Simulating Phishing Email Processing with Instance-Based Learning and Cognitive Chunk Activation

Matthew Shonman, Xiangyang Li, Haoruo Zhang, and Anton Dahbura

Johns Hopkins University Information Security Institute, Baltimore 21218, USA
{mshonma1,xyli,zhanghaoruo,antondahbura}@jhu.edu

Abstract. We present preliminary steps applying computational cognitive modeling to research decision-making of cybersecurity users. Building from a recent empirical study, we adapt Instance-Based Learning Theory and ACT-R's description of memory chunk activation in a cognitive model representing the mental process of users processing emails. In this model, a user classifies emails as phishing or legitimate by counting the number of suspicious-seeming cues in each email; these cues are themselves classified by examining similar, past classifications in long-term memory. When the sum of suspicious cues passes a threshold value, that email is classified as phishing. In a simulation, we manipulate three parameters (suspicion threshold; maximum number of cues processed; weight of similarity term) and examine their effects on accuracy, false positive/negative rates, and email processing time.

Keywords: Phishing, Cognitive Modeling, Chunk Activation.

1 Introduction

While reliable estimates vary, hundreds of millions of phishing emails at a minimum are sent every year [4]. Despite the gradual emergence of automated anti-phish defenses, human judgment remains a significant, and typically sole, means for distinguishing legitimate emails from malicious attacks. Many research efforts identify personality traits and informational cues that relate to processing legitimate and suspicious emails, quantifying their impact on user performance through empirical studies [8-9, 11].

Computational cognitive modeling offers an additional route to study this process. Computational models, including ACT-R and SOAR [2, 7], describe the underlying psychological operations producing human behaviors in physiological movement and problem solving. Compared to "black box" or "product theory" models of phishing and security behavior, which mainly describe correlation between inputs and outputs, these models can offer greater insights into the interactions between a user and a task environment. In addition to predicting potential issues, such as errors in decision making or delays in reaching a task goal, these techniques enable researchers to diagnose plausible causes, based on emerging cognitive conditions, and discern an appropriate remedy.

This work builds from two significant research efforts: the sole notable study using computational cognitive modeling to examine cybersecurity decision-making [6] and a recent empirical study of users making phishing classifications [12]. Our model, based upon Instance-Based Learning Theory (IBLT) and the chunk activation mechanism of ACT-R, portrays users as drawing upon past memory "instances" to determine whether individual cues (such as the email title or presence of poor grammar) are themselves suspicious. Users decide whether a cue is suspicious by matching it with a single previously-encountered cue stored as a memory instance, called a chunk; this "activated" instance is selected from many others according to ACT-R's chunk activation calculation, with the "winner" having the highest value. If a sufficient number of cues in an email are deemed suspicious, the email is classified as phishing.

This simulation study examines the influence of three control parameters on classification accuracy and task completion time:

- *Suspicion threshold*. This term denotes the minimum number of suspicious cues detected before the user marks an email as phishing.
- *Maximum cues processed*. This term denotes the total number of cues per email that a user would likely inspect.
- *Weight of the similarity term in chunk activation calculation*. The ACT-R memory activation formula considers the recency and frequency of an instance's past retrievals and its similarity to the cue currently under consideration.

Our work strives to make several contributions. It is the first study to apply computational cognitive modeling to user security behavior in phishing, widening this important research subject. It presents a comprehensive model of email processing that significantly extends the IBL model integrating cognition chunk activation. Moreover, we are aiming at a systematic effort that also examines data from an empirical study for comparison and validation.

2 Related Work

2.1 Chunk Activation in ACT-R and Instance Based Learning Theory (IBLT)

Anderson [2] proposed the ACT-R cognitive architecture in 1993. In this model, declarative knowledge is stored as discrete "chunks" in long-term memory. ACT-R models a process through which information in memory is retrieved if selected as relevant to a present situation. Relevant information chunks are selected according to an activation value calculation equation, simplified in Kaur et al. [6] as:

$$A_i = B_i + Sim_i + \varepsilon_i. \quad (1)$$

B_i represents a base-level activation, combining the recency and frequency of a chunk's prior retrievals. Sim_i denotes the association or similarity between a chunk and the current situation. ε_i is a random noise term to model imperfection in human cognition. This activation process forms a core component of our own model.

In further detail, for the *i*th memory chunk (equations drawn from Kaur et al. [6]):

$$B_i = \ln\left(\sum_{t_i \in \{1,\ldots,t-1\}} (t - t_i)^{-d}\right) \quad (2)$$

$\{1, \ldots, t-1\}$ represents the set of past activation times for the given chunk. $(t - t_i)$ represents the lapse between current time t and a given past activation time t_i. Decay term d has a default value of 0.5. Our study used relative time, omitting duration units.

$$Sim_i = \sum_{l=1}^{k} P_l * M_{li} \quad (3)$$

P_l is a weight with value -0.01. M_{li} represents the raw similarity score comparing the lth information attribute with the situation represented by the chunk.

$$\varepsilon_i = s * \ln\left(\frac{1-\eta_i}{\eta_i}\right) \quad (4)$$

η_i is drawn from a uniform random distribution between 0 and 1 exclusive. Weight s has a default value of 0.25. 90% of ε_i values lie between ±0.736.

Gonzales et al. [3] developed instance-based learning theory (IBLT) to describe a learning process linked to dynamic decision-making. Experiences are stored in memory as instances with three components: situation (relevant environmental cues), decision (action taken in response to a situation), and utility (post-hoc evaluation of a decision). In order to determine the appropriate action for a current situation, the model considers the utilities of past actions taken in response to similar situations.

2.2 Cognitive Modeling and Computer Security

A limited range of research has thus far applied cognitive modeling to enhance the study of computer security. Veksler et al. [10] discuss several potential uses of cognitive modeling in cybersecurity contexts, such as comparing the effects of training strategies on users and understanding the psychology of attackers, defenders and users to facilitate security improvements and predict human errors. However, this work offers few specifics on implementing its proposals. Veksler and Buchler [9] present three simulations demonstrating that techniques such as model tracing and dynamic parameter adjustment allow computational cognitive models, in the context of social security games, to outperform normative game theory in predicting and responding to cyber attackers. Similarly, Jones et al. [5] describe the use of cognitive agents developed with the Soar architecture to improve training simulations for cyber operators. These agents can consider goals and context in attack and defensive scenarios; they also exhibit generative mechanisms to produce new tactics and learn from experience.

2.3 Computational Cognitive Modeling of Security Decision Making

The sole published work on *computational* cognitive modeling of individual users in computer security is the simulation research of Kaur et al. [6]. Their method draws upon IBLT to describe the behavior of a security analyst determining whether a series of network events constitutes a cyberattack. In this model, situation information is represented as a series of attributes denoting particular details of a network event, including the network location, alert, and operation result. Security analysts classify individual

events as threat or non-threat by examining the selected chunk from a past similar experience in memory. A counter for each event sequence increments for each new event judged as a threat. When the counter surpasses a set threshold, the entire sequence is classified as a cyberattack. Each event under consideration is compared to all instances stored in memory. The instance with the highest ACT-R activation score is retrieved from memory, with its utility used to classify the event under consideration.

Our model differs from the above study in several ways:

- In the Kaur et al. simulation, all incidents (event sequences) are attacks (although individual normal events are present). In our model, an email (the equivalent of a full event sequence) may be either phishing or normal.
- In Kaur et al., the analyst continues to score events in a sequence until either the preset number of threat events have been classified, or all events in the sequence have been classified without triggering the suspicion counter. We change this by adding a maximum-cues-considered parameter.
- In Kaur et al., all network events are of identical structure and all decisions draw upon a single shared pool of memory instances. In our model, cues are of various types and decisions for a given type draw only upon memories of that same type.
- Kaur et al. held the suspicion threshold constant. We vary suspicion threshold as one experimental parameter.
- Moreover, we have conducted an empirical study [12] that collects data from real users, providing rich information for further assistance to the modeling effort.

3 A User Experiment of Email Classification

The user task described fully in Zhang et al. [12] provides context for our current work. Study participants were directed to classify 40 randomly-ordered emails as “keep” or “suspicious.” 20 emails were legitimate and the remaining 20 were phishing. All phishing emails were link-based attacks.

3.1 Condition-Based User Study Task Sets

Two independent variables were manipulated, each with two levels. Participants were randomly assigned to one of four experimental conditions: (1) Multitasking with Incentive; (2) No-Multitasking with Incentive; (3) Multitasking with No Incentive; and (4) No Multitasking and No Incentive.

Multitasking participants answered 20 sets of questions in an online survey system while completing the email sorting task in Roundcube, a webmail system. Each question set was presented for a maximum of two minutes; participants could manually advance to the next question set after one minute elapsed. Thus, multitasking participants had 40 minutes at most to complete both tasks. For the no multitasking condition, participants were given 30 minutes to complete only the email sorting task.

For the incentive conditions, participants could earn additional monetary compensation based on the number of correctly sorted emails. For Condition 1 participants, extra

money earned depended on accuracy of both the email sorting and multitasking tasks. For the no incentive conditions, participants received no additional compensation.

3.2 Email Design and Phishing Cues

All 40 emails were created from real emails with personally identifiable information modified. Phishing emails were derived from a semi-random sample of emails in Cornell University's "Phish Bowl" database (it.cornell.edu/phish-bowl). Legitimate emails were derived from emails received by the research team.

We defined a series of cues, contained within the email, implying whether those emails are legitimate or phishing. Crucially, legitimate emails may contain individual suspicious cues, such as misspellings or an absent greeting, while phishing emails may contain enough non-suspicious cues to seem legitimate. However, phishing emails on average contained more suspicious cues than did legitimate emails, providing a path to accurate classification. The original cue definitions are in Zhang et al. [12]; for the present simulation, the "URL Hyperlink" cue was expanded to encompass two other link cues, while a "Subject" cue was added, for a total of 13 cues shown in Table 1.

Table 1. Phishing Cue Definitions.

Cue Type	Cue Definition
Branding/Logos	Does the email contain company branding and/or logos?
Overall Design	Does the overall email quality appear poor?
Suspicious Sender Name	Does the subject line appear suspicious?
Subject	Does the subject line direct the receiver to take an action?
Lack of Signer Details	Does the email provide sender information beyond a name?
Generic Greeting	Is the email greeting absent/not addressed to the individual?
URL Hyperlink (possibly multiple cues per email)	Scored according to presence or absence of two attributes: • Does the hyperlink text suggest a webpage different from the true link? • Does the hyperlink website match the email sender?
Spelling/Grammar	Does the text contain any spelling/grammar mistakes?
Time Pressure	Does the email request include a deadline?
Threatening Language	Does the email threaten a negative consequence if instructions unfollowed?
Emotional Appeal	Does the email elicit a sympathetic or otherwise emotional response?
Too Good to be True Offer	Does the email present a too-good-to-be-true offer?
Personal Information	Does the email request personal information?

3.3 Experiment Results

Out of 205 participants recruited through Amazon Mechanical Turk, 177 progressed through the full study, with 146 classifying all 40 emails in the given time. Participants were randomly assigned into the four experimental condition groups.

For email sorting accuracy, analysis of variance (ANOVA) testing indicated a significant effect of condition on email classification accuracy, using significance level α

at 0.05. Overall multitasking significantly worsened subjects' sorting accuracy, but incentive alone made no difference in either multitasking or no-multitasking cases. Significant differences were also present between phishing sorting error rates for conditions 1, 2, and 3. However, there was no significant difference between conditions for legitimate email sorting error rates.

Average email processing time was calculated for each email for every condition. Multitasking and incentive showed opposite effects: multitasking reduced users' processing time, while the incentive increased this value. Spending more time on individual emails did not always guarantee better sorting accuracy. For instance, condition 1 participants spent more time per email compared to those in condition 3, without increased accuracy. Although these participants were more "carefully" sorting emails, switching between tasks seemed to pose a challenge.

4 A Cognitive Model of Phishing Judgment Process

4.1 Model Design

Our study greatly extended the Kaur et al. model to fit the single-task design of the above phishing empirical study. Users classify an email by evaluating the email's individual cues (Table 1) as "threat" or "non-threat." The model maintains a counter variable for every email, which increments by one for each cue judged as threat. An email is classified as phishing when the number of cues so judged passes a threshold level.

The cue current in processing is classified according to the long-term memory chunk with the highest activation score at that moment. Chunks in long-term memory represent past email cues for which the email nature (phishing/non-phishing) is known, and contain the following parameters:

- *Cue type*. One of the 13 different types of cues.
- *Attribute score.* Attributes are coded 0 if the question (Table 1) is answered "No," and 1 otherwise. Hyperlink cues feature two attributes, while all others have one.
- *Utility.* The utility value is 0 if the email associated with this past cue was normal; 1 for phishing.

For this simulation study, long-term memory was populated with chunks derived from all 572 cues drawn from the 40 emails. This produced a memory store containing 40 chunks per cue type (one per source email) except for the hyperlink type; the emails contained 0-13 hyperlinks each, all encoded as distinct chunks. In this way, the 40 emails represent the "real" population distribution of cue chunks associated with legitimate and phishing emails, an assumption that we can re-examine and change in future simulations.

In a departure from the Kaur et al. model, not all cues are processed for each email. As more cues are classified, the likelihood increases that even normal emails will be scored as phishing (since normal emails tend to contain cues that are similar to those contained in phishing emails). In order to balance the likelihood of phishing and normal classifications, the model featured a parameter determining the maximum number of

cues which would be classified per email, separate from the suspicion threshold. When this number is reached, the email is immediately scored as normal if the suspicion threshold has not been crossed.

Cues are visited in an order that combines fixed steps and random elements. Expert input and observation through a pilot study suggest that email readers tend to view the following elements in sequence: limited text visuals, sender, subject, greeting, and "story" text. As a result, the model visits the six cues analogous to these elements (the first six cues in Table 1) in a linear order. Because no inherent order emerges for the remaining cues, their order is not fixed. All these cues are stored together in the user's working buffer. For this stage of cue processing, all memory chunks corresponding to these seven cue types are likewise pooled together; the memory chunk being activated determines which cue will be processed next. Once a cue has been processed, chunks of that type are skipped over for future rounds of processing. This sequence resets for each new email classified.

4.2 Simulation Setting

This study sought general insights as to the experimental parameters' influence on the simulation results; thus, we placed limited focus on the impact of specific parameter settings. The maximum number of cues processed per email varied between 7 and 12. The minimum bound ensured that at least one of the remaining cues beyond the first six (fixed-ordered) cues would be classified, while the maximum bound was selected because one email only had 12 cues, without any hyperlinks. The suspicion threshold varied between 2 and 6, always remaining beneath the maximum-cues-processed parameter. Finally, the ACT-R similarity weight P_l took the values -1, -2, and -3. This variation allowed us to examine the similarity term's interaction with the base-level learning and noise terms. Similarity was calculated as the difference between the respective attribute(s) of the cue under consideration and a memory chunk.

The simulation was coded in Python, with chunk management in long-term memory taken directly from the Python ACT-R source code [1]. The simulation was run 100 times for each of the 90 parameter combinations. Output metrics included number of cues processed per email (analogous to total time spent scoring email), classification accuracy, false negative rate (FNR) and false positive rate (FPR). 95% confidence intervals were computed for all simulation results, with ranges of ±0.425 (maximum cues processed), ±0.025 (accuracy), ±0.072 (FNR), and ±0.064 (FPR).

5 Results

Controlling for maximum cue number, accuracy was generally highest for mid-range suspicion threshold values (usually 4 on a 2-6 range). High maximum cue values, though, defied this trend, continuing to increase for high suspicion threshold values. Greater similarity weights were associated with greater accuracy values, but also with greater variation in accuracy for the maximum-cues-considered parameter at high suspicion thresholds.

The mean cues processed metric is equivalent to the average time spent processing an email. This value increased with both greater suspicion thresholds and greater maximum cue levels. Similarity weight had minimal influence on this output.

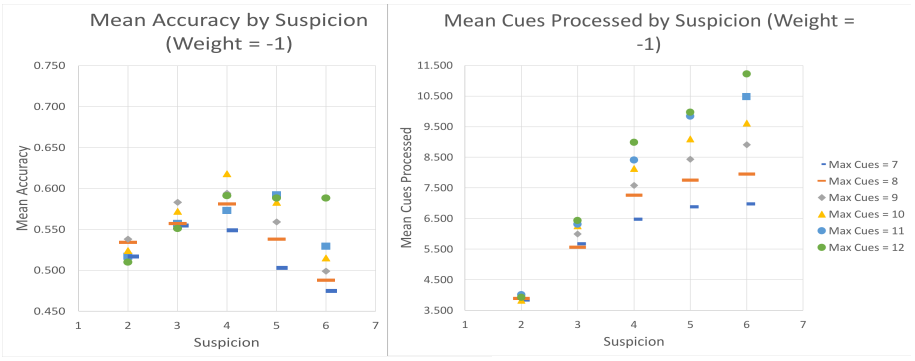

Fig. 1. Mean accuracy (left) and mean cues processed (right) for similarity weight = -1.

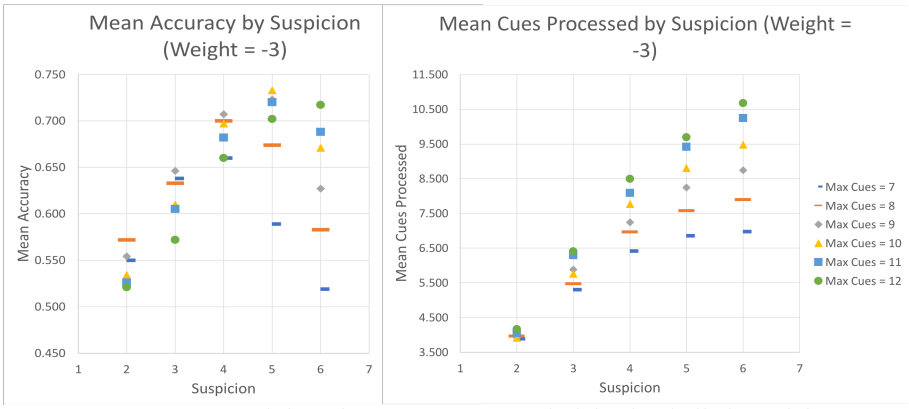

Fig. 2. Mean accuracy (left) and mean cues processed (right) for similarity weight = -3.

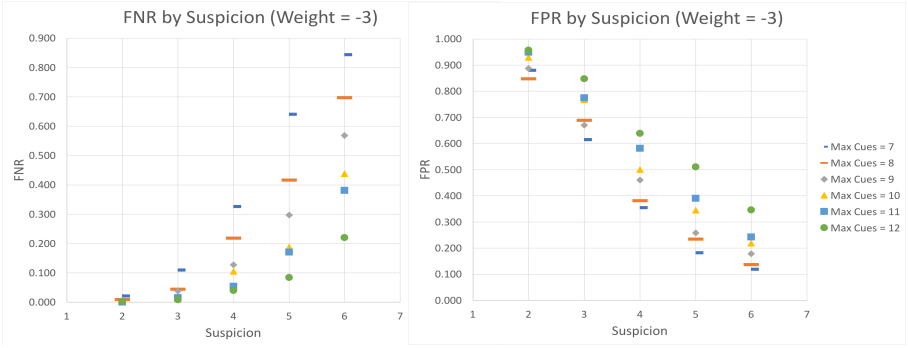

Fig. 3. False negative rate (left) and false positive rate (right) for similarity weight = -3.

FPR tended to decrease as the suspicion threshold was raised. Controlling for suspicion threshold, FPR rose as the maximum cue number increased. Greater similarity weights generally lowered FPR. FNR tended to increase with a greater suspicion threshold. Controlling for suspicion, FNR was generally highest for the lowest maximum cue numbers. Generally, FNR decreased slightly as similarity weight was raised, although some individual points broke with this trend.

6 Analysis

Both phishing and normal emails usually contained some threatening cues, although normal emails had comparatively fewer such cues. Thus, the likelihood of classifying any email as phishing tended to increase with more cues processed, since more opportunities existed for the "user" to encounter threatening cues and surpass the suspicion threshold. This potentially explains FPR and FNR behaviors. FPR tended to increase as the max-cues-processed parameter rose: even if phishing emails were accurately classified at both lower and higher parameter values, normal emails were more likely to be falsely classified at higher values. FPR also tended to decrease as suspicion threshold rose: since normal emails generally contained fewer threatening cues, the suspicion counter was less likely to rise as high for normal emails as for phishing emails. At low suspicion thresholds, this distinction might make little difference in classification rates; for higher thresholds, normal emails were less likely to be falsely classified as phishing.

FNR behavior followed similar principles. When the max-cues-processed parameter decreased, the suspicion counter became less likely to surpass the threshold. Thus, at higher suspicion thresholds, phishing emails were more likely to falsely receive a normal classification.

Accuracy was closely linked to these trends. False positives were more likely at high suspicion thresholds, and false negatives were more likely at low thresholds. Thus, accuracy tended to peak at medium threshold values. Greater similarity weights tended to lower both FPR and FNR, increasing accuracy. Lower weights de-emphasized similarity and increased the recency and frequency effects of past behavior on current decisions; memory instances activated early in a classification round held greater impact on later behavior, increasing the tendency for single instances to be activated for many email cues. This shows the importance of users remaining focused and current, while avoiding internal and external interruptions that might complicate this routine task.

Suspicion threshold and max-cues-processed held expected, positive relationships with the mean number of cues processed per email. Consistent with the user study, spending more time classifying did not correlate with better classification accuracies.

7 Conclusion

This simulation study represents first steps toward a computational cognitive model describing the psychological processes that underlie phishing email classification. Our results imply one initial conclusion: accuracy generally improved when similarity was emphasized over recency and frequency. This suggests that successful security analysts

should adopt a strategy that pays more attention to the current state than to details of emails recently encountered. Our manipulations of the maximum cue and suspicion threshold parameters provide additional insights into the decision-making process for this classification problem.

Computational cognitive modeling can offer powerful insights into the mindsets of cybersecurity operators, but few empirical studies have explored this topic. We plan further research to follow this simulation, beginning with fitting our model to population subgroups from the user study. Future work might explore the effect of phishing content changing over time, as phishing senders adapt their techniques to target skeptical recipients. We hope that our work combining simulation and empirical data will continue to enhance the study of human security informatics

8 Acknowledgement

This work is supported under the National Science Foundation Award No. 1544493.

References

1. ACT-R Software. http://act-r.psy.cmu.edu/software/
2. Anderson JR (1995). ACT: A Simple Theory of Complex Cognition. American Psychologist, 51(4), 355-365.
3. Gonzalez C, Lerch JF, Lebiere C (2003). Instance-based learning in dynamic decision making. Cognitive Science 27, 591-635.
4. Gudkova D, Vergelis M, Shcherbakova T, Demidova N (2018). Spam and phishing in 2017. Securelist. https://securelist.com/spam-and-phishing-in-2017/83833. Accessed 8 Oct 2018.
5. Jones RM, O'Grady R, Nicholson D, Hoffman R, Bunch L, Bradshaw J, Bolton A (2015). Modeling and Integrating Cognitive Agents Within the Emerging Cyber Domain. Paper presented at Interservice/Industry Training, Simulation, and Education Conference 2015.
6. Kaur A, Dutt V, Gonzalez C (2013). Modelling the Security Analyst's Role: Effects of Similarity and Past Experience on Cyber Attack Detection. Proceedings of the 22nd Annual Conference on Behavior Representation in Modeling and Simulation.
7. Laird J (2012). The Soar Cognitive Architecture. Cambridge, MA: MIT Press.
8. Molinaro K, Bolton ML (2018). Evaluating the applicability of the double system lens model to the analysis of phishing email judgments. Computers & Security, Volume 77, Pages 128-137, 2018. doi:10.1016/j.cose.2018.03.012
9. Veksler VD, and Buchler N (2016). Know Your Enemy: Applying Cognitive Modeling in Security Domain. Presented at 38th Annual Meeting of the Cognitive Science Society, Philadelphia, 2016.
10. Veksler VD, Buchler N, Hoffman BE, Cassenti DN, Sampie C, Sugrim S (2018). Simulations in Cyber-Security: A Review of Cognitive Modeling of Network Attackers, Defenders, and Users. Frontiers in Psychology 9. DOI: 10.3389/fpsyg.2018.00691
11. Vishwanath A, Harrison B, Ng YJ (in-press). Suspicion, Cognition, Automaticity Model (SCAM) of Phishing Susceptibility. Communication Research.
12. Zhang H, Singh S, Li X, Dahbura A, Xie M (2018). Multitasking and Monetary Incentive in a Realistic Phishing Study. British Human Computer Interaction Conference, 2018.