[5660] Investigating the Effect of Phishing Sophistication on Phishing Reporting

PRELIMINARY DECISION: accept

Summary of Reviews

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

Reviews

Review 1

TOTAL SCORE: 1

Overall evaluation: 1 (weak accept)

Reviewer's confidence: 3 (medium)

===== Brief paper summary =====

The paper presents a controlled experiment on the detection and reporting of phishing emails based on 446 users recruited using Amazon Mechanical Turk. The paper defines a set of factors (Technical, Contextual, Language and Tone, and Layout) that can be combined to generate different phishing emails to determine their effects on detection and reporting rates.

===== Strengths =====

-The paper tries to get insights into the reporting of phishing emails. It is clear that the phishing email itself cannot motivate the reporting and we need to find new approaches to deal with that.

-On the other hand, it is interesting to see that the set of users that reported a phishing email is not a subset of the users that detected the email. This is an important point that should be taken into account to avoid overloading the IT staff.

-In general a good level of details on the implementation of the experiment and motivations.

===== Weaknesses =====

-Lack of information about how the users interacted with the phishing email provided in the controlled experiment. See the detailed comments.

-I found the explanation of the OR a bit confusing and inconsistent throughout the paper. See detailed comments.

===== Detailed comments for the author(s) =====

-Baseline email and controlled experiment: It is unclear how much the experiment allows the users to interact with the email to spot if it was a phishing email or not. This is critical in particular for the baseline email (TCLgLy). Depending on how the phishing email is provided there is (not) a way to determine if it is legitim or not. TCLgLy has an address legitim but is “spoofed”. The only *observable* difference from the original email is that the email contains a link. Is the email provided as a picture? In this case, there is no way to check if the email is legitim or not because one cannot identify that the email is spoofed (e.g. by looking at the failure of authentication) nor verify that the link points to a fake domain (an email with a link by default is not a phishing email). Potentially, this can explain why T has a high effect because it is not possible to discern if the email is phishing or not (at least for TCLgLy). If this is not the case, the paper should clearly state it and provide more information regarding e.g. the landing domain in the link and how the users were allowed to identify the spoofed email. I am not saying that all users will look at the DMARC field or the preview of the link to detect if this is a phishing email, but still, this should be allowed otherwise the experiment with the baseline email is biased.

-OR: I found confusing the presentation of the OR, in particular, due to the two ways in which the results are reported in Tab. 3 and 5a. When computing the OR one computes the ratio between the odds of reporting the outcome (e.g. 'Likely' in Q5) in the groups subject to the treatment (e.g. T=low) and the odds of the control (T=high). If the OR is > 1 it means that the treatment increases the odds of observing more 'Likely', < 1 vice versa. It is not clear why you have in Tab.3 OR>1 and in Tab.5 OR<1 given you're measuring always the same outcome ('Likely' credible/legitimate). This is probably because you swapped the treatment (i.e. from low to high and vice versa).

-Reporting and motivation: As you discussed, the results showing that the reporting rate is higher than in previous studies is probably due to the intention to report and not actual reporting. Indeed a higher report rate can be achieved by making the reporting mechanism easier ([A] is an example). I think that the context in which the experiment is carried out can play a role too. For example, one can be more worried about reporting a wrong phishing email to its IT staff than just reporting a wrong email to Amazon.

===== Minor comments for the author(s) =====

-Authors could be interested in [A] in which they also investigated reporting in an organization and reporting sustainability

-Section 5.1: "An odds ratio greater than one indicates that the odds decrease…"

[A] Daniele Lain, Kari Kostiainen, Srdjan Čapkun. Phishing in Organizations: Findings from a Large-Scale and Long-Term Study, IEEE S&P 2022

Review 2

TOTAL SCORE: -1

Overall evaluation: -1 (weak reject)

Reviewer's confidence: 3 (medium)

---- OVERALL EVALUATION ----

This paper investigates the relationship of phishing sophistication on reporting thereof by surveying subjects (n=446) after being presented with a hand-crafted phishing email differing on four factors, i.e., technical, contextual, language, layout.

Strengths

+ The paper describes its methodology and analyses very thoroughly.

+ Well-structured and readable paper.

Weaknesses

+ Awkward framing in title, abstract and introduction of the concept 'sophistication', whilst in reality the authors try to measure believability in relation to reporting.

+ Very limited ethics discussion, where one is needed given the study design.

Comments for the authors

The paper in its currents forms lacks a conceptual basis - i.e., the authors lay down there conceptual framework in the introduction, but confuse readers by mixing concepts (sophistication, believability, etc.). In essence the authors present the results of a vignette study without proper context or connecting it to the vast body of earlier work. Moreover, the questionnaire seems to be leading for participants - as I am assuming that participant didn't know they were part of an experiment. That is, first asking participants on their willingness to report, followed by asking questions about the legitimacy of the emails. If the authors would have tested this design, against a design that would reverse this order, they would have been able to elaborate on this choice.

My most prominent concern however is, the lack of an elaborate ethics discussion. Under what premise are participants recruited, do they know they are part of a study etc. Without such discussion and transparency it is hard to value the contribution the authors claim to make.

Review 3

TOTAL SCORE: 1

Overall evaluation: 1 (weak accept)

Reviewer's confidence: 3 (medium)

---- OVERALL EVALUATION ----

The author(s) conducted a controlled experiment to test whether the level of sophistication in phishing email affects detection and reporting behaviors. To do so, the author(s) developed eight experimental scenarios where four factors were varied to create variations in the level of sophistication. Based on current literature, the author(s) identified four factors that affect the believability of emails as technical (e.g., sender’s email address), contextual (e.g., alignment of the email content with the context of the receiver), language and tone (e.g., language and word choices that are appropriate with receiver), and layout (e.g., visual cues of emails such as logo). The main research question focuses on the relationship between believability of phishing emails with detection and reporting behaviors, which are further examined in four parts. These includes testing the relationships between: (a) variations in the four factors and believability of phishing e-mails, (b) believability and detection of phishing e-mails, (c) believability and report of phishing emails, and (d) detection and report of phishing emails. The author(s) recruited respondents from Amazon Mechanical Turk (MTurk). The final sample included 446 respondents out of 500 in the initial sample due to unintelligible responses, repeated respondents, or extremely short completion time. Respondents were randomly assigned into a scenario. The baseline phishing email was modified from an authentic email from Amazon MTurk. The results show that the technical factor played the largest role in the believability of phishing e-mails, detection, as well as reporting behaviors. Results also show the correlation between detection and reporting behavior to be more complex. For example, 18.3% of respondents would report an email that was perceived as legitimate while 47.6% of respondents who detected an email as phishing would not report. Overall, the article addresses a current knowledge gap and contribute to the debate on the issue of knowledge and behavior. The findings highlight that having knowledge or awareness does not necessarily translate to behavior, and this has implications on current effort in phishing prevention as asking employees to report a test phishing email may not be the most effective manner of assessment. The research is informative and insightful.

1) In Section 2 on "Sophistication of Phishing Attacks", the author(s) mentioned that persuasion techniques significantly correlated with victimization but have no impact on believability. Although that may be true for some cases, such statement requires more support and evidence as the cited literature (Ref. 15) was referring to social engineering and spear-phishing while the experiment resembled more general phishing campaigns. I would suggest referencing the article by Greene and colleagues as well (Ref. 16).

2) In Section 4.1, the author(s) did a great job in excluding questionnaires completed in an extremely short amount of time or repeated respondents to avoid possible non-human participants and double counting, respectively.

3) In Section 4.2, the author(s) describe a factor is high if it is present in the test email and low if absent. That is a binary approach to measurement whereas high and low would imply scale. The author(s) should consider changing the high/low description to others such as present/absent or experiment/base. Other than that, the author(s) provided detailed and easy-to-follow description on the developing process for experiment conditions.

4) In Section 4.3 the author(s) stated the inclusion of neutral responses as negative responses. The justification was later provided in Section 6.1 in the discussion on internal validity. I wonder if it is possible for the author(s) to compare rational from respondents with neutral responses versus those from respondents with the "unlikely" or "extremely unlikely" responses. This exploratory comparison could provide some justification, especially since it is not seen in existing literature.

5) In Section 4.6, the author(s) discussed the ethical aspect of the study and stated that respondents received monetary compensation. The author(s) should consider stating the actual amount of compensation provided to respondents.

6) In Section 5, the author(s) presented on findings on the different sub-research questions with great details and explanation. The inclusion of mathematical equations and graphs helped readers in understanding the results. One suggestion would be to move the discussion on rationale from Section 6 to Section 5.4 to highlight the complexities in detection and reporting behavior.

7) For Section 6.1, I appreciated the author(s) reflection on possible threats to validity to the study. One suggestion would to rename this section to "Limitations" and reframe some of these threats as possible directions for future research. For example, for external validity, the author(s) can encourage future scholars to replicate this research with various samples or other context (e.g., organizational context) to improve on the generalizability of the research design and findings.

Review 4

TOTAL SCORE: 1

Overall evaluation: 1 (weak accept)

Reviewer's confidence: 4 (high)

---- OVERALL EVALUATION ----

This paper presents a controlled experiment, designed as an online survey with MTurk workers, on the connection of plausibility of an email to its detection as phishing, and to its reporting rate.

Strengths:

- Potentially useful idea
- Careful study design and data analysis
- Detailed discussion of threats to validity

Weaknesses:

- The explanation of the term “reporting” might have been insufficient, as over 18% of people who thought the email was legitimate still wanted to report it
- Absence of qualitative analysis of reasons for reporting, and of the corresponding RQ, which makes study design deviate from RQs
- Missing related work

I like the idea of this paper. Furthermore, the paper is very well structured and written. RQs, statistical analysis and results are well explained. Nevertheless, my opinion is that the paper misses an important part of its potential.

# Explanation of reporting:

Fig.3: “Q1 Email users may choose to either report or not report for various reasons. How likely are you to report this email to Amazon?” On p. 4, the paper states: "Before answering the questionnaire, the subjects were informed what ‘reporting’ refers to in the context of phishing attacks.".

I think that the wording and layout of this explanation should be discussed in the paper. It seems to be important, as a sizeable portion of participants misunderstood this question. The paper states on p. 8: “The rationale given to report an email while considering it legitimate illustrates that reporting as a concept is often misunderstood by individuals.” This misunderstanding could be due to study design, not to the fact that people really misunderstand the concept. Therefore, this should be noted in "Threats to Validity".

# Absence of qualitative RQ and analysis:

I’m unable to understand why Q2 (reasons for reporting or not reporting) was asked. It does not correspond to any of the RQs in p. 2, which are purely quantitative. Either reasons are important – then the corresponding RQ should be asked. Or the reasons are not important – then Q2 is not needed.

Unfortunately, the participants in this paper are treated as a black box: everything that they did is explained using statistics. Yet, the real reasons for there are hiding in plain sight in Discussion. This paper needs an additional qualitative RQ and the corresponding qualitative analysis. This would very much enrich the findings, and provide good ground for further work on reporting.

# Missing related work

The most important missing paper is quite new, so maybe it was not known at the time of writing. It presents, among other things, a system for reporting phishing emails and its evaluation:

Phishing in Organizations: Findings from a Large-Scale and Long-Term Study
https://arxiv.org/abs/2112.07498

There are some other missing references, for example:

This paper considers, among other things, reasons for reporting and not reporting by experts:

Wash, R. (2020). How experts detect phishing scam emails. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1-28.

This paper considers interplay of persuasive techniques and content of phishing emails:

Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., & Ebner, N. C. (2019). Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. ACM Transactions on Computer-Human Interaction (TOCHI), 26(5), 1-28.

This paper reports factors of recognizing or not recognizing a phishing email which are quite similar to Table 1:

Benenson, Z., Gassmann, F., & Landwirth, R. (2017, April). Unpacking spear phishing susceptibility. In International conference on financial cryptography and data security workshops (pp. 610-627). Springer, Cham.

#Smaller remarks:

p. 1: “In this light, the sophistication of phishing emails can be characterized along three dimensions: realism, relevance and persuasiveness.” How were these three factors determined?

The emergence of four factors in Table 1 is also not clear form the text. Were they systematically derived from literature analysis? If not, this might mean that some additional factors are missing, which should be mentioned in Threats to Validity.

p. 4: “and layout (Lg)” – I guess it’s a typo: language, not layout?