What makes phishing emails hard for humans to detect?

Kuldeep Singh[1], Palvi Aggarwal[1], Prashanth Rajivan[2] and Cleotilde Gonzalez[1]
[1] Dynamic Decision Making Laboratory, Carnegie Mellon University, Pittsburgh, USA
[2] Department of Industrial and Systems Engineering, University of Washington, Seattle, WA

This research investigates the email features that make a phishing email difficult to detect by humans. We use an existing data set of phishing and ham emails and expand that data set by collecting annotations of the features that make the emails phishing. Using the new, annotated data set, we perform cluster analyses to identify the categories of emails and their attributes. We then analyze the accuracy of detection in each category. Our results indicate that the *similarity* of the features of phishing emails to benign emails, play a critical role in the accuracy of detection. The phishing emails that are most similar to ham emails had the lowest accuracy while the phishing emails that were most dissimilar to the ham emails were detected more accurately. Regression models reveal the contribution of phishing email's features to phishing detection accuracy. We discuss the implications of these results to theory and practice.

INTRODUCTION

Phishing attacks continue to be a significant threat to cybersecurity, despite the fact that cyberdefense technologies advance at an increasing rate (*Human Factor Report*, 2019). One of the main explanations for this conundrum is that humans continue to be an important vulnerability, and research to address our understanding of human detection decisions of phishing emails has not advanced as rapidly as it is required (Gonzalez et al., 2014).

Phishing is one of the most successful forms of deception and persuasion in the cyber world, because it takes advantage of social engineering and psychological techniques that exploit *human* weaknesses (Jagatic et al., 2007; Rajivan & Gonzalez, 2018). These human weaknesses include: our almost inevitable tendency to rely on our own memory and experience to make decisions (Gonzalez et al., 2003, 2014), our limited and often biased attention towards items that are "attention catching" (Kumaraguru et al., 2007), and our tendency to believe that things that look similar have similar effects (Tversky, 1972, 1977). These cognitive human factors result in human cognitive biases, which unfortunately, attackers seem to master quite well (Rajivan & Gonzalez, 2018). Since the success of phishing attacks rely on the exploitation of end-user's cognitive and psychological weaknesses, it becomes essential to understand the detection capabilities, decision making, and cognitive biases of end users who respond to phishing emails (Canfield et al., 2016).

Human factors and behavioral cybersecurity researchers have devoted a good amount of research to determine ways to train end-users to detect phishing emails (Kumaraguru et al., 2007; Singh et al., 2019; Jensen et al., 2017). While some of the phishing training solutions have been reasonably successful in increasing people's awareness of phishing attacks (Egelman et al., 2008), this type of research is mostly reactive to the superficial effects of the attacker's intentions. The underlying source of the problem is the lack of understanding of the human weaknesses and how those drive the human decisions to "click" on the wrong email. The current research contributes to expanding our understanding of the human detection decisions of phishing emails, investigating the attributes of emails that are characteristic on phishing behavior.

This research relies on initial findings from (Rajivan & Gonzalez, 2018), who investigated the strategies that adversaries use in phishing attacks. In the current study we use the emails generated by (Rajivan & Gonzalez, 2018), both phishing and benign (i.e. ham) emails, to investigate the relationship between the features of those emails and identification accuracy by the end-user and to explore the categories these emails form according to those features.

Our research also builds on recent findings derived from the creation of a cognitive model for email phishing detection (Cranford et al., 2019). The cognitive model relies on Instance-Based Learning Theory (IBLT) (Gonzalez et al., 2003), a theory of dynamic decisions from experience. According to IBLT, decisions are made by generalizing across past experiences, or instances, that are *similar* to the current situation. Typically, instances represent the features of the decision, the action taken, and the outcome of that decision. The cognitive model (Cranford et al., 2019) makes it clear that the similarities of the attributes (slot values in the instance) can make some memories more or less easy to retrieve. However, determining what features humans rely on to make phishing decisions and the role of the similarity among those features, is necessary to improve any computational model of phishing detection.

Indeed, similarity has been a major determinant of decision making more generally (Tversky, 1977), findings suggest that options that are similar will lead to lower decision accuracy, compared to decisions among options that are more dissimilar (Tversky, 1972). In fact, cyber attackers use this strategy in typosquatting, for example. They

take advantage of the similarity between words in URL names, so that fake websites go unnoticed. Typosquatting includes domain typos that allow attackers to access a list of different emails, receiver name typo, and misspellings URLs (e.g. including "eboy.com" instead of "ebay.com") in the URL address to misdirect end-users on malicious link (Szurdi & Christin, 2017). Such techniques make the phishing emails look similar to end-user's regular emails and they tend to match the features of any email with their experience features.

In summary, the research above suggests that the features of the phishing emails and the similarity between the features of the phishing and ham emails will be a key predictor of detection accuracy.

METHOD

Phishing Detection Dataset

To analyze the influence of email level features on end-user phishing detection, we leveraged a dataset from past psychological experiments on phishing decision making (Singh et al., 2019; Rajivan & Gonzalez, 2018). These experiments were designed to understand the cognitive mechanism crucial to human learning and decision making in the context of phishing attacks. The dataset from these experiments contained responses from 818 participants. Each participant in this dataset responded to 60 unique emails, making a binary decision on whether the email presented to them was phishing or not. A response was considered accurate if the participant correctly detected a phishing email, and incorrect if the participant classified a phishing email as a ham email. The 60 emails presented to each participant were randomly drawn from a larger corpus of 239 unique emails that included 186 phishing emails and 53 ham emails.

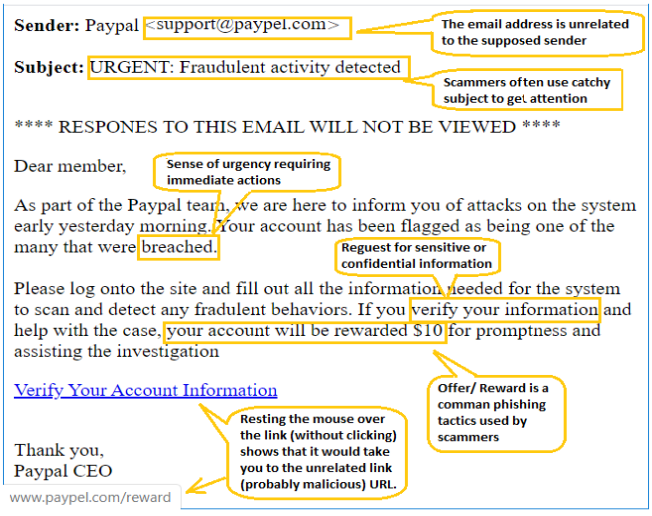

Figure 1: Phishing email features

Depending on the experiment condition, some of the participants received a higher proportion of phishing emails. Please refer to these publications for additional details about these experiment and the distribution of phishing emails between conditions (Singh et al., 2019; Rajivan & Gonzalez, 2018).

The dataset from these experiments contained, in total, 48,080 unique (participant-email pair) responses, which includes 19,598 responses to phishing emails, 28,482 responses to ham emails. Each row in this dataset consists of a participant's response to a given email (phishing or ham). For each response, we had access to the original text of the email presented to the participant, the ground truth (i.e., whether the email was indeed a phishing or a ham email), accuracy of the participant response, and six email level features indicating whether that email text contained attributes usually associated with phishing emails, such as a suspicious link. Figure 1 shows an example email that is tagged with the six features indicative of phishing attack.

Table 1: Email Features

Feature	Description
Sender mismatch	The sender mismatch feature is present if the email sender identity is different from what it pretends. It may be because of spoofed display of the name or domain or because of misspelled words in an email address
Request credentials	The request credentials feature is present if there is a request for personal and sensitive or confidential information in the email
Urgent	The Urgent feature is present if the content of the email creates a sense of urgency, fear, or threat which is a common technique in phishing emails
Offer	Offering a prize, reward, or help is a common phishing tactic. This feature is present if this type of offer is included the email text
Suspicious subject	A subject line of an email can be suspicious if the subject line depicts urgency, fear, threat, offer or reward
Link mismatch	A mismatch between the content of the email and the actual link may indicates this is a phishing email. Also, phishers may use an IP address instead of a URL to requests personal information. If this kind of link exist in an email is said to have a link mismatch

Feature Classification and inter-rater reliability Three researchers with expertise in cybersecurity and phishing attacks, independently classified all 239 emails

in the dataset based on six predefined categories. These categories represent email level features that are usually evident in a phishing email (Kumaraguru et al., 2009). See Table 1 for description of the six email level features used to classify the emails.

After classifying the emails independently, we used Cronbach's alpha (Cronbach, 1951), to measure inter-rater reliability. The alpha value on the classifications between the three raters was 0.88 representing high consistency. Further, for the purposes of analysis, it was essential to arrive at a consensus on a final classification for each feature, for each email. So, disagreements in classifications were discussed and resolved to reach agreement on all the classifications. This process resulted in a classification of 0/1 for each of the six features, for each email. A '0' classification for a feature indicated that the email did not have the feature according to raters. For example, ham emails would not have a sender mismatch or a suspicious link whereas a phishing email would include one or more of these features in it.

RESULTS

Accuracy Distribution

When a participant correctly classified a phishing email as phishing (Hit), the response was accurate, otherwise it was inaccurate (Miss). We calculated the mean detection accuracy per email averaging across all human responses to that email.

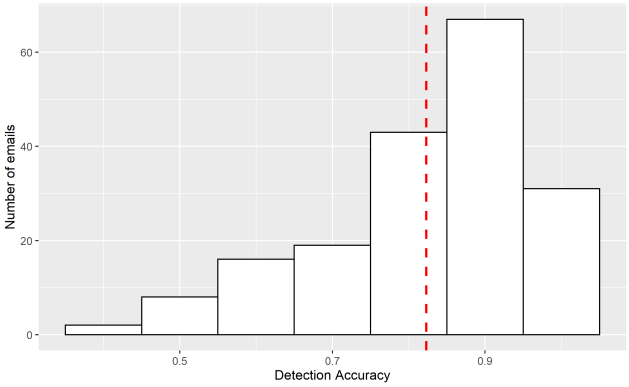

Figure 2: Distribution of Detection accuracy for phishing emails

The distribution of phishing emails (N=186) shown in Figure 2, suggests a negative skew with a mean of 0.82 and median 0.85. Thus, there is a good number of emails with low accuracy, that make it interesting to discover what are the features that make the emails more or less difficult to detect by humans.

k-means Clustering

We used the k-means clustering method (MacQueen et al., 1967) to classify 239 emails (186 phishing and 53 ham) into seven (7) clusters (the optimal number according to the Gap Statistic (Tibshirani et al., 2001)), using the six email feature coding (Table 1). The k-means clustering technique used dimension reduction to find principal components to visualize the data into cluster. Figure 3 show the the resulting two principal components along with their respective correlations, variables within component and their respective weights. The two principal components are arranged in a decreasing order of variance (i.e. $Var(PA1) \geq Var(PA2)$). Similarly, the variables are arranged in decreasing order of correlation within each principal component.The first principal component encompasses five features (i.e. urgent, request credentials, link mismatch, suspicious subject and sender mismatch). The second component encompasses one of the six variables (i.e. offer as shown in Figure 3). The first components ($PA1$) accounted for 41.5% of the variance in given data and the second component ($PA2$) accounted for 21.4% of the variance in the given data. The weights can be characterized as correlations between the variables and the component.

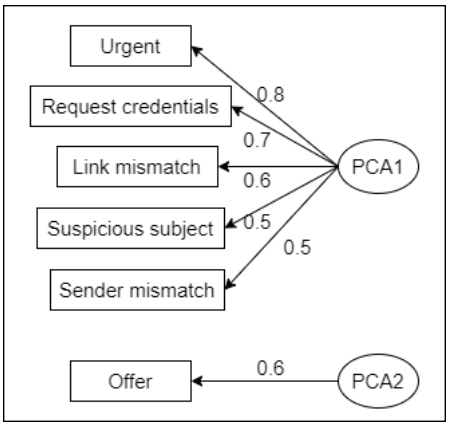

Figure 3: Principal Components and their contributing features

The results of the clustering analysis are shown in Figure 4 and the two dimensions in the plot represent two principal components explained above. Each email belongs to the cluster with the nearest cluster center.

Exploring the emails in the 7 clusters, we observe that cluster4 contains 54 emails, and 53 of them are ham emails (only one phishing email was part of this cluster). Thus, the k-mean clustering clearly separated these emails as belonging to the same category according to the email features. We will refer to cluster4 as the *ham* cluster and the smallest cluster (cluster7) have 15 emails and the largest cluster (cluster1) have 57 emails. The other clusters2,cluster3, cluster5, and cluster6 have cluster size of 38,23,18, and 34 emails respectively. The six clusters resulting from the k-means analyses were composed of all phishing emails.

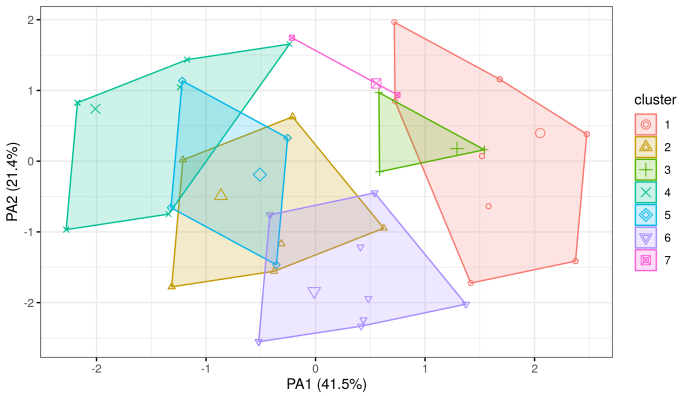

Figure 4: Email Clusters classified by k-means clustering

Phishing Accuracy per Cluster

In order to test the similarity hypothesis on detection accuracy, we calculated the distance between each cluster's center to the ham cluster's center. Figure 5 presents the accuracy of each phishing email, in each of the 6 phishing emails clusters, sorted according to their distance to the ham email cluster. The red line represents the mean accuracy of the cluster.

Figure 5: Detection accuracy of phishing emails in different clusters where each column represent the a cluster and the column title represent cluster distance from ham cluster.

The similarity effect is readily obvious: The emails in the clusters that are closer to the ham email cluster had lower accuracy, compared to the clusters that were farther away from the ham email cluster. Our interpretation is that the overlapping features of the phishing emails to ham emails made the phishing emails more difficult to detect. For example, cluster2 is the nearest ($distance = 1.06$) to the ham cluster and it also has the lowest detection accuracy (72%); while cluster1 is the farthest ($distance = 1.96$) from the ham cluster and it has the highest detection accuracy(90%).

Regression Analysis

A logistic regression model with Accuracy as the dependent variable and the six email features as independent variables helped determine the effects of the email features on the accuracy of detection in phishing emails. The regression model include Mturk id and email id as an error term to indicate the margin of error in the model:
$\text{Accuracy} \sim SenderMismatch + RequestCredentials + SubjectSuspicious + Urgent + Offer + LinkMismatch + (1|MturkId) + (1|emailId)$

This regression model was run with the overall data set (Overall), Cluster 1 (the cluster with the largest distance to ham emails), and Cluster 2 (the cluster with the shortest distance to the ham emails). Table 2 shows the results. Subject suspicious and urgency features were always present in the Cluster 1 data and sender mismatch, request credentials and link mismatch features were always present in the Cluster 2 date (i.e., the regression dropped these variables).

The Overall model suggest that request credentials, subject suspicious, urgent, offer and link mismatch are the most predictive features of accuracy. The odds ratio for the overall model in Table 2 suggest that the presence of request credentials, subject suspicious, urgent, offer and link mismatch increase the the chance of detecting phishing by 1.92, 1.40, 1.95, 2.20 and 1.80 times respectively compared to when these features are not present. The subject suspicious feature does not have a significant effect on detection accuracy.

The Cluster 1 model suggest that only offer is the significant predictive feature of accuracy in this cluster. With the presence of offer, the chance of detecting phishing increase by 3.74 times compare to when offer is unavailable. The Cluster 2 model also suggest that only presence of offer significantly increase the chance of detecting phishing by 2.39 times compare to its absence.

DISCUSSION

Our results demonstrate that the features of phishing emails and their similarity to the ham emails are reliable predictors of the human detection accuracy of a phishing email. The more similar phishing emails are to ham emails the lower the accuracy of detection is. This result demonstrate the similarity effect in a phishing applied decision context (Tversky, 1977, 1972; Gonzalez et al., 2003).

Our cluster and regression results show that phishing features such as request credentials, urgency, and offer in emails are the most predictive of detection accuracy. Whereas, other features (e.g., sender mismatch) are less predictive of detection accuracy. Furthermore, the presence of an offer (i.e., offer of a prize, a reward, or help) was a common predictive feature of accurate detection across clusters and overall.

Table 2: Regression Table

	Overall			Cluster 1			Cluster 2		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	1.16	0.62 – 2.16	0.641	9.17	1.67 – 50.30	**0.011**	1.93	1.35 – 2.76	**<0.001**
sender_mismatch	1.26	0.96 – 1.66	0.100	1.46	0.92 – 2.29	0.105			
request_credentials	1.92	1.39 – 2.63	**<0.001**	1.01	0.50 – 2.04	0.977			
subject_suspicious	1.40	1.01 – 1.94	**0.046**				0.91	0.36 – 2.30	0.837
urgent	1.95	1.38 – 2.77	**<0.001**				1.64	0.93 – 2.87	0.086
offer	2.20	1.56 – 3.09	**<0.001**	3.74	1.84 – 7.59	**<0.001**	2.39	1.35 – 4.24	**0.003**
Link_Mismatch	1.80	1.03 – 3.13	**0.038**	1.08	0.22 – 5.29	0.922			
Observations	19598			6073			3915		
Marginal R^2 / Conditional R^2	0.076 / 0.362			0.047 / 0.356			0.048 / 0.297		

These findings have important theoretical implications: we can improve current cognitive models of phishing detection (Cranford et al., 2019) to include the important features and to measure the similarity of phishing to ham emails. These results also have important practical implications: since humans are aware of some features (i.e., "catchy" offers, the urgency tone, and the request of credentials) and they are less aware of others (i.e., sender mismatch, the analyses of the subject in an email, and the detection of link mismatch), new training feedback should concentrate on these features to improve the accuracy of phishing email detection. The kmeans clustering algorithm used for calculating the cluster is generally used for non-categorical data, and have some limitation on categorical data. In future, we will explore other clustering methods to improve the classification of emails. We will also consider text similarity and features of the text such as authoritativeness, fear, misspelling, emotions etc., to increase the prediction accuracy of our models.

ACKNOWLEDGEMENTS

This research was sponsored by the Army Research Office and accomplished under MURI Grant Number W911NF-17-1-0370.

REFERENCES

Canfield, C. I., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human factors*, *58*(8), 1158–1172.

Cranford, E. A., Lebiere, C., Rajivan, P., Aggarwal, P., & Gonzalez, C. (2019). Modeling cognitive dynamics in end-user response to phishing emails.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297–334.

Egelman, S., Cranor, L. F., & Hong, J. (2008). You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1065–1074).

Gonzalez, C., Ben-Asher, N., Oltramari, A., & Lebiere, C. (2014). Cognition and technology. In *Cyber defense and situational awareness* (pp. 93–117). Springer.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Human factor report. (2019, Nov). Retrieved from `https://www.proofpoint.com/us/resources/threat-reports/human-factor`

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*(10), 94–100.

Jensen, M. L., Dinger, M., Wright, R. T., & Thatcher, J. B. (2017). Training to mitigate phishing attacks using mindfulness techniques. *Journal of Management Information Systems*, *34*(2), 597–626.

Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M. A., & Pham, T. (2009). School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th symposium on usable privacy and security* (p. 3).

Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L. F., & Hong, J. (2007). Getting users to pay attention to anti-phishing education: evaluation of retention and transfer. In *Proceedings of the anti-phishing working groups 2nd annual ecrime researchers summit* (pp. 70–81).

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

Rajivan, P., & Gonzalez, C. (2018). Creative persuasion: A study on adversarial behaviors and strategies in phishing attacks. *Frontiers in psychology*, *9*, 135.

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez, C. (2019). Training to detect phishing emails: Effects of the frequency of experienced phishing emails. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 63, pp. 453–457).

Szurdi, J., & Christin, N. (2017). Email typosquatting. In *Proceedings of the 2017 internet measurement conference* (pp. 419–431).

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological review*, *79*(4), 281.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.