Using Signal Detection Theory to Measure Phishing Detection Ability and Behavior

Casey Canfield
Carnegie Mellon University
Engineering and Public Policy
Pittsburgh, PA 15213
caseycan@cmu.edu

Baruch Fischhoff
Carnegie Mellon University
Social and Decision Science & EPP
Pittsburgh, PA 15213
baruch@cmu.edu

Alex Davis
Carnegie Mellon University
Engineering and Public Policy
Pittsburgh, PA 15213
alexdavis@cmu.edu

1. INTRODUCTION

Phishing attacks are a pervasive problem for individuals, corporations and critical infrastructure. They are among the top ten vectors for cyber threats, particularly for cyber-espionage [1]. Given the difficulty of screening for phishing emails, system operators rely, in part, on human judgment to limit vulnerability. However, at present, human behavior is typically considered the weakest part of cybersecurity strategies [2].

Existing models of phishing susceptibility focus on the role of attention to cues such as the URL and sender [3]. Here, we extend those models to consider the vigilance of individuals targeted by such attacks, using signal detection theory (SDT). SDT characterizes users' performance in terms of their *discrimination ability* (d'), for differentiating between legitimate and phishing emails, and their *decision threshold* (c)*,* for treating an uncertain email as phishing [4]. In applying it to phishing, we build on work by Kumaraguru et al. [5].

The present study demonstrates a method for estimating individual users' discrimination ability and decision thresholds as inputs to future analyses. It illustrates the method by assessing the impact of individual- and task-level variables identified by behavioral research, looking separately at how users decide whether an email is legitimate and what behavior follows those detection decisions.

2. METHOD

2.1 Participants

We recruited 152 participants from U.S. Amazon Mechanical Turk (mTurk) users. Participants were paid $5 and took 50 minutes to finish on average. According to self-reports, 58% were female and 45% had at least a Bachelor's degree. The mean age was 30 years old, with a range of 19 to 59. All owned a computer and 66% used Gmail as their preferred email client. The most common browser was Chrome (63%) and the most common operating system Windows 7 (48%).

2.2 Procedure

Participants (1) received phishing training, (2) evaluated 40 emails as phishing or not, and (3) answered individual difference items – in that order. The training was a comic strip designed by Kumaraguru et al. [5]. For the email task, participants examined 20 legitimate and 20 phishing emails, on behalf of Kelly Harmon, an employee at the fictional Soma Corporation, about whom they received a brief description. All stimuli were designed to mimic the Gmail format.

Phishing emails were adapted from public archives and descriptions in news articles. Legitimate emails were adapted from personal emails and general descriptions. The order of the emails was randomized for each participant. For each email, participants completed two tasks: (1) a behavior task where participants selected an action for the email and (2) a detection task where participants indicated if the email was phishing or not. The order of the behavior and detection tasks was randomly assigned. In addition, we assessed individual difference measures including confidence, perceived consequences, computer knowledge, computer behavior, computer incidents, time spent on the task, and demographics.

3. RESULTS AND DISCUSSION

3.1 Signal Detection Theory Parameters

Both signals (phishing emails) and noise (legitimate emails) can be represented as distributions of stimuli that vary in terms of the decision variable (here, their suspiciousness). Discrimination ability (*d'*) measures the distance between the signal and noise distributions. As *d'* increases, the distributions are further apart and signal and noise are perceived as more distinct. The decision threshold, *c*, is measured in terms of distance from where the distributions intersect. The point of intersection is a decision threshold of 0, indicating no bias toward identifying stimuli as signals or noise. In the context of phishing, a more negative decision threshold indicates a tendency to call uncertain stimuli phishing. We will call negative decision thresholds "cautious" (recognizing that trying to avoid rejecting legitimate emails is also a form of caution).

Performance was more accurate and more cautious on the behavior task, compared to the detection task. For the detection task, the average discrimination ability (d'_D) was 0.96 (*SD*=0.64). The average decision threshold (c_D) was 0.32 (*SD*=0.46), indicating that participants had to be suspicious before treating a message as phishing. These parameters are equivalent to a hit rate of 56% and a false alarm rate of 21%. In the context of this experiment, this represents approximately 8 successful phishing attacks and 4 false alarms per person.

For the behavior task, participants demonstrated high perceptual sensitivity (d'_B= 1.41, *SD*=0.71), responding differently for phishing and legitimate messages. Their actions tended to show little bias (c_B=-0.03, *SD*=0.64). This is equivalent to a hit rate of 77% and a false alarm rate of 25%. In the context of this experiment, this represents approximately 4 successful phishing attacks and 5 false alarms per person. In a paired t-test, the behavior task showed higher discrimination ability ($d'_D = 0.96$ vs. $d'_B = 1.41$, $t(151) = 7.23$, $p < .001$, $d = 0.66$) and lower decision threshold ($c_D = 0.32$ vs. $c_B = -0.03$, $t(151) = 7.81$, $p < .001$, $d = 0.61$).

Figure 1a shows participants’ behavioral responses based on whether they judged a message to be phishing or legitimate in the detection task. For example, almost all said that they would (appropriately) “click link or open attachment” or “reply” for messages that they perceived as legitimate and that they would “report as spam” or “delete” messages that they perceived as phishing. Although they sometimes acted cautiously with messages that they perceived as legitimate (e.g., checking the link or sender), they rarely chose to “click link or open attachment” for emails they perceived as phishing. Figure 1b shows the same actions as a function of whether the messages were actually legitimate or phishing. Their behavior almost always reflected appropriate or cautious actions, given their perceptions. However, their imperfect detection ability meant that such conditionally appropriate behavior still produced responses inappropriate for the message.

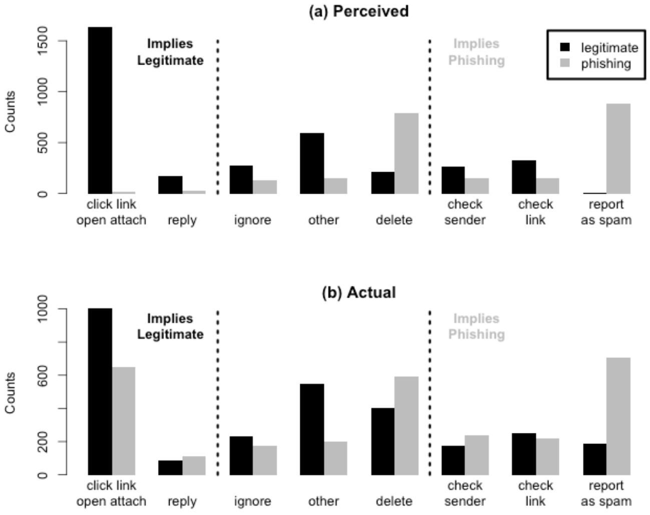

Figure 1. Counts of actions selected based on (a) perceived and (b) actual type of email.

3.2 Sensitivity to Task and Individual Variables

We used regression to identify variables that predict d’ and c for the detection and behavior tasks. We manipulated three task variables: (1) knowledge of the base rate, (2) order of tasks, and (3) use of personal greetings in emails. Of the three variables, only use of personal greetings had a significant effect. A personal greeting (e.g. “Dear Kelly” rather than “Dear Customer”) was associated with lower discrimination ability (d’) and a lower or more cautious decision threshold (c) on the behavior task.

We also assessed the influence of individual differences. For the detection task, improved ability (higher d’) was associated with higher confidence. In addition, participants were more cautious (lower c) when they perceived worse consequences of being successfully phished. For the behavior task, participants had higher ability (d’) and were more cautious (lower c) when they perceived worse consequences. We also observed greater caution (lower c) when participants were older, had more friends and family who had experienced negative computer incidents (e.g. identity theft), and were less confident.

This study has several limitations. Participants assessed artificial emails based on a persona rather than their own emails. Other variables that may influence phishing ability were not controlled or studied here. For example, participants might pay less attention to their real email than they did in this study because they are rushing, using a mobile device, or otherwise distracted.

4. CONCLUSION

Our results suggest two primary conclusions. First, SDT is a useful framework for distinguishing between discrimination ability and decision bias in the context of phishing detection. Participants here knew which actions were appropriate for phishing emails, but they still struggled to distinguish between phishing and legitimate emails. Second, participants used somewhat different decision-making strategies for the detection and behavior task, showing greater sensitivity to context cues for the latter. Future work should assess causality to determine whether manipulating perceived consequences is an effective intervention.

The quantification of performance provided by SDT could, potentially, improve the management of computer systems by providing analytically sound metrics. For example, an organization could use these measures to predict system vulnerability in a risk model, estimate natural variation in a longitudinal study, and evaluate interventions, such as training, incentives, or task restructuring, in a randomized control trial.

5. ACKNOWLEDGMENTS

Casey Canfield was supported by the National Science Foundation Graduate Research Fellowship (1121895) and the Carnegie Mellon Bertucci Fellowship. We thank Wändi Bruine de Bruin, Stephen Broomell, and Lorrie Cranor for comments on study concept. In addition, we thank the Energy & Behavior Group as well as the CyLab Usable Privacy and Security Laboratory (CUPS) for comments on preliminary findings.

6. REFERENCES

[1] Symantec. 2014. Internet Security Threat Report. Technical Report. Retrieved from http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf.

[2] Cranor, L. F. 2008. A Framework for Reasoning About the Human in the Loop. UPSEC. 8, 1-15.

[3] 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), 576–586.

[4] Macmillan, N. A., and Creelman, C. D. 2004. Detection theory: A user's guide. Psychology Press, New York.

[5] Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., and Hong, J. 2010. Teaching Johnny not to fall for phish. ACM Transactions on Internet Technology, 10 (2), 1–31.