Proposals for Holistic Security in Preventing Email Phishing Attacks

Brad Miller

Department of Electrical and Computer Engineering, University of Auckland

bmil852@aucklanduni.ac.nz

Abstract

In recent years, email phishing attacks have been a consistent, costly threat to businesses, users, and organizations. A number of approaches exist to mitigate this threat – none of which have been entirely successful in stopping the tidal wave of attacks. This paper analyses three of these approaches: authentication protocols, machine learning classifiers, and security warnings – comparing and contrasting the strengths and weaknesses of each approach first in isolation, and then holistically. As an outcome of our analysis, we make a number of recommendations for future work in each area, with the intention of better leveraging the function of each individual approach when considered as part of a larger system.

1.0 Introduction

Email phishing attacks involve fraudulent attempts to deceive users for malicious reasons, such as distributing malware or obtaining sensitive information. Despite best efforts to counteract these attacks, email phishing has reportedly been involved in the costly leakage of billions of user records in recent years [1].

This paper compares and contrasts three distinct approaches to securing against email phishing attacks. The first is the use of email authentication protocols which aim to prevent *email spoofing*, where an attacker impersonates

another entity in order to gain a victim's trust. The second is the use of client side warnings which aim to protect users from potential phishing emails through visual security indicators. The third is the use of machine learning techniques which aim to prevent phishing attacks by identifying emails that display particular sets of potentially malicious features.

While email providers typically use one or more of these (and other) approaches in conjunction to prevent phishing attacks, research in each area tends to consider the operation of a given approach in partial or complete isolation, and as a result the function of the security system as a whole can be neglected. By comparing and contrasting the strengths and weaknesses of three different approaches holistically, this paper hopes to shed light on how each method can complement and inform the others. As an outcome of the analysis, some potential suggestions are made for future research in each area, which focus on how the different approaches could function better in conjunction, rather than individually.

It should be noted that the three approaches we focus on here are not a complete set of those used to mitigate phishing attacks. To keep our discussion contained, we will consider other approaches (e.g. counteracting phishing websites, or training users to recognise malicious emails) out of scope.

2.0 Three Approaches

The research domain in the area of phishing is particularly large. To limit our scope, we focus particularly on the following three papers in our discussion:

- Hang Hu, and Gang Wang. End-to-End Measurements of Email Spoofing Attacks (2018). In *Proceedings of the 27th USENIX Security Symposium*.
    - We make reference to this paper to discuss **Email Authentication Protocols** and **Email Security Warnings**.

- Hang Hu, Peng Peng, and Gang Wang. (2017). Towards the adoption of anti-spoofing protocols. Available: https://arxiv.org/abs/1711.06654v3 Accessed 10 October 2018.
    - We make reference to this paper to discuss **Email Authentication Protocols.**

- Aviad Cohen, Nir Nissim, and Yuval Elovici (2018). Novel set of general descriptive features for enhanced detection of malicious emails using machine learning methods. In *Expert Systems With Applications*.
    - We make reference to this paper to discuss **Machine Learning Techniques for Detecting Malicious Emails.**

We consider these works to merit technical discussion that is suitable for the purposes of this paper , and to sufficiently illustrate the approaches they cover. Other publications may be cited where necessary to support certain ideas, but the major focus will remain on these three works.

The following sections (2.1, 2.2, and 2.3) briefly introduce each of the approaches in question, making note of the key strengths and weaknesses in each method in order to give the background information necessary for discussion on how the approaches can work better as a whole.

2.1 Email Authentication Protocols

Email Spoofing is a key part of phishing attacks, where a fraudulent entity poses as someone else to mislead victims. Unfortunately, the message transport protocol underlying email communication, Simple Mail Transfer Protocol (SMTP), provides no verification of sender. This means providers must implement extensions to SMTP, such as Sender Policy Framework (SPF), DomainKeys Identified Mail (DKIM), and Domain-based Message Authentication, Reporting and Conformance (DMARC) to authenticate emails.

In theory, such protocols can provide strong protection against spoofing. However, in practice, the adoption rate is low. In 2018, it was found that only 44.9% of the Alexa top 1 million domains have published a valid SPF record, and only 5.1% have a valid DMARC record [2]. This is slightly higher than in 2015 [3], but still low enough that the protection the protocols offer against spoofing is limited – very few email providers reject incoming email (or even place a warning on emails) if they have failed these protocol checks [2], making it easy for spoofed emails to enter the user inbox.

Multiple explanations have been offered for the low adoption of these protocols. One possibility is that technical limitations in the protocols prevent widespread adoption – the common email use cases of mail forwarding and mailing lists frequently cause SPF, DKIM, and DMARC failure [4]. Another potential problem is that these protocols lack the critical mass necessary to be effective – since adoption is low, there is no penalty to domains that do not publish an SPF/DKIM/DMARC record, as their emails will typically not be discriminated against [4]. Email administrators also perceive that the protocols are difficult to deploy, and do not offer a significant direct benefit compared to the cost of setting them up [4].

2.1 Email Security Warnings

Email security warnings are visual indicators that warn users of emails that may be forged or contain malicious content. These warnings intend to prevent or reduce the occurrence of phishing attacks by modifying user behaviour, advising caution when an email has failed its authentication requirements, or contains suspicious content (e.g. a blacklisted URL or known malware attachment).

In theory, placing warnings on potential phishing emails can be an alternative to outright filtering – email availability can remain high, and the final decision on which emails to interact with can be left up to the user. In practice, however,

many email providers do not place any warnings on potential phishing emails, such as those which fail their authentication requirements [2]. One reason for this might be for fear of 'crying wolf' – because many domains do not implement authentication protocols, emails which fail authentication requirements are often benign. Raising a warning on all of these emails could be counterproductive, as users tend to ignore warnings which are raised too often due to 'alarm fatigue' [5].

Some research also suggests that existing warnings may not be extremely effective. For example, in one study involving 488 participants, it was found that while a visual security indicator slightly reduced the chance an email user would click on a phishing URL, the effect the security cue had was statistically insignificant in a 'real-world' setting [2].

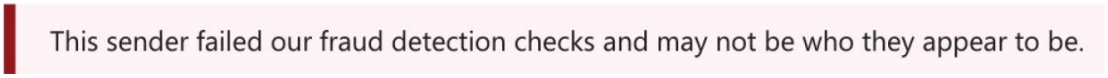

Figure 1: An example of a security warning in the hotmail.com email interface.

2.3 Machine Learning Techniques for Detecting Malicious Emails

Another technique used by email providers to mitigate phishing attacks is filtering suspicious emails using machine learning models. These approaches use a number of features extracted from the content, headers, and attachments of an email as input to a machine learning model. These models have proven to possess a capability for identifying emails which appear malicious, with recent studies touting a classifier with a true positive rate of 0.947, and false positive rate of 0.03 (AUC=0.929) [6].

A drawback of the machine learning approach is that the models it produces are typically difficult to verify. The actual performance of a machine learning classifier, and the true positive/false positive rates of detection it might provide are difficult to ensure, as the data used for training may or may not be

representative of malicious emails in the real world. Additionally, machine learning classifiers may become out of date over time as phishing emails evolve - though it should be mentioned academics in the area typically assure us that the models, and the features used to build them can be modified to account for this [6].

	Authentication Protocols	Security Warnings	ML Methods
Strengths	Strong theoretical capability to prevent spoofing	Can reduce the need for outright filtering, thereby improving email availability	Proven capability to detect malicious emails with a fairly high rate of accuracy
Weaknesses	Low adoption rate, Technical Limitations	Not provably effective, Overuse can result in warnings being ignored	Difficult to verify, Real world performance hard to ensure

Figure 2: Key Strengths and Weaknesses of Each Approach

Figure 2 shows a few key strengths and weaknesses of each approach, which we will refer back to in considering how the individual methods work in conjunction.

3.0 Proposals for Holistic Security

Here, ‘holistic’ security refers to considering a number of elements involved in securing against email phishing attacks as an integrated, interconnected whole. Each of the constituent parts we have mentioned make up an important section of a larger system intended to safeguard against email phishing attacks. By considering how these parts interrelate and work together, we intend to bring up some ideas for future work in each area.

A simplified representation of how the three approaches discussed currently interact is shown in figure 3 below.

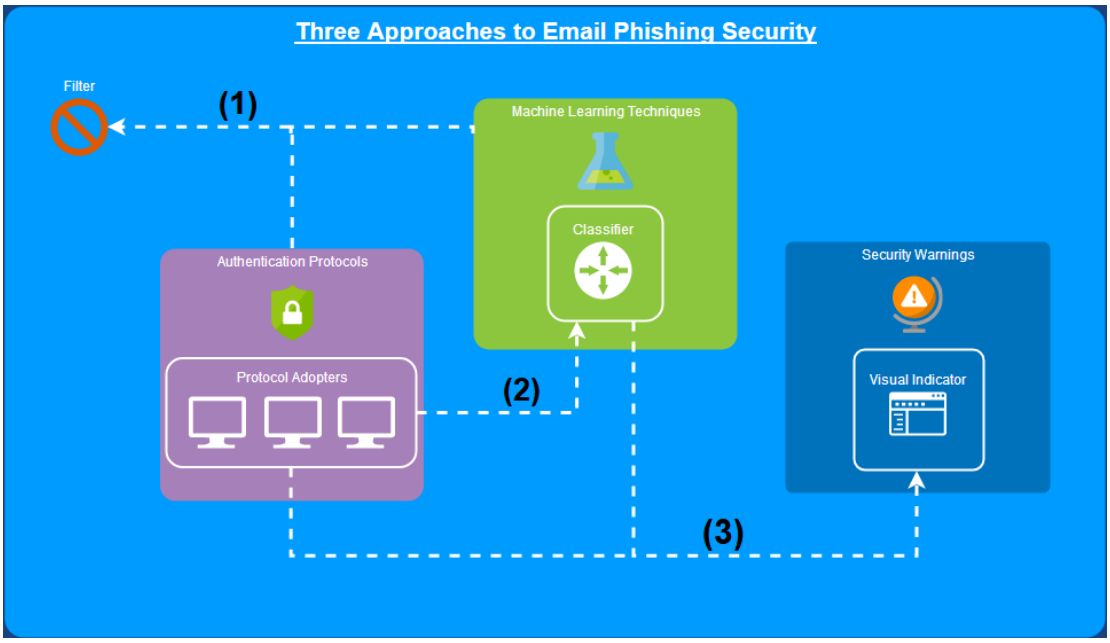

Figure 3: Simplified Representation of Approaches At Present

In the current integrated system, emails may be filtered out before ever reaching the user inbox, if they fail their authentication checks (though this does not always or even *usually* cause an email to be filtered [2]), or if they are detected as being potentially malicious by a machine learning classifier, represented by **(1)**. Some machine learning classifiers (though not all), use information from the authentication checks as one of the input features for their trained model [6], represented by **(2)**. Finally, emails which are not filtered out arrive in the user inbox, and the results of authentication checks and/or the machine learning model may inform the email client about whether or not to display a warning in the user interface (how the actual decision here is made varies wildly between email providers [2]). This is represented by **(3)**.

We now suggest potential areas for future work in the integrated system that better leverage the strengths and weaknesses of each approach when considered holistically. This theoretical system is illustrated in Figure 4.

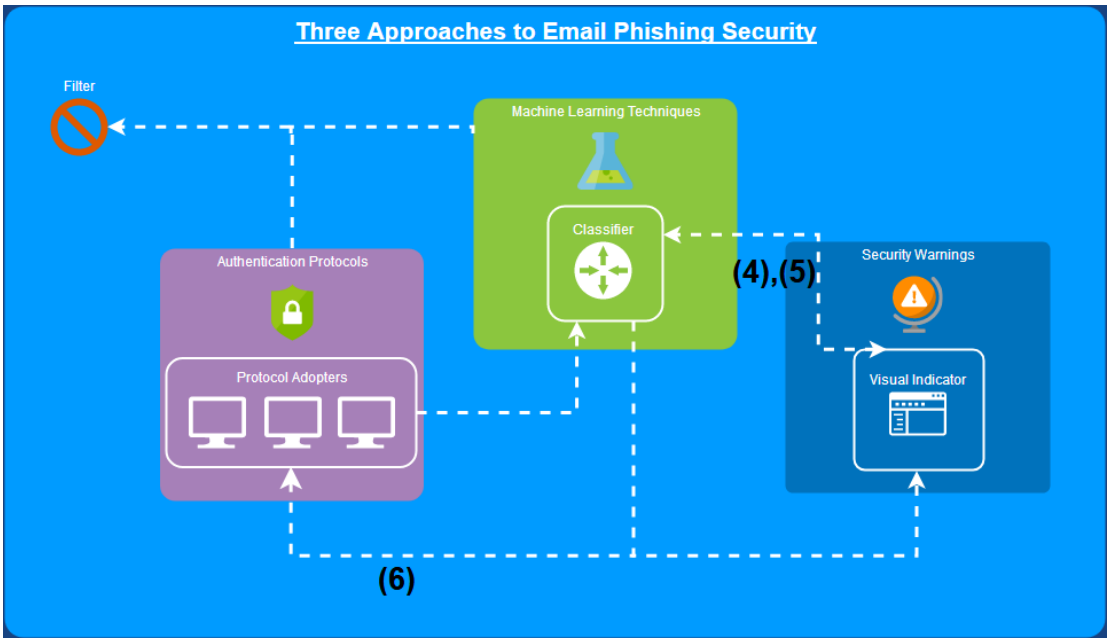

Figure 4: Theoretical Modifications to the Integrated System

The first potential area for consideration is a modification to the way the machine learning model communicates with the client interface when deciding whether or not to display a warning, represented by **(4)**. Machine learning classifiers used in phishing detection models (such as Random Forest, used in [6]), can typically provide an estimate of the certainty of their predictions. We suggest that this certainty estimate be passed to the client interface to allow further reasoning about whether or not to display a warning. When the classifier is highly certain that an email is malicious, the user interface could take the opportunity to display a more obvious warning message than usual, for example by using attractors, which can improve the effectiveness of security warnings [7] – *this could leverage the machine learning model to mitigate two of the major weaknesses in security warnings*: by reducing the risk of crying wolf due to only raising a serious alarm when more certain, and by potentially improving the effectiveness of the warnings themselves.

With this improvement to security warnings, an opportunity also arises for potential changes to how the classifiers themselves are optimized. This

opportunity is represented by **(5)**, and involves a possibility for classifiers to slightly relax their requirement for a very low false positive rate (FPR), allowing optimization to learn more towards a higher true positive rate (TPR). We do not suggest that the classifier optimization is changed completely to ignore FPR, but could instead aim for higher TPR, since the risk associated with a resulting higher FPR would now be mitigated by the reduced impact of 'crying wolf' associated with the suggested change **(4)**. *This could leverage security cues to mitigate one of the major weaknesses in machine learning models:* By reducing the impact of a classifier mistake (when making a false positive) an unverifiable machine learning model becomes slightly less risky to deploy. Security warnings - especially if made more effective - can as a kind of 'backstop' to these classifiers (like airbags in a self driving car!).

The final area for consideration, **(6)** is a possibility for improving user adoption of authentication protocols. This involves a suggestion for email interface designers to learn from adoption of HTTPS [8], and apply a consistent 'trusted' icon for emails with verified sender domains to contrast with security warnings. *This could leverage security cues to mitigate one of the major weaknesses in authentication protocols:* The low adoption rate may be improved by providing extra incentive for domains to implement SPF/DKIM/DMARC. If the adoption rate improves enough to provide a 'critical mass' for these protocols, the outcome might be especially effective. This suggestion should probably be considered 'softer' than the others, as it is based in part on analysis of the user study involving domain administrators conducted in [4], which included only 9 participants.

It should be noted that these suggestions differ slightly from standard ideas for future work in each respective area, which typically focus on how the individual approaches can be improved on their own. We do not consider our suggestions to contradict these ideas for future work, but instead believe that the

recommendations we make could be pursued in parallel, thereby improving protection against email phishing attacks both holistically and atomistically.

4.0 Limitations

The major limitation of this discussion is that we perform no detailed technical analysis of the feasibility of proposals made for each approach. For example, while making optimization changes to machine learning classifiers which relax FPR in favor of TPR is theoretically possible using the novel measure of 'Integrated Detection Rate' proposed in [6], what the actual application of such a change might entail is far outside the scope of this paper.

Additionally, this paper has covered only three approaches to securing against email attacks, which is not a complete set of those in use. A true integrated system includes other approaches, such as training users to recognize phishing emails, and the way these approaches function may partly affect the analysis we have just performed.

Finally, not all of the weaknesses identified in individual approaches were addressed by our suggestions. For example, we do not provide a mitigation to the technical limitations surrounding email forwarding and mailing lists in authentication protocols (which could be a large reason for their low adoption [4]).

5.0 Conclusion

In this paper, we have analysed three major approaches to mitigating email phishing attacks. In doing so, we have highlighted the individual strengths and weaknesses of different approaches, and discussed how when considered together, the methods involved can in some way complement the others. As an outcome of our analysis, we have made a few suggestions for future work in the different areas that may enable the approaches to perform better as a holistic unit.

References

[1] Vector, Inc., "2017 Data Breach Investigations Report," 2017. [Online]. Available: https://www.verizonenterprise.com/resources/reports/2017_dbir_en_xg.pdf. [Accessed 10 October 2018].

[2] H. Hu and G. Wang, "End-to-End Measurements of Email Spoofing Attacks," in *Proceedings of 27th USENIX Security Symposium*, Baltimore, MD, USA, 2018.

[3] I. D. Foster, J. Larson, M. Masich, A. C. Snoeren, S. Savage and K. Levchenko, "Security by Any Other Name: On the Effectiveness of Provider Based Email Security," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, Denver, Colorado, 2015.

[4] H. Hu, P. Peng and G. Wang, "Towards the adoption of anti-spoofing protocols," 2017. [Online]. Available: https://arxiv.org/abs/1711.06654v3 . [Accessed 10 October 2018].

[5] K. Krol, M. Moroz and M. A. Sasse, "Don't work. Can't work? Why it's time to rethink security warnings," in *7th International Conference on Risks and Security of Internet and Systems (CRiSIS)*, Cork, Ireland, 2012.

[6] A. Cohen, N. Nissim and Y. Elovici, "Novel set of general descriptive features for enhanced detection of malicious emails using machine learning methods," *Expert Systems With Applications,* vol. 110, pp. 143-169, 2018.

[7] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs and S. Schechter, "Your attention please: designing security-decision UIs to make genuine risks harder to ignore," in *Proceedings of the Ninth Symposium on Usable Privacy and Security*, Newcastle, United Kingdom, 2013.

[8] A. P. Felt, R. Barnes, A. King, C. Palmer, C. Bentzel and P. Tabriz, "Measuring HTTPS Adoption on the Web," in *26th USENIX Security Symposium*, Vancouver, British Columbia, 2017.