PhishED: Automated contextual feedback for reported Phishing

Citation for published version:
<https://www.usenix.org/conference/soups2022/presentation/jenkins-poster>

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Symposium on Usable Privacy and Security

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
PhishED:
Automated contextual feedback for reported Phishing

Adam Jenkins  
University of Edinburgh

Nadin Kökciyan  
University of Edinburgh

Kami Vaniea  
University of Edinburgh

Abstract
When people receive malicious emails that claim to be from a trusted entity like their bank, they can be frightened and uncertain about how safe it is to ignore or delete the message. The uncertainty is hard on users and it may lead them to engage in unsafe actions like clicking on links "just to check." In this work we look at how to provide quick and accurate support to people who report phishing so that they can confidentially take appropriate action. For this, we will build a phishing-advice tool, PhishEd, that allows people to report malicious emails that they encounter and get automatically generated advice in response that is contextual to the suspicious email. The advice is meant to both help them make an informed decision about the reported email as well as provide some education to help them in handling future malicious emails better.

1 Introduction
Phishing is one of the most effective vectors for tricking users into providing sensitive information, such as account details, to attackers who then use the information to harm users, organizations, and society at large. Recent attacks, such as the Colonial Pipeline shutdown in the US can be attributed to phishing [16]. In the UK, 86% of businesses reported receiving phishing attacks [8] and while 76% of UK users believe that they can recognize and avoid suspicious links in emails [5], our own research shows that only about 8% can accurately read a URL [1]. The staggering £53.7 million lost in the UK in 2020 to impersonation fraud also attests to the scale of the phishing problem faced by business and users [9].

Employer-provided training tends to be delivered either up-front [19] or given after the employee has fallen for a mock phishing attack [14]. These approaches do work, but they are very employer-centric in that employers choose the time and content of the training. Delivering the same training to everyone also forces a compromise between being comprehensive and containing detailed actionable guidance [11]. For example, training commonly advises users to “look at the destination of links” [18] but does not have the space to explain how to do so. Users in this situation have virtually no agency or control and lack the ability to shape their own learning to their needs. Users also experience phishing as an infrequent unexpected event that is mixed in with other activities. Phishing by its nature is designed to appear urgent or threatening. A user receiving such a message may be worried or scared about what might happen if they were to ignore it. Currently their only options are to delete the potential phish (scary) or report it and ask for advice, which may take some time or even not come at all [2].

In this project, we propose a novel phishing-advice tool, PhishEd, which accepts reports from users of potential phishing emails, uses artificial intelligence (AI) to parse out contextual phishing features, and quickly responds back to the user with advice that uses the content of the reported email to explain the reasoning or frame the decision the user needs to make. For example, a reported phishing might contain “HMRC” (HM Revenue and Customs - UK Government department for collection of taxes) but have links leading to Dropbox. The auto response would inform the user that the email is not from HMRC, using technical features such as DomainKeys Identified Mail (DKIM) signatures [7], and highlight that the links lead instead to Dropbox, which HMRC would never use. It would also provide examples from the email to evidence these claims. The objectives of this project are: (i) help users confidently make safe security decisions, (ii) harness “teachable moments” to demonstrate how to detect fraud and relevant technical skills, and (iii) encourage
end-users to want to continue report phishing in the future.

2 Background

Contextual Phishing Feature

In this project, our focus is using AI to provide phishing advice while making the reasoning process transparent to the user. In other words, PhishEd leverages AI’s ability to extract and reason about contextual features of phishing in support of user decision making.

Example contextual phishing features that we identified in our previous work [2, 3] are as follows:

- **Contextual keywords** A phishing email includes a set of keywords that may reveal the context of the message. For example, if a reported email uses terms like “shutdown”, “email”, and “account”, the user may think that such a message was sent through an organization. Similarly, organization keywords like "PayPal", "HMRC", and "Dropbox" can indicate the implied organizational context of the email.

- **Uniform Resource Locators (URLs)** A phishing email is designed to trick users to click on URLs included in the email. Such URLs can be analyzed in conjunction to other contextual factors to provide contextual guidance. For example, if an email contains the "PayPal" keyword but the URLs do not lead to PayPal’s official website, that information can be provided to the user.

- **Email headers** include meta-information about the sender. For example, the From address can be checked against DKIM signatures or whitelists of organization domains.

  After examining the contextual features of a message, the phishing-advice tool will provide a content-based advice as a response to the user’s reported email. For example: “This email is not from the organization’s IT department which will only ever send email from the XXX@XXX.com email address.”, while providing further justifications about the sender’s identity, the content of the message and so on.

Embedded Training and Education

Education and training materials have long been considered as interventions for reducing the risks of phishing attacks, and hence have received interest from the academic community [12]. Training for end-users is available in range of formats and products. One of the most common and popular is that of embedded training, where simulated phishing attacks are conducted within the working contexts of user [14]. Although this form of training has been shown to be highly effective, implementing such training can take considerable effort to construct and monitor simulated phish and their results [2]. Additional interventions have been proposed such as the use of Serious Games [6, 21] or responsive solutions such as chat-bots [4].

Advice and training will routinely advise reporting suspicious emails [17], however research indicates that users may not do so due to a lack of confidence surrounding their abilities to spot legitimate phish [15]. The majority of phishing intervention research has focused on training users to identify phish, however this project intends to extend research by investigating the potential “teachable moment” once an suspicious email has been reported [10]. To our knowledge, the PhishEd project is the first investigation to focus solely on developing interventions for providing contextual phishing advice on demand.

3 Design: Template & System

**Nutrition Labels** We have taken initial inspiration from the work of Kelly et al. who developed “privacy nutrition labels”, essentially condensing privacy policies and their information into food nutrition label like format [13]. This approach was successfully adapted to the phishing domain by Althobaiti et al. [3] who developed phishing report interface to explain phishing features to humans, as depicted in Figure 3.

**Masters Project** A student Master thesis projects was conducted as part of this ongoing work, which focused on creating initial designs for the PhishEd system. Zhang used a user-centered approach to inform their designs, implementing a prototype, gaining user feedback and further iterating to create a final mock design [22] shown in Figure 4.

**NEAT & SPRUCE** We plan to improve on our template designs by integrating Microsoft’s security warning design principles, NEAT (Figure 1) and SPRUCE (Figure 2). This guidance was developed to be a part of Microsoft’s product teams’ development process, even providing wallet-sized cards and lessons on integrating the advice in practice [20]. We use this guidance to provide us with an initial design space for generating contextual advice templates for discussion in our planned focus groups (Section 4).

**Artificial Intelligence (AI)** AI has historically focused on the automatic detection of phishing emails using machine learning approaches, such as classification, that promise high accuracy. In the phishing context, it is not trivial to reach 100% accuracy due to the changing nature of the problem (e.g., attackers change their strategies constantly) and the high costs of false positives. Our focus is using AI to provide phishing advice while making the reasoning process transparent to the user. PhishEd leverages AI’s ability to extract and reason about contextual features of phishing in support of user
decision making. We will use Natural Language Processing techniques such as Named-Entity Recognition or Semantic Role Labelling to capture contextual features; and unsupervised machine learning approaches such as clustering to group emails together based on extracted features. AI approaches will aid in identifying ongoing phishing campaigns, and our campaign-based advice templates will be used to provide feedback to the user when they report potential phish.

4 Future Work: Iteration and Deployment

To achieve the goals of PhishEd, we will employ a range of methods to aid in the iterative design process, and the final evaluation of the designs.

Focus Groups & Design Workshops Once further ideation and design alternatives have been completed we will take
these initial designs and templates to end-users for user-driven iterative design. The project aims to identify the types and structure of phishing advice that users desire when reporting suspicious emails, by using the contextual information contained within the suspected phish. These studies will be conducted with members of our institution, including both staff and students from a range of departments and backgrounds to generate a representative participant user group.

**Lab Study** Upon completing our initial designs we will conduct a small lab-based user study using a mock email inbox and a set of tasks designed to encourage engagement with our reporting system. The lab study is focused on investigating the legitimacy of our early designs and prototypes, while also being used to identify and usability concerns and the suitability of our proposed solution within the working contexts of users.

**Longitudinal Study** We are currently working closely with an organization to incorporate our proposed system into their phishing reporting systems. We will develop an Outlook add-in to integrate PhishEd to Microsoft Outlook, where users will be able to report phish easily. This add-in will be deployed for the users who participate in our study. We will monitor the usage of the reporting tool and also measure the tool’s impact on the participants’ experiences when reporting and handling suspicious emails.

**Acknowledgments**

This work was funded in part by the UKRI Strategic Priorities Fund via the REPHRAIN Research Centre, and a Google Research Award. The authors would also like to thank Zeyu Zhang for his contributions to the project and appreciate the continual feedback of the TULiPS research group.

**References**

