Phishing Attacks Over Time: A Longitudinal Study

Emergent Research Forum paper

Dave Yates
University of Denver
dave.yates@du.edu

Albert L. Harris
Appalachian State University
harrisal@appstate.edu

Abstract

In this paper we examine phishing emails received over a thirteen-year period and evaluate how they have changed on a number of characteristics. Using a dual-path model of persuasion, we categorize some characteristics as central (such as persuasiveness) and some as peripheral (such as message appearance), and hypothesize that both types of characteristics should be more prominent as phishing attacks have evolved and matured. Surprisingly, results show phishing emails are not more sophisticated over time. We comment on these results, discuss implications for IT security research, and describe future research directions.

Keywords

Phishing, e-mail, longitudinal analysis, SpamAssassin, credibility.

Introduction

Phishing attacks have plagued users and organizations for decades. Phishing emails appear in users' inboxes and are designed to entice users to respond, click a link to a fake website, or disclose sensitive information such as passwords (Hong 2012). These emails range from banal messages promising wealth, to threats from tech support to lock out accounts, to highly sophisticated spoofs of legitimate websites such as banks. The phishing methodology uses techniques to entice or persuade a user to take action that ultimately leads to further exploitation and compromise.

Previous studies have shown that phishing attacks originate from all areas of the globe and similarly target global users (Hong 2012). Some experts believe phishing has evolved from being the effort of individual hackers targeting specific networks to more organized, widespread attacks, in some cases state-sponsored. Recent phishing schemes include, according to the United States Federal Bureau of Investigation schemes related to elections, holiday shopping, and natural disasters. As techniques evolve, are users more at risk? Attacks may become more sophisticated over time (Kumaraguru et al. 2007), but so too do users become more wary of phishing attacks and automated anti-phishing tools also become more capable. Ramesh et al. (2014) and Hamid and Abawajy (2014) describe new techniques for detecting phishing attacks before they reach users. Similarly, Gowtham and Krishnamurthi (2014) explain how anti-phishing techniques must evolve as attack vectors evolve, noting that phishing attacks are often zero-warning and short lived. Despite new techniques, the onus is often on the user to identify a phishing attack (Vishwanath et al. 2011).

Researchers have found that users respond to or resist phishing emails based on the level of persuasiveness in the attacks, combined with reputation mechanisms and other cues that might fool a user into believe an attack was actually a valid request (Dhamija et al. 2006). Wang et al. (2009), James et al. (2013), and Wright et al. (2014) identify a host of factors based on dual-path models of persuasion and influence. Over time, it is unclear whether users’ increasing awareness to these techniques is lagging, keeping pace with, or exceeding attackers’ attempts to innovate and expand their attacks. A great majority of phishing attacks are automatically detected and filtered out before users seem them; conversely, usually the only successful attacks that are reported are those with spectacular results which may or may not be indicative of the success of phishing emails overall. Thus as phishing attacks have
Phishing Attacks Over Time

Phishing attacks succeed by causing a user to accomplish some response action advantageous to the attacker, such as responding to an email with personal information, clicking on a web link that leads to a fictitious website or automatically installs malware, or taking other actions which compromise the user (Anti-Phishing Working Group). Some phishing emails never reach users – automated detection tools such as Spam Assassin (http://spamassassin.apache.org) classify and eliminate phishing attacks that meet certain criteria so users never see them. These tools typically work by assessing a blend of characteristics and assigning a score or likelihood measure that the message is a phishing attack. When compared to a threshold, those messages above the threshold are eliminated and the rest are passed through to the user. Once in the hands of the user, the message is subject to a subjective evaluation of credibility (Wang et al. 2009). Those messages deemed credible for whatever reason are acted upon; those that are not are ignored or deleted. Depending on the attack, users may have multiple opportunities to decide a message is an attack and not a credible action – they may first click on a link, have to enter personal information, or download a file (Burns et al. 2013). However we are most concerned about the initial response action the user takes, and thus their first assessment of credibility when they see the phishing message.

Prior research has examined credibility of phishing attacks from one of two main perspectives. The first governs what might be called the ‘technological’ side of phishing attacks, namely the originating and target domains, message and website formats, detection heuristics, and automated detection and classification (Gowtham and Krishnamurthi 2014; Hamid and Abawajy 2014; Ramesh et al. 2014; Zhang et al. 2006). The second perspective, which is more pertinent to this research, is the user-interaction perspective, namely the characteristics of phishing emails themselves that users encounter (James et al. 2013; Wang et al. 2009). Users typically encounter phishing attacks that have passed through phishing detection mechanisms to land in a users’ email inbox. Once there, users are presented with phishing emails in the same manner as they might receive legitimate messages from friends, colleagues, and trusted organizations. At this point, the credibility of the phishing email becomes the determining factor as to whether or not a user will respond to the attack (Wright et al. 2014).

Users evaluate phishing attack credibility based on two categories of factors. The first is message appearance qualities (Wang et al. 2009), such as the visual appeal of the message, whether it appears to be from a legitimate and known source (evidenced by formatting, company logos, etc.), whether it seems legitimate on the surface (correct addressing, grammar, and scope), and whether there is a clear response action such as a URL link to activate. Wang et al. explain that such cues activate automated or peripheral persuasion mechanisms (Petty and Cacioppo 1986) that cause users to form a snap judgment as to the credibility of the message and whether or not they should comply with the request. Similarly, Wright et al. (2014) explain phishing attacks in terms of social influence mechanisms (Cialdini 2009), activating what Stanovich and West (2000) describe as System 1 thinking (automated response).

The second category of factors are evaluative judgments a user makes regarding the level of persuasiveness of the message, based on the credibility of the sender, the urgency and justification of the action in the message, and the penalty for not complying (Wang et al. 2009). These factors activate more deliberate or ‘central’ persuasion according to Petty and Cacioppo’s (1986) model of persuasion, or what Stanovich and West (2000) refer to as System 2 (deliberate and analytic) thinking. While Petty and
Cacioppo and Stanovich and Wests’ frameworks are similar, they describe different cognitive functions (the former a dual-path method of persuasion, the latter two ways individuals respond to influence); yet when applied to phishing attacks both describe similar sets of characteristics. For consistency, we will adopt Petty and Cacioppo’s (1986) terminology of ‘peripheral’ to describe the first category of factors, and ‘central’ to describe the second. We posit that as attackers learn and improve their techniques, peripheral and central characteristics of persuasion should be manifest in more sophisticated techniques embedded in phishing emails.

Researchers believe both types of characteristics are important, and describe ways in which they might evolve. Therefore, our goal is to evaluate whether or not phishing attacks have become more sophisticated over time, employing improved techniques for both peripheral and central response, and whether or not the overall credibility of the attacks have increased as well. We summarize these goals in the following hypotheses:

\[ H_1: \text{More recent phishing attacks will be rated as more credible} \]

\[ H_2: \text{More recent phishing attacks will have greater peripheral characteristics} \]

\[ H_3: \text{More recent phishing attacks will have greater central characteristics} \]

**Method and Analysis**

To evaluate the changes in phishing attacks over time, one of the authors collected over 1000 phishing messages over a multi-year period, from 2002-2014. In this preliminary analysis we compared a group of the earliest to a group of the latest (most recent) emails. Note that all of these emails successfully made it through spam filters to the user’s inbox; however, most messages were assigned a ‘spam score’ by the SpamAssassin tool installed on the University email server. Thus the sample is biased from the point of view of phishing attack originators, since some messages were automatically filtered out; however from a user’s perspective it is representative of the population of phishing emails that are received and acted upon. The messages were hand-coded for a number of peripheral and central characteristics, following Wang et al.’s (2009) description of phishing email characteristics. These characteristics are shown in Table 1.

<table>
<thead>
<tr>
<th><strong>Peripheral Characteristics</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of typos (TYPOS)</td>
<td>A count of misspellings or out of place words in the message text</td>
</tr>
<tr>
<td>Number of grammatical mistakes (GRAMMAR)</td>
<td>A count of grammatical mistakes such as fragments, improper references, or improper word choice</td>
</tr>
<tr>
<td>Spoof of a legitimate source (SPOOF)</td>
<td>Whether the message was a spoof of a legitimate source, such as Paypal or Citibank.</td>
</tr>
<tr>
<td>Link for response action (LINK)</td>
<td>Whether the message included a hyperlink to a website for the user to click as a response action</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Central Characteristics</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorized sender (SENDER)</td>
<td>Whether or not the message appears to originate from a specific and known sender</td>
</tr>
<tr>
<td>Individually addressed (ADDRESSED)</td>
<td>Whether or not the message was addressed to the user</td>
</tr>
</tbody>
</table>
---|---
Justification (JUSTIF) | Whether the message included a 'carrot' (incentive) or stick (threat) as the persuasive mechanism (Wang et al. 2009). Coded on a 1-5 scale.

Table 1: Phishing Message Characteristics

To gauge source credibility (CRED) we employed a composite measure suggested by McCroskey and Teven (1999) that assesses competence, trustworthiness, and goodwill. Source credibility was coded on a 1-5 scale for each component measure then averaged. As an objective measure of sophistication we employed the SpamAssassin score assigned to each message by the email server. SpamAssassin scores range from 0-10; but those above the threshold value of 6 were blocked so actual scores range from 0-6. SpamAssassin uses rule sets and a genetic algorithm to detect keywords or phrases, originators, and/or formats that are consistent with questionable email that might harm the recipient. While spam includes emails that might not be phishing attacks (such as aggressive marketing ploys), conversely all phishing attacks would be classified as spam by the tool. Thus the SpamAssassin score is a useful and representative measure of the strength of the phishing attack.

Messages were coded by a graduate MBA student at a private western-US university, who had no prior information security experience. Prior to coding, the student was trained and provided a scoring rubric, following Denzin and Lincoln (2003). A portion of the responses was independently recoded for inter-rater reliability and a random sample of messages was further reviewed by an undergraduate student using a Q-sort strategy to order messages by certain characteristics. While these checks continue as more data is coded, preliminary results indicate acceptable reliability in the coding methodology. Future analyses will include a third coder and calculation of Fleiss’ Kappa to demonstrate suitable inter-rater reliability. This study includes data from 157 messages, 106 from the timeframe 2002-2006, and 51 from 2014. These messages represent the first 106 and last 51 in the dataset by the date received.

Table 2 provides descriptive statistics, broken out into earlier vs. later messages. The analysis was conducted in SPSS using analysis of variance (ANOVA) and for the binomial (i.e. 0/1 variables), logistic regression.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Earlier Messages (n=106)</th>
<th>Later Messages (n=51)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYPOS</td>
<td>Mean 0.73, Std. Deviation 1.593</td>
<td>Mean 1.08, Std. Deviation 1.51</td>
</tr>
<tr>
<td>GRAMMAR</td>
<td>Mean 0.02, Std. Deviation 1.855</td>
<td>Mean 1.08, Std. Deviation 1.44</td>
</tr>
<tr>
<td>SPOOF</td>
<td>Mean 0.82, Std. Deviation 0.385</td>
<td>Mean 0.71, Std. Deviation 0.46</td>
</tr>
<tr>
<td>VISAPPEAL</td>
<td>Mean 3.90, Std. Deviation 0.93</td>
<td>Mean 3.24, Std. Deviation 0.98</td>
</tr>
<tr>
<td>LINK</td>
<td>Mean 0.86, Std. Deviation 0.352</td>
<td>Mean 0.63, Std. Deviation 0.49</td>
</tr>
<tr>
<td>SENDER</td>
<td>Mean 0.75, Std. Deviation 0.56</td>
<td>Mean 0.80, Std. Deviation 0.77</td>
</tr>
<tr>
<td>ADDRESSED</td>
<td>Mean 0.67, Std. Deviation 0.44</td>
<td>Mean 0.80, Std. Deviation 0.55</td>
</tr>
<tr>
<td>URGENCY</td>
<td>Mean 3.35, Std. Deviation 1.54</td>
<td>Mean 2.80, Std. Deviation 1.41</td>
</tr>
<tr>
<td>PENALTY</td>
<td>Mean 2.85, Std. Deviation 1.68</td>
<td>Mean 2.45, Std. Deviation 1.55</td>
</tr>
<tr>
<td>JUSTIF</td>
<td>Mean 3.37, Std. Deviation 1.30</td>
<td>Mean 2.65, Std. Deviation 1.21</td>
</tr>
<tr>
<td>SPAMASSASSIN</td>
<td>Mean 3.41, Std. Deviation 1.68</td>
<td>Mean 3.24, Std. Deviation 4.42</td>
</tr>
<tr>
<td>CRED</td>
<td>Mean 3.45, Std. Deviation 1.11</td>
<td>Mean 2.69, Std. Deviation 1.11</td>
</tr>
</tbody>
</table>

Table 2: Descriptive Statistics
To test hypothesis 1 we conducted an independent samples t-test ANOVA using data from 2002-2006 as the first series and data from 2014 as the second series. We looked for significant differences in SpamAssassin scores and Source Credibility between the two series. For SpamAssassin scores, results indicated that the difference in scores from the earliest to the latest messages was not significant (t = 0.337, p = 0.74). For Source Credibility, we found that the later messages were significantly less credible (t = 3.98, p < 0.001), contrary to expectations. Thus H1 was not supported, and analysis yielded the unexpected result that later messages were judged less credible than earlier messages.

To test hypotheses 2 and 3, we conducted an independent samples t-test ANOVA for the continuous peripheral variables and logistic regression for the binary variables. Of the peripheral and central characteristics, three showed significant differences between the earlier and later message sets. Urgency (t = 2.11, p < 0.05), Justification (t = 3.30, p = 0.001), and presence of a Link ($\chi^2 = 10.107, p < 0.001$) were all significantly LOWER in the later group than the earlier group, contrary to expectations. The other characteristics did not significantly differ among messages spanning 2002-2006 and 2014.

**Discussion and Conclusion**

It is surprising that the only significant differences found between source credibility and the peripheral and central characteristics were in the negative direction, meaning that the later messages showed weaker persuasive elements than did the earlier messages. These results indicate that either phishing detection algorithms are doing a better job at filtering out the more sophisticated attacks, which is likely, or that phishing attackers are relying on the same techniques they always have and do not have to innovate to yield results. The wealth of research on educating users about phishing (Burns et al. 2013; Sheng et al 2010) suggests that users are still highly susceptible to even basic phishing attacks.

These are preliminary results – a more comprehensive analysis of the complete dataset should yield more comprehensive and concrete analysis regarding each of the different phishing attack characteristics. Yet, these results offer researchers interesting conclusions for future research. First, while it is unclear from these data how many people actually respond to phishing attacks, it is clear that the attack vectors (i.e. the characteristics) are not changing dramatically. Thus, education efforts ought to be ultimately successful in countering phishing attacks. Since new users join organizations every day this education is a constant, not one-time effort, however. Similarly, users might be able to detect patterns in even highly persuasive phishing messages and thus avoid them, despite their credibility. Thus higher credibility does not necessarily mean greater exploitation. It might, counter intuitively, make credible phishing messages easier to identify and avoid.

In summary, more research is required to better understand how these characteristics have changed as phishing evolves. The data point to a wealth of interesting avenues to explore. The more researchers understand the relative contribution of each of these characteristics toward overall credibility, the easier it will be to prevent future phishing attacks.

**REFERENCES**


