Association for Information Systems
AIS Electronic Library (AISeL)

WISP 2015 Proceedings

Pre-ICIS Workshop on Information Security and Privacy (SIGSEC)

Winter 12-13-2015

A Study on Resolution Skills in Phishing Detection

Yuan Li
Columbia College

Jingguo Wang
University of Texas, Arlington

Raghav Rao
State University of New York, Buffalo

A Study on Resolution Skills in Phishing Detection

Yuan Li
Columbia College, Columbia, South Carolina, USA {yli@columbiasc.edu}

Jingguo Wang
University of Texas, Arlington, Texas, USA {jwang@uta.edu}

H. Raghav Rao
State University of New York, Buffalo, New York, USA {mgmtrao@buffalo.edu}

ABSTRACT

This study examines resolution skills in phishing email detection, defined as the abilities of individuals to discern correct judgments from incorrect judgments in probabilistic decision-making. An illustration of the resolution skills is provided. A number of antecedents to resolution skills in phishing email detection, including familiarity with the sender, familiarity with the email, online transaction experience, prior victimization of phishing attack, perceived self-efficacy, time to judgment, and variability of time in judgments, are examined. Implications of the study are further discussed.

Keywords: Phishing detection, confidence, accuracy, resolution, calibration

INTRODUCTION

The ability of individuals to correctly recognize phishing emails from genuine emails is of critical importance in today's networked society. Traditionally, such ability has been studied as detection accuracy, i.e., the likelihood of an email being correctly classified as either a phishing email or a genuine business email (Vishwanath et al., 2011; Wright and Marett, 2010). There are two limitations in this line of research. First, a person's ability to recognize a lie is different from his or her ability to recognize a truth (Alba and Hutchinson, 2000), so that

accuracy alone is not sufficient to reflect such difference. Second, as judgmental confidence plays an important role in driving individual behavior (Berger, 1992), its influence should not be ignored. A focus on judgmental accuracy as well as on confidence (Keren, 1997) can help address the issues.

Studies on judgmental abilities from the accuracy-confidence perspective typically examine two types of abilities: calibration and resolution (Keren, 1997; Liberman and Tversky, 1993). Calibration refers to the correspondence between confidence and accuracy in judgments; resolution, on the other hand, refers to the ability of a person to discern correct judgments from incorrect judgments (Bjorkman, 1992). They reflect different judgmental abilities and are not interchangeable (Keren, 1997; Sharp et al., 1988). As overconfidence (and thus miscalibration) is common in probabilistic judgments and in phishing detection (Hong et al., 2013; Kumaraguru et al., 2007), we suggest that a study on resolution skills may bring new insights into understanding individuals' phishing detection abilities. In the next sections, we first present a conceptual basis of the resolutions skills in phishing detection, and then develop hypotheses to examine the antecedents to resolution skills. An empirical test of the hypotheses is reported, followed by a brief discussion on the implications of the study.

RESOLUTION SKILLS IN PHISHING DETECTION: A CONCEPTUAL BASIS

In judgment under uncertainty, a person's judgmental confidence, in addition to accuracy, is often an important factor to consider (Karen, 1997). Confidence represents the strength with which a person believes that a specific statement, opinion, or decision is the best possible (Peterson and Pitz, 1988; Tang et al., 2014). It is measured as a subjective probability, such as 100% denoting very confident and 50% denoting by chance. Although the full-range measure (i.e., 0-100%) of confidence is also in use in literature, the half-range measure (i.e., 50-100%) is

logically more reasonable and statistically more reliable (Juslin and Olsson, 1997). We adopt the half-range measure in our study but cite the full-range measure wherever applicable.

Judgmental research usually presents a series (*N*) of questions to subjects, and for each question, the subjects are asked to choose the correct answer and indicate the confidence (denoted by f_n for the n-th question) in the answer (Palmer et al., 2013; Weber and Brewer, 2004). The simplest scenario is to have two possible answers for each question, such as "Is this a genuine email or a phishing email?" Accuracy of the judgment is then measured based on its outcome (denoted by d_n for the n-th judgment). Following Yates (1982), we set d_n to 1 for a correct judgment and 0 for an incorrect judgment.

The Resolution Index (RI) is a metric to measure resolution skills (Baranski and Petrusic, 1994). First, the *N* judgments are categorized into *J* categories based on the confidence levels (i.e., the f_n scores), with each category representing a similar level of confidence. Six categories, including 50-59%, 60-69%, 70-79%, 80-89%, 90-99%, and 100%, are commonly used. The mean accuracy in each category is then calculated, and RI is derived from the equation:

$$\text{RI} = \left(\frac{1}{N}\right) \sum_{j=1}^{J} N_j (\bar{d}_J - \bar{d})^2$$

where N_j is the number of judgments in the j-th category, $\bar{d}_J$ is the mean accuracy in the j-th category, and $\bar{d}$ is the overall accuracy of the *N* judgments. A higher RI score means a stronger judgmental ability (Baranski and Petrusic, 1994). The squared difference in the equation suggests that RI captures the ability of a person to "differentiate instances when an event is going to occur from those when it is not (Yaniv et al., 1991, p. 612)," since the score can be maximized when $\bar{d}_J$ equals 1 or 0. In other words, judgments with the same outcome (correct or not) are assigned to the same confidence category, and that "maximum resolution occurs if correct responses and incorrect responses are sorted into different categories (Bjorkman, 1994, p. 4)."

We adapt an example from Yaniv et al. (1991) to illustrate the resolution skills in phishing email detection and its difference from the overall accuracy. Three persons (A, B, and C) make judgments on 50 emails; their confidence levels and accuracy are shown in Table 1. To be consistent with Yaniv et al. (1991), we use the full-range measure of confidence in the example. For simplicity reason, we assume the three persons assign the same confidence to the same email judgment, although the accuracy differs in the following emails: # 24, #25, #36, and #37. Due to the length of the paper, other email judgments (such as #1-#20), which are identical between the persons, are not shown on the table; the information is available upon request.

Email #	Confidence (f_n)	Accuracy (d_n)		
		A	B	C
21	40%	1	1	1
22	40%	1	1	1
23	40%	1	1	1
24	40%	1	1	0
25	40%	1	0	0
26	40%	0	0	0
27	40%	0	0	0
28	40%	0	0	0
29	40%	0	0	0
30	40%	0	0	0
	$\bar{d}$	0.420	0.420	0.420

Email #	Confidence (f_n)	Accuracy (d_n)		
		A	B	C
31	60%	1	1	1
32	60%	1	1	1
33	60%	1	1	1
34	60%	1	1	1
35	60%	1	1	1
36	60%	0	1	1
37	60%	0	0	1
38	60%	0	0	0
39	60%	0	0	0
40	60%	0	0	0
	RI	0.0396	0.0436	0.0556

Table 1. An illustration of the resolution skills in phishing detection

It can be seen from the table that all three persons make the same percentages of correct judgments ($\bar{d}$=.42 or 42%), so that overall accuracy, as mentioned above, is not a good indicator of their varied judgmental abilities. The RI suggests that C exhibits better judgmental ability than A and B, which is reflected in the raw data: compared to the 40% confidence level, the 60% confidence level means that the subject should be more accurate at this level, which is true for C only. For A and B, both exhibit less accuracy at this level. On the other hand, A and B make more correct judgments than C at the 40% confidence level, despite the fact that these judgments

are associated with lower confidence (as compared to 60%). In other words, in judging the emails, C is more capable of discerning correct judgments from incorrect judgments than A and B. This example illustrates the value of using RI to capture such ability.

ANTECEDENTS TO RESOLUTION SKILLS: A RESEARCH MODEL

The Probabilistic Mental Model (PMM) theory (Gigerenzer et al., 1991) suggests that in making a judgment, a person first attempts the Local Mental Model (LMM), such as judging an email that he or she has seen before. If a LMM cannot be employed (i.e., the email was not seen before), the PMM is constructed for probabilistic judgment. The PMM puts the problem into a larger context that connects the specific structure of the task with a probability structure of a corresponding reference class stored in the long-term memory, and the reference class includes information cures related to the problem. For example, if a person receives an email from an online vendor asking to validate personal information, the person may recall similar emails (i.e., the reference class) from other vendors and assess the likelihood that this kind of requests was normally made.

A critical condition for accurate judgment is that the subjective probabilities of the person must equal the corresponding ecological probabilities of the reference class (Keren, 1997). It requires the subject to have general knowledge of the reference class in the task. As different people have different knowledge or information cues about phishing attack, their abilities in judging the emails (i.e., the resolution skills) would differ. In this study, we examine a few factors, including familiarity with the sender, familiarity with the email, online transaction experience, prior victimization of phishing attack, perceived self-efficacy in phishing detection, time to judgment, and variability of time in judgments, for their impacts on resolutions skills. These factors are recognized from prior literature on phishing detection.

First, perceived familiarity with the email sender may influence resolution skills. In fact, the same look-and-feel in emails (such as business names and logos) may deceive people, as they are less likely to respond to an unknown company but may react more favorably to a familiar company (Li, 2013). Perceived familiarity with the email sender causes people to feel relaxed and therefore lower their level of concerns, leading to poor judgment. This explains why millions of unsuspecting consumers and business users of Dun and Bradstreet have been targeted by phishing emails with the company's name and logo (www.bbb.org/blog/2013/07/dun-bradstreet-reports-phishing-email-scam/). We hypothesize:

H1: Perceived familiarity with the email senders is negatively associated with resolution skills in phishing detection.

Perceived familiarity with the email, on the other hand, helps to boost the resolution skills. This can be explained by the Local Mental Model (or LMM) mentioned above. If a person saw the email before, he or she may have developed the mental model about the email, so the person can retrieve the mental model for direct judgment (with 100% confidence) instead of judging based on external information cues (such as the reference class) that may lead to misjudgment. For example, the notorious Nigerian phishing scam has been known by many people, and its damage has been dwindling. Therefore, we hypothesize:

H2: Perceived familiarity with the emails is positively associated with resolution skills in phishing detection.

Online transaction experience reflects one's general knowledge on online business environment and communication. Eastin and LaRose (2000) suggest that Internet use is positively related to a person's ability in accomplishing online tasks. Online experience helps users to accumulate information cues about potential online threats, therefore expanding the

potential reference class to judge new threats. In general, Internet savvy subjects are more capable of discerning phishing emails than amateurs (Wright and Marett, 2010). We hypothesize:

H3: Online transaction experience is positively associated with resolution skills in phishing detection.

Prior victimization of phishing attack provides more specific information about its risk. Similar to the familiarity with emails, prior victimization helps the individual to build Local Mental Models (or LMM) about the phishing threat, enabling the individual to identify similar attacks directly. Although it's possible that the mental model may not be readily available in memory (i.e., the bad experience may have slipped out of one's mind), it is less common since the damage it causes may be too severe to forget. Therefore we hypothesize:

H4: Prior victimization of phishing attack is positively associated with resolution skills in phishing detection.

Perceived self-efficacy in phishing detection reflects one's perceived ability in recognizing phishing emails. Stone (1994) suggests that the initial, "first-impression" efficacy judgments are biased toward overestimates of one's ability. Further research also shows that self-efficacy may increase the likelihood of committing logic errors in analytic games and leads to overconfidence (Vancouver et al., 2002). Such a positive self-efficacy expectation is likely to increase post-decision perceptions of, but not necessarily the actual, performance in a cognitively complex task such as phishing detection. Therefore, we hypothesize:

H5: Perceived self-efficacy in phishing detection is negatively associated with resolution skills in phishing detection.

How much cognitive effort (such as time) is spent on processing emails also influences the resolution skills in phishing detection. Time to judgment, defined as the time spent to make

the judgment, is a major factor to consider, because people generally avoid effortful reasoning and thought process, so that the lack of decision effort leads to poor judgment (Payne, 1976). In terms of phishing detection, it means people are unable or unwilling to consider all information cues in the emails and rash to decisions to save efforts. Because of this, their resolution skills in phishing detection may be weakened. On the other hand, more time spent on information processing may improve judgmental outcome, so that we hypothesize:

H6: Time to judgment in phishing detection is positively associated with resolution skills in phishing detection.

The allocation of cognitive effort in dealing with emails also influences resolution skills. When resources, such as time, are constrained and therefore unevenly allocated, proper identification of information cues in the judgmental task may not occur and false beliefs may be developed (Alba and Hutchinson, 2000). Variability in attention allocation influences subsequent judgments (Keren, 1997) as when a person deals with many emails daily, which is particularly true in the era of information overload. Greater variability in time is associated with the use of heuristic rules that requires less or incomplete information and over-relies on selected information (Payne, 1976), leading to misjudgment. Therefore, we hypothesize:

H7: Variability of time in phishing detection is negatively associated with resolution skills in phishing detection.

A few control variables are also included in the study, including: number of emails received every day, easiness of the email judgments, education, gender, and age.

RESEARCH METHOD AND RESULTS

We conducted an online experiment via the Qualtrics Research Suite to study individuals' abilities in phishing detection. The use of the research suite allows us to reach more online users

and collect responses from a diverse sample. This study is a part of a larger research project that investigates phishing email detection.

In the experiment, each subject made judgments of 16 emails randomly chosen from a pool of 50 phishing and genuine emails; different subjects received different sets of emails due to the random selection. The experiment proceeded as follows. First, each subject completed items measuring perceived self-efficacy of phishing detection, and then was presented with the 16 emails. We used images of the emails for this study, following practice in related studies (e.g., Vishwanath et al., 2011; Wang et al., 2012). The emails were presented to the subject sequentially, and for each email the subject was asked whether it was legitimate or not (time elapsed to make the judgment was recorded online), and how confident he or she was (with a number between 50 and 100). For each email, the subject was also asked whether he or she had received or seen it before, and how familiar he or she was with the business entity in the email. After all the judgments, the subject completed the rest of the measurements including the demographic information. A total of 600 responses were collected.

Table 2 shows the measurements of the independent variables. As the unit of analysis is at the individual level, we aggregated familiarity with email senders and familiarity with emails, respectively, to measure a subject's overall familiarity with senders and with emails. Several other constructs in this table were measured with aggregated scores as well to reflect the same unit of analysis. Time to judgment in each email was recorded automatically, as mentioned above. The total judgmental time spent on the 16 emails was then calculated to measure time to judgment. Following prior literature in behavioral decision making (Brucks, 1985; Payne, 1976; Stone, 1994), the coefficient of variation (CV) of the time spent on the emails was calculated to measure variability of time in judgments.

Independent variables	Measurement methods	Further treatment
Familiarity with the email senders	How familiar do you think you are with the business entity indicated in the email?	Averaged across the 16 emails to form a single indicator
Familiarity with the emails	Have you personally received or seen this particular email before this survey?	Averaged across the 16 emails to form a single indicator
Online transaction experience	1) buying products or services online with a credit card, a debit card, or a payment service such as PayPal; 2) accessing bank accounts (such as checking, saving, mortgage) online; 3) paying bills (such as electronic, utility, credit cards, loans) online; and 4) buying and selling stocks or mutual funds online.	The four items were summed to form a single indicator
Prior victimization of phishing attack	1) someone used or attempted to use your credit cards without permission; 2) someone used or attempted to use your accounts such as your wireless phone account, bank account or debit/check cards without your permission; and 3) someone used or attempted to use your personal information without permission to obtain new credit cards or loans, run up debts, open other accounts, or commit other frauds.	The three items were summed to form a single indicator
Perceived self-efficacy in phishing detection	1) I can recognize phishing emails, and 2) I can differentiate phishing emails from legitimate ones	The two items were summed to form a single indicator
Time to judgment	Recorded online	Total time spent on the 16 emails was calculated
Variability of time in judgments	The coefficient of variation (CV) of the time spent on the emails was calculated	

Table 2 Measurement of independent variables

The dependent variable RI was calculated based on the method described above (see the Equation). Of the control variables, easiness of the email judgments was measured by first calculating for each email the percentage of individuals who made correct judgment on that email, and then averaging across the sixteen emails that a subject worked on. We did this because, as mentioned above, each subject handled a different set of emails randomly drawn from the pool, so that it could happen that a set of emails was more challenging than another set.

Other control variables were measured using single items.

Of the 600 subjects, 8 failed to do the experiment and were excluded from the sample, resulting in 592 valid observations. We first measure the resolution skills (using RI) and compare to overall accuracy and confidence. As shown in Table 3, overconfidence is apparent in this experiment, as the mean confidence (0.8089) is much higher than the mean accuracy (0.6670). The RI score, although very small, is normal: Baranski and Petrusic (1994) point out that the resolution score above .10 is rarely encountered.

	Accuracy	Confidence	RI
Mean	0.6670	0.8089	0.0414
Std. Dev.	0.1504	0.1213	0.0339
C.I. (5%)	0.3750	0.5658	0
C.I. (95%)	0.8750	0.9744	0.1085

C.I. – Confidence Internal; Sample size = 592

Table 3 Measurements of phishing detection abilities

We then ran Multiple Liner Regressions to test the hypotheses. Variance Inflation Factors (VIFs) were calculated to detect potential multicollinearity, with no significant issues noticed. The results are reported in Table 4. Familiarity with the email senders has a negative impact on RI, supporting H1. Familiarity with emails has no significant effect on RI, rejecting H2. Online transaction experience is positively related to RI, supporting H3. Prior victimization of phishing attack is positively related to RI, supporting H4. Self-efficacy in phishing detection has no significant impact on RI, rejecting H5. Time to judgments has no significant impact on RI, thus rejecting H6. Finally, variability of time in phishing detection is negatively related to RI, providing support to H7. Of the control variables, none has a significant impact on RI.

	RI		
	β	t-value	Sig.
(Constant)		1.909	0.057
Familiarity w/ the senders (H1)	-0.1	-2.167	0.031
Familiarity with emails (H2)	0.048	1.086	0.278

Online experience (H3)	0.15	3.383	0.001
Prior victimization (H4)	0.1	2.393	0.017
Self-efficacy (H5)	-0.022	-0.536	0.592
Time to judgments (H6)	0.078	1.816	0.07
Variability of time (H7)	-0.13	-2.987	0.003
# of emails received daily	0.043	1.066	0.287
Easiness of judgments	-0.037	-0.901	0.368
Age	-0.082	-1.92	0.055
Gender (male=0, female=1)	0.049	1.189	0.235
Education	-0.01	-0.244	0.807
R Square	0.057		
Adjusted R Square	0.037		

Table 4 Results of Multiple Liner Regressions

CONCLUDING REMARKS

The study has a couple of potential contributions. First, we illustrate that the resolution skill differs from judgmental accuracy, and in some cases it is a better indicator of one's judgmental ability in dealing with phishing emails. Nevertheless, both the resolution skills and the overall accuracy are still low among our subjects, suggesting that more efforts are needed to understand the cause of their weak phishing detection abilities. This calls for more research on this type of ability and ways to improve resolution skills.

Second, the antecedents to resolution skills suggest that researchers and practitioners may design corresponding mechanisms to enhance people's judgmental abilities in phishing detection. For instance, both online experience and prior victimization enhance resolution skills, meaning that education of the threat of phishing attacks can help people become aware of the risks and improve their judgmental abilities. Variability of time is a major threat to resolution skills. This suggests that people should learn to manage their time properly in handling emails to avoid judgmental bias. For instance, Sharp et al. (1988) show that performance feedback can help boost resolution skills, which may be embedded in training programs.

REFERENCES

Alba, J. W., and Hutchinson, J. W. 2000. "Knowledge Calibration: What Consumers Know and What They Think They Know," Journal of Consumer Research (27:2), pp.123–156.

Baranski, J. V. and Petrusic, W. M. 1994. "The calibration and resolution of confidence in perceptual judgments," Perception and Psychophysics (55), pp. 412-428.

Berger, I.E., 1992. "The Nature of Attitude Accessibility and Attitude Confidence: A Triangulated Experiment," Journal of Consumer Psychology, 1(2), pp. 103-123.

Bjorkman, M. 1992. "Knowledge, calibration, and resolution: a linear model," Organizational Behavior and Human Decision Processes (51), pp. 1-21.

Bjorkman, M. 1994. "Internal cue theory: calibration and resolution of confidence in general knowledge," Organizational Behavior and Human Decision Processes (58), pp. 386-405.

Brucks, M. 1985. "The effects of product class knowledge on information search behavior," Journal of Consumer Research (12:1), pp. 1–16.

Eastin, M. S., and LaRose, R. 2000. "Internet Self-Efficacy and the Psychology of the Digital Divide," Journal of Computer-Mediated Communication 16(1).

Gigerenzer, G., Hoffrage, U. and Kleinbolting, H. 1991. "Probabilistic mental models: A Brunswikian theory of confidence," Psychological Review (98), pp. 506-528.

Hong, K.W., Kelley, C.M., Tembe, R., Murphy-Hill, E., and Mayhorn, C.B. 2013. "Keeping Up With The Joneses: Assessing Phishing Susceptibility In An Email Task," Meeting of the Human Factors and Ergonomics Society.

Juslin, P., and Olsson, H. 1997. "Thurstonian and Brunswikian Origins of Uncertainty in Judgment: A Sampling Model of Confidence in Sensory Discrimination," Psychological Review (104:2), pp. 344–366.

Keren, G. 1997. "On the calibration of probability judgments: Some critical comments and alternative perspectives," Journal of Behavioral Decision Making (10), pp. 269-278.

Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L. and Hong, J. 2007. "Getting Users to Pay Attention to Anti-Phishing Education: Evaluation of Retention and Transfer," APWG eCrime Researchers Summit, October, 4-5, 2007, Pittsburgh, PA, USA.

Li, Y. 2013. "The impact of disposition to privacy, website reputation and website familiarity on information privacy concerns," Decision Support Systems (57), pp. 343–354.

Liberman, V. and Tversky, A. 1993. "On the Evaluation of Probability Judgments: Calibration, Resolution, and Monotonicity," Psychological Bulletin (114:1), pp.162-173.

Palmer, M.A., Brewer, N., Weber, N. and Nagesh, A. 2013. "The Confidence-Accuracy Relationship for Eyewitness Identification Decisions: Effects of Exposure Duration, Retention Interval, and Divided Attention," Journal of Experimental Psychology: Applied (19:1), pp. 55–71.

Payne, J. W. 1976. "Task complexity and contingent processing in decision making: An information search and protocol analysis," Organizational Behavior and Human Performance 16(2), pp. 366–387.

Peterson, D. K., and Pitz, G. F. 1988. "Confidence, uncertainty, and the use of information," Journal of Experimental Psychology: Learning, Memory, and Cognition (14:1), pp. 85–92.

Sharp, G.L., Cutler, B.L. and Penrod, S.D. 1988. "Performance Feedback Improves the Resolution of Confidence Judgments," Organizational Behavior and Human Decision

Processes. 42, pp. 271-283.
Stone, D. N. 1994. "Overconfidence in Initial Self-Efficacy Judgments: Effects on Decision Processes and Performance," Organizational Behavior and Human Decision Processes (59:3), pp. 452–474.
Tang, F., Hess, T. J., Valacich, J. S., and Sweeney, J. T. 2014. "The Effects of Visualization and Interactivity on Calibration in Financial Decision-Making," Behavioral Research in Accounting (26:1), pp. 25–58.
Vancouver, J. B., Thompson, C. M., Tischner, E. C., and Putka, D. J. 2002. "Two studies examining the negative effect of self-efficacy on performance," Journal of Applied Psychology (87:3), pp. 506–516.
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), pp. 576–586.
Wang, J., Herath, T., Chen, R., Vishwanath, A., and Rao, H. R. 2012. "Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email," IEEE Transactions on Professional Communication (55:4), pp. 345–362.
Weber, N., and Brewer, N. 2004. "Confidence–Accuracy Calibration in Absolute and Relative Face Recognition Judgments," Journal of Experimental Psychology: Applied (10:3), pp. 156–172.
Wright, R. T., and Marett, K. 2010. "The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived," Journal of Management Information Systems (27:1), pp. 273–303.
Yaniv, I., Yates, J.F., and Smith, J.E.K. 1991. "Measures of discrimination skill in probabilistic judgment," Psychological Bulletin (110:3), pp. 611-617.
Yates, J. F. 1982. "External correspondence: Decomposition of the mean probability score," Organizational Behavior and Human Performance (30), pp. 132-156.