An Analysis of User Behaviors in Phishing eMail using Machine Learning Techniques

Yi Li[1], Kaiqi Xiong[1] and Xiangyang Li[2]
[1]*Intelligent Computer Networking and Security Lab, University of South Florida, Tampa, U.S.A.*
[2]*Johns Hopkins University, Baltimore, U.S.A.*

Keywords: User Behaviors, Phishing Email, Machine Learning, Amazon Mechanical Turk.

Abstract: Understanding user behaviors plays an important role in security situation assessments and computer system operations. There are very challenging and limited studies on email user behaviors. To study user behaviors related with phishing emails, we design and investigate an email test platform to understand how users behave differently when they read emails, some of which are phishing. We used a set of emails including phishing emails from the real world. We collect experimental data including participants' basic background information, time measurement, and their answers to survey questions. We first check whether or not factors such as intervention, phishing types, and incentive mechanisms play a major role in user behaviors when phishing attacks occur. We then evaluate the significance of each attribute with a performance score. The performance score is a metric demonstrating how a user makes a correct judgment on phishing while phishing attacks occur. We propose a machine learning framework, which contains attribute reduction and 10-fold cross-validation, to predict the performance of a user based on our collected data.

1 INTRODUCTION

Phishing is an online identity theft, which is disguised in an email and other messages to deceive victims into providing their login credentials and personal information (Gupta et al., 2016). It is prevalent today since current growing Internet techniques heavily involve the sensitive information of users. Therefore, more and more personal computers and mobile device users are exposed to phishing attacks. Many researchers have studied phishing attack problems where many solutions have been proposed to detect phishing attacks at different levels (Chin et al., 2018). However, there are only a very few studies on understanding how users' behavior can contribute to susceptibility to phishing. By understanding users' behaviors on phishing attacks, we can determine how to educate users so that they can be prevented from phishing attacks better.

Because of the inhomogeneity of users' network security education levels, users are susceptible to phishing attacks at different degrees (Goel et al., 2017). Although security and usability experts claim that a computer system should not rely on users' behavior, researchers have found that phishing attack are directly correlated with user behavior factors (Williams and Li, 2017). Thus, an important security prevention method is to educate users to adapt better security behaviors, where user behavior education refers to teaching Internet users about phishing awareness and defense techniques. Education-based approaches usually offer online information or educational games (Rakhra and Kaur, 2018).

In this research, we aim at studying user behavior factors, such as intervention, phishing type, and a monetary incentive, to understand how a user behaves during phishing email attacks and what mechanism may prevent a user from being a victim of such attacks. Here, intervention is defined as a mechanism that helps users be aware of the phishing attacks more easily by modifying phishing types to make them appear more obvious (Yang et al., 2015). A monetary incentive is introduced to motivate users to pay attention to phishing attacks (Brase, 2009).

The goal of this study is first to understand how user behaviors are correlated to phishing through an analysis of the collected experimental data and then to develop a model to predict how likely a user will be a victim based on the user's profile and behaviors. For this purpose, we explore to answer the following challenging questions in this research:

1. How intervention can affect user behaviors?

Li, Y., Xiong, K. and Li, X.
An Analysis of User Behaviors in Phishing eMail using Machine Learning Techniques.
DOI: 10.5220/0008119805290534
In *Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019)*, pages 529-534
ISBN: 978-989-758-378-0

2. Which phishing type is more harmful than others?
3. How can a monetary incentive affect a user's behavior and sorting?
4. How accurately can we predict the performance of a user on email sorting based on user profiles and behaviors?

To answer the above questions, we propose two study designs, on-site study design and online study design. In both study designs, participants are asked to conduct a pre-setup experiment on our testbed, where each participant first read a number of emails and then sort them into either "phishing" or "non-phishing." We introduce a performance score to record the total number of the correctnesses of a participant's sorting.

The rest of the paper is organized as follows. In the section 2, we present related work on phishing emails. In the section 3, we introduce the designs of our two studies. In the section 4, we present our machine learning framework. We evaluate the results of our studies in the section 5. Finally, we conclude our findings in the section 6.

2 RELATED WORK

As phishing becomes a more and more popular attack vector, email has been the most common way to conduct phishing attacks (Pande and Voditel, 2017). Vishwanath et al. (Vishwanath et al., 2011) discovered that most phishing emails are peripherally processed and the decisions made by individuals are usually based on simple clues embedded in an email message. They also found that if the email contains urgent information, the user will typically ignore other clues that could potentially help detect the deception.

Vishwanath et al. (Vishwanath et al., 2016) later conducted an experiment to examine the factors for phishing susceptibility and they found that an individual email habit was an important factor for phishing susceptibility. They found that those people with entrenched email habits tended to be more susceptible to phishing attacks. This is due to their habits that as soon as a notification arrives, they are likely to open it though they do not realize that they are opening it.

Interventions can be utilized for better understanding user behavior in phishing susceptibility when existing studies also have consequently focused on training individuals to better detect fraudulent emails (Burns et al., 2013). Liang et al. (Liang and Xue, 2010) demonstrated the effectiveness of warning interfaces with two groups, one control group that had no warnings for phishing attack, and another group that had warnings. They recruit nine participants in total, where eight of them are fell for the attack. After experiments, most of the participants claim that they did not notice the warning and some don't even know what it means. Further, many of the participants admit that they don't know the meaning of phishing.

Many existing studies have shown that people are vulnerable to phishing for the following reasons (Vishwanath et al., 2011). Many users do not trust security indicators on the websites (Wu et al., 2006). Attackers can easily replicate legitimate websites since people usually judge a website by how a website looks and how they feel about it (Harrison et al., 2016). Although some users are aware of phishing, the information does not contribute to detect or prevent phishing attacks (Parsons et al., 2016). Nowadays, machine learning techniques have been applied to detect the phishing emails (Smadi et al., 2015).

Muthal et al. (Muthal et al., 2017) has recently studied on user behaviors in phishing attacks with incentive and intervention. They conducted a three-round experiment where participants distinguish phishing emails from normal emails. In our study, we follow closely from their experience but do more analysis. We not only study how user behaviors will affect phishing attack outcomes but also try to predict how users will perform based on their behaviors and background.

3 THE STUDY DESIGN

In our study, we use emails obtained from the real world with some necessary modifications for personal information protection. Phishing emails were derived from a semi-random sample of emails in "Phish Bowl" database (Database, 2018). Normal emails were derived from legitimate emails received by the research team.

3.1 Phishing Types

One purpose of this research is to study which type of phishing attacks is more malicious to user. There are three types of phishing attacks used in our study:
1. A Suspicious Sender's Email Address: Nowadays, people are flooded with emails and tend to pay less attention of the sender's email address. They usually only look at the sender's name, neglect of the email address, or just catch a glimpse of the email address. The scammer has a high chance to replicate the email address. For example, the letter 'l' and the number '1' are very similar. Therefore, the scammer could

utilize this feature to create a fake 'we11sfargo' domain name rather than 'wellsfargo.'

2. Suspicious Links or Attachments: The links could be manipulated through using similar characters or misspelling issues. For example, a link contains the word 'directdeposit' could be misspelled as 'direct-depost.' The suspicious attachments can be disguised as the pdf file, exe file, or other types of files.

3. Malicious Email Contents: This type of phishing is quite tricky. At first glance, the email content seems normal to most of people. However, this kind of phishing attacks contains suspicious contents. For example, the contents may have several grammar issues or the icon of popular social networks is faked. They are very hard to notice if the user is not familiar with those popular social medias or if the user is not a native English speaker.

3.2 On-site Study Design

In this study design, its email sorting task consists of three rounds and each round is preloaded with 20 emails. There are 5 legitimate emails and 3 different phishing types, where each type includes 5 emails. In the second round, the intervention is introduced to the participants based on their performance of the first round. We recruit 40 participants to perform this task. During the experiment, the participants are asked to differentiate the phishing emails from legitimate emails. After the tasks in three rounds, participants are required to take a survey in the lab, where their backgrounds are asked.

3.2.1 Participant Recruitment

The IRB had been approved before we started to recruit participants (The approval number is: Pro00026240.) In the on-site study, participants are students as we recruited them on campus. We have recruited 40 participants at our university to perform this user study. The average age of the participants is about 23 years old while the participants' ages range from 18 to 38. Among 40 participants, 18 are female and 22 are male. The distribution of the participants is shown in Table 1.

Table 1: Participants Basic Information Distribution.

Gender		*Age*	
Female	45%	18∼ 20	25%
Male	55%	20∼30	67.5%
		30∼38	7.5%
Education			
Undergraduate	65%	Ph.D.	20%
Graduate	10%	Faculty	5%

We introduce a monetary incentive in our study. It is designed to answer the third question in section 1. We want to study whether the monetary incentive will affect a user's performance or not.

3.2.2 Survey

The survey was carried out after three rounds of email sorting tasks. We used an online survey platform to record the answers from participants. This survey contains 30 questions and is mainly about the background of participants, such as, age and some general questions about their experience and habits of using social medias. There were also some questions related to the email sorting tasks they just took. The example of the survey questions is shown as follows: For instances, "Have you taken any cybersecurity courses?" and "I briefly looked at the sender/source of the emails." are two examples of the survey questions. This survey can better help us understand user behaviors regarding to phishing attacks.

3.3 Online Study Design

Although by performing the on-site study, we can sufficiently answer the first three questions mentioned in the section 1, it is not sufficient for us to thoroughly understand user behaviors regarding phishing attacks. This is due to the limitation of demographic diversity and the number of participants recruited, etc. Therefore, we propose the online study design. It can sufficiently help us to answer the question of how accurately can we predict the performance based on user behavior.

3.3.1 Participant Recruitment

Participants are recruited from Amazon Mechanical Turk (MTurk) (MTurk, 2018). We recruited 90 participants in total for this online study. The distribution of the participants is shown in Table 2.

Table 2: Participants Basic Information Distribution.

Gender		*Age*	
Female	39%	20∼ 35	65.6%
Male	61%	36∼50	26.7%
		51∼61	7.7%
Highest Education			
High School	13.3%	Master	6.7%
College	77.8%	Doctorate	2.2%

We keep the monetary incentive mechanism in the online study since it is an useful feature/attribute for predicting the performance of user behavior regarding phishing attacks.

4 MACHINE LEARNING FRAMEWORK

The goal of using a machine learning method is to predict a user's performance regarding phishing attacks, whether or not the user can do well or poorly. Hence, we divided the performance score into two different classes, *Good* and *Poor*, based on the average score. We apply machine learning approaches to predict the overall performance (performance score). We build 4 different machine learning models, Decision Tree-J48, Naive Bayes (NB), Support Vector Machine (SVM), and Multilayer Perceptron (MP). Since we have 119 attributes but only 90 datasets, we first perform a stepwise attribute reduction. To select the best attributes for the ablove machine learning models, we first perform a Pearson-correlation coefficient analysis to observe the importance of each single attribute. We then fit our data into a linear regression model to evaluate all the attributes.

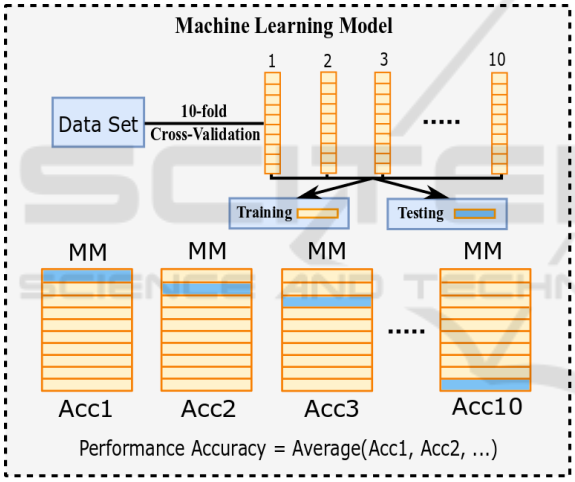

Figure 1: Machine learning model with 10 fold cross-validation.

We also propose to use the method of 10 fold cross-validation to precisely predict the performance of each participant, as shown in Figure 1, where MM is short for the machine learning model. The original dataset is randomly partitioned into 10 equal size subdatasets. Of the 10 subdatasets, a single subdataset is retained as the validation data for testing the model and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (i.e., 10 folds), with each of the 10 subdatasets used exactly once as the validation data. The 10 results from the folds can then be averaged to produce a single estimate. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. Besides the 10 fold cross-validation, we also utilize the similar idea of cross-validation for the attributes. Suppose we have m attributes and each time we randomly select n attributes to feed into our cross-validation machine learning model. This process will be running in k times, where $m = k \times n$.

We have m attributes in total after attribute selection. Then we randomly choose n attributes to do the cross-validation training by applying machine learning model. The next step is to calculate the performance accuracy. This process can be running k times. These k performance accuracies are averaged to form one final accuracy. We use 10-fold cross-validation to do the training and validation. Each time we will obtain an accuracy, and this will be done 10 times. The average accuracy is calculated by averaging all the accuracies.

5 EVALUATION

In this section, we analyze and identify what factors may make a significant impact on a phishing attack outcome and the evaluation of our machine learning models will be presented. The evaluation of intervention, phishing types, and a monetary incentive are using the dataset from the on-site study, while the evaluation of performance prediction is using the dataset from the online study.

Table 3: Email Round Score and Time.

Attributes (Mean)	*Round1*	*Round2*	*Round3*	*R2-R1*	*R3-R2*
Phish_Score	10.18	11.6	10.02	1.42	-0.15
Total_Score	14.2	15.23	14.25	1.025	0.05
Phish_Time(s)	437.58	413.45	433.25	-24.13	19.8
Total_Time(s)	630.88	600.4	568.35	-30.48	-32.05

5.1 Intervention Evaluation

To answer the first question in the section 1, we calculate mean phishing score, mean total score, mean total processing time and mean phishing processing time. The result is shown in Table 3. The intervention is introduced in the second round based on the performance of the participant from the first round. The full score of phishing score is 15 and the total score is 20. As shown in the table 3, the second round has been slightly improved compare with the first round. The mean time used in the second round is also lesser than the first round. However, in the third round, the performance score has decreased and even worse compared with the first round.

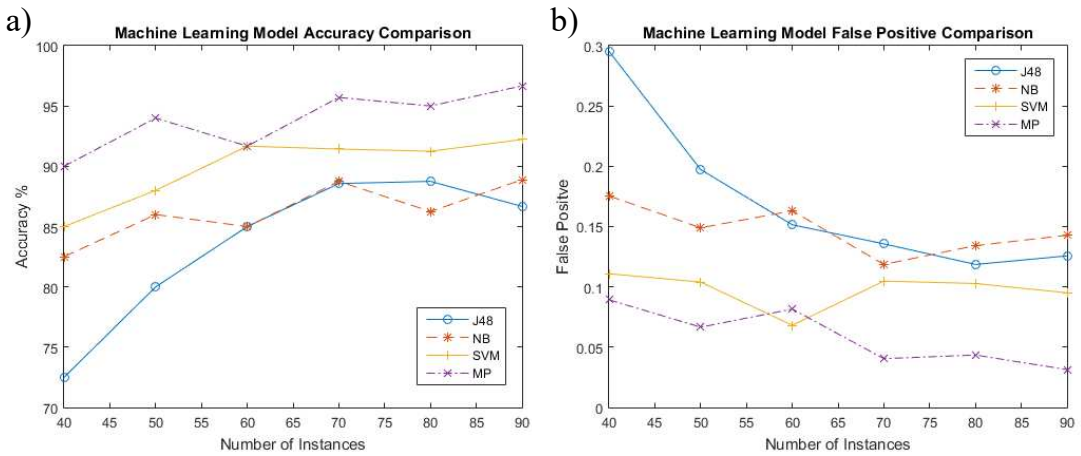

Figure 2: (a) Accuracy of each machine learning model with different number of instances. (b) Evaluation of false positive rate for different machine learning models.

5.2 Phishing Type Evaluation

We analyze the performance score and time for different types of phishing attacks. The second question in section 1 can be answered in Table 4. Type 1 phishing attack contains a suspicious sender's email address, type 2 phishing attack has suspicious links or attachments, and type 3 phishing attack contains malicious contents. The mean score (full score is 15) and mean time are calculated by taking average of all 40 participants' score and time of different phishing types. The intervention frequency describes the total times of a certain type phishing intervention introduced in the task. We can see from the table that type 1 phishing has the lowest score and it has been used the most as an intervention. This implies that the type 1 phishing is more harmful compared to the other two types. In addition, it is not hard to see that the score is in inversely proportion to intervention frequency. Thus, intervention is a suitable attribute that can be used in our machine learning models.

Table 4: Different Types of Phishing Score and Time.

Phishing Type	*Mean Score*	*Mean time(s)*	*Frequency*
Type 1	9.5	447.425	17
Type 2	11.35	431.875	8
Type 3	10.95	404.975	15

5.3 Monetary Incentive

The next question is whether a monetary incentive affects the performance and total processing time. We calculate the mean total performance score, mean phishing performance score, mean total processing time, and mean phishing processing time of all 40 participants. The result is shown in Table 5. Control means that there is no monetary incentive and Incentive represents that this group will get a monetary incentive. We can see from the table that the group with incentive has a higher performance score than the group who doesn't. Furthermore, the incentive group tends to spend more time than the control group. Therefore, incentive is also a useful attribute regarding a phishing outcome.

Table 5: Monetary Incentive Analysis.

Condition	*Phish_Score*	*Total_Score*	*Phish_Time*	*Total_time*
Control	30.1	42.65	1148.95	1580.5
Incentive	33.5	44.7	1419.6	2018.75

5.4 Performance Prediction Evaluation

After applying stepwise attribute selection, we chose 16 attributes in our machine learning framework. Now, we evaluate the performance accuracies of our machine learning framework. The accuracy in the following analysis is the final accuracy by averaging 10 folds accuracies. Figure 2 (a) shows the relationship between accuracy and number of instances, which means the number of participants because we treat each participant as an instance. We can see as the number of instances increases, the accuracy is also increasing. Among them, Multilayer Perceptron has the best accuracy, which is 93.84% in average. When using all 90 instances, the accuracy reaches 96.67% for Multilayer Perceptron. SVM has the second best accuracy; the average accuracy for SVM is 89.93%. In addition, when using all 90 instances, it has the best accuracy, which is 92.22%. The average accuracies for Naive Bayes and J48 are 86.23% and 83.58%, respectively.

6 CONCLUSIONS AND FUTURE WORK

We have studied how users behave when they encounter phishing email attacks. To the best of our knowledge, this is the first comprehensive and quantitative investigation of how users react in email checking and reading that have become an integral part of our daily life. We have designed two studies, on-site study and online study. We have applied statistical methods to analyze our on-site dataset and explore the answers to the questions on how intervention, phishing types, and a monetary incentive affect user behaviors when phishing attacks are encountered. Our analysis have showed that participants with intervention and a monetary incentive perform better than other cases. Phishing type 1, suspicious senders' email addresses, tends to be more harmful to users compared to other two phishing types. We have further developed machine learning techniques with the 10-fold cross-validation to analyze the data collected in the online study. We have analyzed the best attributes used in our machine learning framework. By choosing 16 attributes, we have achieved the user performance prediction accuracies of 86.67%, 88.89%, 92.22%, and 96.67% for J48, Naive Bayes, SVM, and Multilayer Perceptron, respectively.

In daily-life scenarios, we tend to deal with many other things while checking our emails; thus, we plan to investigate a multitasking experiment platform to understand how multitasking will affect the behavior of a user accordingly.

ACKNOWLEDGEMENT

We would like to thank NSF for partially sponsoring the work under grants #1620868, #1620871, #1620862, and #1651280. We also thank the JHU team that provided the data used in this project.

REFERENCES

Brase, G. L. (2009). How different types of participant payments alter task performance. *Judgment and Decision Making*, page 419.

Burns, M. B., Durcikova, A., and Jenkins, J. L. (2013). What kind of interventions can help users from falling for phishing attempts: a research proposal for examining stage-appropriate interventions. In *HICSS*.

Chin, T. J., Xiong, K., and Hu, C. (2018). Phishlimiter: A phishing detection and mitigation approach using software-defined networking. In *IEEE Access*.

Database, P. B. (Accessed Sept. 2018). [online]. Available: https://it.cornell.edu/phish-bowl.

Goel, S., Williams, K., and Dincelli, E. (2017). Got phished? Internet security and human vulnerability. *Journal of the Association for Information Systems*, 18(1):22.

Gupta, S., Singhal, A., and Kapoor, A. (2016). A literature survey on social engineering attacks: Phishing attack. In *ICCCA*, pages 537–540.

Harrison, B., Svetieva, E., and Vishwanath, A. (2016). Individual processing of phishing emails: How attention and elaboration protect against phishing. *Online Information Review*, 40(2):265–281.

Liang, H. and Xue, Y. (2010). Understanding security behaviors in personal computer usage: A threat avoidance perspective. *JAIS*.

MTurk, A. M. T. W. (Accessed Sept, 2018). *[Online]. Available: https://www.mturk.com/mturk/welcome*.

Muthal, S., Li, S., Huang, Y., Li, X., Dahbura, A., Bos, N., and Molinaro, K. (2017). A phishing study of user behavior with incentive and informed intervention. In *Proceedings of the National Cyber Summit*.

Pande, D. N. and Voditel, P. S. (2017). Spear phishing: Diagnosing attack paradigm. In *WiSPNET*, pages 2720–2724. IEEE.

Parsons, K., Butavicius, M., Pattinson, M., Calic, D., Mccormac, A., and Jerram, C. (2016). Do users focus on the correct cues to differentiate between phishing and genuine emails? *arXiv preprint:1605.04717*.

Rakhra, M. and Kaur, D. (2018). Studying user's computer security behaviour in developing an effective antiphishing educational framework. In *ICISC*. IEEE.

Smadi, S., Aslam, N., Zhang, L., Alasem, R., and Hossain, M. (2015). Detection of phishing emails using data mining algorithms. In *SKIMA*, pages 1–8. IEEE.

Vishwanath, A., Harrison, B., and Ng, Y. J. (2016). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research*.

Vishwanath, A., Herath, T., Chen, R., Wang, J., and Rao, H. R. (2011). Why do people get phished? testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, 51(3).

Williams, N. and Li, S. (2017). Simulating human detection of phishing websites: An investigation into the applicability of the act-r cognitive behaviour architecture model. In *CYBCONF*, pages 1–8. IEEE.

Wu, M., Miller, R. C., and Garfinkel, S. L. (2006). Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM.

Yang, W., Chen, J., Xiong, A., Proctor, R. W., and Li, N. (2015). Effectiveness of a phishing warning in field settings. In *Proceedings of the Symposium and Boot-camp on the Science of Security*, page 14. ACM.