The (Relative) Impact of Email Cues on the Perceived Threat of Phishing Attacks A User Perspective on Phishing Deceptiveness

DECISION: Accept

AUTHORS: Pavlo Burda, Maria Eleni Kokkini, Luca Allodi and Nicola Zannone

Summary of Reviews

- Review 1: 1 (3)
- Review 2: 3 (3)
- Review 2: 2 (4)

Reviews

Review 1

Total score: 1

Overall evaluation: 1

Reviewer's confidence: 3

This paper delves into the intricate relationship between phishing cues and the perceived deceptiveness of phishing emails. Through an experiment with 74 participants, the study aims to unravel how various cues in phishing emails influence users' perceptions of deception effectiveness.

The experiment, conducted on Amazon Mechanical Turk, consisted of two tasks. In the first task, participants were tasked with ranking crafted phishing emails impersonating MTurk based on their perceived deceptiveness. The findings shed light on several intriguing insights.

The study reveals that phishing attempts incorporating persuasion principles, such as Persuasion with Authority and Commitment & Consistency, along with closely mimicked sender impersonation (e.g., logo, link), were perceived as more deceptive by participants. Notably, while persuasion cues exerted the strongest influence on perceived deceptiveness, impersonation cues significantly impacted participants' expectations regarding the effectiveness of the phishing attack.

One of the notable conclusions drawn from the research is the significant impact of impersonation, email etiquette, and persuasion on enhancing user perceptions of email deceptiveness. Furthermore, the study highlights that impersonation cues specifically contribute positively to the expected effectiveness of a phishing attack.

Overall, the paper provides valuable insights into the psychological mechanisms underlying users' perceptions of phishing emails. By identifying the specific cues that influence perceived deceptiveness and expected effectiveness, the study offers valuable implications for cybersecurity awareness and education initiatives. However, it is essential to acknowledge the limitations of the study, such as the relatively small sample size and the use of a specific platform for participant recruitment.

Review 2

Total score: 3

Overall evaluation: 3

Reviewer's confidence: 3

This paper describes an experiment to test factors contributing to the success of phishing email detection. The authors performed an experiment on MTurk with 74 participants who were asked to rank several example emails in effectiveness.

Using MTurk created severe limitations on the experiment, the subjects could only see an image of the email, and not the email itself. Furthermore, the subjects could be affected by response fatigue, even though the experimenters would randomise the order, this may still be present, and the subjects also knew they were looking at phishing emails, affecting their responses. Finally, as the authors note, MTurk participants are not a reflection of US population, and from this experimental setup it is hard to see what kind of population is actually represented.

Nonetheless, this experiment provides some interesting indications for phishing and especially towards the factors of perceived deceptiveness and expected effectiveness.

Looking forward to seeing this presented.

Review 3

Total score: 2

Overall evaluation: 2

Reviewer's confidence: 4

This paper presents a technically very well designed survey that asks users to compare deceptiveness and effectiveness of 8 phishing emails. The statistical analysis is done well, as far as I can see (not being a statistics expert). This paper seems to investigate a very small detail of an already very thoroughly investigated problem of phishing susceptibility, such that a workshop is an

appropriate venue for it. The results are interesting, even if not ground-breaking. I do not see any reasons not to accept this paper, if it makes some changes as indicated below.

1) The most important limitation of this paper is that, like in virtually all quantitative phishing papers before, the cues in emails were designed solely form the researchers' perspective and not tested beforehand with the users. Thus, as the paper notes in the Discussion, the relevance cue "backfired".

Yet, what is relevance? This should be defined in the present paper. I would say that relevance means how the email fits the life context of the user, and I agree that in general, relevance is difficult to incorporate into the generic phishing attack. However, it is possible to incorporate it into a mildly targeted phishing campaign, for example at students at Christmas time, as was done by Benenson et al. (reference below). If this is done well, the relevance ques become one of the most prominent factors in user decisions.

2) Mostly quantitative studies are cited in the paper, and the Related Work section does not distinguish between quantitative and qualitative studies (e.g., Reference 56 Parsons et al.). Yet, some qualitative studies investigated the "user perspective", which is stated as missing in the literature on page 2: "the user perspective is generally not considered", for example:

Benenson et al.: Unpacking spear phishing susceptibility

Wash: How Experts Detect Phishing Scam Emails

Nthala & Wash: How Non-Experts Try to Detect Phishing Scam Emails

3) Please tone down the "user perspective" statements in this study, as it is not possible to know user perspective without actually ASKING the users. This study is quantitative, which is absolutely okay, but this means that the perspective or researchers (who define hypotheses) is considered, and not the perspective of users, which would need a more qualitative approach. This is also why the relevance ques "backfired": the backfired from the researchers' perspective, as the users were not asked how relevant the cue might be for them. Although users were asked to explain why they would or would not have been deceived by the emails, their answers are not considered in the study.

4) Generally, I find asking people whether they would have been deceived by an email very difficult, as an honest answer should always be "it depends". People just cannot objectively estimate their own susceptibility, and this should be reflected in the Threats to Validity. As a side note, I think this estimation might be possibly more connected to the individual differences on the Big Five scale or something similar than to the actual susceptibility.