Phish Finders: Analyzing Patterns in Email and Website Scams

(2024)

Thomas Baratta[1] *& William Murphy*[2]

Student[1]*, Software Engineering, Junior, Florida Gulf Coast University*
Student[2]*, Software Engineering, Senior, Florida Gulf Coast University*

Introduction

Amid rising cybersecurity threats, companies and organizations face the challenge of effectively safeguarding their systems from phishing attacks. With malicious emails and websites targeting employees, organizations need to provide their employees with cyber security training that adequately prepares them for real world attacks. To support organizations in their focus on employee training, our research utilizes data generated from a crowdsource initiative where volunteers tested their ability to effectively detect malicious emails and websites.

Phishing attacks represent a significant and evolving threat to organizations worldwide, posing substantial risks to private information, financial assets, etc. In recent years, the volume of these attacks have increased which imposes an urgent understanding of such attacks to develop countermeasures. This poses the question “how can organizations improve their cybersecurity training?” This paper addresses the imperative for organizations to enhance their defenses against phishing through cybersecurity training protocol. More specifically, it investigates patterns in email and website scams, using volunteer science to identify which platform (email or websites) the volunteers struggled with more. Furthermore, it explores the tendency of volunteers that overlooked certain cues with aim to inform recommendations for cybersecurity training strategies. Our findings indicate that the volunteers tend to overlook abstract ‘appeal to’ cue types along with a much higher chance of being victimized within website images. Given the deceptive nature of phishing threats and their potential ramifications, this study was constructed to equip employees with the knowledge and skills necessary to mitigate these risks effectively.

In our research a set of images was provided to crowd source volunteers comprising both malicious and trustworthy content. The volunteers were tasked with identifying whether the provided images were malicious or not. If the image was deemed malicious, they were then asked to identify the different cue types that led to their decision. Our data revealed that crowd volunteers did a relatively god job overall at identifying the maliciousness of an image but showed a high victimization rate within websites images indicating websites can be more deceptive than emails. The volunteers were able to identify cues they frequently encountered, such as "incorrect spelling and grammar," "invalid domain or sender," and "potential malicious link" phishing cues. Additionally, our data showed that cues that didn't occur as frequently were more likely to be overlooked by our volunteers, such as "appeal to Authority" and "appeal to Greed" phishing cues. This data lead us to recommend that companies and organizations should train their employees with more caution on these overlooked cue types along with websites in specific given their deceptive nature.

In the realm of cybersecurity, the ability to accurately identify and mitigate malicious content from a human perspective is crucial, however, advances in AI (artificial intelligence) have shown promising results that can aid humans in this process. A recent study, *"Phishing Identification: Citizen Science Volunteers vs. ChatGPT,"* conducted by anonymous authors, investigated the comparative effectiveness of human volunteers and AI in detecting malicious content. Their findings revealed that AI, exemplified by ChatGPT, yielded a lower victimization rate compared to humans but tended to struggle with 'appeal to' cue types. AI can excel in identifying concrete phishing cues, such as suspicious URLs and improper spelling, providing a robust first layer of protection. Conversely, humans are better equipped to detect more abstract phishing cues, such as nuanced language and context that might appeal to emotional or psychological triggers. By integrating AI's analytical strengths with human intuition and contextual understanding, a synergistic approach can be developed. This collaboration allows for a more comprehensive detection system, where AI offers a second perspective that enhances human vigilance and vice versa. While our research does not delve into AI, it focuses primarily on human detection of malicious content. However, based on the findings of *"Phishing Identification: Citizen Science Volunteers vs. ChatGPT,"* we believe integrating AI with human efforts can provide a valuable second layer of support in detecting malicious content.

Methods

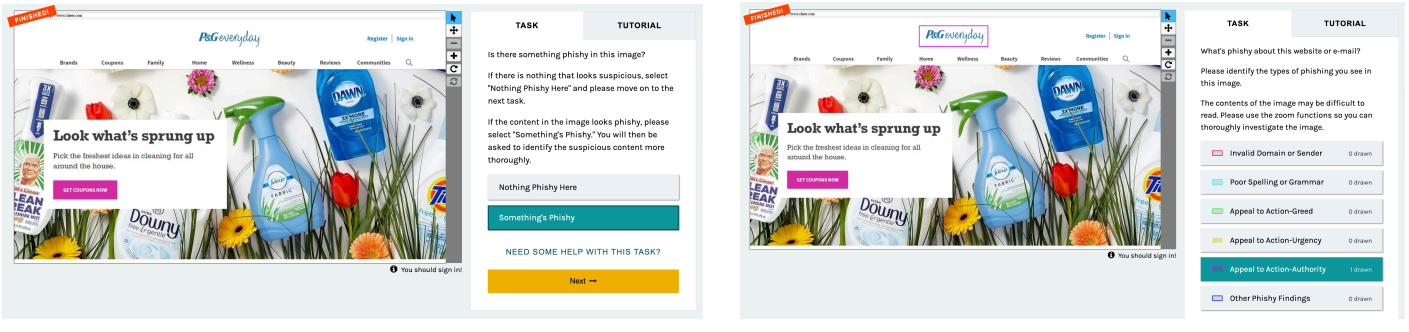
This study employed a data-driven approach, analyzing data from the Phish Finders project on the Zooniverse crowdsourcing platform. Phish Finders is a project that

utilizes open source volunteers to evaluate malicious and non-malicious images to gather data. The project's goal was to create a synthetic environment that accurately reflects real-world situations to assess and identify phishing indicators. Volunteers were shown images depicting both trustworthy and malicious content through a generic web browser or email interface, mirroring the scenarios individuals encounter in their daily online activities.

Within this platform, volunteers were asked to view and evaluate trustworthy and malicious images of email and website content, and to rate them as “phishy” or “not phishy”. If the volunteer were to label the image as malicious, they were asked to label the “phishy” part of the image with a bounding box. These bounding boxes were labeled with a variety of five different malicious cue types. These cue types included “Domain or Sender”, “Spelling or Grammar”, “Malicious Links”, and appeals used to extract information - “Appeal to Greed”, “Appeal to Authority”, “Appeal to Greed” (Refer to Figure 1). If the participants deemed an image malicious but weren’t satisfied with the five cues listed above, they were given the option to select “Other Phishy Findings” within their annotated bounding box.

The dataset given in this project was comprised of 30 “gold standard” images that were carefully evaluated by cybersecurity professionals along with over 1800 other images. The data includes a participant id, session time, a boolean variable to specify if the image was malicious, along with labels and bounding boxes drawn. Our approach focused on analyzing and filtering this data specifically from the subset of gold standard images. By concentrating on these meticulously curated images, we aimed to extract meaningful patterns within the data provided by the participants. Our objective was to identify and extract valuable data to inform and enhance employers' cybersecurity training programs. Through this targeted analysis, we sought to provide insights that could empower organizations to better prepare their employees in recognizing and mitigating phishing threats effectively.

Figure 1. **Left:** This shows the perspective of a Zooniverse volunteer generating an answer for the given image. **Right:**This shows the perspective of a Zooniverse volunteer labeling the image with a cue type bounding box.



Findings

Our findings were derived from three key aspects: evaluating the volunteers' performance in detecting malicious content, identifying which platform had a higher victimization rate between email and website images, and analyzing their performance with different cue types within malicious images. By breaking down the results in these areas, we gained a comprehensive understanding of the volunteers' abilities and the factors influencing their detection accuracy.

Volunteer Performance

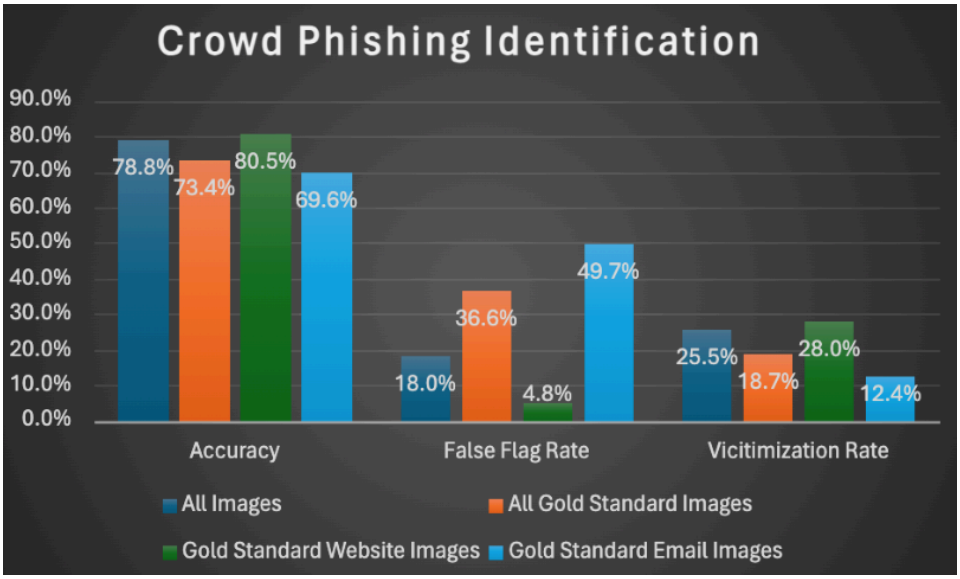
To measure the overall performance the volunteers had when evaluating malicious and non-malicious images, we broke down their results by measuring their accuracy rate, false flag rate and victimization rate. Their accuracy rate tells us how well they were able to identify whether the image was malicious or non malicious when asked “is there something phishy in this image?” (Refer to the top image in Figure 1). False flag rates refer to a “false alarm” where volunteers label an image as malicious when in reality the image is non malicious. Victimization rates tells us how frequently a volunteer has been “fooled” by a malicious image. This occurs when a volunteer labels an image as non-malicious when the image is malicious.

Table 1. Volunteer Performance Confusion Matrix: Identifying malicious content within an image

		User inputs “Something’s Phishy”	User inputs “Nothing Phishy Here”
Data	Malicious Image	Hit (True Positive)	Miss (False Negative)
	Non-Malicious Image	False Alarm (False Flag)	Correct Rejection (True Negative)

Within our data, we found that the volunteers performed relatively well in identifying the true nature of the images. Within the gold standard subset, 73.4% of volunteers were able to accurately determine the true nature of a given image compared to 78.8% across all images (Refer to Figure 2).

Figure 2. This graphs shows the accuracy, false flag and victimization rates across email and websites images.



Emails vs. Websites

When comparing the victimization rates between email images and website images, our findings revealed a notable difference. The victimization rate for website images was more than twice as high as that for email images, with website images having a victimization rate of 28.0% compared to 12.4% for email images within the gold standard subset. This substantial disparity suggests that websites pose a higher threat to users, indicating that users are more likely to be fooled by malicious content on websites.

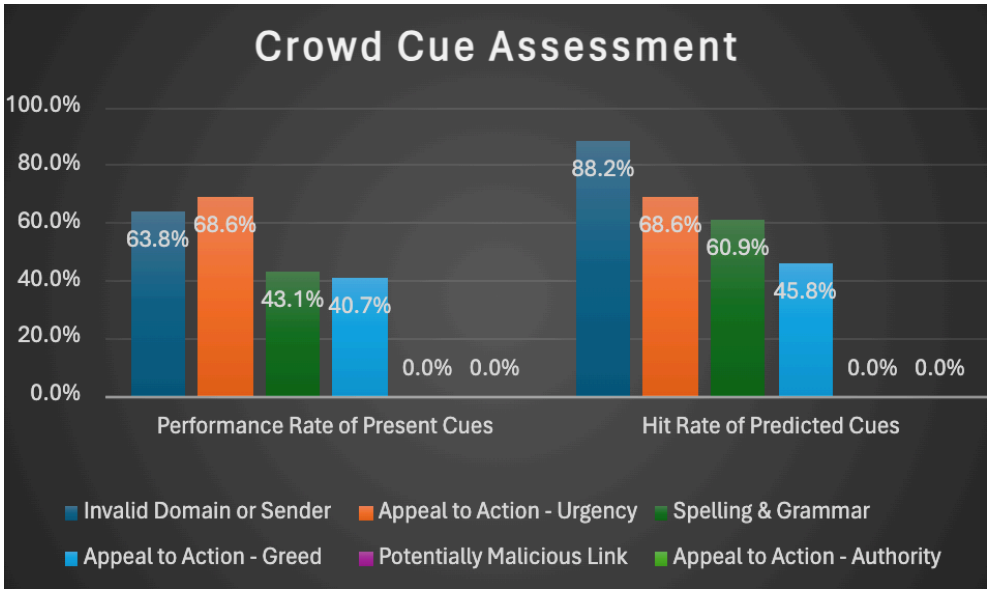
Moreover, the volunteers demonstrated a conservative approach when evaluating email images, and a very aggressive approach for website images. This led to a high false flag or "false alarm" rate for emails and a very low false flag rate for websites. The false flag rate for email images within the gold standard subset was 49.7% compared to a 4.8%

for website images, indicating that volunteers frequently labeled non-malicious email images as malicious while for website images, the volunteers were much more aggressive and labeled majority of the images as non-malicious. This high false flag rate highlights the volunteers' cautiousness and suggests that the perceived risk associated with email content may cause volunteers to be overly cautious in their evaluations. Conversely, the low false flag rate for websites indicates that users need to take more precaution with websites, as many more volunteers were fooled by malicious content within website images.

Cue Type Performance & Analysis

When analyzing the cue types in our data set, we examined how accurately the volunteers identified cue types within malicious content. On the left side of Figure 3, we analyzed the volunteers' performance in identifying malicious cue types present in images. For example, within the gold standard images where the 'invalid domain or sender' cue type was present, 63.8% of the volunteer responses correctly identified it. On the right side, we evaluated the volunteers' overall performance across all gold standard images. For example, across all cue types inputted by volunteers in the gold standard images, 60.9% of the 'spelling or grammar' cue type inputs were correctly identified.

Figure 3. This graphs shows the volunteers' performance at identifying cue types.



In Figure 3, you will notice a 0% value for both 'potentially malicious link' and 'appeal to authority.' Our dataset contained no instances of images with the 'potentially malicious link' cue type, resulting in a 0% value. Regarding the 'appeal to authority' cue type, this is a true zero. Shockingly, none of the volunteers in the gold standard subset correctly identified an 'appeal to authority' cue type.

Based on this graph, we can see that the volunteers had a very good hit rate for identifying the 'invalid domain or sender' cue type, with a rate of 88.2%. Additionally, the volunteers were effective at identifying the 'appeal to urgency' cue type within malicious content, with a rate of 68.6%. This suggests that volunteers were more successful in identifying the 'appeal to urgency' and 'invalid domain or sender' cue types, while they struggled significantly with the 'appeal to authority' cues, achieving a 0% success rate.

Discussion

The findings from our research provide significant insights into the efficacy of volunteers in identifying phishing cues in email and website images, with aim to provide valuable implications for cybersecurity training programs. Overall, the volunteers performed well in detecting malicious content, but they struggled with identifying specific cue types and discerning phishing attempts in website images.

One notable observation was the volunteers' difficulty with identifying the 'appeal to greed' cue type. The 'appeal to' cue types, such as 'appeal to greed' and 'appeal to authority,' are more abstract and subjective compared to other cues. These cues require deeper analysis and perspective thinking, which explains why they were often overlooked. This finding suggests that users need to develop a better understanding of the perspective required when evaluating emails and websites to recognize the more abstract cues frequently employed in malicious content. Previously, we mentioned the involvement of AI within detecting malicious content and its potential strength at detecting concrete cues along with its weakness at identifying abstract cues. We believe training employees to become better at detecting abstract cue along with using AI to detect those obvious and more concrete cues can provide a more efficient and effective solution to detecting malicious content. This training can be done by emphasizing a more creative and perspectival approach by training your employees to evaluate malicious abstract cue types similar to our volunteer experiment.

To our surprise, our research revealed that volunteers had more difficulty identifying phishing attempts on websites compared to emails. This was primarily due to the volunteers' conservative approach to email images, resulting in a high false flag rate for emails. While the exact reason for this conservative approach is unclear, we can

speculate that the deceptive user interfaces scammers create can make malicious websites harder to identify. Scammers can design websites to closely mimic legitimate company sites, making it easy to mistake a malicious website for a genuine one. These deceptive websites often contain 'appeal to authority' cues, further emphasizing the importance of understanding and recognizing these abstract cues. In contrast, emails are harder for scammers to convincingly impersonate since they lack the ability to create a misleading interface, reducing the potential for deception.

The findings suggest that improving cybersecurity training programs to address these challenges is crucial. Training should focus on enhancing users' ability to recognize abstract phishing cues, particularly in website contexts. By fostering a deeper understanding of the nuanced cues that scammers use and emphasizing the complementary roles of human intuition and technological tools, organizations can better prepare their employees to defend against phishing threats effectively.

Limitations and Future Work

There are several limitations within our study that need to be addressed to provide a comprehensive understanding of phishing detection and improve future research. Firstly, the sample size of our research is relatively small. We conducted part of this research using the gold standard subset of images, which contained a sample size of only 30 images. While these images were meticulously curated by cybersecurity professionals, a larger sample size would enhance the reliability and generalizability of our findings. Future studies should aim to include a more extensive and diverse dataset to validate and expand upon our results.

Another limitation is the skewed nature of our database. As noted in Figure 3, our dataset did not include any images with a ‘malicious link’ cue type. This absence could impact the comprehensiveness of our analysis, as malicious links are a common phishing tactic. Including a wider variety of cue types in future datasets will allow for a more thorough evaluation of volunteers' ability to detect different forms of phishing attempts.

Additionally, our study focused on static images rather than dynamic content. Real-world phishing attempts often involve interactive elements, such as clickable links and dynamic user interfaces, which can influence the detection process. Future research should consider incorporating dynamic content to more accurately reflect the complexity of real-world phishing scenarios.

Lastly, the integration of artificial intelligence (AI) and machine learning (ML) in phishing detection was not extensively covered in our study. While human judgment

remains crucial, leveraging AI and ML could significantly enhance detection accuracy and efficiency. Future research should investigate hybrid models that combine the strengths of human intuition with the analytical capabilities of AI, exploring how these systems can complement each other in identifying complex phishing cues. Look for "*Phishing Identification: Citizen Science Volunteers vs. ChatGPT"* in our references for more information on the capabilities of AI.

While our study provides valuable insights into the efficacy of volunteers in detecting phishing cues, addressing these limitations in future research will help to refine and improve cybersecurity training programs. By expanding sample sizes, diversifying datasets, incorporating dynamic content and exploring the collaborative effect with humans and artificial intelligence, future studies can build on our findings to develop more robust and effective strategies for combating phishing threats.

Conclusion

In summary, our research highlights the capabilities and challenges of using volunteers to detect phishing cues in email and website images. Despite their overall effectiveness in identifying malicious content, volunteers struggled with more abstract cue types specifically with 'appeal to' cues and found website images particularly challenging. These findings underscore the need for comprehensive cybersecurity training that addresses nuanced cues. Additionally, we mention the collaboration of AI to detect concrete cues and emphasize the importance of creative and perspectival thinking in phishing detection, particularly for cues that require deeper analysis and contextual understanding. By addressing the identified limitations and building on our findings, future research can enhance the effectiveness of cybersecurity measures and better prepare individuals to recognize and mitigate phishing threats.

References

Grover, A., Hale, M., Ahuja, V., & Rosser, H. (2021). Phish Finders. *Zooniverse*. (https://www.zooniverse.org/projects/holliekrosser/phish-finders)

Ahuja, V., Rosser, H., Grover, A., & Hale, M. (2022). Phish Finders: Improving Cybersecurity Training Tools Using Citizen Science. **ICIS*. (https://aisel.aisnet.org/icis2022/security/security/1)

Anonymous Authors. (2024). Phishing Identification: Citizen Science Volunteers vs ChatGPT. *In . ACM, New York, NY, USA, 15 pages*. (https://www.doi.org/XXXXXXX.XXXXXXX)

Chertoff, M. (2023, April 13). Cyber Risk Is Growing. Here's How Companies Can Keep Up. *Harvard Business Review*. (https://hbr.org/2023/04/cyber-risk-is-growing-heres-how-companies-can-keep-up)

Yampolskiy, A. (2023, February 15). What does 2024 have in store for the world of cybersecurity. *World Economic Forum*. (https://www.weforum.org/agenda/2024/02/what-does-2024-have-in-store-for-the-world-of-cybersecurity/)