and industry efforts, people continue to fall for phish. Users have difficulty identifying phishing cues without training, and automated phish detection will always require mechanisms for users to override system decisions. Warnings can play a crucial role in providing this safety valve while providing the appropriate cues for humans to identify and avoid phishing links.

We evaluated the effects of different phishing warning features aimed at making both suspect features of a phishing link and the warning itself more salient (placement, activation) and nudging people away from clicking a suspicious link (forced attention). Our controlled online experiment showed that providing a warning near a link is more effective at reducing phishing CTR than a banner warning at the top of an email. In addition, banner warnings force users to search for links by hovering while not providing enough contextual clues to identify a phishing link once a user hovers. While phishers will continue to adapt their methods to circumvent security processes, our results indicate how improving the placement, interactivity, and contextual cues of phishing warnings can help reduce cognitive burden in identifying and avoiding phishing attacks.

ACKNOWLEDGMENTS

This research was partially funded by the University of Michigan School of Information and the Institute of Museum and Library Services (RE-01-15-0086-15). The authors thank all study participants for their efforts, as well as the CHI reviewers for their extensive and constructive feedback. The authors would also like to thank Akshay Potnis and the University of Michigan's ITS department for their support and contributions in early stages of this project.

REFERENCES

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. 2017. Nudges for privacy and security: understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)* 50, 3 (2017), 44.

[2] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness.. In *USENIX security symposium*, Vol. 13.

[3] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. 2015. Your location has been shared 5,398 times!: A field study on mobile app privacy nudging. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems.* ACM, 787–796.

[4] Bonnie Brinton Anderson, C Brock Kirwan, Jeffrey L Jenkins, David Eargle, Seth Howard, and Anthony Vance. 2015. How polymorphic warnings reduce habituation in the brain: Insights from an fMRI study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* ACM, 2883–2892.

[5] Anti Phishing Working Group (APWG). 2016. Global Phishing Report 2015-2016. http://docs.apwg.org/reports/APWG_Global_Phishing_Report_2015-2016.pdf

[6] Anti Phishing Working Group (APWG). 2018. *Phishing Activity Trends Report Q1 2018.* Technical Report.

[7] Lujo Bauer, Cristian Bravo-Lillo, Lorrie Cranor, and Elli Fragkaki. 2013. *CMU Warning Design Guidelines.* Technical Report CMU-CyLab-13-002. Carnegie Mellon University.

[8] André Bergholz, Jan De Beer, Sebastian Glahn, Marie-Francine Moens, Gerhard Paaß, and Siehyun Strobel. 2010. New filtering approaches for phishing email. *Journal of computer security* 18, 1 (2010), 7–35.

[9] Cristian Bravo-Lillo. 2014. Improving Computer Security Dialogs: An Exploration of Attention and Habituation.

[10] Cristian Bravo-Lillo, Lorrie Cranor, Saranga Komanduri, Stuart Schechter, and Manya Sleeper. 2014. Harder to Ignore? Revisiting Pop-Up Fatigue and Approaches to Prevent It. In *10th Symposium On Usable Privacy and Security (SOUPS 2014).* USENIX Association, Menlo Park, CA, 105–111. https://www.usenix.org/conference/soups2014/proceedings/presentation/bravo-lillo

[11] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. 2011. Bridging the Gap in Computer Security Warnings: A Mental Model Approach. *IEEE Security & Privacy* 9, 2 (2011), 18–26.

[12] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. 2013. Your attention please: designing security-decision UIs to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security.* ACM, 6.

[13] José Carlos Brustoloni and Ricardo Villamarín-Salomón. 2007. Improving security decisions with polymorphic and audited dialogs. In *Proceedings of the 3rd symposium on Usable privacy and security.* ACM, 76–85.

[14] Madhusudhanan Chandrasekaran, Krishnan Narayanan, and Shambhu Upadhyaya. 2006. Phishing E-Mail Detection Based on Structural Properties. (2006), 7.

[15] Lorrie Faith Cranor. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.* 10 (2012), 273.

[16] Rachna Dhamija, J Doug Tygar, and Marti Hearst. 2006. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems.* ACM, 581–590.

[17] Julie S Downs, Mandy Holbrook, and Lorrie Faith Cranor. 2007. Behavioral response to phishing risk. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit.* ACM, 37–44.

[18] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 1065–1074.

[19] Adrienne Porter Felt, Alex Ainslie, Robert W Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. 2015. Improving SSL warnings: Comprehension and adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* ACM, 2893–2902.

[20] Adrienne Porter Felt, Robert W Reeder, Hazim Almuhimedi, and Sunny Consolvo. 2014. Experimenting at scale with google chrome's SSL warning. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems.* ACM, 2667–2670.

[21] Ian Fette, Norman Sadeh, and Anthony Tomasic. 2007. Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web.* ACM, 649–656.

[22] Kristin Firth, David A. Hoffman, and Tess Wilkinson-Ryan. 2017. Law and Psychology Grows Up, Goes Online, and Replicates. (2017).

[23] Gmail Help Forum. 2018. Suspicious link issue! - Google Product Forums. https://productforums.google.com/forum/#!msg/gmail/h_yeYefHFWk/Fj8o4q3HAQAJ

[24] Simson Garfinkel and Heather Richter Lipford. 2014. Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust* 5, 2 (2014), 1–124.

[25] Anti Phishing Working Group. 2017. *Global Phishing Survey: Trends and Domain Name Use in 2016.* https://docs.apwg.org/reports/APWG_Global_Phishing_Report_2015-2016.pdf

[26] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. 2014. Using Personal Examples to Improve Risk Communication for Security & Privacy Decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14).* ACM, New York, NY, USA, 2647–2656. https://doi.org/10.1145/2556288.2556978

[27] Jason Hong. 2012. The state of phishing attacks. *Commun. ACM* 55, 1 (2012), 74–81.

[28] The Radicati Group Inc. 2017. Email Statistics Report, 2017-2021. http://www.radicati.com/wp/wp-content/uploads/2017/01/Email-Statistics-Report-2017-2021-Executive-Summary.pdf

[29] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. 2009. A nutrition label for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security.* ACM, 4.

[30] Soyun Kim and Michael S Wogalter. 2009. Habituation, dishabituation, and recovery effects in visual warnings. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 53. Sage Publications Sage CA: Los Angeles, CA, 1612–1616.

[31] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. 2009. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security.* ACM, 3.

[32] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2010. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)* 10, 2 (2010), 7.

[33] Phish Labs. 2017. 2017 Phishing and Threat Intelligence Report. https://pages.phishlabs.com/rs/130-BFB-942/images/2017%20PhishLabs%20Phishing%20and%20Threat%20Intelligence%20Report.pdf

[34] Elmer Lastdrager, Inés Carvajal Gallardo, Pieter Hartel, and Marianne Junger. 2017. How Effective is Anti-Phishing Training for Children?. In *Symposium on Usable Privacy and Security (SOUPS).*

[35] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction.* Morgan Kaufmann.

[36] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. 2011. Does domain highlighting help people identify phishing sites?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2075–2084.

[37] Gang Liu, Guang Xiang, Bryan A Pendleton, Jason I Hong, and Wenyin Liu. 2011. Smartening the crowds: computational techniques for improving human verification to fight phishing scams. In *Proceedings of the Seventh Symposium on Usable Privacy and Security.* ACM, 8.

[38] Samuel Marchal, Giovanni Armano, Tommi Grondahl, Kalle Saari, Nidhi Singh, and N Asokan. 2017. Off-the-Hook: An Efficient and Usable Client-Side Phishing Prevention Application. *IEEE Trans. Comput.* (2017).

[39] Samuel Marchal, Kalle Saari, Nidhi Singh, and N. Asokan. 2015. Know Your Phish: Novel Techniques for Detecting Phishing Sites and their Targets. *CoRR* abs/1510.06501 (2015). arXiv:1510.06501 http://arxiv.org.offcampus.lib.washington.edu/abs/1510.06501

[40] Federal Bureau of Investigation Internet Crime Complaint Center. 2017. 2017 Internet Crime Report. (2017), 29.

[41] Kenneth Olmstead and Aaron Smith. 2017. Pew Research Center. http://www.pewinternet.org/2017/03/22/what-the-public-knows-about-cybersecurity/

[42] Kenneth Olmstead and Aaron Smith. 2017. Pew Research Center. http://www.pewinternet.org/2017/01/26/americans-and-cybersecurity/

[43] Sameer Patil, Roberto Hoyle, Roman Schlegel, Apu Kapadia, and Adam J. Lee. 2015. Interrupt Now or Inform Later?: Comparing Immediate and Delayed Privacy Feedback. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* ACM, New York, NY, USA, 1415–1418. https://doi.org/10.1145/2702123.2702165

[44] Sameer Patil, Roman Schlegel, Apu Kapadia, and Adam J. Lee. 2014. Reflection or Action?: How Feedback and Control Affect Location Sharing Decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14).* ACM, New York, NY, USA, 101–110. https://doi.org/10.1145/2556288.2557121

[45] PhishTank. 2018. *Join the fight against phishing.* Technical Report. https://www.phishtank.com/

[46] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015).* 1–17.

[47] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. [n. d.]. The Emperor's New Security Indicators. In *Security and Privacy, 2007. SP'07. IEEE Symposium On* (2007). IEEE, 51–65. http://ieeexplore.ieee.org/abstract/document/4223213/

[48] UC Berkeley Information Security and Policy. 2018. Phishing | Information Security and Policy. https://security.berkeley.edu/resources/phishing

[49] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security - SOUPS '07.* ACM Press, Pittsburgh, Pennsylvania, 88. https://doi.org/10.1145/1280680.1280692

[50] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. 2009. Crying Wolf: An Empirical Study of SSL Warning Effectiveness.. In *USENIX Security Symposium.* 399–416.

[51] Symantec. 2018. Internet Security Threat Report.

[52] Janice Y Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti. 2011. The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research* 22, 2 (2011), 254–268.

[53] Verizon. 2018. *2018 Data Breach Investigations Report.* Technical Report 11th Edition.

[54] Melanie Volkamer, Karen Renaud, and Paul Gerber. 2016. Spot the Phish by Checking the Pruned URL. *Information and Computer Security* 24, 4 (Oct. 2016), 372–385. https://doi.org/10.1108/ICS-07-2015-0032

[55] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. 2017. User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. *Computers & Security* 71 (2017), 100 – 113. https://doi.org/10.1016/j.cose.2017.02.004

[56] Jaclyn Wainer, Laura Dabbish, and Robert Kraut. 2011. Should I open this email?: inbox-level cues, curiosity and attention to email. In *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM, 3439–3448.

[57] Rick Wash and Molly M. Cooper. 2018. Who Provides Phishing Training?: Facts, Stories, and People Like Me. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* ACM Press, Montreal QC, Canada, 1–12. https://doi.org/10.1145/3173574.3174066

[58] Liu Wenyin, Ning Fang, Xiaojun Quan, Bite Qiu, and Gang Liu. 2010. Discovering Phishing Target Based on Semantic Link Network. *Future Generation Computer Systems* 26, 3 (March 2010), 381–388. https://doi.org/10.1016/j.future.2009.07.012

[59] Colin Whittaker, Brian Ryner, and Marria Nazif. 2007. Large-Scale Automatic Classification of Phishing Pages. (2007), 14.

[60] Michael S Wogalter, Vincent C Conzola, and Tonya L Smith-Jackson. 2002. based guidelines for warning design and evaluation. *Applied ergonomics* 33, 3 (2002), 219–230.

[61] IBM X-Force. 2018. IBM X-Force Threat Intelligence Index 2018. https://www.ibm.com/security/data-breach/threat-intelligence

[62] Guang Xiang, Jason Hong, Carolyn P Rose, and Lorrie Cranor. 2011. Cantina+: A feature-rich machine learning framework for detecting phishing web sites. *ACM Transactions on Information and System Security (TISSEC)* 14, 2 (2011), 21.

[63] Weining Yang, Jing Chen, Aiping Xiong, Robert W. Proctor, and Ninghui Li. [n. d.]. Effectiveness of a Phishing Warning in Field Settings. ACM Press, 1–2. https://doi.org/10.1145/2746194.2746208

[64] Yue Zhang, Jason I. Hong, and Lorrie F. Cranor. 2007. Cantina: A Content-Based Approach to Detecting Phishing Web Sites. In *Proceedings of the 16th International Conference on World Wide Web - WWW '07*. ACM Press, Banff, Alberta, Canada, 639. https://doi.org/10.1145/1242572.1242659