What Makes People Click the Link? Studying a Watering Hole Attack in the Wild

anonymous

ABSTRACT

Messages connected to cyber attacks typically utilize various tactics to induce the intended target to click on the malicious link within the message. We carried out a Watering Hole attack in the wild to investigate whether susceptibility to click on the link in an attack message is dependent on a person's individual characteristics as well as variations in the nature of the message. In our relatively tech savvy sample, we found that gender or individual personality traits did not influence susceptibility to the attack messages. However, we found that different messages resulted in deceiving people to different extents. Our results suggest that analyzing message contents to detect common types of cues and appeals could provide improved detection of attack messages with malicious intent. Moreover, personalized solutions tailored to an individual would benefit from focusing on technical knowledge and efficacy rather than other personal characteristics of the individual.

Author Keywords

Attack messages; online scams; security; cybercrime; personality.

CCS Concepts

•Security and privacy → Social aspects of security and privacy; •Human-centered computing → Empirical studies in HCI;

INTRODUCTION

Across the world, the Internet has become an integral part of people's everyday lives. Recently, the Federal Communications Commission (FCC) declared broadband Internet a public utility [10]. The global growth in Internet usage has led to a corresponding increase in crimes, scams, and other malicious activities that target Internet users. Cyber attacks continue to grow exponentially with individuals targeted via various kinds of messages. While phishing is one of the most well-known examples of such attacks, there is a large variety of online tactics that attempt to victimize people, ranging from some relatively benign as clickbait (that attempts to draw eyeballs to ad-laden pages) to something egregious such as tricking a person to install malicious software. Despite this variety, the manner in which the attacks are operationalized typically involves a common element, viz., making the targeted individuals take the action desired by the attacker. Usually, the desired action is activating the attack by clicking a link within a message. Such messages may be delivered in a variety of ways, such as email, instant messages, online advertisements, compromised Web pages, etc.

Although the link as presented in the message appears to be legitimate, it leads to a site controlled by the attacker. For example, the link sometimes takes individuals to a Web site that looks and feels authentic and requests sensitive information such as name, password, date of birth, social security number, etc. If the victim provides the information, the attacker makes use of it for fraudulent purposes, such as identity theft, unauthorized money transfers, routing illicit activities/content via the victim's account, etc.. Alternatively, the site may ask the victim to install malware disguised as software serving a legitimate purpose. Persuading an individual to act on the contents of the bogus message involves the use of social engineering and exploitation of common human behavioral tendencies. The motivations provided by attackers to get victims to click the malicious link vary and include incomplete account information, password reset, work related issues, business opportunities, potential financial gain, software upgrade, etc.

Many of these nefarious messages exploit socio-psychological factors as well as the lack of security-relevant technical knowledge of today's average Internet user. In order to make the attack more effective, the attacker may pose as a known or trusted entity or provide a service or benefit that entices the intended targets while masking the true intentions of the attacker. Consider, for example, the recent controversy over the discovery that Cambridge Analytica collected of the data of nearly 50 million Facebook; the massive data collection was enabled by an App whose ostensible purpose was to provide a personality assessment. Interestingly, in this particular case, the spread of the deceptive App was amplified via the crowdwork platform Amazon Mechanical Turk (AMT) by creating a task that provided a micropayment that incentivized AMT workers to install the App in question. In the same vein, we designed and operationalized a real-world watering hole attack that: (i). developed a trusted relationship with the intended victims via engagement unrelated to the attack, and (ii). delivered messages that required the victims to click a link that would activate the attack, and (iii). utilized AMT as the platform for the delivery.

An important step in developing tools and techniques that defend against such attacks is understanding the factors that

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.XXXXXXX

make attack messages successful in persuading the intended targets to click on malicious links. Such an understanding can provide insight to help mitigate the risk posed by these attacks. To this end, we make the following contributions based on deployment of the attack we designed:

- Carry out a (benign) watering hole attack "in the wild" with a large and diverse sample.
- Investigate the effectiveness of three types of emotional appeals in eliciting a response from the target individual.
- Examine the impact of individual personality traits on susceptibility to attack messages.

Next, we discuss related work that provides the background that led to our research questions. We then describe the method we used to carry out an attack that addresses the posed research questions. We analyze the responses to our attack based on contents of the attack messages and characteristics of the targeted individuals. We proceed to discuss the implications of the work for improving resilience against real-world attacks followed by promising directions for future work.

RELATED WORK

A particularly widespread attack, which aims to obtain sensitive information by impersonating a trusted party, is known as 'phishing.' The attack is typically carried out by sending an email that contains a link to a spoofed Web site. A phishing email message appears to be from a legitimate source, such as a bank or a system administrator. The contents of the message present a situation that requires the recipient to take action by clicking the included link. Hong [24] provides a detailed overview of the various types of phishing attacks and their operationalizations.

A recent study [25], albeit with a sample of undergraduate students, found that most participants misclassified phishing emails despite reported confidence in the ability to detect phishing attacks. Several researchers have sought to understand why various attacks are successful in deceiving people to such an extent. The underlying factors uncovered in these studies fall into the following categories:

Education / Training: It has been shown that education and training helps people recognize fraudulent emails and Web sites [28, 29, 41, 42]. Yet, Caputo [8] found that over 30% of the participants clicked on at least one phishing link and 10% clicked on all three phishing emails, even though training was administered between the emails. Further, Sheng et al. [41] found that women were more likely to fall for phishing despite greater exposure to anti-phishing education. Similarly, Kajzer et al. [27] report that the efficacy of information security awareness messages depends on the personality traits of the recipient. These findings indicate that the individual characteristics discussed above may influence the effects of security education. Importantly, the results underscore the need to complement security education and training with additional tools and techniques personalized to the individual.

Usability / User Interface: As Hong [24] points out, the usability of interfaces of email clients and Web browsers provides few effective cues that can aid the average person in judging the legitimacy of a message and associated links. Similar issues affect the user experience for security advice [22] and warning messages, which are often ignored or dismissed, especially if they are passive [15].

Content: Dhamija, Tygar, and Hearst [13] found that visual deception tricks used by attackers could fool even the savviest of users. Attack messages often also deceive people by utilizing social techniques, such as conveying urgency or appealing to their sense of greed [24]. Further, Tsow and Jakobsson [45] found that such techniques are more effective when they combine authenticity-inspiring design features with 'narrative strength' of the message in making the recipient connect with the contents. Blythe and Petrie and Clark [7] carried out a literary analysis outlining how attackers attempt to boost the narrative strength of the message content via strategies such as pastiche, interpellation, premise formulation, and call to action. In a large-scale online study that presented various scam and non-scam scenarios, Heartfield and Loukas [21] found that people's ability to detect attacks differed greatly based on the details of the scenario.

Individual characteristics: An individual's likelihood of falling for an attack has been reported to be affected by a number of personal characteristics. In particular, studies show that vulnerability to attacks is influenced by a lack of technical knowledge and skills, such as differentiating between a browser's 'chrome' and content areas [13], reasoning related to decisions about sharing sensitive information [14], etc. On a related note, several studies have found women to be more vulnerable to phishing [7, 19, 20, 25, 41], possibly owing to lower technical knowledge [41]. Sheng et al. [41] further found that younger participants were more susceptible to being phished. Additionally, Hong et al. [25] found that the ability to identify non-legitimate emails was affected by dispositional trust [32] and personality traits. Halevi et al. [19] report that clicking on an attack link was correlated with the personality trait of neuroticism, and Halevi et al. [20] found it correlated with conscientiousness, but only for women.

In our research, we focused on message content and individual characteristics, with the goal of generating insight that can inform the other two aspects mentioned above, viz., usability and education.

While many of the studies mentioned above have provided useful findings and insight, they suffer from a number of limitations. In particular, several of the studies are based on relatively small samples with large proportions of students. Moreover, many of the studies were conducted in the lab and/or involved fictitious scenarios. In contrast, our study utilizes a large sample drawn from a more general population to carry out an investigation based on an 'in the wild' realistic attack.

A traditional phishing attack consists of an attacker emailing the *same* message in bulk to a large number of individuals. Defensive techniques against such tactics have gained in effectiveness and are often able to flag and filter a large proportion of bulk phishing emails, either stopping them from ever being

seen by the recipient or alerting the user via easily noticed warnings that flag the message as malicious. To counter these measures, attackers have evolved their phishing tactics to follow a technique termed as 'spear phishing.' Spear phishing is a form of phishing in which attack email messages are customized to the recipient. For example, a spear phishing email may address the recipient by his or her name and include other information that references the recipient's trusted social or professional contexts [4, 46]. In addition to automated techniques, spear phishing could also involve semi-automated or manual approaches in which the attacker invests time and effort in establishing and building a trusted relationship with the intended target prior to deploying the spear phishing attack.

Recent data shows that spear phishing attacks have reached epidemic proportions, with no clear solution [5, 26]. In addition to targeting private individuals, spear phishing attacks are increasingly targeted at organizations through their employees, including top level executives [33, 34]. A successful spear phishing attack puts organizations at risk of huge damages via not only direct financial losses but also compromise of sensitive customer and/or employee data and theft of intellectual property [39]. For example, an employee in the Payroll Department of Snapchat compromised the identities of current and former employees in response to a spear phishing message from an attacker impersonating as the Chief Executive Officer [40], and an employee of an investment firm in Michigan transferred nearly half a million dollars to a scammer in response to a spear phishing email [31]. A recent survey by the security solutions provider Cloudmark found that spear phishing is recognized by organizations as one of the top security concerns [9].

The greater success rate of spear phishing over traditional phishing points at the customized targeting of the phishing emails as one aspect that makes it effective. However, little is known regarding the relative success rates of different kinds of attack messages (e.g., request for installing a software update vs. warning for resetting an account password). To investigate this aspect, we formulated the following research question:

RQ1: Is an individual's susceptibility to being deceived influenced by the content of the attack message?

Specifically, we compared three attack messages in terms of their effectiveness in deceiving the recipient. In this regard, we extend the work of Oliviera et al. [38] who reported that people's tendency to be deceived by attack messages is influenced by a combination of the persuasive strategy and life domain (financial, health, ideological, legal, security, and social) connected with the message content. As detailed in the Method section, unlike Oliviera et al.'s messages that, while addressed specifically to the individual recipient (like in our study), originated from fake accounts on mainstream services, such as Gmail, our attack messages appeared to originate from a party trusted by the recipient. Moreover, Oliviera et al.'s results are derived from repeated measures with a self-selected sample that could be biased due to the privacy-invasive study procedures required to support the deceptive cover story for the study. Our between-subjects study overcomes these limitations by reporting on a more generalized sample without repeated measures.

Prior research has found that individuals vary in their susceptibility to Internet attacks in general [36] and phishing in particular [6, 8, 19, 20, 25, 41]. To this end, Modic and Anderson [35] provide a susceptibility-to-persuasion scale as a psychometric tool to evaluate an individual's proclivity to be persuaded by an online message. We delve deeper into this aspect by examining the impact of an individual's personality traits on the response elicited by various types of attack messages. The most widely used measures of personality traits are based on five theoretical dimensions known as the Big Five personality traits or the five factor model (FFM) [11]. These five traits are: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. The role of these personality traits in decision making is a growing research area [3, 12, 23, 43]. In particular, personality traits have been shown to play an important role in users' online behavior [2, 16, 17, 30, 37]. Therefore, we posed our next research question as:

RQ2: Is an individual's susceptibility to attack messages affected by his or her personal characteristics?

As mentioned above, a few prior studies [19, 20, 25, 27, 36] provide indications that personality traits might affect how individuals respond to attacks. However, Hong et al. [25] and Modic and Lea [36] did not carry out a 'real' attack. Hong et al.'s findings are based on a fictitious scenario presented to undergraduate students in the laboratory while Modic and Lea's results are derived from self-reported reactions to scenarios of common Internet attacks from a sample with a large proportion of undergraduate and postgraduate students. We overcome these shortcomings by examining the relationship between personality traits and real-life behavior. The studies by Halevi et al. [19, 20] did carry out realistic attacks. However, each study used a single message on relatively small and homogeneous samples. Moreover, one of the studies [19] was carried out with a sample of psychology undergraduates. In contrast, our study includes a realistic investigation of multiple types of attack messages with a large and diverse sample of individuals.

By addressing the above two research questions, we aim to fill important gaps in the literature by comparing the effectiveness of specific attack strategies and attempting to explain individual variation in susceptibility to being deceived by examining its relationship with an individual's personal characteristics.

METHOD

In order to tackle the research questions, we conducted a study in which we carried out a watering hole attack on participants recruited via a task advertised via the Amazon Mechanical Turk (AMT) crowd work platform. Figure 1 shows the overall organization of the study. As outlined in Figure 1, the study involved the following steps:

1. Questionnaire on individual characteristics:
The first step in the study was advertised as an AMT task of responding to a questionnaire on computer usage and personal characteristics. In order to limit the effect of culture,

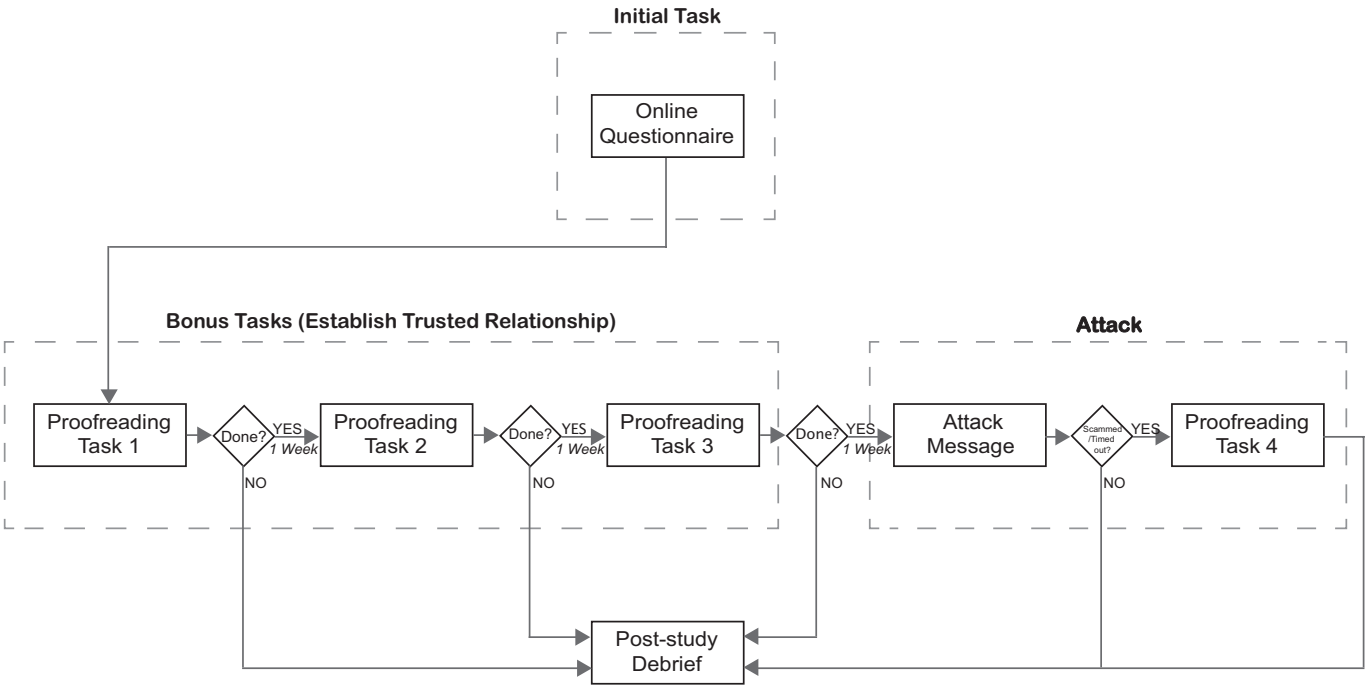

Figure 1. Overall organization and flow of the various components of study.

we restricted participation to AMT workers based in the US and further sought only those who had a high (99%) rate of successful task completion. Upon accepting the task, participants were asked to visit the questionnaire page hosted on a server at our university. Upon visiting the questionnaire URL, participants were provided information about the study and asked to proceed if they consented to participation. To avoid priming, the study information as well as the questionnaire URL made no mention of computer security or cyber attacks. Those who consented proceeded to answer the questionnaire which included questions on demographics and computer usage along with items intended to measure the Big Five personality traits [11, 17], viz., Extraversion, Neuroticism, Conscientiousness, Openness, and Agreeableness. The questionnaire also included items to measure knowledge and efficacy regarding computer-mediated communication (CMC) taken from Spitzberg's CMC Competence scale [44]. We further embedded two 'attention check' questions within the questionnaire in order to flag non-attentive respondents. The questionnaire did not ask for personally identifiable information; all responses were anonymous. Upon completion, participants were provided with a randomly generated code to enter on AMT to validate completion of the questionnaire task. Participants were paid $0.75 for completing the task.

2. Proofreading Tasks:

In order to set the stage for the watering hole attack by developing an ongoing connection with individuals that required them to use the same site repeatedly, we utilized the AMT functionality that allows sending follow-up tasks to those who have completed a previous task from a requester. Specifically, we devised a 'proofreading' task that required an individual to read a text paragraph and indicate the number of writing mistakes found in the text. Such types of tasks are commonly encountered on AMT. One week after a respondent answered the initial questionnaire, we emailed him or her via AMT to seek participation in the proofreading task (see Figure 2). The email included a link to the proofreading task hosted on the same server as the questionnaire. Those who completed the first proofreading task received the next proofreading task one week after completion. Similarly, those who completed the second task were emailed a third proofreading task a week later. Participants were paid $0.50 for each completed proofreading task.

3. Watering Hole Attack:

We deemed everyone who responded to the three proofreading tasks as having become used to visiting our site to perform our tasks, thus serving as a suitable target for a targeted attack. We carried out the attack by emailing these individuals a fourth proofreading task via AMT. Like the email messages for the previous three tasks they had completed, the message for the fourth task also pointed them to a URL on our server. However, unlike the first three proofreading tasks, the Web page at the URL for the fourth task first displayed an attack message instead of the proofreading task.

Anyone visiting the URL for the fourth task was randomly shown one of the three messages. As can be seen in Figure 3, each message included a clickable 'Download' link. Anyone who clicked on the link in the messages was considered to have fallen for the attack. While a real attack would result in a click leading an individual to a malicious site or software, our study posed no such harm as the link was non-functional; it merely kept the participant on the same Web page. One minute after being presented with the attack message, those who had not closed the browser window or tab were automatically

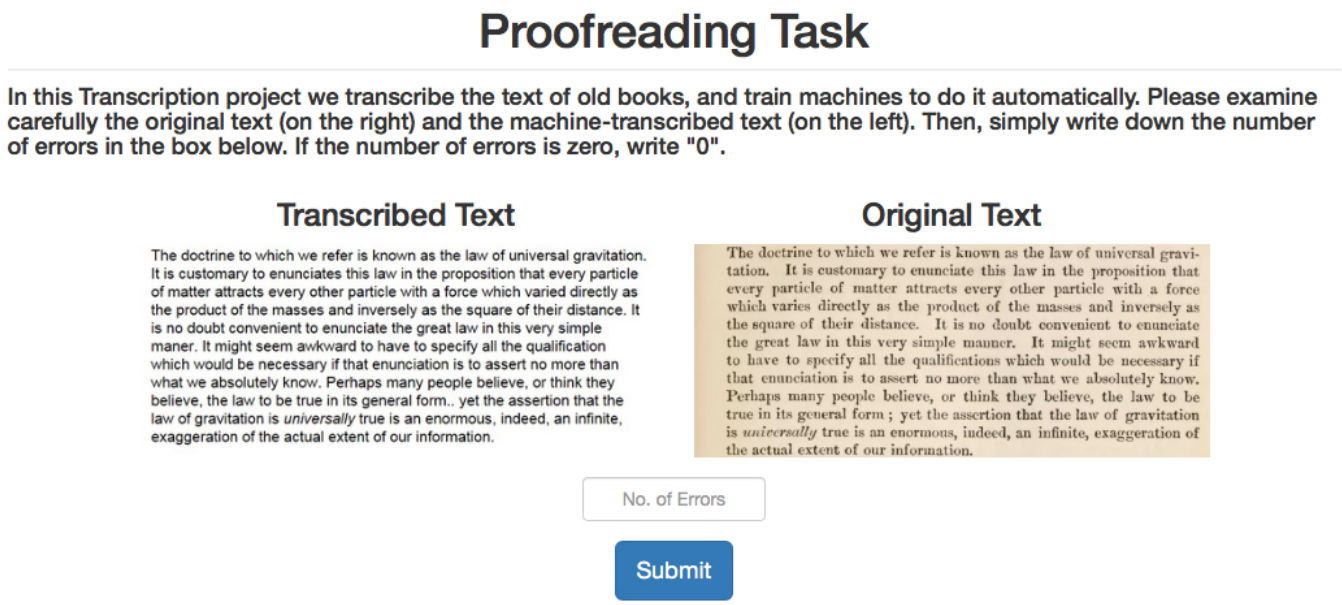

Figure 2. One of the proofreading tasks used to establish a trusted relationship with the participants.

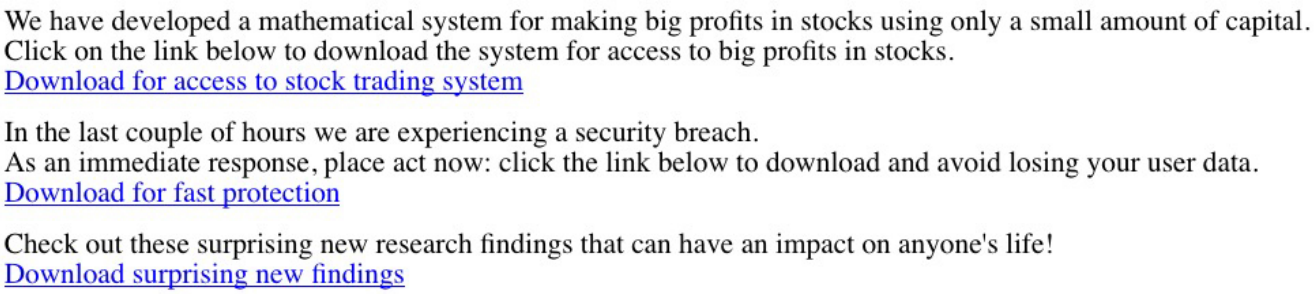

Figure 3. Screenshots of the three attack messages.

redirected to the fourth proofreading task regardless of whether they had clicked on the link in the attack message. At this time, participants could complete the fourth proofreading task just as they completed the earlier three. Those who completed the task despite not clicking on the link in the attack message were considered to have partially fallen for the attack because their response suggested that they ignored the implications of the earlier suspicious message. At the conclusion of the study, we sent everyone a debriefing message revealing and explaining the true purpose of the study.

The study was conducted for a period of 7 months from November 2015 to June 2016. We received 847 responses to the initial questionnaire task, out of which we retained the 808 that successfully passed attention checks. Of these, 266 completed all of the first three proofreading tasks and received the attack email. Ultimately, 235 (142 female, 97 male, with ages ranging from 18 to 70+) visited the attack page and saw one of the three attack messages.

FINDINGS

Of the 239 individuals who were targets of our attack across the three message types, 97 (40.6%) clicked on the 'Download' link on the attack page, and a further 35 (14.6%) completed the fourth proofreading task that was presented after the attack page timed out upon not receiving a click. As mentioned in the Method section, we treated the former as falling *fully* for the attack and the latter as falling *partially* for the attack. Overall, more than half of the participants (55.2%) were fully or partially deceived by our attack, thus underscoring the general effectiveness of deploying an attack in a manner that exploits a trusted channel. For instance, spear phishing tactics attempt to deceive people precisely by spoofing a trusted channel.

The 239 participants were distributed roughly equally across the three types of messages; 75 were shown the *security breach* message, 88 saw the *profit system* message, and the remaining 76 received the *research findings* message (see the Method section for the text of the messages). Table 1 shows the response patterns for each of the messages broken down by whether an individual clicked the link in the attack message. Those who did not click the link are further separated into those who

completed the proofreading task that was eventually shown after the attack page timed out. It can be clearly seen in Table 1 that the distribution of responses differed for each of the three messages. The (*research findings*) message was the most successful in luring people into clicking the link while the (*security breach*) message was the least effective; the scamming success of the (*profit system*) message was roughly in between these two extremes.

For each participant, we computed the scores for each of the Big Five personality traits, viz., openness, agreeableness, extraversion, conscientiousness, and neuroticism, as measured by our questionnaire items. Overall, the sample showed high levels of openness, agreeableness, and conscientiousness and low levels of extraversion. Levels of neuroticism were mostly concentrated around the middle. We calculated the scores for CMC Knowledge and CMC Efficacy for each respondent by averaging the corresponding items from the questionnaire. Most of our respondents rated their CMC knowledge and efficacy to be high (means and medians for both were greater than 4 on a 5 point scale). This should be unsurprising since the population of AMT workers can be expected to be more technically savvy than the average population. It is however noteworthy that a sizable proportion of our respondents fell for the attack despite higher technical competence (see Table 1).

To address the research questions outlined earlier, we examined two binary outcome variables:

- *scammed*: This variable indicates whether a participant was deceived by the attack, either by clicking the attack link or by completing the subsequent task. Essentially, this variable divides the sample into two parts: those who did not fall for the attack (i.e., the second column of Table 1) and the rest.

- *scammed-strong*: This variable indicates whether a participant clicked the link on our attack pages. In this case, we excluded those participants who did not click the link but completed the subsequent task; the excluded participants are already covered under the *scammed* variable and their numbers are relatively small (see the fourth column of Table 1), so we can safely remove them from consideration to maintain *scammed-strong* as a binary variable as well as to avoid statistical limitations posed by their relatively small numbers.

Since both outcome variables are binary, we employed Binary Logistic Regression (BLR) for each of the two cases. The size of the sample exposed to our attack messages along with the proportions of individuals deceived into clicking each of the messages provide sufficient statistical power to employ BLR to investigate the research questions we outlined in Section 2. The BLR used *message-type*, the five personality traits, *CMC-knowledge*, CMC-efficacy as the predictor variables along with two demographic variables (viz., *age-group* and *gender* and *errors* reported by our participants in the initial three proofreading tasks. In order to proceed systematically in the order of importance/significance of the independent variables pertaining to our research questions, we followed a stepwise approach. Such an approach is common when refining a model in successive steps based on priority ordering of variables/hypotheses. We started with *message-type* as the first step, added personality traits in the second step, *CMC-knowledge* and *CMC-efficacy* as the third step, and the demographic variables and the *errors* as the fourth step. At each step, we retained variables from the prior step only if they exhibited statistically significant effects, and we tested for interaction effects among the predictor variables, wherever needed.

Both regressions confirmed that *message-type* is a statistically significant predictor of an individual's susceptibility to being scammed ($p < 0.001$). Table 2 shows regression results for the variable *scammed*; regression results for the *scammed-strong* variable are similar. Surprisingly, we found that all of the personality traits as well as *CMC-knowledge* and *CMC-efficacy* were unrelated to an individual's response to the attack. Further, unlike prior studies in the literature, we did not observe any effects of gender or age on the tendency to fall for the attack, either fully or partially. In previous studies that reported a gender effect, there were differences between the technical knowledge among the two genders. In our case, there is no statistical differences between the two genders regarding reported CMC knowledge. This suggests that technical knowledge appears to be a stronger indicator for susceptibility to attacks than demographic factors such as age or gender. The *errors* reported by our participants in the proofreading tasks were also statistically unrelated to how participants reacted to the attack. In summary, our findings indicate that the content of a message is the primary driver behind a person's susceptibility to being deceived by an attack message; demographics and personality traits seem to have no influence, at least for a tech savvy population with high CMC knowledge and efficacy.

DISCUSSION

Although we carried out a realistic attack, there are a number of important details that must be discussed as they may have potentially impacted our results as well as their interpretation and generalizability. First, the URL included in our emails directed participants to a server hosted at our university. While it is not inconceivable for servers at legitimate organizations to be compromised into hosting attacks, it does differ from attacks that direct individuals to non-legitimate URLs not belonging to the organization or the party being impersonated. Moreover, we do not know the extent to which our participants were mindful of checking the legitimacy of the URL. On a related note, the attack email was received via a trusted channel (viz., the AMT platform) which might have influenced the degree of trust in the legitimacy of the message. Second, compared to the average population, our sample of highly rated AMT workers is notably more knowledgable and skilled with CMC. That said, it should be concerning that even reasonably tech savvy individuals exhibited relatively high levels of vulnerability to a targeted attack; the situation is likely worse for the general population. Third, individuals who participate in research studies in general, and AMT workers in particular, may be more curious compared to the average population, thus more likely to respond to cues that exploit curiosity. Moreover, the message that targeted curiosity referred to 'research findings' which might have evoked greater interest from our sample as it is composed of those

Table 1. Responses to different attack messages.

Message Type	NOT *scammed* (Did not click link)	*Scammed*	
		(Clicked link) *strong*	(Completed Task) *weak*
Security breach	47 (62.7%)	18 (24.0%)	10 (13.3%)
Profit system	36 (40.9%)	35 (39.8%)	17 (19.3%)
Research findings	24 (31.6%)	44 (57.9%)	8 (10.5%)

who choose to participate in research studies. Fourth, AMT workers may not always be fully diligent and attentive when performing tasks, especially if their primary intention is to maximize the payment received per unit time. While we included 'attention check' questions in the initial questionnaire, there were no explicit attention checks during the proofreading tasks and the attack. The large ranges and variances in the number of errors reported for each of the proofreading tasks may indicate varying levels of attentiveness, and we do not know whether this variability impacted the responses to the attack messages. It should however be noted that only those who passed the attention checks in the original questionnaire proceeded to the latter parts of the study. Thus, egregiously inattentive participants are likely not included in our results.

Our results highlight that individuals are vulnerable to attacks at multiple levels. Some of the participants who did not click the attack link regardless completed the proofreading task presented after the attack page timed out. We do not know why these participants did not click the link until the page timed out. While the lack of response could indicate that the participants did not find the message content interesting or relevant, another possible reason could be that they correctly deemed the attack messages to be suspicious. In the latter case, completing the proofreading task presented after the attack may be indicative of the inability to extrapolate the heuristic used to identify potential attacks and link it to subsequent interactions with the potentially malicious entity. In our study, once the first suspicious page was encountered, an individual should have been cautious of any other pages on the server and refrained from clicking any links or buttons on other pages as well. It is conceivable that those who completed the post-attack proofreading task were more focused on getting paid for the task. If so, their behavior affirms previous research showing many individuals engage in risky online behavior to obtain even the smallest of monetary reward [18].

A possible explanation for the observed variation in response rates across the three attack messages might pertain to the specifics of the message contents. For instance, those who did not perceive the messages to be useful or relevant may not click the attack link regardless of other facets of the content. As we noted, we observed the lowest rates of vulnerability for the message that mentioned a security breach. A possible explanation for this finding might pertain to the details mentioned the message. Specifically, the message referred to 'user data.' Since our participants did not maintain any data on our servers, they may have treated the message as suspicious and/or inapplicable owing to its lack of fit with expectations [6] and/or lower 'narrative strength' [45]. On the other hand, our participants were most likely to be deceived by the message regarding research findings.

Hong [24] describes that attack messages often employ appeals to various emotions in order to elicit a response from the target. Differences in emotional content across our messages could be another potential explanation for the observed differences across the messages. To evaluate this aspect, we carried out a post-hoc supplementary study in which we asked AMT workers to rate the content of each message for its emotional appeal. Specifically, the respondents used a 1–5 scale to indicate the extent to which the message appealed to six common emotions (viz., anger, greed, envy, urgency, curiosity, and fear), with 1 indicating that the message does not appeal to the emotion at all and 5 indicating that the message appeals to the emotion perfectly. The respondent saw each of the three messages in Figure 3 in random order. The order of the six rated emotions was also randomized. Respondents were paid $0.25 for participating in the study. We screened potential participants by following the approach of Allcott and Gentzkow [1], asking explicitly whether they commit to providing thoughtful answers. Only those who answered in the affirmative qualified to participate.

Table 3 shows the average ratings for each emotion for each of the three attack messages across 101 valid responses. As Table 3 indicates, the three messages do indeed differ in emotional content with the messages for security breach, profit system, and curiosity received high average ratings of 4.5, 3.4, and 4.1 for urgency, greed, and curiosity, respectively.

It can be observed in Table 3 that emotions are complex; as such emotional appeals in naturalistic messages are not crisp and may involve presence of other emotional cues besides the dominant one. For instance, the message regarding the security breach may induce fear, the one mentioning the profit system may engender curiosity, and the one on research findings may be perceived as urgent. Further research with more messages instantiating different emotions is needed to unpack the impact of specific emotional appeals. For instance, a follow up study with messages designed to target each emotion could further unpack these aspects.

Our underlying intention was to mimic a real-world attack. As such, our attack was targeted specifically at those who had completed our prior tasks on AMT and would have treated us as a trusted party based on their prior interactions with us. At the same time, we needed to construct our attack to ensure practical feasibility and reasonable believability while complying with AMT's Terms of Service and platform constraints. These operational constraints did necessitate a few deviations from a 'typical' in-the-wild attacks. First, the attack email

Table 2. Results of a stepwise binomial logistic regression for the variable *scammed*.

	scammed		
Variable	**Odds ratio**	**95% CI**	**p**
STEP 1:			
baseline: Message type = *securitybreach*			
Message type = *profitsystem*	2.4246	(1.2966, 4.6030)	0.0060**
Message type = *researchfindings*	3.6369	(1.8747, 7.2261)	0.0002***
STEP 2:	*χ^2(Step1, Step 2) = 11.1800, Degrees of freedom = 15*		*0.7397*
baseline: Message type = *securitybreach*			
Message type = *profitsystem*	6.0304	(0.0239, 1662.6504)	0.5240
Message type = *researchfindings*	7.6872	(0.0168, 4054.2426)	0.5160
Openness	0.9786	(0.5389, 1.7993)	0.9430
Conscientiousness	0.9831	(0.4340, 2.2375)	0.9670
Extraversion	0.8972	(0.5346, 1.4812)	0.6730
Agreeableness	1.2647	(0.5951, 2.7785)	0.5460
Neuroticism	1.0354	(0.5203, 2.0303)	0.9190
Message type = *profitsystem*: Openness	0.9990	(0.4289, 2.2990)	0.9980
Message type = *researchfindings*: Openness	0.6370	(0.2535, 1.5049)	0.3160
Message type = *profitsystem*: Conscientiousness	0.7022	(0.2504, 1.9380)	0.4950
Message type = *researchfindings*: Conscientiousness	1.0976	(0.3885, 3.0776)	0.8590
Message type = *profitsystem*: Extraversion	0.9506	(0.4789, 1.8927)	0.8840
Message type = *researchfindings*: Extraversion	1.7059	(0.7292, 4.1930)	0.2280
Message type = *profitsystem*: Agreeableness	1.0019	(0.3577, 2.7832)	0.9970
Message type = *researchfindings*: Agreeableness	1.0646	(0.3595, 3.1236)	0.9090
Message type = *profitsystem*: Neuroticism	1.1841	(0.4487, 3.1787)	0.7340
Message type = *researchfindings*: Neuroticism	0.8086	(0.3015, 2.1818)	0.6720
STEP 3:	*χ^2(Step1, Step 3) = 1.8195, Degrees of freedom = 6*		*0.9355*
baseline: Message type = *securitybreach*			
Message type = *profitsystem*	8.2538	(0.0651, 1234.9358)	0.3980
Message type = *resesarchfindings*	2.8753	(0.0196, 469.0485)	0.6780
CMC-efficacy	0.7892	(0.2668, 2.3704)	0.6650
CMC-knowledge	1.6047	(0.4973, 5.3852)	0.4310
Message type = *profitsystem*: CMC-efficacy	1.5113	(0.3616, 6.2852)	0.5670
Message type = *resesarchfindings*: CMC-efficacy	2.0281	(0.4375, 9.5472)	0.3640
Message type = *profitsystem*: CMC-knowledge	0.5008	(0.0979, 2.4662)	0.3980
Message type = *resesarchfindings*: CMC-knowledge	0.5372	(0.1097, 2.4900)	0.4320
STEP 4:	*χ^2(Step1, Step 4) = 8.7205, Degrees of freedom = 6*		*0.1899*
baseline: Message type = *securitybreach*			
baseline: Gender = *female*			
baseline: Age-group = *18–29*			
Message type = *profitsystem*	2.5407	(1.3303, 4.9422)	0.0052**
Message type = *resesarchfindings*	3.4321	(1.7345, 6.9523)	0.0005***
Gender = *male*	1.0463	(0.5923, 1.8571)	0.8763
Age-group = *30–49*	0.9319	(0.4765, 1.8161)	0.8358
Age-group = *50+*	1.9240	(0.8205, 4.6187)	0.1360
Errors: Task 1	0.9961	(0.9085, 1.1190)	0.9389
Errors: Task 2	1.0117	(0.8798, 1.1660)	0.8710
Errors: Task 3	1.2049	(0.9267, 1.5804)	0.1688

Statistical significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 3. Average ratings for the extent to which each attack message appeals to a given emotion on a 1–5 scale (1 = Message does not appeal to the emotion at all and 5 = Message appeals to the emotion perfectly).

Rated Emotion	*Content of Attack Message*		
	Security breach	Profit system	Research findings
Urgency	4.5	2.2	2.5
Greed	1.3	3.4	1.3
Curiosity	2.7	3.3	4.1
Fear	3.7	1.2	1.3
Anger	2.1	1.1	1.1
Envy	1.1	1.7	1.4

was not personalized to the individual even though it was sent using AMT's targeted email functionality. Second, the attack message was not contained within the attack email itself but presented after clicking the link in the email. That said, links in phishing emails often take the target to page that demands further action, such as providing login credentials or personal information. In our case, the solicited action was clicking to download content. Third, the attack page was hosted on a university server. While the logistics of our watering hole attack may differ somewhat from attacks observed in the wild, it shares core aspects that are common to several common attacks, such as phishing, spear phishing, etc. As such, we believe our findings can not only generalize to real-world attacks but also cover a broad variety of tactics. Regardless of the extent to which our attack matches those observed in the real world, it should be noted that we report findings based on *comparison* across the different study conditions. As the operational elements of the attack were exactly the same across these conditions, the observed effects can be safely attributed only to differences in the message content.

Finally, the use of deception in our study involves considerations regarding ethical aspects of conducting such studies. We needed to use deception since the study clearly could not have been carried out otherwise. However, we minimized potential harm to the participants with the use of a non-functioning attack link. Further, at the end of the study, we debriefed all participants by revealing the true goal of the study. With these mitigating measures, we deemed that the potential harms of our deception were negligible, especially in relation to the potential benefits of the findings for developing more effective defenses against online attacks. The study protocol details, including these ethical considerations, were reviewed and approved by our university's Institutional Review Board (IRB).

IMPLICATIONS

Our results hold several important implications for efforts to defend against online attacks:

Resources for defense measures could derive more benefit from enhanced techniques for analyzing message content rather than attempting to personalize the solutions based on demographics or personality characteristics of the recipient. As a result, the same solution may be effectively deployed across a broad spectrum of individuals, thus enabling it to scale without needing to handle the complexities of myriad customized configurations. In this regard, content analyses that detect sentiment and emotion could potentially enhance current automated techniques.

In addition to the automated sentiment detection proposed above, user training as well as user interface (UI) functionalities could be designed to defend against emotional appeals. Such efforts could for instance be modeled on training and UI features that have been successful in getting users to recognize the importance of treating email attachments with caution in order to avoid getting infected by viruses or malware. Promoting caution against overt emotional manipulation could also have the side benefit of helping users avoid clickbait sites with low signal-to-noise ratios.

Our findings could be applied in several ways to enhance user training and education pertaining to attack detection and protection. First, users must be cautioned that once any site exhibits suspicious activity, subsequent interactions with that site should be deemed insecure even when they appear legitimate. Second, users must be taught to handle potentially suspicious cases via 'alternate channel' verification, such as a phone call or instant message to a relevant trusted party. Third, users need to recognize that a legitimate site could potentially turn malicious if it has been compromised, so if something seems 'out of the ordinary' then merely verifying the site's URL or SSL certificate for authenticity may not be sufficient. In addition to explicitly scheduled training, efforts to elevate attack related technical knowledge and skills could also target delivery at contextually opportune moments, such as right after someone falls for a 'mock' attack.

Our findings are contrary to those of prior studies that reported influence of gender and personality traits on tendency to be deceived by cyber attacks. These differences could potentially be attributed to study deployment and/or sample characteristics. For instance, several of the past studies involved simulated scenarios with predominantly student samples. In contrast, our study utilized deception to carry out a realistic attack with a reasonably large sample drawn from a non-student population. The differences could also be indicative of long-term learning effects across the population that could have neutralized the influences of gender and personality as the general population gains more familiarity and experience in dealing with cyber attacks. Seeking refined explanations for these differences needs further longitudinal investigations of realistic attacks across diverse populations. In this regard, several companies are known to carry out internal 'security drills' that include spear phishing attacks. While such drills are important, they do not go far enough because they focus only on detecting whether an employee successfully avoided the mock attack, without a deeper examination of individual characteristics, such as demographics, personality, technical skills, job functions, etc. Moreover, results from a given company might not generalize outside the particular firm. Therefore, we suggest a complementary approach that includes studies from work as well as non-work settings.

LIMITATIONS

As discussed in the Findings and Discussion sections, our sample of AMT workers differs from the average population in a

number of ways, such as technical skills, attitudes toward research studies, etc. Therefore, generalization of these findings to the general population needs further verification via replication with other samples. In particular, future studies should be designed to test the effects of specific technical skills by including those with low as well as high tech efficacy. The specifics of our attack deployment, such as the individual applicability and relevance of the message contents, the pretext of a 'research study,' and the monetary compensation involved in our tasks, could have affected our results and could limit their generalizability to typical attacks encountered in-the-wild. As a result, there is a need to verify the results via in more realistic scenarios (for instance, simulated attacks carried out internally in organizations).

CONCLUSION

We carried out a realistic targeted watering hole attack on individuals recruited via Amazon Mechanical Turk for an unrelated study. Nearly 40% of our participants clicked on the attack link, despite reporting high levels of knowledge and efficacy related to online interactions. Our findings suggest that different kinds of message content and emotional cues vary in their ability to deceive people. On the other hand, we found no differences in susceptibility to the attack across different ages, genders, and personality traits. Therefore, when dealing with a reasonably tech savvy population, such as knowledge workers, customizing defenses based on demographics and personality may be unnecessary. Instead, cyber defense resources could be utilized to augment intrusion detection filters to include automated sentiment analyses of message content to flag strong emotional signals. It may also be beneficial to raise awareness, knowledge, and experience via periodic internal deployment of realistic attacks that provide appropriate post-attack training. For the less technically savvy home users, the situation is likely worse. Further research is needed for unpacking the interplay between attack susceptibility and specific technical skills that can help detect attacks.

ACKNOWLEDGMENTS

Anonymous

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. *Social media and fake news in the 2016 election*. Technical Report. National Bureau of Economic Research.

[2] Yair Amichai-Hamburger and Gideon Vinitzky. 2010. Social network use and personality. *Computers in human behavior* 26, 6 (2010), 1289–1295.

[3] Eduardo B Andrade and Dan Ariely. 2009. The enduring impact of transient emotions on decision making. *Organizational Behavior and Human Decision Processes* 109, 1 (2009), 1–8.

[4] Vidur Apparao. 2016. What you need to know about spear phishing. http://betanews.com/2016/05/06/how-spear-phishing-works-and-how-to-avoid-it/. (2016).

[5] Richard Barber. 2016. Modern Spear Phishing is a Security Wake-Up Call. http://ww2.cfo.com/cyber-security-technology/2016/04/modern-spear-phishing-security-wake-call/. (April 2016).

[6] Zinaida Benenson, Freya Gassmann, and Robert Landwirth. 2017. Unpacking Spear Phishing Susceptibility. In *Financial Cryptography and Data Security*, Michael Brenner, Kurt Rohloff, Joseph Bonneau, Andrew Miller, Peter Y.A. Ryan, Vanessa Teague, Andrea Bracciali, Massimiliano Sala, Federico Pintore, and Markus Jakobsson (Eds.). Springer International Publishing, Cham, 610–627.

[7] Mark Blythe, Helen Petrie, and John A. Clark. 2011. F for Fake: Four Studies on How We Fall for Phish. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 3469–3478. DOI: http://dx.doi.org/10.1145/1978942.1979459

[8] D. D. Caputo, S. L. Pfleeger, J. D. Freeman, and M. E. Johnson. 2014. Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy* 12, 1 (Jan 2014), 28–38. DOI: http://dx.doi.org/10.1109/MSP.2013.106

[9] Cloudmark. 2016. Survey Reveals Spear Phishing as a Top Security Concern to Enterprises. https://blog.cloudmark.com/2016/01/13/survey-spear-phishing-a-top-security-concern-to-enterprises/. (January 2016).

[10] Federal Communications Commission. 2015. FCC adopts strong, sustainable rules to protect the Open Internet. https://apps.fcc.gov/edocs_public/attachmatch/DOC-332260A1.pdf. (February 2015).

[11] PT Costa Jr and Robert R McCrae. 1992. Neo personality inventory–revised (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual. *Odessa, FL: Psychological Assessment Resources* (1992).

[12] Caroline Davis, Karen Patte, Stacey Tweed, and Claire Curtis. 2007. Personality traits associated with decision-making deficits. *Personality and Individual Differences* 42, 2 (2007), 279–290.

[13] 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.

[14] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. 2006. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security*. ACM, 79–90.

[15] 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 (CHI '08)*. ACM, New York, NY, USA, 1065–1074. DOI: http://dx.doi.org/10.1145/1357054.1357219

[16] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 149–156.

[17] Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist* 48, 1 (1993), 26.

[18] Jens Grossklags and Alessandro Acquisti. 2007. When 25 Cents is Too Much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information. In *WEIS 2007: Workshop on the Economics of Information Security*.

[19] Tzipora Halevi, James Lewis, and Nasir Memon. 2013. A Pilot Study of Cyber Security and Privacy Related Behavior and Personality Traits. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13 Companion)*. ACM, New York, NY, USA, 737–744. DOI: http://dx.doi.org/10.1145/2487788.2488034

[20] Tzipora Halevi, Nasir Memon, and Oded Nov. 2015. Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks. http://ssrn.com/abstract=2544742. (January 2015).

[21] Ryan Heartfield, George Loukas, and others. 2016. Predicting the performance of users as human sensors of security threats in social media. *International Journal on Cyber Situational Awareness (IJCSA)* 1, 1 (2016).

[22] Cormac Herley. 2009. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop (NSPW '09)*. ACM, New York, NY, USA, 133–144. DOI: http://dx.doi.org/10.1145/1719030.1719050

[23] Benjamin E Hilbig. 2008. Individual differences in fast-and-frugal decision making: Neuroticism and the recognition heuristic. *Journal of Research in Personality* 42, 6 (2008), 1641–1645.

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

[25] Kyung Wha Hong, Christopher M. Kelley, Rucha Tembe, Emerson Murphy-Hill, and Christopher B Mayhorn. 2013. Keeping Up With The Joneses Assessing Phishing Susceptibility in an Email Task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 57. SAGE Publications, 1012–1016.

[26] Tom N Jagatic, Nathaniel A Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social phishing. *Commun. ACM* 50, 10 (2007), 94–100.

[27] Mitchell Kajzer, John D'Arcy, Charles R Crowell, Aaron Striegel, and Dirk Van Bruggen. 2014. An exploratory investigation of message-person congruence in information security awareness campaigns. *Computers & security* 43 (2014), 64–76.

[28] 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 (SOUPS '09)*. ACM, New York, NY, USA, Article 3, 12 pages. DOI: http://dx.doi.org/10.1145/1572532.1572536

[29] 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.

[30] Eric Langstedt. 2011. *An examination of the Five Factor Model Personality Traits as predictors of online social network use*. University of Connecticut.

[31] Mike Martindale. 2016. Troy investment company hacked; $495K stolen. http://www.detroitnews.com/story/news/local/oakland-county/2016/05/03/troy-investment-company-hacked/83879240/. (May 2016).

[32] Stephanie M Merritt and Daniel R Ilgen. 2008. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 2 (2008), 194–210.

[33] Matthew Mesa. 2016. Phish Scales: Malicious Actor Combines Personalized Email, Variety of Malware To Target Execs. https://www.proofpoint.com/us/threat-insight/post/phish-scales-malicious-actor-target-execs. (April 2016).

[34] Trend Micro. 2012. Spear-Phishing Email: Most Favored APT Attack Bait. https://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-spear-phishing-email-most-favored-apt-attack-bait.pdf. (2012).

[35] David Modic and Ross J Anderson. 2014. We will make you like our research: The development of a susceptibility-to-persuasion scale. Available at: https://ssrn.com/abstract=2446971orhttp://dx.doi.org/10.2139/ssrn.2446971. (2014).

[36] David Modic and Stephen E. G. Lea. 2012. How neurotic are scam victims, really? The big five and Internet scams. Available at: https://ssrn.com/abstract=2448130orhttp://dx.doi.org/10.2139/ssrn.2448130. (September 2012).

[37] Anne Myers and Christine Hansen. 2011. *Experimental psychology*. Cengage Learning.

[38] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. 2017. Dissecting Spear Phishing Emails for Older vs Young Adults: On the Interplay of Weapons of Influence and Life Domains in Predicting Susceptibility to Phishing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 6412–6424. DOI: http://dx.doi.org/10.1145/3025453.3025831

[39] Thomas Pfeiffer, Heike Theuerling, and Michaela Kauer. 2013. Click Me If You Can! In *Human Aspects of Information Security, Privacy, and Trust*. Springer, 155–166.

[40] Jon Russell. 2016. Snapchat employee data leaks out following phishing attack. http://techcrunch.com/2016/02/29/snapchat-employee-data-leaks-out-following-phishing-attack/. (February 2016).

[41] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 373–382.

[42] 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, New York, NY, USA, 88–99. DOI: http://dx.doi.org/10.1145/1280680.1280692

[43] Mark Snyder, Allen M Omoto, and Dylan M Smith. 2009. The role of persuasion strategies in motivating individual and collective action. *The political psychology of democratic citizenship* (2009), 125–150.

[44] Brian H. Spitzberg. 2006. Preliminary Development of a Model and Measure of Computer-Mediated Communication (CMC) Competence. *Journal of Computer-Mediated Communication* 11, 2 (2006), 629–666. DOI: http://dx.doi.org/10.1111/j.1083-6101.2006.00030.x

[45] Alex Tsow and Markus Jakobsson. 2007. Deceit and Deception: A Large User Study of Phishing. https://www.cs.indiana.edu/ftp/techreports/TR649.pdf. (2007).

[46] Kim Zetter. 2015. Hacker Lexicon: What Are Phishing and Spear Phishing? https://www.wired.com/2015/04/hacker-lexicon-spear-phishing/. (April 2015).